WELLCOME GENOME CAMPUS ADVANCED COURSE

Functional Genomics and Systems Biology

15-24 June 2016

Held at
Wellcome Genome Campus Advanced Courses Laboratory
Wellcome Trust Sanger Institute
Wellcome Genome Campus
Cambridge, UK

STAFF

Advanced Courses Manager

Dr Rebecca Twells
(email: advancedcourses@wellcomegenomecampus.org)
The Wellcome Trust, Hinxton, Cambridge

Course Instructors

Dr Ioannis Ragoussis *(e mail: Ioannis.ragoussis@mcgill.ca)*
McGill University, Montreal, Canada

Dr Tom Freeman *(e mail: Tom.Freeman@roslin.ed.ac.uk)*
Roslin Institute, University of Edinburgh

Dr Andrew Fraser *(e mail: andyfraser.utoronto@gmail.com)*
University of Toronto, Canada

Dr Anton Enright *(e mail: aje@ebi.ac.uk)*
European Bioinformatics Institute, Cambridge

Dr Jyoti Choudhary *(e mail: jc4@sanger.ac.uk)*
Wellcome Trust Sanger Institute, Cambridge

Course Assistants

Dr Mark Barnett
Roslin Institute, University of Edinburgh

Mr Daniel Bode
University of Dundee

Dr Carme Camps
The Wellcome Trust Centre for Human Genetics, Oxford

Dr Mat Davis
European Bioinformatics Institute, Cambridge

Dr Mercedes Pardo
Wellcome Trust Sanger Institute, Cambridge

Dr Mark Spensley
University of Toronto, Canada

Dr Stijn van Dongen
European Bioinformatics Institute, Cambridge

Mr David Wang
McGill University, Montreal, Canada

Dr Tim Regan
Roslin Institute, University of Edinburgh

Mr Dunarel Badescu
McGill University, Montreal, Canada

Advanced Courses Team	**Yvonne Thornton**	Advanced Courses Administrator
	Julie Ormond	Advanced Courses Laboratory Manager
	Darren Hughes	Advanced Courses Programme Officer
	Nicola Stevens	Advanced Courses Assistant Administrator
	Kate Waite	Advanced Courses Scientific Administrator
	Pamela Black	Advanced Courses Education Officer
	Martin Aslett	Advanced Courses IT Manager

Wellcome Trust Advanced Courses
URL: **www.wellcomegenomecampus.org/coursesandconferences**
email: **advancedcourses@wellcomegenomecampus.org**

ACKNOWLEDGEMENTS

We are very grateful to the following companies for their support in the loaning of equipment, gifts of consumables and/or technical advice.

Company	Web site	UK Telephone
Fluidigm	https://www.fluidigm.com	+33 1 60 92 42 40
Leica	www.leica-microsystems.com	01908 246 246
Olympus	www.olympus.co.uk/microscopy	01702 445160

Speakers

We would like to thank the following for giving seminars during the course:
Dr Marc-Emmanuel Dumas
Dr Mike Quail
Dr John Marioni
Dr Chris Yau
Professor Judith Steen
Professor Donal O'Carroll
Dr Geoff Scopes

Functional Genomics and Systems Biology

15-24 June 2016

Seminar Programme

Thursday 16 June
12:00 C303
'TBC'
Dr Marc-Emmanuel Dumas
Imperial College London, UK

Friday 17 June
12:00 C302
'Library Preparation for NGS'
Dr Mike Quail
Wellcome Trust Sanger Institute, UK

Monday 20 June
12:00 C302
'Using single-cell genomics to study early development'
Dr John Marioni
European Bioinformatics Institute, UK

Tuesday 21 June
12:00 C302
'Statistical uncertainty and why we should care'
Dr Chris Yau
Wellcome Trust Centre for Human Genetics, Oxford, UK

Wednesday 22 June
12:00 C302
'Interfacing the Proteome and Transcriptome'
Professor Judith Steen
Harvard Medical School, USA

Thursday 23 June
12:00 C303
'Non-coding solutions to developmental challenges'
Professor Donal O'Carroll
University of Edinburgh, UK

Functional Genomics 2016 Timetable

	Wednesday 15-Jun-16	Thursday 16-Jun-16	Friday 17-Jun-16	Saturday 18-Jun-16	Sunday 19-Jun-16	Monday 20-Jun-16	Tuesday 21-Jun-16	Wednesday 22-Jun-16	Thursday 23-Jun-16	Friday 24-Jun-16	
8.30		Briefing	Briefing	Intro to Pathway & Network Databases Anton Enright Discovery Zone	Briefing	Briefing	R & BioConductor Anton Enright	Introduction to Mass Spectrometry data analysis	**MS data analysis** Quant proteomics talk Quant MS data	Course Review Discovery Zone	8.30
9.00		IT Room	Discovery Zone		Discovery Zone	Discovery Zone	IT Room	IT Room			9.00
9.30		Lab	Lab	Lab	Lab	Lab	Lab	Gel Discussion			9.30
10:00								Tea/Coffee SF Lobby			10:00
10:30		Tea/Coffee SF Lobby	Tea/Coffee Discovery Zone	Tea/Coffee Discovery Zone	Tea/Coffee Discovery Zone	Tea/Coffee Discovery Zone	Tea/Coffee SF Lobby	Course microarray data anlysis / R/Bioc Anton Enright IT Room	IT Room	Tea/Coffee Discovery Zone	10:30
11:00		Lab	Lab	Lab	R & Bioconductor Practical Anton Enright	Lab	Lab		Analysis of Course Data Tom Freeman/Anton Enright	Course Review	11:00
11:30								Overvieew of worm data Mammalian RNAi	IT Room		11:30
12:00		Seminar **Marc-Emmanuel Dumas**	Seminar **Mike Quail**			Seminar **John Marioni**	Seminar **Chris Yau**	Seminar **Judith Steen**	Seminar **Donal O'Carrol**	Lunch	12:00
12:30		C303	C302		IT Room	C302	C302	C302	C303	Restaurant	12:30
13:00		Lunch Participant Talks	Lunch Participant Talks	Lunch Participant Talks	Lunch Participant Talks	Lunch Instructor Talks	Lunch Instructor Talks	Lunch Instructor Talks	Lunch Instructor Talks	Departure	13:00
13:30		C303	C303	C303	C303	C302	C303	C303	C303		13:30
14:00	Registration Conference Centre	Discussion	Discussion	Discussion	Discussion	Discussion	Discussion	Discussion	Discussion		14:00
14:30		Lab	Lab	Lab	Lab	Lab	BioLayout Practical Tom Freeman	Course microRNA Seq Analysis Anton Enright	Analysis of Course Data Anton Enright		14:30
15:00	Sanger Tour Gp 1					Geoff Scopes Affymetrix Discovery Zone					15:00
15:30	Fluidigm Tour Gp 2	Tea/Coffee SF Lobby		Tea/Coffee DZ	Tea/Coffee DZ			IT Room	IT Room		15:30
16:00	Fluidigm Tour Gp1 Sanger Tour Gp 2	Lab	Tea/Coffee DZ Lab	Lab	Lab	Lab	Tea/Coffee SF Lobby RNA Seq Analysis Anton Enright	Tea/Coffee SF Lobby Open Session Data Analysis	Tea/Coffee SF Lobby mRNA Seq data anlysis Anton/Jiannis/Dunarel		16:00
16:30	Introduction Lead Instructor WTAC Discovery zone					BioLayout Lab					16:30
17:00		Introduction to Microarray Analysis Tom Freeman					IT Room	Analysis of particiapant data			17:00
17:30	EBI Overview					Practical Tom Freeman	RNA -seq Tutorial		IT Room		17:30
18:00	Discovery zone			Supper	Introduction to BioLayout Practical Tom Freeman				Tom Freeman/Anton Enright IT Room		18:00
18:30	WTAC Lab Orientation	IT Room	Lab	Restaurant	IT Room	MiRNA detection talk Jiannis Ragoussis	IT Room	IT Room	Free Time		18:30
19:00	BBQ	Supper	Supper	Coach to Cambridge	BBQ & Beer Tasting	Supper	Supper	Supper	Pre-dinner drinks New Space Bar		19:00
19:30	Restaurant	Restaurant	Restaurant		Restaurant	Restaurant	Restaurant	Restaurant			19:30
20.00				Punting	Free Evening	Introduction to Network Analysis TBA Tom/Stijn/Matt		BioLayout Practical Tom Freeman	Course Dinner Restaurant		20:00
20.30											20:30
21:00								IT Room			21:00
21:30											21:30
22.00				Coach returns 23:00							22:00
	Wednesday 15-Jun-16	Thursday 16-Jun-16	Friday 17-Jun-16	Saturday 18-Jun-16	Sunday 19-Jun-16	Monday 20-Jun-16	Tuesday 21-Jun-16	Wednesday 22-Jun-16	Thursday 23-Jun-16	Friday 24-Jun-16	

INTRODUCTION

The "classical" genetics approach to disease gene identification leads to a positional cloning or positional candidate gene approach. The positional cloning and functional analysis of genes involved in human inherited diseases or animal models of disease is a multistep process involving a variety of experimental and computer based techniques. The initial steps are focused on finding a chromosomal position for the disease or mutant gene using chromosome abnormalities such as deletions and translocations and/or genetic linkage analysis in families segregating the disease phenotype. Once the chromosome position is identified further steps are needed to refine this position using a combination of both physical and genetic mapping. The integrated physical and genetic maps are fairly well defined in the public domain using yeast artificial chromosome (YAC) clones, E coli based artificial chromosome clones (BACs, PACs), radiation hybrid panels and human genomic sequence (for example http://www.broad.mit.edu/resources.html). When a disease or mutant gene can be narrowed in position to about 2-3 Mb then transcriptional maps are required to identify candidate genes that can be tested for mutations in affected individuals. For human diseases and the animal models for which the DNA sequence is available, the cataloging of genes is done by mining databases and using appropriate websites such as Ensembl http://www.ensembl.org/. Further verification of the database information may be required and this involves RT-PCR or RACE experiments.

The use of bioinformatics, as more information is now available in the public domain databases, allows many of the steps in positional cloning and functional analysis to be accomplished *in silico*.

In complementation to genetic studies, another way of deriving candidate genes is to perform microarray experiments and compare disease versus healthy control states or different and identify genes that show differential expression patterns [1]. More recently approaches have been developed that bring us several steps further towards the goal of understanding the function of all genes and these are based on mutagenesis, RNA interference and the identification of protein-protein interactions.

Once the correct disease gene is identified then functional analysis begins to determine the tissue and developmental expression of the gene, the sub cellular localisation of the encoded protein and interacting proteins in the functional pathway. Further functional studies can be performed by expressing the disease gene either as cDNA or large genomic constructs in cell-lines or transgenic mice allowing further experimental analysis of function [2].

All of the approaches required for the identification and functional analysis of disease or mutant genes are fairly straightforward yet require the use of a variety of techniques and experimental systems. The aim of this course and protocol manual is to introduce these methods to researchers who have a basic molecular biology background.

Historical perspective of gene identification

So far expressed sequences have been identified and mapped through the large scale genomic DNA sequencing projects. The human genome sequence is now available and the sequence status of several chromosomes is regarded as "complete", [3-6]. Although "known" genes and ESTs have been positioned on the sequence, the identification of additional potential exons can be achieved using computer algorithms. The most successful analysis programs are GRAIL (Uberbacher & Mural, 1991) which uses a multi-sensor/neuronal network approach to identify potential coding sequences and BlastX (Gish & States, 1993) which directly translates nucleotide sequences and performs homology searches in protein databases. A combination of programs is commonly used (for example NIX at HGMP), which combine results derived from different algorithms and indicating the position of low complexity sequences, potential exons, sequence database matches, polyadenylation sites, putative promotor sequences etc. This type of analysis is very powerful but "wet" experiments may still be important for obtaining a complete transcript map (Tripodis *et al.,* 2000), even now when the genome sequence is regarded as annotated for all known genes but not necessarily all expressed sequences [7, 8]. Systematic studies are underway in order to produce a concise collection of full length cDNA clones for all main mammalian species [9].

Functional studies

In principle we can define two types of systems that are used for functional studies: "Closed" systems that investigate a defined set of known genes based on a working hypothesis and "open" systems where prior knowledge of the gene sequence is not required. An approach such as serial analysis of gene expression (SAGE) or more recently the application of high throughput sequencing technologies [10] represents an open system, while a microarray of cDNAs or oligonucleotides a closed system for monitoring gene expression (for method description see under "microarray" below). For functional analysis a systematic ENU mutagenesis approach [11] or a yeast two hybrid library screen represent open systems, while an RNAi approach [12] or microarray based yeast two-hybrid screen [13] represent closed systems even at high throughput.

Expression profiling

In addition to the analysis of individual genes, more global methods are now available to analyse gene expression. One strategy is to sequence multiple ESTs from each cell type or tissue of choice followed by sequence and bioinformatic analyses. This can be done using short oligonucleotide tags for each gene constructed into libraries that can be sequenced to a higher redundancy than EST sequencing. This has been called Serial Analysis of Gene Expression (SAGE) [14]. Random oligonucleotides can be used to prime cDNA synthesis in pools for a gel based comparison of gene expression termed "differential display" and its subsequent improvements (Liang and Pardee 1992, Liang *et al.*, 1993, Bauer *et al.*, 1993).

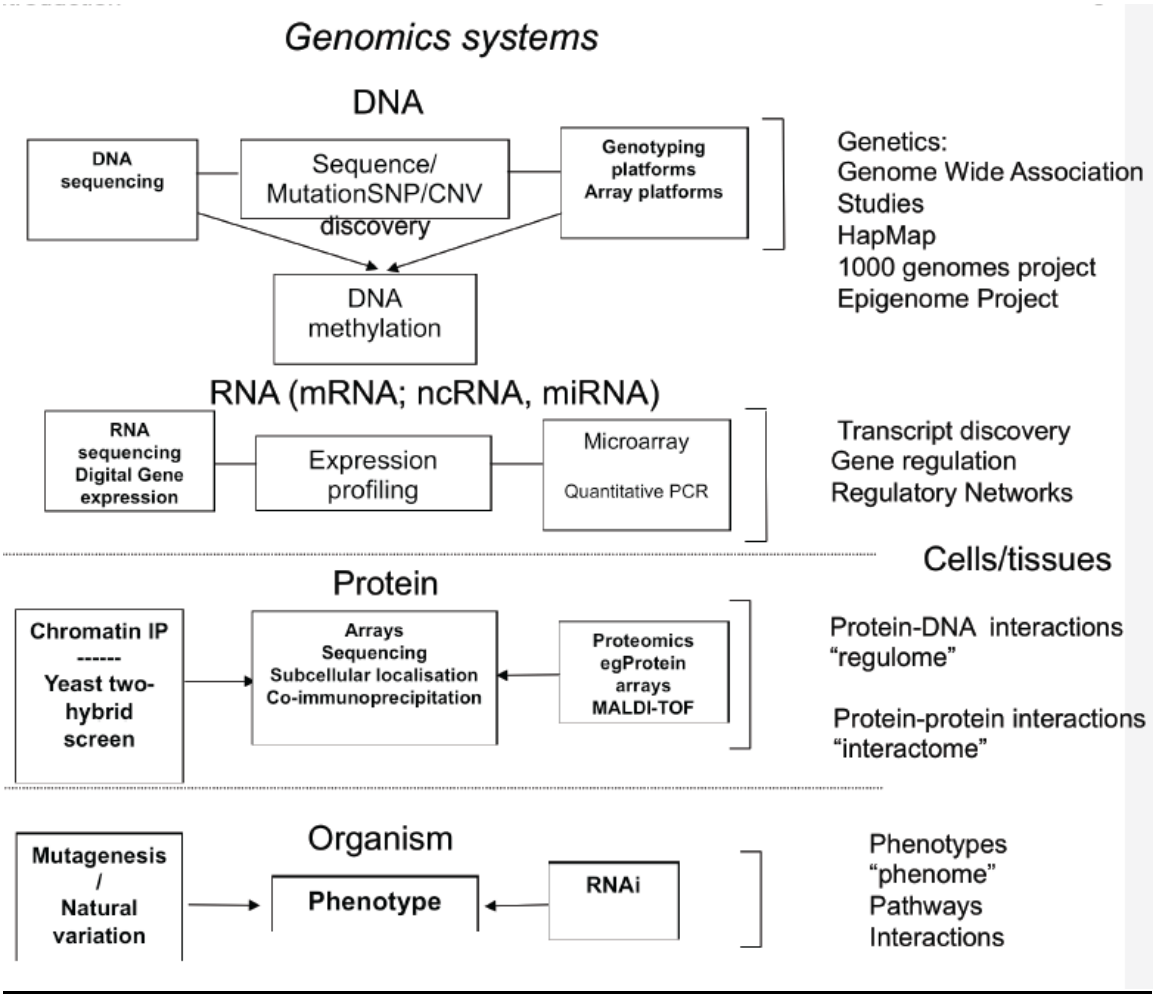


Figure 1: Open and closed system for functional genomics

Oligonucleotide arrays on "DNA chips " are also used to analyse gene expression after hybridisation with cDNA derived from mRNA isolated from cells or tissues of interest (for review see Hoheisel 1997**,** for description of the Affymetrix genechip see [15]. Oligonucleotide arrays have been used to analyse sequence variability in the human mitochondrial genome (Chee *et al.*, 1996) and human normal colon cancer tissue [16]. The other array system spots cDNA or DNA onto glass substrates followed by differential hybridisation of fluorescently labeled DNA samples. This has been used to analyse the yeast genome (Schalon *et al.*, 1996; DeRisi *et al.*, 1997), and expression patterns in human tissues or disease states (DeRisi *et al.*, 1996) as well as changes in DNA copy number in cancer cell lines (DeRisi *et al.,* 2000**).** More recently microarray data from different species have been used to identify evolutionary conserved, co-regulated genes that can be clustered into functional modules [17, 18]. However, despite big advances, the analysis of

microarray data is still problematic with significant discrepancies between different methods and across platforms [19]. One reason for these discrepancies is the lack of sensitivity of microarrays leading to the unreliable detection of low level transcripts (i.e. less than 10 copies/cell) [20] [21]. Recently robust methods for the analysis of commercial microarrays, such as Affymetrix's GeneChips have been developed [22], while new types of comprehensive arrays, as produced by Illumina allow for additional validations [23]. Affymetrix has now produced arrays that interrogate individual exons and are ideal for the discovery of tissue specific splicing events [24, 25]

Validation or more detailed analysis can be performed using quantitative RT-PCR techniques (reviewed in Bustin, S.A.: Meaningful Quantitation of mRNA using real time PCR (2004). in PCR Technology: Current Innovations Ed. T Weissensteiner, CRC press, p225-233) while more recently advanced, high-throughput competitive PCR methods have been developed [26].

The recent emergence of high throughput and low cost sequencing technologies [27]pioneered by 454 technologies followed by Solexa and Applied Biosystems is having a major impact on expression profiling applications and functional genomics in general[28]. The ability to generate 100 Mbp to 1 Gbp of sequence in a single instrument run allows the identification and quantitation of transcripts as well as the discovery of splice variants and polymorphisms [29, 30]. The technology had a particularly high impact in small RNA discovery and characterization [31] [32, 33] [34].

Typically several millions of 25-150bp reads are generated from mRNAs and mapped to the genome in order to identify transcribed sequences. Relative quantification is achieved by expressing the data as reads per kilobase per million reads (RPKM). Further refinements of the method include the generation of reads that retain transcript orientation information [35, 36]. A further advance has come in the form of protocols suitable for low input and degraded samples [41]. More importantly, it is now possible to sequence a transcriptome *de novo* opening new ways in the study of uncharacterized organisms. (reviewed in [42] .

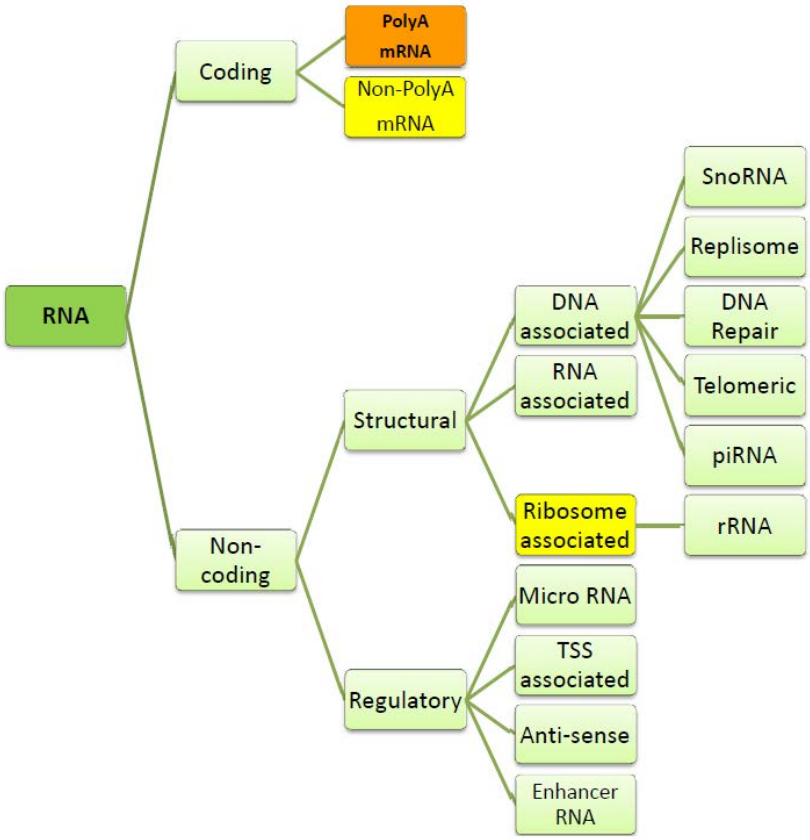


Figure 2: Species of RNAs. In order to identify all one needs to combine total RNA sequencing with small fraction sequencing. In typical polyA mRNA sequencing or microarray experiments only the polyAmRNA fraction (orange) is analyzed.

Current microarrays supplied by Affymetrix , directional mRNA-seq protocols suitable for single cell transcriptome analysiswill be used in the course providing the opportunity to get experience with the two main platforms used today.In additions methods for producing libraries containg both polyA and non polyA transcripts will be applied. In addition a bioinformatics analysis of data is included in terms of tutorials and hands on practicals.

Single Cell transcriptomics

Analysis at the resolution of single cells is the new frontier in genomics and proteomics. It aims to investigate cellular heterogeneity by analyzing individual cells leading to accurate representation of cell-to cell variation instead of the stochastic average resulting from bulk measurements. This level of resolution is important in order to advance our understanding in biomedical fields such as cancer, prenatal diagnosis, aging and neuroscience, as well as in basic biological processes such as development and immunity.

Towards this aim, ultra sensitive methods that can capture transcripts from from single cells [37] [38] [39] have been developed. These have been combined with microfluidic or droplet based technologies to capture and analyse single cells [40] [70]
In addition, bespoke Bioinformatic analysis tools had to be developed. They required new approaches in order to estimate biological variants and separate from technical noise. The application of RNA spikes became a good method to account for the latter and were incorporated in standard methodology. [71][72]

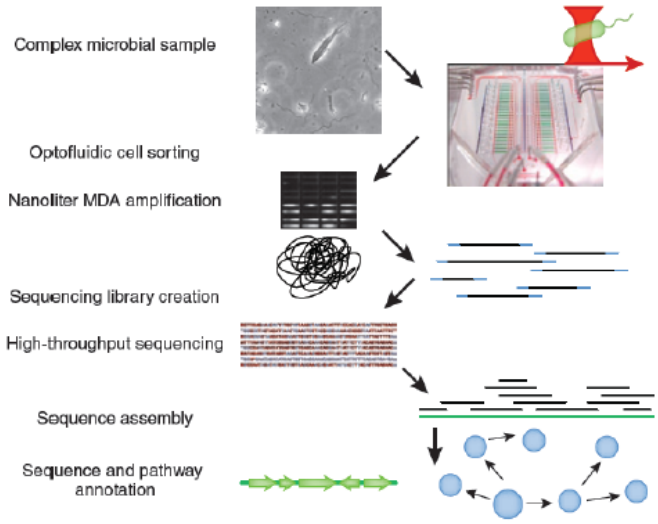


Figure 3: Single cell genomics is a new field. It can be applied to a wide field of applications and an example is shown above, whereby an approach involving isolation of microorganisms and analysis using NGS is shown.

During the course we will capture single cells from an experiment using human breast cancer MCF-7 cells with a microfluidic device available by Fluidigm (C1) and perform cDNA synthesis followed by RNA-seq. Analysis will also be performed in order to identify cell subpopulations.

MicroRNAs

MicroRNAs were first discovered in plants, drosophila and C.elegans in connection with the discovery of RNA interference [43-45]. miRNAs are small RNA molecules with a length of approximately 22 nucleotides. There are now thousands of miRNAs discovered, 450 of which are identified in man, while a considerable number are highly conserved, for example about one third of the human micrRNAs have homologs in C. elegans (http://www.sanger.ac.uk/Software/Rfam/mirna/). The microRNa sequences are deposited imiRbase http://microrna.sanger.ac.uk/, containing nearly 5000 sequences today.

miRNAs are generated from an original Pol II transcribed pre-miRNA mRNA, whereby a hairpin structure is formed, and used to generate the short miRNA transcripts.
A schematic representation of the miRNA genearation and function is given over page:

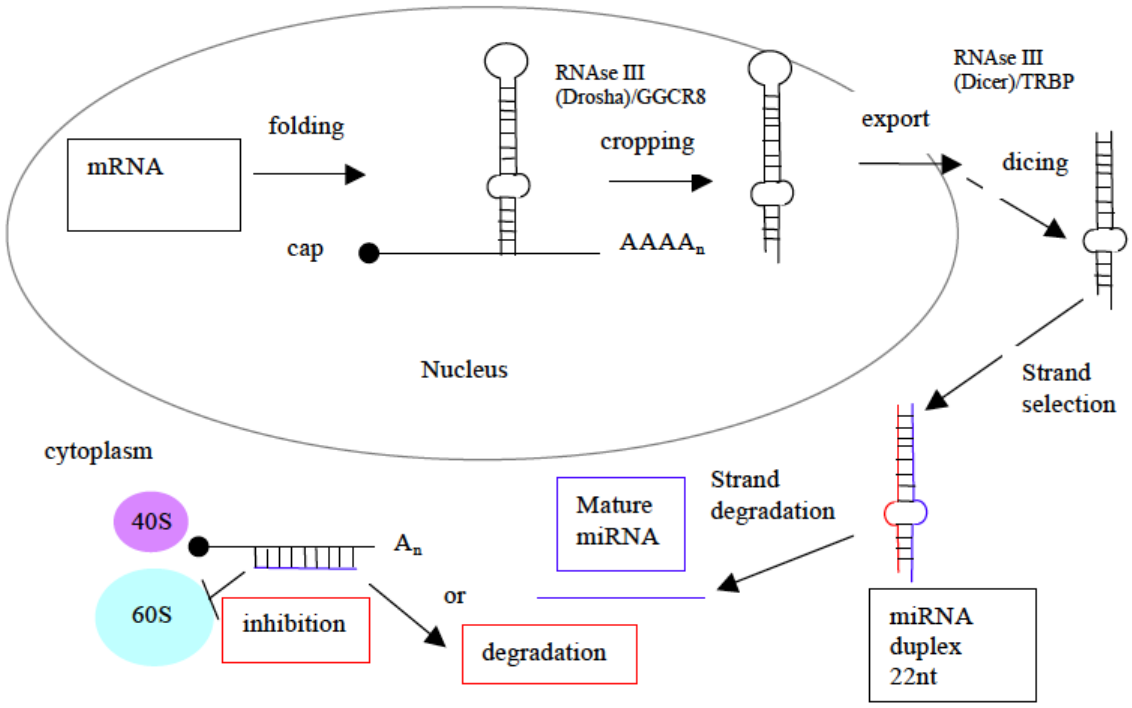


Figure 4: miRNAs are generated from mRNAs, folding to form hairpin loops. These are recognized and cropped (aprox. length 70nt; and 2nt 3' overhang) by the microprocessor complex (RNAse III Drosha and GCR8, a double strand RNA binding protein), exported from the nucleus to the cytoplasm by exportin-5. Cytoplasmic RNAse III Dicer with the RNA binding protein TRBP(trans-activation responsive RNA binding protein) produces an RNA duplex. The strand is separated and usually one strand is degraded, while the other one forms the miRNA. miRNAs are then involved in either silencing the translation of mRNAs or lead to their degradation.

miRNAs are involved in the regulation of gene expression through a number of ways, including the degradation of transcripts as well as inhibition of the translation machinery following the contact of the miRNA with the 3'UTR of the mRNA. This contact is imperfect, and many 3'UTRs of mammalian genes are able to form duplexes with known miRNAs and it is estimated that up to 30% of animal kingdom genes may be regulated by miRNAs [46-50]. In a pioneering effort lead by Alan Bradley at the WTSI, a mouse model for studying the function of a microRNA (microRNA-155) was generated, demonstrating its multiple roles in immune functions [51].

<u>In this course we will sequence miRNAs to arrays containing sequences corresponding to the set of known human miRNAs.</u>

RNAi

RNAi is a method that allows the epigenetic inactivation of genes through the cleavage of mRNA mediated through the interference of homologous double stranded RNA molecules. The dsRNA is initially either introduced into the cytoplasm, or generated from endogenous transcripts or taken up by the cell. Subsequently the dsRNA is cut by Dicer into siRNAs that are involved in silencing mechanisms similar to miRNAs. For review see [52]

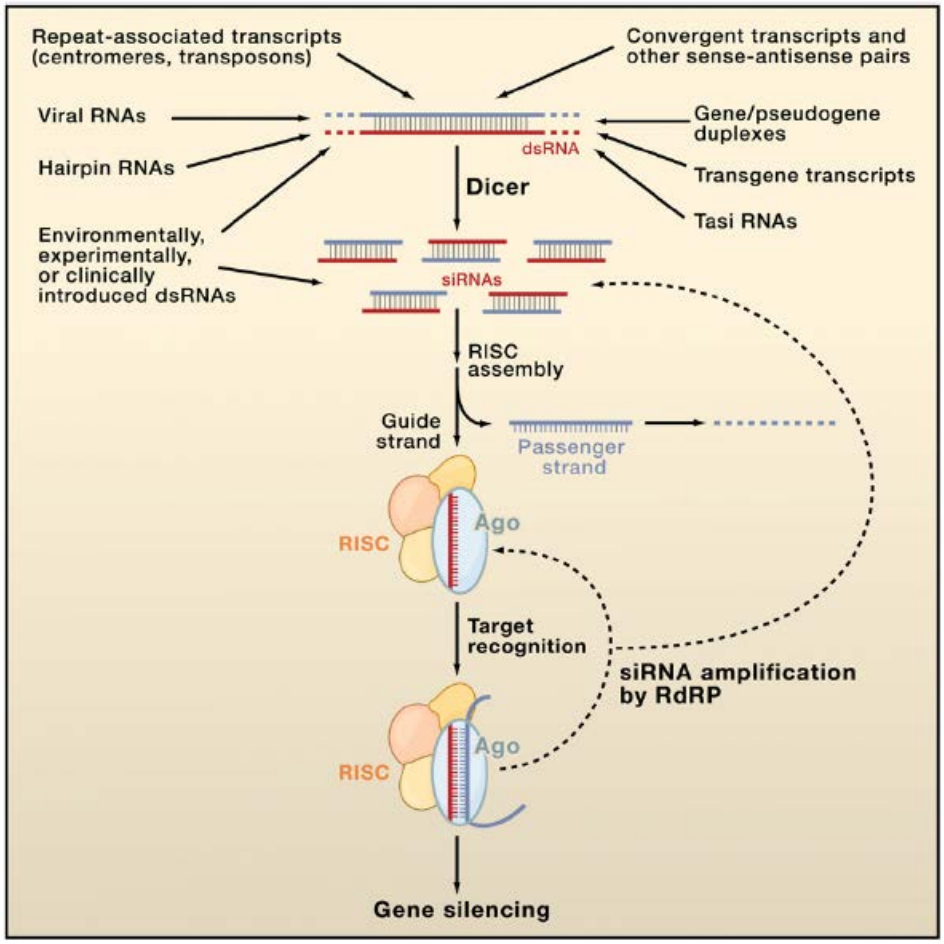


Figure 5: Sources of siRNAs and their action (From Carthew and Sontheimer, Cell 2009)

The main purpose of these mechanism is defense against viruses Initially applied systematically on C. elegans [53], the technique works in a variety of organisms [52] including mammalian and plant cells [53-56]. It has also a great potential as therapeutic approach in the Clinic [57-58]. Recently expression vectors have been developed directing the synthesis of short hairpin RNAs (shRNAs) that act as short interfering RNA (siRNA)-like

molecules [59]. Their action is similar to miRNAs, apart from the fact that normally they form perfect base pairing to the target transcript. This way a stable suppression of gene expression can be achieved allowing large-scale loss of function genetic screens in mammalian systems [12]. Systematic pathway studies have been now demonstrated for C. elegans [60], while the method can be combined with protein interaction studies [61] in order to study both gene networks and whole pathways.
Experiments using C. elegans as a model organism will be performed in this course.

Practical computational biology

There is an exponentially expanding body of knowledge of biological sequences. The databases of sequences double in size every year but the problems associated with such large bodies of data also increase. Particularly now the human genome sequence has been completed [3].

Biologists must now acquire the skills to navigate through this data and to use the computer programs to extract the maximum amount of data from their sequences, avoiding the various pitfalls that can befall them.

The Internet and the WWW has resulted in a profusion of sites which offer various forms of analysis of sequences.
Apart from sequence analysis computational biology is now facing the challenge of integrating information generated by high-throughput techniques including not only sequence information but sequence variation, gene expression profiles, protein-DNA and protein-protein interactions as well as protein modification and metabolites. This information is used to construct interaction or regulatory networks in what is now a systems biology [62]. In systems biology the data produced and published by the different disciplines are extracted and put into databases and organized in pathways using a number of algorithms [63]. The most predominant databases are KEGG (http://www.genome.jp/kegg/) and GO. These represent bioinformatics resources that aim at achieving "a complete computer representation of the cell, the organism and the biosphere" (KEGG), and thus allow researchers to identify pathways involved in phenotypes from for example expression data or predict the effect of perturbations in a given system. GO (http://www.geneontology.org/) is a gene ontology database aimed at attributing functions to genes and relationships between genes and functions. Several bioinformatics tools for data mining have been developed and are available at these sites [64-66]

In the laboratory sessions you will work through a number of practical examples on how to use some of these sites to find genes and predict their function.

Yeast two-hybrid analysis

To identify interacting protein, specific antibodies can be used to immunoprecipitate complexes formed with the protein of interest, by cross-linking or by co-fractionation using chromatography. In addition, the yeast two hybrid system is a powerful method to identify interacting proteins in vivo. This system utilizes two expression plasmids in yeast, one of which carries a hybrid gene encoding a DNA-binding domain fused to protein X and the second, an activation domain fused to protein Y (Fields and Song, 1989, Fields and Sternglanz, 1994). The interaction of protein X and protein Y bring together the DNA binding domain with the activation domain resulting in transcription of a reporter gene which allows simple selection in yeast. This system has been extended to clone genes that encode DNA binding proteins, to find peptides that bind to proteins and as a drug screening strategy . Many variations have been developed allowing the detection of protein-DNA, RNA-protein and small molecule-protein interactions [66-68]. The method can be scaled up for whole genome level studies as demonstrated in the model organism C. elegans [68].

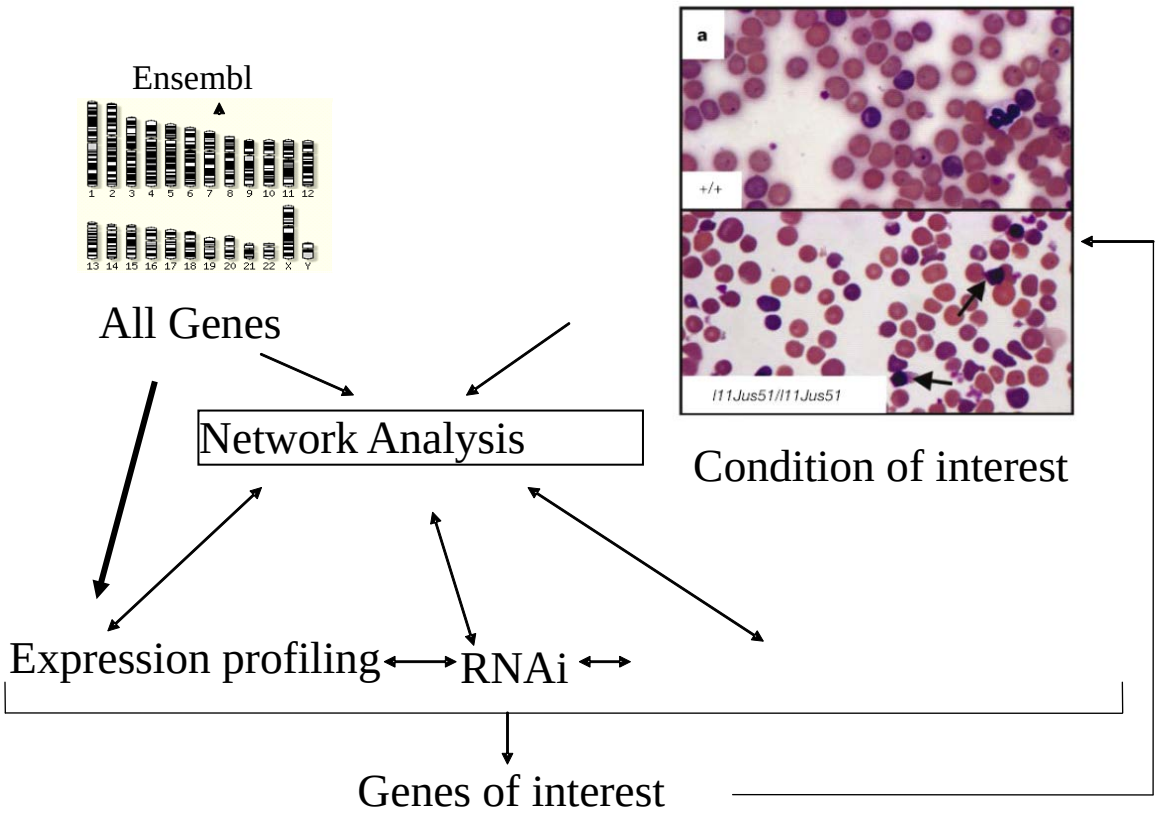


Figure 6: Summary diagram showing the relationship between the methods applied in the course. The genes of interest can be selected using bioinformatics approaches leading to hypotheses, which can be tested using experimental techniques. At the same time the high-

throughput techniques can be used in a non-hypothesis driven, systematic way to generate genes of interest. The haematological condition figure is from [69].

References

[1] D.A. Stephan, Y. Chen, Y. Jiang, L. Malechek, J.Z. Gu, C.M. Robbins, M.L. Bittner, J.A. Morris, E. Carstea, P.S. Meltzer, K. Adler, R. Garlick, J.M. Trent, M.A. Ashlock, Positional cloning utilizing genomic DNA microarrays: the Niemann-Pick type C gene as a model system, Mol Genet Metab, 70 (2000) 10-18.
[2] E.D. Carstea, J.A. Morris, K.G. Coleman, S.K. Loftus, D. Zhang, C. Cummings, J. Gu, M.A. Rosenfeld, W.J. Pavan, D.B. Krizman, J. Nagle, M.H. Polymeropoulos, S.L. Sturley, Y.A. Ioannou, M.E. Higgins, M. Comly, A. Cooney, A. Brown, C.R. Kaneski, E.J. Blanchette-Mackie, N.K. Dwyer, E.B. Neufeld, T.Y. Chang, L. Liscum, D.A. Tagle, et al., Niemann-Pick C1 disease gene: homology to mediators of cholesterol homeostasis, Science, 277 (1997) 228-231.
[3] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, Pan, Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium, Nature, 409 (2001) 860-921.
[4] P. Deloukas, L.H. Matthews, J. Ashurst, J. Burton, J.G. Gilbert, M. Jones, G. Stavrides, J.P. Almeida, A.K. Babbage, C.L. Bagguley, J. Bailey, K.F. Barlow, K.N. Bates, L.M. Beard, D.M. Beare, O.P. Beasley, C.P. Bird, S.E. Blakey, A.M. Bridgeman, A.J. Brown, D. Buck, W. Burrill, A.P. Butler, C. Carder, N.P. Carter, J.C. Chapman, M. Clamp, G. Clark, L.N. Clark, S.Y. Clark, C.M. Clee, S. Clegg, V.E. Cobley, R.E. Collier, R. Connor, N.R. Corby, A. Coulson, G.J. Coville, R. Deadman, P. Dhami, M. Dunn, A.G. Ellington, J.A. Frankland, A. Fraser, L. French, P. Garner, D.V. Grafham, C. Griffiths, M.N. Griffiths, R. Gwilliam, R.E. Hall, S. Hammond, J.L. Harley, P.D.

Heath, S. Ho, J.L. Holden, P.J. Howden, E. Huckle, A.R. Hunt, S.E. Hunt, K. Jekosch, C.M. Johnson, D. Johnson, M.P. Kay, A.M. Kimberley, A. King, A. Knights, G.K. Laird, S. Lawlor, M.H. Lehvaslaiho, M. Leversha, C. Lloyd, D.M. Lloyd, J.D. Lovell, V.L. Marsh, S.L. Martin, L.J. McConnachie, K. McLay, A.A. McMurray, S. Milne, D. Mistry, M.J. Moore, J.C. Mullikin, T. Nickerson, K. Oliver, A. Parker, R. Patel, T.A. Pearce, A.I. Peck, B.J. Phillimore, S.R. Prathalingam, R.W. Plumb, H. Ramsay, C.M. Rice, M.T. Ross, C.E. Scott, H.K. Sehra, R. Shownkeen, S. Sims, C.D. Skuce, M.L. Smith, C. Soderlund, C.A. Steward, J.E. Sulston, M. Swann, N. Sycamore, R. Taylor, L. Tee, D.W. Thomas, A. Thorpe, A. Tracey, A.C. Tromans, M. Vaudin, M. Wall, J.M. Wallis, S.L. Whitehead, P. Whittaker, D.L. Willey, L. Williams, S.A. Williams, L. Wilming, P.W. Wray, T. Hubbard, R.M. Durbin, D.R. Bentley, S. Beck, J. Rogers, The DNA sequence and comparative analysis of human chromosome 20, Nature, 414 (2001) 865-871.

[5] I. Dunham, N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, R. Bruskiewich, D.M. Beare, M. Clamp, L.J. Smink, R. Ainscough, J.P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow, K.N. Bates, O. Beasley, C.P. Bird, S. Blakey, A.M. Bridgeman, D. Buck, J. Burgess, W.D. Burrill, K.P. O'Brien, et al., The DNA sequence of human chromosome 22, Nature, 402 (1999) 489-495.

[6] M. Hattori, A. Fujiyama, T.D. Taylor, H. Watanabe, T. Yada, H.S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.K. Choi, E. Soeda, M. Ohki, T. Takagi, Y. Sakaki, S. Taudien, K. Blechschmidt, A. Polley, U. Menzel, J. Delabar, K. Kumpf, R. Lehmann, D. Patterson, K. Reichwald, A. Rump, M. Schillhabel, A. Schudy, W. Zimmermann, A. Rosenthal, J. Kudoh, K. Schibuya, K. Kawasaki, S. Asakawa, A. Shintani, T. Sasaki, K. Nagamine, S. Mitsuyama, S.E. Antonarakis, S. Minoshima, N. Shimizu, G. Nordsiek, K. Hornischer, P. Brant, M. Scharfe, O. Schon, A. Desario, J. Reichelt, G. Kauer, H. Blocker, J. Ramser, A. Beck, S. Klages, S. Hennig, L. Riesselmann, E. Dagand, T. Haaf, S. Wehrmeyer, K. Borzym, K. Gardiner, D. Nizetic, F. Francis, H. Lehrach, R. Reinhardt, M.L. Yaspo, The DNA sequence of human chromosome 21, Nature, 405 (2000) 311-319.

[7] P. Kapranov, S.E. Cawley, J. Drenkow, S. Bekiranov, R.L. Strausberg, S.P. Fodor, T.R. Gingeras, Large-scale transcriptional activity in chromosomes 21 and 22, Science, 296 (2002) 916-919.

[8] S. Cawley, S. Bekiranov, H.H. Ng, P. Kapranov, E.A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A.J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, T.R. Gingeras, Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs, Cell, 116 (2004) 499-509.

[9] D.S. Gerhard, L. Wagner, E.A. Feingold, C.M. Shenmen, L.H. Grouse, G. Schuler, S.L. Klein, S. Old, R. Rasooly, P. Good, M. Guyer, A.M. Peck, J.G. Derge, D. Lipman, F.S. Collins, W. Jang, S. Sherry, M. Feolo, L. Misquitta, E. Lee, K. Rotmistrovsky, S.F. Greenhut, C.F. Schaefer, K. Buetow, T.I. Bonner, D. Haussler, J. Kent, M. Kiekhaus, T. Furey, M. Brent, C. Prange, K. Schreiber, N. Shapiro, N.K. Bhat, R.F. Hopkins, F. Hsie, T. Driscoll, M.B. Soares, T.L. Casavant, T.E. Scheetz, M.J. Brown-stein, T.B. Usdin, S. Toshiyuki, P. Carninci, Y. Piao, D.B. Dudekula, M.S. Ko, K. Kawakami, Y. Suzuki, S. Sugano, C.E. Gruber, M.R. Smith, B. Simmons, T. Moore, R. Waterman, S.L. Johnson, Y. Ruan, C.L. Wei, S. Mathavan, P.H. Gunaratne, J. Wu, A.M. Garcia, S.W. Hulyk, E. Fuh, Y. Yuan, A. Sneed, C. Kowis, A. Hodgson, D.M. Muzny, J. McPherson, R.A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madari, A.C. Young, K.D. Wetherby, S.J. Granite, P.N. Kwong, C.P. Brinkley, R.L. Pearson, G.G. Bouffard, R.W. Blakesly, E.D. Green, M.C. Dickson, A.C. Rodriguez, J. Grimwood, J. Schmutz,

R.M. Myers, Y.S. Butterfield, M. Griffith, O.L. Griffith, M.I. Krzywinski, N. Liao, R. Morin, D. Palmquist, A.S. Petrescu, U. Skalska, D.E. Smailus, J.M. Stott, A. Schnerch, J.E. Schein, S.J. Jones, R.A. Holt, A. Baross, M.A. Marra, S. Clifton, K.A. Makowski, S. Bosak, J. Malek, The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC), Genome Res, 14 (2004) 2121-2127.
[10] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg, Genome sequencing in microfabricated high-density picolitre reactors, Nature, 437 (2005) 376-380.
[11] S.D. Brown, R. Balling, Systematic approaches to mouse mutagenesis, Curr Opin Genet Dev, 11 (2001) 268-273.
[12] K. Berns, E.M. Hijmans, J. Mullenders, T.R. Brummelkamp, A. Velds, M. Heimerikx, R.M. Kerkhoven, M. Madiredjo, W. Nijkamp, B. Weigelt, R. Agami, W. Ge, G. Cavet, P.S. Linsley, R.L. Beijersbergen, R. Bernards, A large-scale RNAi screen in human cells identifies new components of the p53 pathway, Nature, 428 (2004) 431-437.
[13] P. Uetz, Two-hybrid arrays, Curr Opin Chem Biol, 6 (2002) 57-62.
[14] V.E. Velculescu, L. Zhang, B. Vogelstein, K.W. Kinzler, Serial analysis of gene expression, Science, 270 (1995) 484-487.
[15] R.J. Lipshutz, S.P. Fodor, T.R. Gingeras, D.J. Lockhart, High density synthetic oligonucleotide arrays, Nat Genet, 21 (1999) 20-24.
[16] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc Natl Acad Sci U S A, 96 (1999) 6745-6750.
[17] J.M. Stuart, E. Segal, D. Koller, S.K. Kim, A gene-coexpression network for global discovery of conserved genetic modules, Science, 302 (2003) 249-255.
[18] Z. Bar-Joseph, G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.B. Gordon, E. Fraenkel, T.S. Jaakkola, R.A. Young, D.K. Gifford, Computational discovery of gene modules and regulatory networks, Nat Biotechnol, 21 (2003) 1337-1342.
[19] G.L. Miklos, R. Maleszka, Microarray reality checks in the context of a complex disease, Nat Biotechnol, 22 (2004) 615-621.
[20] M.D. Kane, T.A. Jatkoe, C.R. Stumpf, J. Lu, J.D. Thomas, S.J. Madore, Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays, Nucleic Acids Res, 28 (2000) 4552-4557.
[21] S.E. Choe, M. Boutros, A.M. Michelson, G.M. Church, M.S. Halfon, Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset, Genome Biol, 6 (2005) R16.
[22] K. Shedden, W. Chen, R. Kuick, D. Ghosh, J. Macdonald, K.R. Cho, T.J. Giordano, S.B. Gruber, E.R. Fearon, J.M. Taylor, S. Hanash, Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data, BMC Bioinformatics, 6 (2005) 26.

[23] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, P. Pavlidis, Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms, Nucleic Acids Res, 33 (2005) 5914-5923.
[24] P.J. Gardina, T.A. Clark, B. Shimada, M.K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams, Y. Turpaz, Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array, BMC Genomics, 7 (2006) 325.
[25] T.A. Clark, A.C. Schweitzer, T.X. Chen, M.K. Staples, G. Lu, H. Wang, A. Williams, J.E. Blume, Discovery of tissue-specific exons using comprehensive human exon microarrays, Genome Biol, 8 (2007) R64.
[26] C. Ding, C.R. Cantor, A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS, Proc Natl Acad Sci U S A, 100 (2003) 3059-3064.
[27] S.T. Bennett, C. Barnes, A. Cox, L. Davies, C. Brown, Toward the 1,000 dollars human genome, Pharmacogenomics, 6 (2005) 373-382.
[28] A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, K. Zhao, High-resolution profiling of histone methylations in the human genome, Cell, 129 (2007) 823-837.
[29] S.J. Emrich, W.B. Barbazuk, L. Li, P.S. Schnable, Gene discovery and annotation using LCM-454 transcriptome sequencing, Genome Res, 17 (2007) 69-73.
[30] B. Wold, R.M. Myers, Sequence census methods for functional genomics, Nat Methods, 5 (2008) 19-21.
[31] C. Lu, K. Kulkarni, F.F. Souret, R. MuthuValliappan, S.S. Tej, R.S. Poethig, I.R. Henderson, S.E. Jacobsen, W. Wang, P.J. Green, B.C. Meyers, MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant, Genome Res, 16 (2006) 1276-1288.
[32] E.A. Glazov, P.A. Cottee, W.C. Barris, R.J. Moore, B.P. Dalrymple, M.L. Tizard, A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach, Genome Res, (2008).
[33] R.D. Morin, M.D. O'Connor, M. Griffith, F. Kuchenbauer, A. Delaney, A.L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst, C.J. Eaves, M.A. Marra, Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells, Genome Res, 18 (2008) 610-621.
[34] D.M. Tyler, K. Okamura, W.J. Chung, J.W. Hagen, E. Berezikov, G.J. Hannon, E.C. Lai, Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci, Genes Dev, 22 (2008) 26-36.
[35] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, A. Soldatov, Transcriptome analysis by strand-specific sequencing of complementary DNA, Nucleic Acids Res, 37 (2009) e123.
[36] T.T. Perkins, R.A. Kingsley, M.C. Fookes, P.P. Gardner, K.D. James, L. Yu, S.A. Assefa, M. He, N.J. Croucher, D.J. Pickard, D.J. Maskell, J. Parkhill, J. Choudhary, N.R. Thomson, G. Dougan, A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi, PLoS Genet, 5 (2009) e1000569.
[37] F. Tang, C. Barbacioru, E. Nordman, B. Li, N. Xu, V.I. Bashkirov, K. Lao, M.A. Surani, RNA-Seq analysis to capture the transcriptome landscape of a single cell, Nat Protoc, 5 516-535.

[38] Picelli, S., A. K. Bjorklund, O. R. Faridani, S. Sagasser, G. Winberg and R. Sandberg (2013). "Smart-seq2 for sensitive full-length transcriptome profiling in single cells." Nat Methods.24056875
[39] Sandberg, R. (2014). "Entering the era of single-cell transcriptomics in biology and medicine." Nat Methods **11**(1): 22-24.
[40} Wu, A. R., N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke and S. R. Quake (2014). "Quantitative assessment of single-cell RNA-sequencing methods." Nat Methods **11**(1): 41-46.
[41] X. Adiconis, D. Borges-Rivera, R. Satija, D.S. Deluca, M.A. Busby, A.M. Berlin, A. Sivachenko, D.A. Thompson, A. Wysoker, T. Fennell, A. Gnirke, N. Pochet, A. Regev, J.Z. Levin, Comparative analysis of RNA sequencing methods for degraded or low-input samples, Nature methods, (2013).
[42] P.A. McGettigan, Transcriptomics in the RNA-seq era, Current opinion in chemical biology, 17 (2013) 4-11.
[43] A.J. Hamilton, D.C. Baulcombe, A species of small antisense RNA in posttranscriptional gene silencing in plants, Science, 286 (1999) 950-952.
[44] S.M. Hammond, E. Bernstein, D. Beach, G.J. Hannon, An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells, Nature, 404 (2000) 293-296.
[45] 5M. Lagos-Quintana, R. Rauhut, W. Lendeckel, T. Tuschl, Identification of novel genes coding for small expressed RNAs, Science, 294 (2001) 853-858.
[46] N.C. Lau, L.P. Lim, E.G. Weinstein, D.P. Bartel, An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans, Science, 294 (2001) 858-862.
[47] L. Jones, Revealing micro-RNAs in plants, Trends Plant Sci, 7 (2002) 473-475.
[48] K.C. Pang, M.C. Frith, J.S. Mattick, Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function, Trends Genet, 22 (2006) 1-5.
[49] V.N. Kim, J.W. Nam, Genomics of microRNA, Trends Genet, 22 (2006) 165-173.
[50] R.W. Carthew, Gene regulation by microRNAs, Curr Opin Genet Dev, 16 (2006) 203-208.
[51] A. Rodriguez, E. Vigorito, S. Clare, M.V. Warren, P. Couttet, D.R. Soond, S. van Dongen, R.J. Grocock, P.P. Das, E.A. Miska, D. Vetrie, K. Okkenhaug, A.J. Enright, G. Dougan, M. Turner, A. Bradley, Requirement of bic/microRNA-155 for normal immune function, Science, 316 (2007) 608-611.
[52] R.W. Carthew, E.J. Sontheimer, Origins and Mechanisms of miRNAs and siRNAs, Cell, 136 (2009) 642-655.
[53] A.G. Fraser, R.S. Kamath, P. Zipperlen, M. Martinez-Campos, M. Sohrmann, J. Ahringer, Functional genomic analysis of C. elegans chromosome I by systematic RNA interference, Nature, 408 (2000) 325-330.
[54] R. Barstead, Genome-wide RNAi, Curr Opin Chem Biol, 5 (2001) 63-66.
[55] I.A. Hope, RNAi surges on: application to cultured mammalian cells, Trends Genet, 17 (2001) 440.
[56] A. Wojtkowiak, A. Siek, M. Alejska, A. Jarmolowski, Z. Szweykowska-Kulinska, M. Figlerowicz, RNAi And Viral Vectors As Useful Tools In The Functional Genomics Of Plants. Construction Of BMV-Based Vectors For RNA Delivery Into Plant Cells, Cell Mol Biol Lett, 7 (2002) 511-522.

[57] D.J. Shuey, D.E. McCallus, T. Giordano, RNAi: gene-silencing in therapeutic intervention, Drug Discov Today, 7 (2002) 1040-1046.
[58] P.J. Paddison, G.J. Hannon, RNA interference: the new somatic cell genetics?, Cancer Cell, 2 (2002) 17-23.
[59] M.A. Martinez, B. Clotet, J.A. Este, RNA interference of HIV replication, Trends Immunol, 23 (2002) 559-561.
[60] P.J. Paddison, A.A. Caudy, R. Sachidanandam, G.J. Hannon, Short hairpin activated gene silencing in Mammalian cells, Methods Mol Biol, 265 (2004) 85-100.
[61] K. Ashrafi, F.Y. Chang, J.L. Watts, A.G. Fraser, R.S. Kamath, J. Ahringer, G. Ruvkun, Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes, Nature, 421 (2003) 268-272.
[62] M. Tewari, P.J. Hu, J.S. Ahn, N. Ayivi-Guedehoussou, P.O. Vidalain, S. Li, S. Milstein, C.M. Armstrong, M. Boxem, M.D. Butler, S. Busiguina, J.F. Rual, N. Ibarrola, S.T. Chaklos, N. Bertin, P. Vaglio, M.L. Edgley, K.V. King, P.S. Albert, J. Vandenhaute, A. Pandey, D.L. Riddle, G. Ruvkun, M. Vidal, Systematic interactome mapping and genetic perturbation analysis of a C. elegans TGF-beta signaling network, Mol Cell, 13 (2004) 469-482.
[63] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, R.A. Young, Transcriptional regulatory code of a eukaryotic genome, Nature, 431 (2004) 99-104.
[64]M. Kellis, The changing face of genomics, Genome Biol, 5 (2004) 324.
[65] P. Khatri, S. Draghici, Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinformatics, 21 (2005) 3587-3595.
[66] P. Khatri, S. Sellamuthu, P. Malhotra, K. Amin, A. Done, S. Draghici, Recent additions and improvements to the Onto-Tools, Nucleic Acids Res, 33 (2005) W762-765.
[67] M. Vidal, P. Legrain, Yeast forward and reverse 'n'-hybrid systems, Nucleic Acids Res, 27 (1999) 919-929.
[68] S. Li, C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, M. Vidal, A map of the interactome network of the metazoan C. elegans, Science, 303 (2004) 540-543.
[69] B.T. Kile, K.E. Hentges, A.T. Clark, H. Nakamura, A.P. Salinger, B. Liu, N. Box, D.W. Stockton, R.L. Johnson, R.R. Behringer, A. Bradley, M.J. Justice, Functional genetic analysis of mouse chromosome 11, Nature, 425 (2003) 81-86.
[70] [Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, Grody WW, Hegde MR, Hoeltge GA, Leonard DG, Merker JD, Nagarajan R, Palicki LA, Robetorye RS, Schrijver I, Weck KE, Voelkerding KV. 2015. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch Pathol Lab Med* **139**(4): 481-493.]
[71] [Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, Baying B , Benes V, Teichmann SA, Marioni JC, Heisler MG. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10**(11): 1093-1095
[72] Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann

SA, Marioni JC, Stegle O. 2015. Computational analysis of cell-to-cell heterogeneity in single -cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* **33**(2): 155-160.]

6. Expression profiling, network analysis and pathway modelling

Affymetrix Expression Profiling Practical

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves should be worn during all practical sessions

Gene Chip WT PLUS Reagent Kit – Cat# 902281

20x hybridisation control buffer-harmful
UDG – Irritant/Corrosive
APE1 - Irritant/Corrosive
5 x TdT – Irritant/Toxic/Carcinogenic – not to be handled by new or expectant mothers
DNA labelling reagent - Irritant
TdT - Irritant/Corrosive
10 x CDNA labelling reagent – Irritant/Corrosive
Second strand enzyme - Irritant/Corrosive
IVT enzyme - Irritant/Corrosive
2nd cycle ss-cDNA Buffer - Irritant/Corrosive
2nd cycle ss-cDNA Enzyme - Irritant/Corrosive
Rnase H - Irritant/Corrosive

GeneChip Hybridisation, Wash and Stain Kit Cat# 900720

Prehybridisation Mix - Irritant
2 x Hybridisation Mix - Irritant
DMSO – Harmful/Irritant
Stain cocktail 1 - Irritant
Stain Cocktail 2 - Irritant
Array Holding Buffer - Irritant
Wash Buffer A - Irritant
Wash Buffer B – Irritant

Ethanol – Highly flammable
Isopropanol – Highly flammable/Irritant

6. Affymetrix Expression Profiling

Introduction

The potential to correlate the genetic makeup of an organism to its biological function has moved into a new era. This change has primarily been driven by the completion of whole genome sequences and by the acquisition of catalogues of all the encoded genes. The gene sequences for the genomes of man, mouse and rat are already known and publicly available and many others are either finished or in progress. For the first time therefore, a near-complete gene catalogue is available for a number of important model organisms, providing an unrivalled opportunity to begin to understand the processes of life at the global genetic level. One approach to identifying and characterising which of the newly identified genes may be relevant to a given biological system or disease is by analysing the expression pattern of the genes, so called expression profiling.

There are now a number of different commercial technology platforms for performing microarray expression the main ones including Illumina, Agilent and Nimblegen. This section details the use of one of the original expression profiling platforms the Affymetrix GeneChip system. The Affymetrix platform now provides a means to perform parallel analyses on thousands of transcripts in a single assay, indeed the every single protein coding transcript encoded in the genome and furthermore every exon of every transcript. The results provide a semi-quantitative assessment of whether the expression of the genes included on the microarrays have been up or down regulated, or remained unchanged. As such, the system provides a powerful approach with which to investigate biological specimens to screen for an alteration in transcription, which may regulate or accompany biochemical or physiological change. The particular chips to be run during the course will be the Mouse 1.1 ST Arrays which have been designed for the new Titan system which can process 96 arrays in a single run. This platform has only recently become available and is designed as a low cost platform capable of analysing every gene in the mammalian genome. The Affymetrix® Human Gene 1.1 ST Array Plates are designed for medium and high-throughput expression profiling on the GeneTitan™ or GeneAtlas Instruments, enabling researchers to process multiple arrays in parallel. Each peg array strip of plate consists of 4, 16, 24 or 96 microarrays and contains the same probe sets as the GeneChip® Human Gene 1.0 ST Cartridge Array. Each array is comprised of more than 770,000 unique 25-mer oligonucleotide probes that interrogate over 28,000 genes. Discovery content, such as transcript regions supported by more speculative sources including Expressed Sequence Tags (ESTs) and gene predictions, are not interrogated by the Affymetrix Mouse Gene 1.1 ST Array Plates.

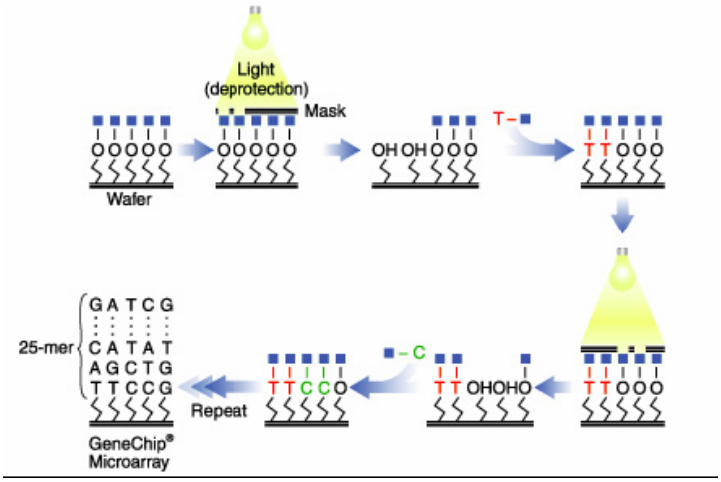
During the course you will take a number of RNA samples through the QC and labelling steps prior to hybridisation to the chips. It is hoped that by so doing you will be able to gets hands on experience of performing this procedure and become familiar with laboratory issues that can affect data quality. We will also compare and contrast the Affymetrix arrays and with other microarray platforms, discussing some of the pros and cons of the different microarray platforms. The practical work will be supplemented by a brief introductory course on microarray expression analysis where some of the basic issues with data quality and the concepts behind data analysis will be explored. We will then go on to explore the data generated on the course using network-based analysis techniques.

The Affymetrix Genechip System

Affymetrix first started developing their technology platform at about the same time as work began on spotted arrays i.e. the mid-90's. However, until about 2000 it was rendered inaccessible to the majority of research groups because of its price. Whilst still relatively expensive, around £100-£300 for each GeneChip, the technology is now within the price range of many academic groups. The platform has the advantage that it requires very little set up time, it being an essentially an 'off-the-shelf' technology and is capable of producing high quality expression data for a wide range of genes and species. It also has the advantage of having a stable format that has been widely accepted and many software tools are available for data analysis. Older GeneChip formats made it possible to analyse all human genes on a single chip and now as already mentioned, it is possible to monitor the expression of every known exon. On the down side, there is no flexibility of what is on a chip and the expense still has the effect of limiting the number of samples that one can realistically process. With improved production facilities and increasing competition, however, prices are coming down and many are now choosing to try out this impressive technology to see what impact it can have on their own work.

GeneChip probe arrays are manufactured using technology that combines photolithography and combinatorial chemistry. Up to 6.5 million different oligonucleotide probes are synthesized on each array. Each oligonucleotide is located in a specific area on the array called a probe cell. Each probe cell contains hundreds of thousands to millions of copies of a given oligonucleotide.

Probe arrays are manufactured in a series of cycles. Initially, a glass substrate is coated with linkers containing photolabile protecting groups. Then, a mask is applied that exposes selected portions of the probe array to ultraviolet light. Illumination removes the photolabile protecting groups enabling selective nucleoside phosphoramidite addition only at the previously exposed sites. Next, a different mask is applied and the cycle of illumination and chemical coupling is performed again. By repeating this cycle, a specific set of oligonucleotide probes is synthesized with each probe type in a known location. The completed probe arrays are packaged into cartridges.



The synthesis of these oligonucleotides on GeneChip microarrays are based on the concept of photolithography:

1. Light is shined through a mask onto a chip that has initial starting strands where the DNA will be built from
2. The mask has specific tiny openings that allow the light to come in contact with the wafer at specific sections (in this diagram there are 5 probes only and each could represent a different feature)
3. Any place where light hits, removes a "protective" group from the strands
4. Free nucleotides (the red T) are washed over the chip and the nucleotides will combine with any strand that had lost its' protective group in the previous step
5. This is then repeated (shine light through a mask, deprotect the strands, add free nucleotides) numerous times until a each strand built is 25 base pairs long

Affymetrix cartridge array (left) and array strip (right)

During the laboratory procedure described in this manual, biotin-labeled RNA or DNA fragments referred to as the "target" are hybridized to the probe array. The hybridized probe array is stained with streptavidin phycoerythrin conjugate and scanned by a GeneChip Scanner. The amount of light emitted at 570 nm is proportional to the bound target at each location on the probe array.

The protocols below refer to the extraction and labelling of eukaryotic RNAs where the amount of total RNA available is >100 ng, the so called small sample protocol. For prokaryotic protocols and protocols using other amounts of RNA please refer to the Affymetrix manual or look elsewhere as a number of different protocols are available.

1.1 RNA Extraction

Isolation of Total RNA – Trizol method

Other RNA extraction methods may be used but this is the one we use routinely in my lab prior to microarray analysis. ***This method is provided only for reference, RNA will be supplied.***

1. Store tissue/cells on dry ice at 70^{o}C until required.
2. Homogenise tissue using Ultra-Turrax tissue homogeniser, adding approximately 1 ml Trizol to 50-100 mg tissues.
3. Incubate at 15-30^{o}C for 5 min.
4. Add 0.2 ml chloroform per 1 ml Trizol and shake for 15 sec.
5. Incubate at 15-30^{o}C for 2-3 min.
6. Centrifuge at <4000 rpm, 2-8^{o}C for 15 min.
7. Transfer the colourless upper aqueous phase (containing RNA) to a fresh screw cap polypropylene centrifuge tube OR an appropriate number of eppendorf tubes.
8. Add 0.5 ml isopropanol per 1 ml Trizol in initial homogenisation step.
9. Incubate at 15-30^{o}C for 10 min.

10. Centrifuge at <4000 rpm, 2-8^{o}C for 10 min, forming a gelatinous pellet.
11. Remove and discard supernatant.
12. Wash the RNA pellet in 1 ml of 75% (v/v) ethanol per 1 ml of Trizol in original prep.
13. Vortex to resuspend pellet, and centrifuge at <4000 rpm, 2-8^{o}C for 5 min.
14. Allow the RNA pellet to dry (air/vacuum dry) briefly, and resuspend in an appropriate volume of DEPC-treated MilliQ or 0.5% SDS.
15. Incubate at 55-60^{o}C for 10 min to ensure total resuspension.
16. Quantify and assess purity of RNA: (conc. = OD_{260nm} x dilution x 40) in ng/µl

i. OD_{260nm} = 1.0 ≅ 40 ng/µl for pure RNA.

ii. $OD_{260nm}/_{280nm}$ = 1.8 - 2.0 for "good" RNA. RNA with an $OD_{260nm}/_{280nm}$ between 1.6 – 1.8 may still be acceptable.

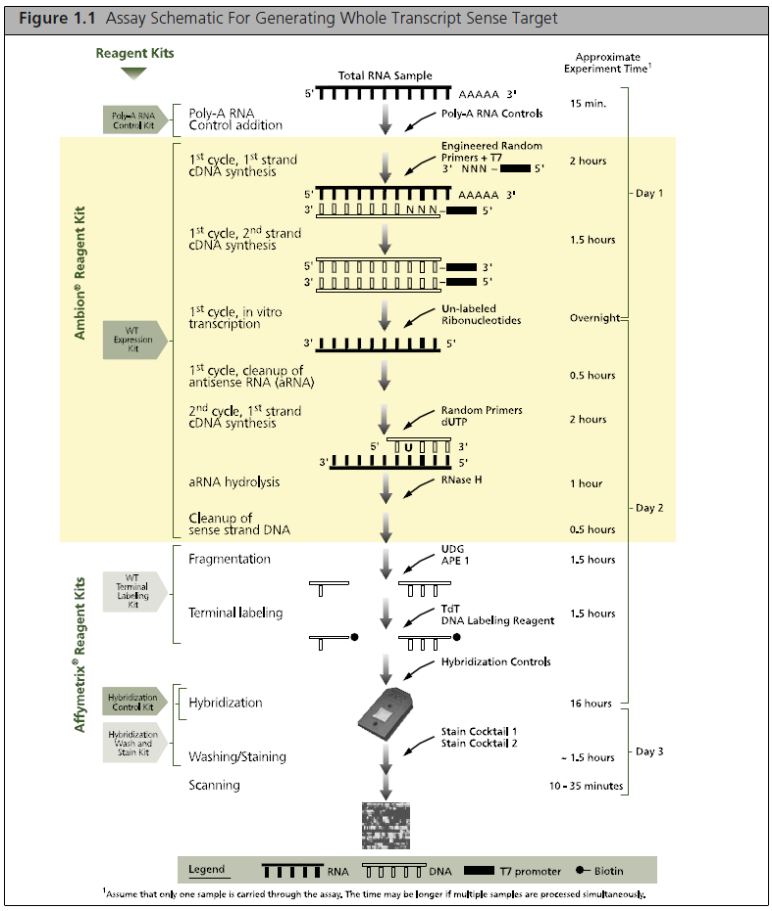
iii. Using non-denaturing agarose gel electrophoresis or an Agilent Bioanalyser you should observe a smear of RNA across 0.5 - >5 kb with "strong" peaks corresponding to the 18S (≅2 kb) and 28S (≅5 kb) ribosomal RNA.

Note

High-quality total RNA can also be successfully isolated from mammalian **cells** (such as cultured cells and lymphocytes) using the RNeasy Mini Kit from QIAGEN.

The Affymetrix WT Plus Reagent Kit

Schematic overview



Health and Safety

For safety and biohazard guidelines, refer to the "Safety" appendix in the *The Ambion® WT Expression Kit Protocol* (PN 4425209). For all chemicals in **bold red** type, read the MSDS and follow the instructions. Wear appropriate protective eyewear, clothing, and gloves.

Before you begin

- Prepare your total RNA according to your laboratory's procedure.
- Determine your input RNA quantity.
- Prepare the Poly-A RNA Controls.
- Evaluate RNA quality by determining its $A_{260/A280}$ ratio. RNA of acceptable quality is in the range 1.7 to 2.1.
- Evaluate RNA integrity by microfluidic analysis or denaturing agarose gel electrophoresis.
- Program your thermal cycler.

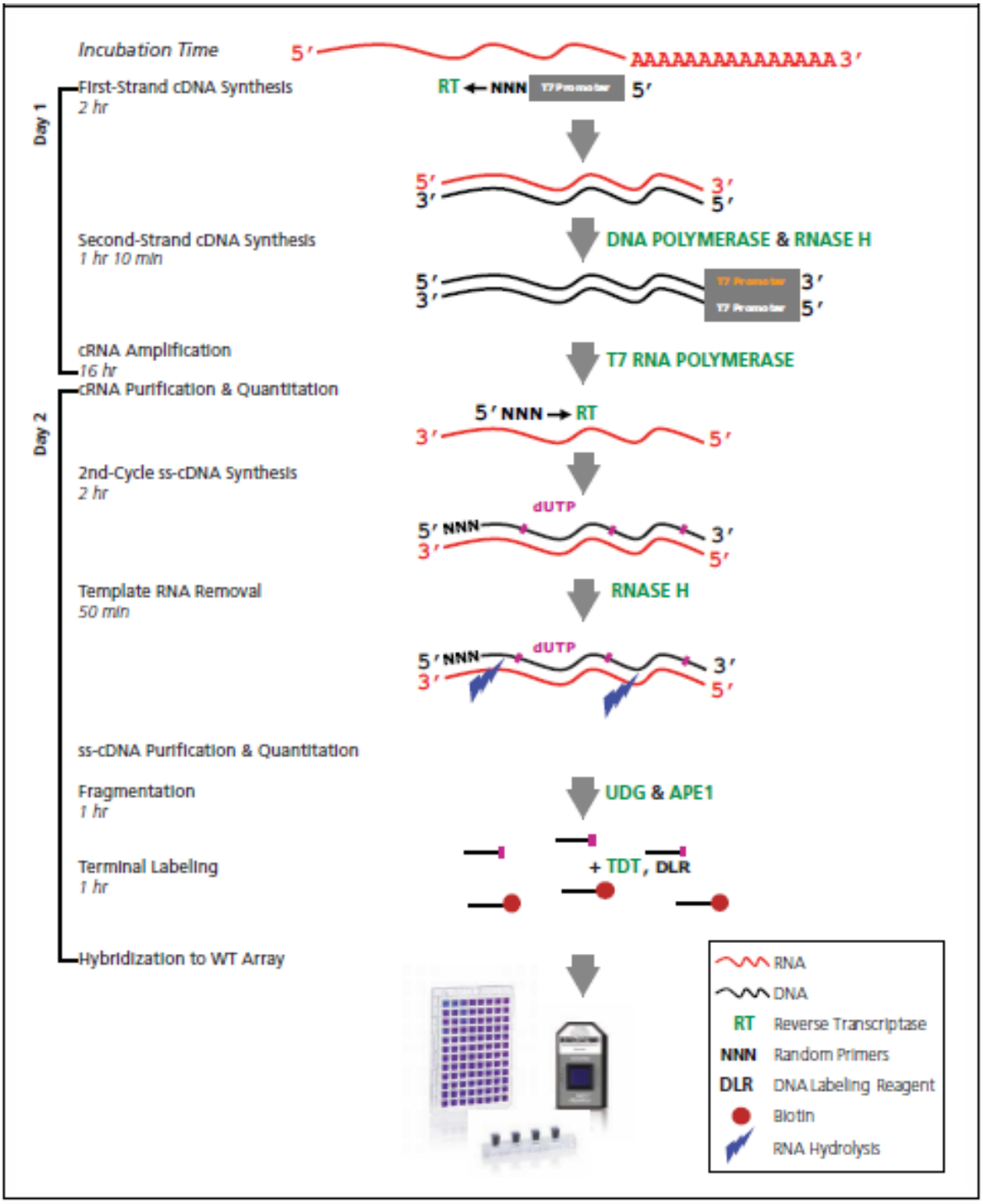
‡ For MJ Research/BioRad thermal cyclers, engage the heated lid.

Thermal Cycler Programs

Program	Heated Lid Temp	Alternate Protocol*	Step 1	Step 2	Step 3	Step 4	Volume
First-Strand cDNA Synthesis	42°C	105°C	25°C, 60 min	42°C, 60 min	4°C, 2 min		10 µL
Second-Strand cDNA Synthesis	RT or disable	Lid open	16°C, 60 min	65°C, 10 min	4°C, 2 min		30 µL
In Vitro Transcription cRNA Synthesis	40°C	40°C oven	40°C, 16 hr	4°C, hold			60 µL
2nd-Cycle Primers-cRNA Annealing	70°C	105°C	70°C, 5 min	25°C, 5 min	4°C, 2 min		28 µL
2nd-Cycle ss-cDNA Synthesis	70°C	105°C	25°C, 10 min	42°C, 90 min	70°C, 10 min	4°C, hold	40 µL
RNA Hydrolysis	70°C	105°C	37°C, 45min	95°C, 5 min	4°C, hold		44 µL
Fragmentation	93°C	105°C	37°C, 60 min	93°C, 2 min	4°C, hold		48 µL
Labeling	70°C	105°C	37°C, 60 min	70°C, 10 min	4°C, hold		60 µL
Hybridization Control	65°C	105°C	65°C, 5 min				Variable
Hybridization Cocktail	99°C	105°C	95°C or 99°C, 5 min	45°C, 5 min			Variable

**For thermal cyclers that lack a programmable heated lid.*

Assay Workflow **WT Plus Amplification and Labeling Process**



DAY 1

1. Synthesize First-Strand cDNA

In this reverse transcription procedure, total RNA is primed with primers containing a T7 promoter sequence. The reaction synthesizes single-stranded cDNA with T7 promoter sequence at the 5' end.

1. Prepare First-Strand Master Mix.

A. On ice, prepare the First-Strand Master Mix in a nuclease-free tube. Combine the components in the sequence shown in the table below. Prepare the master mix for all the total RNA samples in the experiment. Include ~5% excess volume to correct for pipetting losses.

First Strand Master Mix

Component	Volume for One Reaction
First-Strand Buffer	4 µL
First-Strand Enzyme	1 µL
Total Volume	**5 µL**

B. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the mix at the bottom of the tube. Proceed immediately to the next step.

C. On ice, transfer 5 µL of the First-Strand Master Mix to each tube or well.

2. Add total RNA to each First-Strand Master Mix aliquot.

A. On ice, add 5 µL of the total RNA to each (5 µL) tube or well containing the First-Strand Master Mix for a final reaction volume of 10 µL.

B. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the reaction at the bottom of the tube or well, then proceed immediately to the next step.

3. Incubate for 1 hr at 25°C, then for 1 hr at 42°C, then for at least 2 min at 4°C.

A. Incubate the first-strand synthesis reaction in a thermal cycler using the First-Strand cDNA Synthesis program.

B. Immediately after the incubation, centrifuge briefly to collect the first-strand cDNA at the bottom of the tube or well.

C. Place the sample on ice for 2 min to cool the plastic, then proceed immediately to *Synthesize Second-Strand cDNA*

IMPORTANT: Transferring Second-Strand Master Mix to hot plastics may significantly reduce cRNA yields. Holding the First-Strand cDNA Synthesis reaction at 4°C for longer than 10 min may significantly reduce cRNA yields.

TIP: When there is approximately 15 min left on the thermal cycler you may start reagent preparation for Second-Strand cDNA Synthesis

2. Synthesize Second-Strand cDNA

In this procedure, single-stranded cDNA is converted to double-stranded cDNA, which acts as a template for *in vitro* transcription. The reaction uses DNA polymerase and RNase H to simultaneously degrade the RNA and synthesize second-strand cDNA.

IMPORTANT: Pre-cool thermal cycler block to 16°C

1. Prepare Second-Strand Master Mix.

A. On ice, prepare the Second-Strand Master Mix in a nuclease-free tube. Combine the components in the sequence shown in the table below. Prepare the master mix for all the first-strand cDNA samples in the experiment. Include ~5% excess volume to correct for pipetting losses.

Second-Strand Master Mix

Component	Volume for One Reaction
Second-Strand Buffer	18 µL
Second-Strand Enzyme	2 µL
Total Volume	**20 µL**

B. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the mix at the bottom of the tube and proceed immediately to the next step.

C. On ice, transfer 20 µL of the Second-Strand Master Mix to each (10 µL) first-strand cDNA sample for a final reaction volume of 30 µL.

D. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the reaction at the bottom of the tube or well, then proceed immediately to the next step.

2. Incubate for 1 hr at 16°C, then for 10 min at 65°C, then for at least 2 min at 4°C.

A. Incubate the second-strand synthesis reaction in a thermal cycler using the Second-Strand cDNA Synthesis program.

IMPORTANT: Disable the heated lid of the thermal cycler or keep the lid off during the Second-Strand cDNA Synthesis

B. Immediately after the incubation, centrifuge briefly to collect the second-strand cDNA at the bottom of the tube or well.

C. Place the sample on ice, then proceed immediately to *Synthesize cRNA by In Vitro Transcription*

TIP: When there is approximately 15 min left on the thermal cycler you may start reagent preparation for In Vitro Transcription.

3. Synthesize cRNA by In Vitro Transcription

In this procedure, antisense RNA (complimentary RNA or cRNA) is synthesized and amplified by *in vitro* transcription (IVT) of the second-stranded cDNA template using T7 RNA polymerase. This method of RNA sample preparation is based on the original T7 *in vitro* transcription technology known as the Eberwine or RT-IVT method (Van Gelder *et al*., 1990).

IMPORTANT:

Transfer the second-strand cDNA samples to room temperature for ≥ 5 min while preparing IVT Master Mix.

After the IVT Buffer is thawed completely, leave the IVT Buffer at room temperature for ≥ 10 min before preparing the IVT Master Mix.

1. Prepare IVT Master Mix.

NOTE: **This step is performed at room temperature**

A. At room temperature, prepare the IVT Master Mix in a nuclease-free tube. Combine the components in the sequence shown in the table below. Prepare the master mix for all the second- strand cDNA samples in the experiment. Include ~5% excess volume to correct for pipetting losses.

IVT Master Mix

Component	Volume for One Reaction
IVT Buffer	24 µL
IVT Enzyme	6 µL
Total Volume	**30 µL**

B. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the mix at the bottom of the tube, then proceed immediately to the next step.

C. At room temperature, transfer 30 µL of the IVT Master Mix to each (30 µL) second-strand cDNA sample for a final reaction volume of 60 µL.

D. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the reaction at the bottom of the tube or well, then proceed immediately to the next step.

2. Incubate for 16 hr at 40°C, then at 4°C.

A. Incubate the IVT reaction in a thermal cycler using the In Vitro Transcription cRNA Synthesis program.

B. After the incubation, centrifuge briefly to collect the cRNA at the bottom of the tube or well.

C. Place the reaction on ice, then proceed to *Purify cRNA*, or immediately freeze the samples at –20°C for storage.

TIP: STOPPING POINT. The cRNA samples can be stored overnight at –20°C

Day 2

1) Purify cRNA

In this procedure, enzymes, salts, inorganic phosphates, and unincorporated nucleotides are removed to prepare the cRNA for 2nd-cycle single-stranded cDNA synthesis.

Beginning the cRNA Purification

IMPORTANT:

Preheat the Nuclease-free Water in a heat block or thermal cycler to 65°C for at least 10 min.

Mix the Purification Beads thoroughly by vortexing before use to ensure that they are fully dispersed. Transfer the appropriate amount of Purification Beads to a nuclease-free tube or container, and allow the Purification Beads to equilibrate at room temperature. For each reaction, 100 µL plus ~10% overage will be needed.

Prepare fresh dilutions of 80% ethanol wash solution each time from 100% ethanol (Molecular Biology Grade or equivalent) and Nuclease-free Water in a nuclease-free tube or container. For each reaction, 600 µL plus ~10% overage will be needed.

Transfer the cRNA sample to room temperature while preparing the Purification Beads.

NOTE:

Occasionally, the bead/sample mixture may be brownish in color and not completely clear when placed on magnet. In those situations, switch to a different position of magnet on the magnetic stand, a new magnetic stand, or spin out pellets.

This entire procedure is performed at room temperature.

1. Bind cRNA to Purification Beads.

A. Mix the Purification Beads container by vortexing to resuspend the magnetic particles that may have settled.

B. Add 100 µL of the Purification Beads to each (60 µL) cRNA sample, mix by pipetting up and down, and transfer to a well of a U-bottom plate.

TIP:

Any unused wells should be covered with a plate sealer so that the plate can safely be reused.

Use multichannel pipette when processing multiple samples.

C. Mix well by pipetting up and down 10 times.

D. Incubate for 10 min. The cRNA in the sample binds to the Purification Beads during this incubation.

E. Move the plate to a magnetic stand to capture the Purification Beads. When capture is complete (after ~5 min), the mixture is transparent, and the Purification Beads form pellets against the magnets in the magnetic stand. The exact capture time depends on the magnetic stand that you use, and the amount of cRNA generated by *in vitro* transcription.

F. Carefully aspirate and discard the supernatant without disturbing the Purification Beads. Keep the plate on the magnetic stand.

2. Wash the Purification Beads.

A. While on the magnetic stand, add 200 µL of 80% ethanol wash solution to each well and incubate for 30 sec.

B. Slowly aspirate and discard the 80% ethanol wash solution without disturbing the Purification Beads.

C. Repeat Step A and Step B twice for a total of 3 washes with 200 µL of 80% ethanol wash solution.
Completely remove the final wash solution.

D. Air-dry on the magnetic stand for 5 min until no liquid is visible, yet the pellet appears shiny. Additional time may be required. Do not over-dry the beads as this will reduce the elution efficiency. The bead surface will appear dull, and may have surface cracks when it is over-dry.

3. Elute cRNA.

A. Remove the plate from the magnetic stand. Add to each sample 27 µL of the preheated (65°C) Nuclease-free Water and incubate for 1 min.

B. Mix well by pipetting up and down 10 times.

C. Move the plate to the magnetic stand for ~ 5 min to capture the Purification Beads.

D. Transfer the supernatant, which contains the eluted cRNA, to a nuclease-free tube.

E. Place the purified cRNA samples on ice, then proceed to *Assess cRNA Yield and Size Distribution*, or immediately freeze the samples at –20°C for storage.

NOTE:

□ Minimal bead carryover will not inhibit subsequent enzymatic reactions.

□ It may be difficult to resuspend magnetic particles and aspirate purified cRNA when the cRNA is very concentrated. To elute the sample with high concentration cRNA, add an additional 10-30 µL of the preheated Nuclease-free Water to the well, incubate for
1 min, and proceed to Step 3B.

TIP: STOPPING POINT. The purified cRNA samples can be stored overnight at –20°C. For long-term storage, store samples at –80°C and keep the number of freeze-thaw cycles to 3 or less to ensure cRNA integrity.

2) Assess cRNA Yield and Size Distribution

Expected cRNA Yield

The cRNA yield depends on the amount and quality of non-rRNA in the input total RNA. Because the proportion of non-rRNA in total RNA is affected by factors such as the health of the organism and the organ from which it is isolated, cRNA yield from equal amounts of total RNA may vary considerably.

During development of this kit, using a wide variety of tissue types, 50 ng of input total RNA yielded 15 to 40 µg of cRNA. For most tissue types, the recommended 100 ng of input total RNA should provide
>20 µg of cRNA.

Determine cRNA Yield by UV Absorbance

Determine the concentration of a cRNA solution by measuring its absorbance at 260 nm. Use Nuclease- free Water as blank. Affymetrix recommends using NanoDrop Spectrophotometers for convenience. No dilutions or cuvettes are needed; just use 1.5 µL of the cRNA sample directly. Samples with cRNA concentrations greater than 3,000 ng/uL should be diluted with Nuclease-free Water before measurement and reaction setup. Use the diluted cRNA as the input to prepare 15 µg cRNA in 2nd cycle cDNA synthesis reaction.

Alternatively, determine the cRNA concentration by diluting an aliquot of the preparation in Nuclease- free Water and reading the absorbance in a traditional spectrophotometer at 260 nm. Calculate the concentration in µg/mL using the equation shown below (1 A_{260} = 40 µg RNA/mL).

A_{260} × dilution factor × 40 = µg RNA/mL

(Optional) Expected cRNA Size Distribution

The expected cRNA profile is a distribution of sizes from 50 to 4500 nt with most of the cRNA sizes in the 200 to 2000 nt range. The distribution is quite jagged and does not resemble the profile observed when using a traditional dT-based amplification kit such as 3' IVT Express kit. This step is optional.

Determine cRNA size distribution using a Bioanalyzer.

Affymetrix recommends analyzing cRNA size distribution using an Agilent 2100 Bioanalyzer, a RNA 6000 Nano Kit (PN5067-1511), and mRNA Nano Series II assay. If there is sufficient yield, then load approximately 300 ng of cRNA per well on the Bioanalyzer. If there is insufficient yield, then use as little as 200 ng of cRNA per well. To analyze cRNA size using a Bioanalyzer, follow the manufacturer's instructions.

TIP: STOPPING POINT. The purified cRNA samples can be stored overnight at –20°C.

3) Synthesize 2nd-Cycle Single-Stranded cDNA

In this procedure, sense-strand cDNA is synthesized by the reverse transcription of cRNA using 2nd- Cycle Primers. The sense-strand cDNA contains dUTP at a fixed ratio relative to dTTP. 15 µg of cRNA is required for 2nd-cycle single-stranded cDNA synthesis.

1. Prepare 15 µg of cRNA.

   On ice, prepare 625 ng/µL cRNA. This is equal to 15 µg cRNA in a volume of 24 µL. If necessary, use Nuclease-free Water to bring the cRNA sample to 24 µL.

   NOTE: High-concentration cRNA samples (> 3000 ng/µL) should be diluted with Nuclease-free Water before measurement and reaction setup. Use the diluted cRNA as the input to prepare 15 µg of cRNA.

2. Prepare cRNA and 2nd-Cycle Primers Mix.

   A. On ice, combine:

   □ 24 µL of cRNA (15 µg)

   □ 4 µL of 2nd-Cycle Primers

   B. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the mix at the bottom of the tube, then proceed immediately to the next step.

3. Incubate for 5 min at 70°C, then 5 min at 25°C, then 2 min at 4°C.

   A. Incubate the cRNA/Primers mix in a thermal cycler using the 2nd-Cycle Primers-cRNA Annealing program.

   B. Immediately after the incubation, centrifuge briefly to collect the cRNA/Primers mix at the bottom of the tube or well.

   C. Place the mix on ice, then proceed immediately to the next step.

4. Prepare 2nd-Cycle ss-cDNA Master Mix.

   A. On ice, prepare the 2nd-Cycle ss-cDNA Master Mix in a nuclease-free tube. Combine the components in the sequence shown in the table below. Prepare the master mix for all the cRNA/ Primers samples in the experiment. Include ~5% excess volume to correct for pipetting losses.

2nd-Cycle ss-cDNA Master Mix

Component	Volume for One Reaction
2nd-Cycle ss-cDNA Buffer	8 µL
2nd-Cycle ss-cDNA Enzyme	4 µL

Total Volume	12 µL

B. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the mix at the bottom of the tube and proceed immediately to the next step.

C. On ice, transfer 12 µL of the 2nd-Cycle ss-cDNA Master Mix to each (28 µL) cRNA/2nd-Cycle Primers sample for a final reaction volume of 40 µL.

D. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the reaction at the bottom of the tube or well, then proceed immediately to the next step.

5. Incubate for 10 min at 25°C, then 90 min at 42°C, then 10 min at 70°C, then for at least 2 min at 4°C.

A. Incubate the 2nd-cycle synthesis reaction in a thermal cycler using the 2nd-Cycle ss-cDNA Synthesis program.

B. Immediately after the incubation, centrifuge briefly to collect the 2nd-cycle ss-cDNA at the bottom of the tube or well.

C. Place the sample on ice and proceed immediately to *Hydrolyze RNA Using RNase H*.

4) Hydrolyze RNA Using RNase H

In this procedure, RNase H hydrolyzes the cRNA template leaving single-stranded cDNA.

1. Add RNase H to each 2nd-cycle ss-cDNA sample.

A. On ice, add 4 µL of the RNase H to each (40 µL) 2nd-cycle ss-cDNA sample for a final reaction volume of 44 µL.

B. Mix thoroughly by gently vortexing. Centrifuge briefly to collect the reaction at the bottom of the tube or well, then proceed immediately to the next step.

2. Incubate for 45 min at 37°C, then for 5 min at 95°C, then for at least 2 min at 4°C.

A. Incubate the RNA hydrolysis reaction in a thermal cycler using the RNA Hydrolysis program.

B. Immediately after the incubation, centrifuge briefly to collect the hydrolyzed 2nd-cycle ss-cDNA
at the bottom of the tube or well.

C. Place the samples on ice and proceed immediately to the next step.

3. Add Nuclease-free Water to each hydrolyzed 2nd-cycle ss-cDNA sample.

A. On ice, add 11 µL of the Nuclease-free Water to each (44 µL) hydrolyzed 2nd-cycle ss-cDNA sample for a final reaction volume of 55 µL.

B. Mix thoroughly by gently vortexing. Centrifuge briefly to collect the reaction at the bottom of the tube or well.

C. Place the sample on ice, then proceed to *Purify 2nd-Cycle Single-Stranded cDNA*, or immediately freeze the samples at –20°C for storage.

TIP: STOPPING POINT. The hydrolyzed ss-cDNA samples can be stored overnight at –20°C.

5) Purify 2nd-Cycle Single-Stranded cDNA

After hydrolysis, the 2nd-cycle single-stranded cDNA is purified to remove enzymes, salts, and unincorporated dNTPs. This step prepares the cDNA for fragmentation and labeling.

Beginning the Single-Stranded cDNA Purification

IMPORTANT:

□ Preheat the Nuclease-free Water in a heat block or thermal cycler to 65°C for at least 10 min.

□ Mix the Purification Beads thoroughly by vortexing before use to ensure that they are fully dispersed. Transfer the appropriate amount of Purification Beads to a nuclease-free tube or container, and allow the Purification Beads to equilibrate at room temperature. For each reaction, 100 µL plus ~10% overage will be needed.

□ Prepare fresh dilutions of 80% ethanol wash solution each time from 100% ethanol (Molecular Biology Grade or equivalent) and Nuclease-free Water in a nuclease-free tube or container. For each reaction, 600 µL plus ~10% overage will be needed.

□ Transfer the cDNA sample to room temperature while preparing the Purification Beads.

NOTE:

□ Occasionally, the bead/sample mixture may be brownish in color and not completely clear when placed on magnet. In those situations, switch to a different position of magnet on the magnetic stand, a new magnetic stand, or spin out pellets.

□ This entire procedure is performed at room temperature.

1. Bind ss-cDNA to Purification Beads.

A. Mix the Purification Beads container by vortexing to resuspend the magnetic particles that may have settled.

B. Add 100 µL of Purification Beads to each (55 µL) 2nd-cycle ss-cDNA sample, mix by pipetting up and down, and transfer to a well of a U-bottom plate.

C. Add 150 µL of 100% ethanol to each (155 µL) ss-cDNA/Beads sample. Mix well by pipetting up and down 10 times.

D. Incubate for 20 min. The ss-cDNA in the sample binds to the Purification Beads during this incubation.

E. Move the plate to a magnetic stand to capture the Purification Beads. When capture is complete (after ~5 min), the mixture is transparent, and the Purification Beads form pellets against the magnets in the magnetic stand. The exact capture time depends on the magnetic stand that you use, and the amount of ss-cDNA generated by 2nd-Cycle ss-cDNA Synthesis.

F. Carefully aspirate and discard the supernatant without disturbing the Purification Beads. Keep the plate on the magnetic stand.

2. Wash the Purification Beads.

   A. While on the magnetic stand, add 200 µL of 80% ethanol wash solution to each well and incubate for 30 sec.

   B. Slowly aspirate and discard the 80% ethanol wash solution without disturbing the Purification Beads.

   C. Repeat Step A and Step B twice for a total of 3 washes with 200 µL of 80% ethanol wash solution.
   Completely remove the final wash solution.

   D. Air-dry on the magnetic stand for 5 min until no liquid is visible, yet the pellet appears shiny. Additional time may be required. Do not over-dry the beads as this will reduce the elution efficiency. The bead surface will appear dull, and may have surface cracks when it is over-dry.

3. Elute ss-cDNA.

   A. Remove the plate from the magnetic stand. Add to each sample 30 µL of the preheated (65°C) Nuclease-free Water and incubate for 1 min.

   B. Mix well by pipetting up and down 10 times.

   C. Move the plate to the magnetic stand for ~ 5 min to capture the Purification Beads.

   D. Transfer the supernatant, which contains the eluted ss-cDNA, to a nuclease-free tube.

   E. Place the purified ss-cDNA samples on ice, then proceed to *Assess Single-Stranded cDNA Yield and Size Distribution*, or immediately freeze the samples at –20°C for storage.

NOTE: Minimal bead carryover will not inhibit subsequent enzymatic reactions.

TIP: STOPPING POINT. The purified ss-cDNA samples can be stored overnight at –20°C. For long-term storage at –20°C, we recommend not to proceed to the fragmentation and labeling reaction and store the samples as ss-cDNA.

6) Assess Single-Stranded cDNA Yield and Size Distribution

Expected Single-Stranded cDNA Yield

During development of this kit, using a wide variety of tissue types, 15 µg of input cRNA yielded 5.5 to 15 µg of ss-cDNA. For most tissue types, the recommended 15 µg of input cRNA should yield > 5.5 µg of ss-cDNA.

Determine Single-Stranded DNA Yield by UV Absorbance

Determine the concentration of a ss-cDNA solution by measuring its absorbance at 260 nm. Use Nuclease-free Water as blank. Affymetrix recommends using NanoDrop Spectrophotometers for convenience. No dilutions or cuvettes are needed; just use 1.5 µL of the cDNA sample directly.

Alternatively, determine the ss-cDNA concentration by diluting an aliquot of the preparation in Nuclease-free Water and reading the absorbance in a traditional spectrophotometer at 260 nm. Calculate the concentration in µg/mL using the equation below (1 A_{260} = 33 µg DNA/mL).

A_{260} × dilution factor × 33 = µg DNA/mL

NOTE: The equation above applies only to single-stranded cDNA

(Optional) Expected Single-Stranded cDNA Size Distribution

The expected cDNA profile does not resemble the cRNA profile. The median cDNA size is approximately 400 nt. This step is optional.

Determine Single-Stranded cDNA Size Distribution Using a Bioanalyzer

Affymetrix recommends analyzing cDNA size distribution using an Agilent 2100 Bioanalyzer, a RNA 6000 Nano Kit (PN5067-1511), and mRNA Nano Series II assay. If there is sufficient yield, load approximately 250 ng of cDNA per well. If there is insufficient yield, then use as little as 200 ng of cDNA per well. To analyze cDNA size using a bioanalyzer, follow the manufacturer's instructions.

TIP: STOPPING POINT. The purified ss-cDNA samples can be stored overnight at –20°C. For long-term storage at –20°C, we recommend not to proceed to the fragmentation and labeling reaction and store the samples as ss-cDNA.

7) Fragment and Label Single-Stranded cDNA

In this procedure, the purified, sense-strand cDNA is fragmented by uracil-DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease 1 (APE 1) at the unnatural dUTP residues and breaks the DNA strand. The fragmented cDNA is labeled by terminal deoxynucleotidyl transferase (TdT) using the Affymetrix proprietary DNA Labeling Reagent that is covalently linked to biotin. 5.5 µg of single- stranded cDNA is required for fragmentation and labeling.

1. Prepare 5.5 µg of ss-cDNA.

   On ice, prepare 176 ng/µL ss-cDNA. This is equal to 5.5 µg ss-cDNA in a volume of 31.2 µL. If necessary, use Nuclease-free Water to bring the ss-cDNA sample to 31.2 µL.

2. Prepare Fragmentation Master Mix.

   A. On ice, prepare the Fragmentation Master Mix in a nuclease-free tube. Combine the components in the sequence shown in the table below. Prepare the master mix for all the ss-cDNA samples in the experiment. Include ~5% excess volume to correct for pipetting

losses.

Fragmentation Master Mix

Component	Volume for One Reaction
Nuclease-free Water	10 µL
10X cDNA Fragmentation Buffer	4.8 µL
UDG, 10 U/µL	1 µL
APE 1, 1,000 U/µL	1 µL
Total Volume	**16.8 µL**

B. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the mix at the bottom of the tube, then proceed immediately to the next step.

C. On ice, transfer 16.8 µL of the Fragmentation Master Mix to each (31.2 µL) purified ss-cDNA sample for a final reaction volume of 48 µL.

D. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the reaction at the bottom of the tube or well, then proceed immediately to the next step.

3. Incubate for 1 hr at 37°C, then for 2 min at 93°C, then for at least 2 min at 4°C.

A. Incubate the fragmentation reaction in a thermal cycler using the Fragmentation program.

B. Immediately after the incubation, centrifuge briefly to collect the fragmented ss-cDNA at the bottom of the tube or well.

C. Place the sample on ice, then proceed immediately to the next step.

4. (Optional) The fragmented ss-cDNA sample can be used for size analysis using a Bioanalyzer. Please see the Reagent Kit Guide that comes with the RNA 6000 Nano LabChip Kit for detailed instructions. The range in peak size of the fragmented samples should be approximately 40 to 70 nt.

5. On ice, transfer 45 µL of the fragmented ss-cDNA sample to each tube or well.

6. Prepare Labeling Master Mix.

A. On ice, prepare the Labeling Master Mix in a nuclease-free tube. Combine the components in the sequence shown in the table below. Prepare the master mix for all the fragmented ss-cDNA samples in the experiment. Include ~5% excess volume to correct for pipetting losses.

Labeling Master Mix

Component	Volume for One Reaction

5X TdT Buffer	12 µL
DNA Labeling Reagent, 5 mM	1 µL
TdT, 30 U/µL	2 µL
Total Volume	**15 µL**

B. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the mix at the bottom of the tube, then proceed immediately to the next step.

C. On ice, transfer 15 µL of the Labeling Master Mix to each (45 µL) fragmented ss-cDNA sample for a final reaction volume of 60 µL.

D. Mix thoroughly by gently vortexing the tube. Centrifuge briefly to collect the reaction at the bottom of the tube or well, then proceed immediately to the next step.

7. Incubate for 1 hr at 37°C, then for 10 min at 70°C, then for at least 2 min at 4°C.

A. Incubate the labeling reaction in a thermal cycler using the Labeling program.

B. Immediately after the incubation, centrifuge briefly to collect the fragmented and labeled ss-cDNA at the bottom of the tube or well.

C. Place the sample on ice, then proceed to *WT Array Hybridization*, or immediately freeze the samples at –20°C for storage.

TIP: STOPPING POINT. The fragmented and labeled ss-cDNA samples can be stored overnight at –20°C. For long-term storage at –20°C, we recommend to store the samples as unfragmented and unlabeled ss-cDNA.

2. Network Visualisation and Analysis of Gene Expression Data using BioLayout *Express*^{3D} - Practical

Getting started with Microarray Gene Expression Analysis

Construction of an Input File

The format of an input file for expression data is given (Box 1). In short, the minimum requirement for the program is one column (the first) of unique gene/probe identifiers followed by columns of data derived from individual samples. The unique identifier column is searchable within the program as well as being used to support hyper-linking via a web search or linked to a specific website, and is also used for display purposes on graphs. It is therefore useful if this column contains information that supports these activities. In general we have found that a concatenation of a gene symbol and probe ID provides a label that is both understandable and specific to the measurement. As mentioned above, BioLayout *Express*^{3D} is not restricted in the type of data that can be loaded and will accept unnormalised, normalised, natural scale, ratiometric or log transformed data. However, for most purposes we would recommend the use of normalised, natural scale data where technical variation has been minimised but contrasts between measurement values are maintained. Columns of data should be placed after the unique identifier but should be ordered according to biological groupings. Whilst the order of the samples does not affect the Pearson correlation value and therefore resultant graph, it does affect the ease with which the expression profile of selected genes can be interpreted.

Input files can be also structured to allow the import and display of more information. Columns of annotation may be placed between the unique identifier and the data columns. This annotation may take the form of a categorization in classes into which transcripts can grouped, and these classes can be displayed on the graph. The program assigns a different display color to each class and colors nodes according to the class they are in. Furthermore there are inbuilt tools that allow selected groups of genes to be mined for over- or under-representation of specific classes (see below: Mining Selected Genes for Over-Representation of Classes). Gene annotation may include GO terms, statistical lists, clusters, gene sets, pathway or protein family membership etc. The import of classes with data therefore allows the overlay of information onto the graph and can provide a powerful aid to graph interpretation.

The number of probes on the array (rows) and samples (columns of data) is limited only by the ability to store the information in RAM, but tests suggest that for most configurations of modern computers it should be possible to work with hundreds of genome-wide arrays i.e. with 20-50,000 probes worth of data. In principle there is no need to filter the data prior to loading in BioLayout *Express*^{3D}. Structure in network graphs is made up of groups of genes whose expression is highly correlated and this is generally the most interesting aspect to any dataset. However one might wish to filter data prior to loading so as to remove low intensity data or to remove genes whose expression does not alter over the experiment i.e. ‘flat line’ data. The need to do this depends on the experimental design and the hypothesis or question being addressed. Low intensity data by its nature is noisy and therefore does not tend to correlate highly with other low intensity data when the sample size is large (>20 arrays). Certain normalisation methods, however, particularly quantile-based methods such as RMA and gcRMA, effectively regularise data by nature of their assigning of expression values to normalised rank values. With

small data sets (<20 arrays), especially where the biology is relatively similar across samples, many genes may not be changing in their expression level and their expression profiles are likely to show a high degree of correlation. This can therefore translate into a large network, even a high correlation cut offs, of essentially genes you are not interested in. Removal of this data or the statistical selection of genes that are likely to be of high interest, makes the graphs more manageable and focused on the research question.

Input files are tab delimited and can be assembled in a text editor or spreadsheet software (e.g. Microsoft Excel). In order to recognise them as such they should to be saved with the extension '.expression'. Examples of expression files are given on the BioLayout *Express*^{3D} website (www.biolayout.org) under the datasets section.

Worked Example

Below we go through an example analysis of microarray expression data. The data used is the Genome Novartis Foundation (GNF) mouse tissue atlas[1]. This analysis was performed on a custom designed Affymetrix GeneChip (named GNF1M) that possessed 36,182 probe sets designed to cover every known mouse gene and a number of alternative transcripts of these. Run across these arrays was RNA derived from 61 different 'tissues' covering most major mouse organs and/or sub-divisions thereof. The dataset has been used and cited widely in a range of genomic investigations including the work of the original paper describing this approach to the analysis of gene expression data and the tool Biolayout *Express*^{3D} [2]. The data is derived from 122 arrays and represents a medium to large data set. However, the complexity of expression patterns that are to be found in this data due to cell/tissue specific gene expression represents a significant challenge to analysis and visualisation of the data. Indeed, it was the complexity of the analysis of this and related data that acted as the catalyst for the development of this tool. The data is also available to query from the GNF's excellent BioGPS site.

Data Import:

1. *Download and install BioLayout Express*^{3D} – The application should run on most PC's, Apple Macs and Linux systems. For PCs using Vista 32 bit or XP operating system or Apple Macs running the Leopard systems we recommend downloading and using the installers available on the *BioLayout Express*^{3D} website (http://www.biolayout.org/). Java 1.6 is included in the PC installer, for Apple Macs if not already installed, it will need to be. For Apple Macs running the Tiger operating system which is only compatible with Java 1.5 there is also an installer available, although this will not contain the latest updates and optimisations. A jar file is also available that will run on all platforms. We recommend opening the jar file using a .bat (PC) or .cmd (Mac) file using the script on the website. To install, go to BioLayout *Express*^{3D} website and select *Downloads*. Download the appropriate installation package or .jar file and install package. Open BioLayout *Express*^{3D}.

2. *Download data* – go to BioLayout *Express*^{3D} website and select *Data sets*. Scroll down to data, click on GNF1M Mouse tissue atlas and download. Decompress (Unzip) file.

3. To open file select *File* ->*Open*. The *File Open* dialog will appear, find and select file and click *Open*. Alternatively a file may be opened by double-clicking on the file assuming the file has an extension that is a recognised (associated) file type and the program is installed (as opposed to running as .jar file), or a file may be dropped into the BioLayout *Express*^{3D} window.

4. The *Load Expression Data* dialog will appear (Fig.1). Generally you will not need to change settings within this window and will be able to go straight to OK. However there are number of options that can be changed:

    4.1 *min correlation* box refers to the correlation threshold above which correlations will be saved. A correlation matrix file can be very large if all correlations i.e. -1 to +1 were saved. For example a microarray of 50,000 probe features requires 1.25 billion calculations, therefore only correlations above a certain value are saved, $r = 0.7$ being the default and suitable for most applications.

    4.2 *corr. metric.* For the work described here (and indeed for most work performed by the authors) the Pearson correlation measure has been used to generate the network graphs from expression data. However, in principle it is possible to construct graphs based on any measure that results in a weighted edge between components and we have also implemented the Spearman rank correlation calculation as a selectable alternative to the Pearson correlation.

    4.3 *data columns start*. The program should recognise a file's structure (unique identifier, class/annotation columns and data columns) automatically, colouring them red, green and blue respectively, based on the input format. However this may not always happen especially if the final column of annotation is a numeric. This dialog allows the user to override the automatic selection.

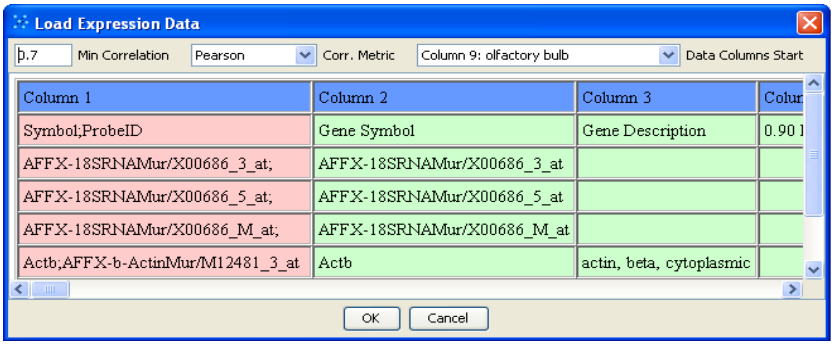


*Figure 1: The **Load Expression Data** dialog*

5. Following loading of the expression data into memory, if this data has not previously been loaded and a correlation matrix file does not already exist (in the same folder) the program will begin calculating a correlation matrix. The number of calculations necessary increases exponentially with the number of rows of the input file. A small file of just several thousand rows of data will therefore be calculated very quickly, this file with >36k probe sets will take a few minutes to perform the ~700 million calculations necessary to construct the matrix. Once calculated however a correlation matrix (e.g. .pearson) file can be used for all future studies (assuming the expression file does not change in name, order or number of probes or samples) and is kept in the same folder.

6. Once the correlation matrix file has been calculated (or an existing one located) the *Expression Graph Settings* dialog will appear. This presents two graphs derived from the data

(Fig. 2). On the left of the dialog box is plotted a graph of the network size vs. correlation threshold for the data. On the *x*-axis is plotted the number of nodes and edges and the *y*-axis the correlation threshold range of the stored values. The two lines of dots represent the number of nodes (pink, lower) and edges (orange, higher) that would be included in the graph across the range of potentially selectable thresholds. The red vertical line denotes the currently selected value (default $r = 0.85$) as determined by the slider at the bottom of the window. The lower the cut-off the larger the graph. A threshold is chosen that balances graph size and complexity. The ideal graph is fully connected as every node has a measured degree of similarity with every other node, defined by correlation. In practice however, such graphs are impractical due to their size (N^2) and because biological relationships of interest normally occur between nodes with very high degrees of correlation. Hence, some form of thresholding is generally desirable. The ideal threshold may be determined empirically by the lay out of different sized graphs but may also determined by the size of graph one is capable of rendering. Depending on your hardware configuration (in particular the graphics card), graphs of up to 20,000 nodes or 3 million edges may be loaded. However, for most configurations we recommend a working limit of half this. Above this size one is more likely to experience issues in rendering. For this exercise we use a setting of $r = 0.95$, resulting in a graph composed of 3,107 nodes 201,877 edges. On the right of the dialog is plotted a graph of the distribution of node degrees (number of edges per node) at the selected threshold. Whilst not of immediate use, it does provide some clues as to the likely graph structure. When you have selected the desired correlation threshold click OK.

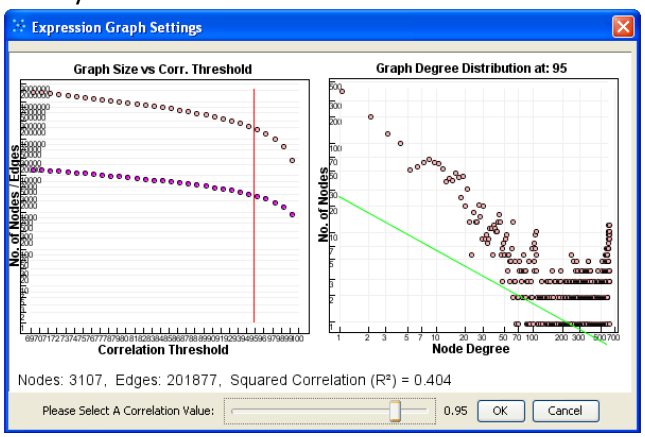


*Figure 2: The **Expression Graph Settings** dialog*

First view of the data

7. *Navigation in the 3D interface – basic controls*. Graphs will first appear in BioLayout *Express*^{3D} window rendered in 3-Dimensional space (Fig. 3). Your default settings (see below) will determine many of the aesthetic properties of the graph but before looking at how these may be changed to suit a user's preferences, we should explore the basic navigation of networks. BioLayout *Express*^{3D} supports the 3D rendering of graphs which provides a fast and intuitive interface to understand the often complex relationships between the entities represented in the graph.

To navigate around the graph:

* *left mouse button* rotates of the current view
* *middle mouse button* allows sideways movement (translation) of the current view

* *right mouse button* for zoom in/out
* holding down *Shift and clicking the left mouse button* on a node makes it the centre of the graph's axis of rotation

In situations where a 3-button mouse is not available, these commands are also available under the 3D menu bar.

Before going on, explore the graph using these commands as they are fundamental to your ability to interface with the data.

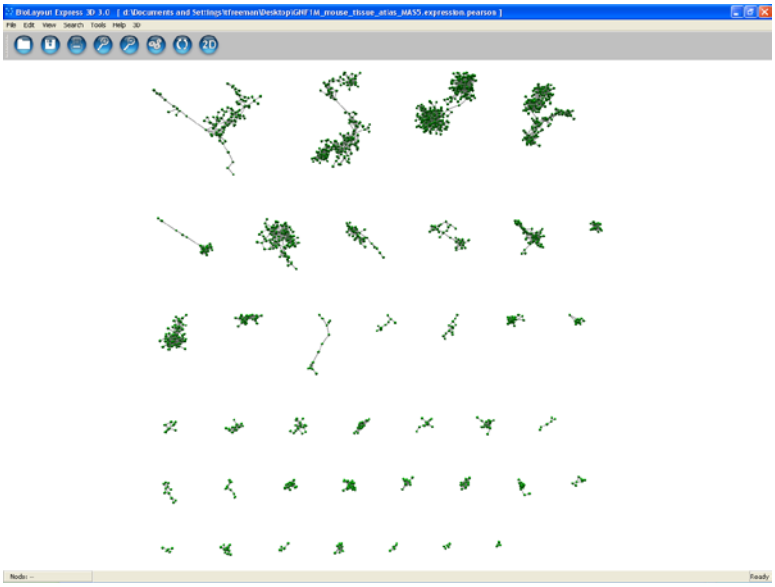


Figure 3: The main BioLayout Express^{3D} graph window

Personalising Your Display

8. When downloaded, a number of settings allow one to specify the aesthetic characteristics of the 2D and 3D interfaces. For a full description of options to change these please consult sections dealing with the *Properties* menu (see Supplementary data). Here we detail just the most important variables, available to change by selection of the *Properties* toolbox by clicking on the icon on the menu bar or selecting it under *Tools* ->*Properties* (Shift+P). This will give you access to 8 tabbed dialog boxes that allow the user to specify a range of options for personalising the aesthetic characteristics of the graphs as well as options for their analysis. The most commonly customised aesthetic characteristics are highlighted below:

9. *Background Colour.* Click on the *General* tab and select *3D Background Colour*. Select any colour from the palette provided. White or black is generally preferable but many other options are provided (Similarly this may be done for the 2D environment).

10. *Tiled Layout.* Under the *Layout* tab you can choose to alter how the graph is laid out. As default graphs are laid out as a *Tiled Layout*. This is to say that individual graph components (groups of interconnected nodes that share no edges with other graph components) are 'tiled' side-by-side, with the component with the greatest diameter being initially placed in the top left hand corner and the smallest component in the bottom right of the tiled graph. If this checkbox

is clicked off then components are laid out in an organic fashion and form a 'cloud' of components. For most applications a *Tiled Layout* provides a more intuitive format.

11. *Minimum Component Size.* Also listed under the *Layout* tab is the option to alter this setting. Expression graphs potentially contain many small components (<5) which are formed through either chance correlations or more often arise due to a redundancy in the targets of certain probe sets. As such they are generally not of interest and take up a lot of the plot space. They may be filtered out by setting this to an appropriate value. Changes to both the *Tiled Layout* and *Minimum Component Size* will only come into affect when new preferences are saved and the data is reloaded.

12. *Edges*. As default edges are coloured to reflect the Pearson correlation which they represent. Hence red edges represent a high correlation measure and blue a low correlation (relative to the range selected for display). Whilst sometimes useful to see, edges with a consistent colour may produce a more aesthetically pleasing graph. To change click on *Edges* tab, select *Colour Edges By* and select *Colour*. Select your edge colour of choice (preferably one that contrasts with the background colour) and hit OK. *Edge Thickness* can also be adjusted here and for larger graphs (>1000 nodes) you might find thinner edges e.g. 0.4 are preferable.

13. In order to ensure that alterations to the graph view are changed for future sessions select *Tools ->Save Preferences* (Alt+P)*.
**For Mac and Linux users the 'Command' key is used an alternative to the Alt key.*

14. We will address other ways to alter the appearance of the graphs later but for now you are ready to explore your data.

Graph Clustering using the Markov Clustering Algorithm (MCL)

15. Network graphs formed from expression data are often large and highly structured. This structure is a direct consequence of co-ordinate gene expression and the graphs provide an excellent interface to display and analyse these relationships. Implemented within BioLayout *Express*^{3D} is the MCL clustering algorithm which represents a powerful approach to dividing graphs non-subjectively into discrete chunks of genes sharing similarities in their expression i.e. clusters. MCL has been shown to compare favourably to other commonly used algorithms in clustering of large graphs [3] and as such it represents a robust state-of-the-art general-purpose clustering algorithm, available also as a stand-alone software package. A full description of the MCL algorithm is provided elsewhere[4]**.** Go to *Properties* (Shift+P) and select the *MCL* tab. The top *Inflation* slider is the most important factor defining the 'granularity' of the clustering. A high inflation value e.g. 4 will result in numerous small but 'clean' clusters whereas a low inflation value of say 1.5 will give fewer but generally less 'clean' clusters. For most purposes we have found that an MCL inflation value of between 1.7 and 2.2 works optimally. One other setting that is commonly used is the *Smallest Cluster Allowed*. In areas where networks are constructed from nodes with sparse connectivity, clusters tend to be small. This can result in the generation of many (hundreds) of small clusters that often surround and connect cliques of high connectivity. Setting *Smallest Cluster Allowed* to a number (usually between 3 and 10) will result in all clusters falling below that number in size being assigned to *No Class*. This means that th[es]e clusters can then be easily filtered away (see: *Removal of Small Clusters).*

16. Select a different MCL inflation values, hit OK and then go to the *Tools* menu and select bottom option *Cluster* Using MCL. A clustering dialog will then appear to mark the progress of the operation. The bigger the graph and the lower the MCL inflation value, the longer it will take. Following clustering, the graph will be re-centred and nodes will appear coloured according to cluster to which they belong. Clusters are numbered according to the number of nodes they contain (the largest cluster will be Cluster 1) and assigned an arbitrary colour. A clustering at a given inflation value will be added to, and displayed as, a new *Class* but will not be added to the input file.

Selecting Nodes and Viewing their Properties

17. There are number of basic ways to select nodes in a graph. Holding down Shift while dragging the mouse with left mouse button held down can be used to select nodes in the graph. Selected genes will be highlighted in the graph by encirclement with a 'cage' (see Fig. 4).

18. Holding down Shift+Alt while dragging the mouse with left mouse button held down can be used to select additional nodes in the graph, without deselecting the previous ones.

19. *Select Nodes Within The Same Class.* If a node belongs to a given class e.g. cluster, it may be selected and the command Ctrl+Alt+S will select nodes within the same class (see also *Selection Menu*).

20. In order to find a specific gene or class of genes (see later) select *Search* from the top menu bar and in the appropriate dialog box type in the gene of interest or select genes of a given class.

21. Ctrl+A selects all nodes in the graph. Various attributes of the selected nodes may now be viewed.

22. Select nodes by one of the means discussed above and open the *Class Viewer* (Ctrl+C). The left-hand side of the window will display the expression profile of the selected nodes scaled to the maximum expression value of those selected and the right hand side lists the identity of the selected genes and shows details of their associated properties/annotation/class membership (see Fig. 4).

23. The view of the expression data may be plotted in *Log Scale*, shown as the mean expression value of those selected (*Selection Mean*) or shown as mean values of individual classes represented in the selection (*Class Mean*) by clicking on the appropriate button above the expression window. The *Rescale* button rescales the data following calculation of the mean. *Grid Lines* may also be applied for better connection between data and sample.

24. The list on the right hand side of the screen by default lists all the selected nodes and displays their unique identifier, number of input and output edges (expression graphs are non-directional so input and output edges are equivalent) and the current class selected. If the graph has just been clustered, this will be the selected class.

25. Click on the box marked *View All Classes.* This will display all *Class* columns in the input file between the unique identifier and the data, plus the results of any clusterings that have been performed during the current analysis session. This may be a lot of columns of information and

their simultaneous display may be undesirable. To select specific columns of interest click on *Choose Columns To Hide* and uncheck those data that you do not wish to display and the display list will be automatically updated.

26. To browse the network based on class membership (this is particularly useful in assessing the results of a cluster analysis), click on *Find By Class*. If you wish to browse the data then select the first class (or cluster) and using *Next Class* one can browse the profile and content of those classes (clusters). This is a very fast way of reviewing the results clustering.

27. Having identified your 'genes of interest' you can edit your selection by deselecting the boxes next to the gene/probe identifier (first column). Clicking *Refresh Selection in Table* will update the table and the expression graph accordingly.

28. To export a list of selected genes as a tab limited file click on *Export List as…* and specify the location and name of the file to be saved.

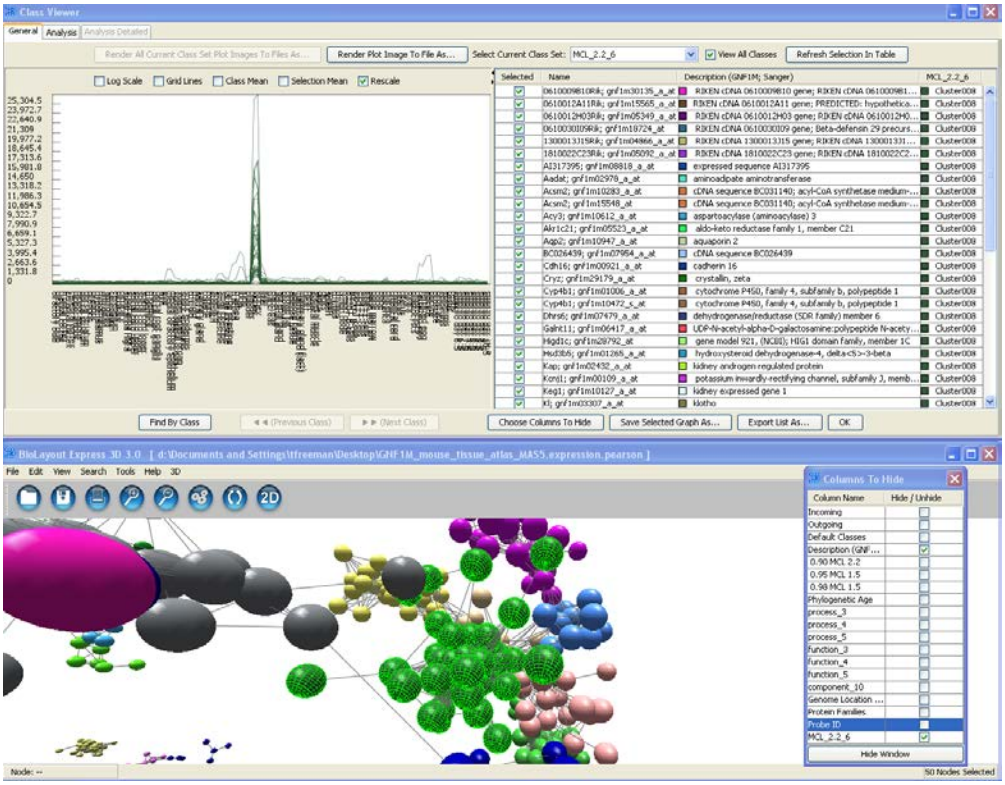


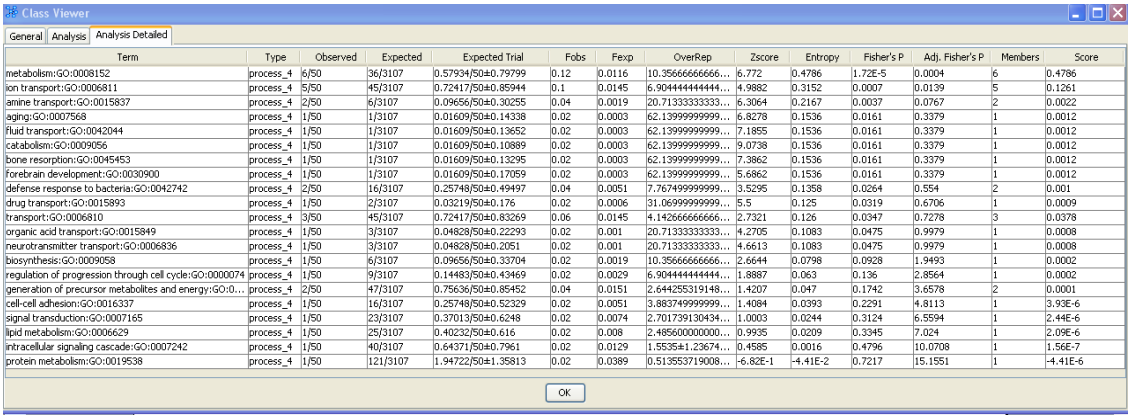
*Figure 4: The **Class Viewer** on top with the main BioLayout Express^{3D} graph window below*

Mining Selected Genes for Over-Representation of Classes

29. A group of selected genes e.g. a cluster may be mined to see if they are enriched for a given annotation class. This approach is widely used elsewhere to explore gene lists in order to query them for their over-representation of certain classes of genes usually relating to gene function (e.g. GO terms)[5] or gene sets (previously derived gene lists)[6]. This functionality is supported in BioLayout *Express*^{3D}.

30. Select genes of interest and open *Cluster Viewer* (Ctrl+C) or select within this window. In the top right hand corner click on the *Analysis* tab. The window will display all the available columns of annotation and class membership (the program does not differentiate between the two) and on the right the accumulated entropy score for each column.

31. Select an annotation class of interest and click the *Details* button at the bottom of the page. Depending on the size of the gene list selected and the number of terms within the selected class this may take a little time to calculate. The *Analysis Detailed* tab will appear and it will be populated with the frequencies of each term in the selected group of genes (Fig. 5. A detailed description all the calculations performed and presented within this window is given in the manual (Appendix 1) but in short, the most informative column is the *Adjusted Fisher's P value* which gives a robust statistical score of the relative representation of each class term in the selected genes relative to the background of the entire chip. Over-representation of specific classes thereby providing clues to the biological significance of the selected gene. Clicking on the *Details For All* button in the *Analysis* tab will examine the relative enrichment for all terms in every class. Whilst potentially useful, the number of calculations necessary to perform this task inevitably makes this process slow. A final cautionary note on mining data in this manner; if you are working with a network derived from filtered data you may have already be working with a biased selection and the mining of specific terms will be compared to this background potentially invalidating the results. Likewise where data includes numerous probe sets designed to the same gene, all of which are likely to share the same annotation, enrichment of terms associated with these genes may appear to be artificially enriched if they are included multiple times in the same selection.


Class Viewer

General | Analysis | Analysis Detailed

Term	Type	Observed	Expected	Expected Trial	Fobs	Fexp	OverRep	Zscore	Entropy	Fisher's P	Adj. Fisher's P	Members	Score
metabolism:GO:0008152	process_4	6/50	36/3107	0.57934/50±0.79799	0.12	0.0116	10.356666666666…	6.772	0.4786	1.72E-5	0.0004	6	0.4786
ion transport:GO:0006811	process_4	5/50	45/3107	0.72417/50±0.85944	0.1	0.0145	6.904444444444…	4.9882	0.3152	0.0007	0.0139	5	0.1261
amine transport:GO:0015837	process_4	2/50	6/3107	0.09656/50±0.30255	0.04	0.0019	20.713333333333…	6.3064	0.2167	0.0037	0.0767	2	0.0022
aging:GO:0007568	process_4	1/50	1/3107	0.01609/50±0.14338	0.02	0.0003	62.139999999999…	6.8278	0.1536	0.0161	0.3379	1	0.0012
fluid transport:GO:0042044	process_4	1/50	1/3107	0.01609/50±0.13652	0.02	0.0003	62.139999999999…	7.1855	0.1536	0.0161	0.3379	1	0.0012
catabolism:GO:0009056	process_4	1/50	1/3107	0.01609/50±0.10889	0.02	0.0003	62.139999999999…	9.0738	0.1536	0.0161	0.3379	1	0.0012
bone resorption:GO:0045453	process_4	1/50	1/3107	0.01609/50±0.13295	0.02	0.0003	62.139999999999…	7.3862	0.1536	0.0161	0.3379	1	0.0012
forebrain development:GO:0030900	process_4	1/50	1/3107	0.01609/50±0.17059	0.02	0.0003	62.139999999999…	5.6862	0.1536	0.0161	0.3379	1	0.0012
defense response to bacteria:GO:0042742	process_4	2/50	16/3107	0.25748/50±0.49497	0.04	0.0051	7.767499999999…	3.5295	0.1358	0.0264	0.554	2	0.001
drug transport:GO:0015893	process_4	1/50	2/3107	0.03219/50±0.176	0.02	0.0006	31.069999999999…	5.5	0.125	0.0319	0.6706	1	0.0009
transport:GO:0006810	process_4	3/50	45/3107	0.72417/50±0.83269	0.06	0.0145	4.142666666666…	2.7321	0.126	0.0347	0.7278	3	0.0378
organic acid transport:GO:0015849	process_4	1/50	3/3107	0.04828/50±0.22293	0.02	0.001	20.713333333333…	4.2705	0.1083	0.0475	0.9979	1	0.0008
neurotransmitter transport:GO:0006836	process_4	1/50	3/3107	0.04828/50±0.2051	0.02	0.001	20.713333333333…	4.6613	0.1083	0.0475	0.9979	1	0.0008
biosynthesis:GO:0009058	process_4	1/50	6/3107	0.09656/50±0.33704	0.02	0.0019	10.356666666666…	2.6644	0.0798	0.0928	1.9493	1	0.0002
regulation of progression through cell cycle:GO:0000074	process_4	1/50	9/3107	0.14483/50±0.43469	0.02	0.0029	6.904444444444…	1.8887	0.063	0.136	2.8564	1	0.0002
generation of precursor metabolites and energy:GO:0…	process_4	2/50	47/3107	0.75636/50±0.85452	0.04	0.0151	2.644255319148…	1.4207	0.047	0.1742	3.6578	2	0.0001
cell-cell adhesion:GO:0016337	process_4	1/50	16/3107	0.25748/50±0.52329	0.02	0.0051	3.883749999999…	1.4084	0.0393	0.2291	4.8113	1	3.93E-6
signal transduction:GO:0007165	process_4	1/50	23/3107	0.37013/50±0.6248	0.02	0.0074	2.701739130434…	1.0003	0.0244	0.3124	6.5594	1	2.44E-6
lipid metabolism:GO:0006629	process_4	1/50	25/3107	0.40232/50±0.616	0.02	0.008	2.485600000000…	0.9935	0.0209	0.3345	7.024	1	2.09E-6
intracellular signaling cascade:GO:0007242	process_4	1/50	40/3107	0.64371/50±0.7961	0.02	0.0129	1.5535±1.23674…	0.4585	0.0016	0.4796	10.0708	1	1.56E-7
protein metabolism:GO:0019538	process_4	1/50	121/3107	1.94722/50±1.35813	0.02	0.0389	0.513553719008…	-6.82E-1	-4.41E-2	0.7217	15.1551	1	-4.41E-6

OK

*Figure 5: The **Analysis Detailed** tabbed option panel of the **Class Viewer***

Other Useful Functions for Expression Analysis

32. *Removal of Small Clusters.* If a value has been entered in the *Smallest Cluster Allowed* box under the *MCL* tab, clusters below the size selected will not be assigned to cluster class i.e. they will be listed as *No Class* for that clustering. Nodes belonging to small clusters often reside at the periphery of a graph and often represent genes only loosely related (perhaps by chance) to the main structure of the graph. If you select a node not assigned to a class (coloured dark blue as default) and *Select Nodes Within The Same Class* (Ctrl+Alt+S), all nodes belonging to this class will be highlighted. To hide these nodes go to *View -> Hide Selected Nodes* (Ctrl+Shift+H). This will effectively give the graph a 'haircut', removing nodes from around the central structure. These nodes may be added back to the graph by selecting *View -> Unhide Nodes* (Ctrl+U).

33. *Filter by edge weight or number of edges.* Graphs can be filtered based on edge weight such that all edges below a set threshold and potentially the nodes that depend on them for their connectivity to the graph, will be removed. Go *Edit -> Filter Edges By Weight* (Cntl+Alt+W) and move the slider bar to desired setting. If *Preview* button is checked then the changes to the graph will be visible and if *Also Hide/Unhide Nodes* box is checked nodes will disappear when they no longer posses any edges above the selected threshold. Similarly graphs may also be filter to remove nodes based on the number of edges they possess. As nodes with a low node degree tend to be on the periphery of the main structure of a graph this has the effect of giving a graph a 'haircut' removing the nodes on the periphery of the main structure. Go *Edit -> Filter Edges By Weight* (Cntl+W) and select the minimum node degree.

34. *Changing the Visual Properties of Nodes*. Make a selection of a given group of nodes and open the *Properties* (Shift+P) window. The window should open directly on the *Nodes* tab where the properties of the selected nodes may be changed (if no nodes are selected this menu will not be available). In this window the *Shape*, *Node Size* and *Transparency* of the selected nodes may be altered. Try altering these settings and then press Apply or OK.

35. *Display of Node Labels.* There are number of options for displaying the names (labels) of nodes. *View -> Show All Labels* (Crtl+Shift+L) will display labels for all nodes and *Show Labels Of Selected Nodes* (Crtl+L). To reverse and hide node labels use either Hide All Labels (Crtl+Alt+Shift+H) or *Hide Labels from Selected Nodes* (Crtl+Alt+H).

36. *Collapsing Nodes*. In order to simplify the view of a selected graph, nodes may be 'collapsed'. The most common use of this function is when dealing with graphs with complex structure containing multiple clusters. Cluster the graph (or select a previous clustering) and then under the *Edit* menu select *Collapse Cluster By Class* (Crtl+Alt+Shift+G). All nodes belonging to a given class will be collapsed into a single node where the diameter (volume) of the node is proportional to the number of nodes in the original class (cluster). The connectivity between clusters will be maintained and size of the spheres may enlarged or reduced for aesthetic reasons using the commands *Ctrl+>*or *Ctrl+<*, respectively. The operation may be reversed by selecting *Edit->Uncollapse All Groups* (Crtl+Alt+Shift+U).

37. *Image Export*. To export of images of expression graphs select *3D* from the top menu and *Render Graph Image to File*. Files will be stored in folder called 'Screenshots' in the same folder from which the data was launched. Files can be saved as .png or .jpg format. A high definition image for presentation purposes can be generated by selecting *Render Hi Res Graph Image To File As…*under the *3D* menu. The resolution of the image can be adjusted using the slide bar found under *Properties->3D Rendering* listed as *High Resolution Image To File Render Option*. If set at high values this operation may take a few seconds to complete.

38. *Saving Graphs.* Entire graphs may be saved as a layout file whereby the plot co-ordinates are stored from the existing graph, such that when a saved graph is reloaded the time taken to layout the graph is omitted. When dealing with large graphs this may represent a considerable saving in load time. If expression data is to be viewed from these graphs, the layout file must always be stored in the same folder. To save an entire graph go to *File->Save Graph As…*(Ctrl+S). Likewise sub-graphs may be saved using *Save Selected Graph As…* and *Save Visible Graph As…* listed under the same *File* menu.

Working with Other Data Formats

BioLayout *Express*^{3D} supports the construction of network graphs from data imported into the tool in a number of standard graph formats (see box 1 for details). Below we describe some of the other possibilities for editing networks in the tool using a GraphML input file to illustrate some of these features. GraphML is an easy-to-use file format for network graphs. It is based on XML and hence is suited as a common denominator for all kinds of services generating, archiving, or processing network graphs.

39. *Generation of GraphML file*. There are now a number of programs that support the import and export of GraphML files. yEd Graph Editor (yFiles, Tubingen, Germany) is a general graph editing tool that we have been using for the construction of diagrams of macrophage signaling and effector pathways[7]. Once created, a GraphML file may be directly opened in Biolayout *Express*^{3D}. For this exercise we will use the macrophage activation pathway[7] (Figure 6a). To download the file, go to the BioLayout *Express*^{3D} website (http://www.biolayout.org/) and select *Downloads*. Scroll down to datasets, click on the pathway in GraphML file format and download. Decompress (Unzip) file.

40. To open the file, select *File ->Open*. The *File Open* dialog will appear. Find and select file and click *Open*. Alternatively a GraphML file may be dropped directly into the BioLayout *Express*^{3D} window.

41. If opened in 3D mode, a graph will be generated as shown in Figure 6b. At present the BioLayout *Express*^{3D} GraphML parser imports all the graph details but will display only the component name in 3D mode. To display component names go to *View -> Show All Labels*.

42. If the 2D button on the menu bar is clicked then the graph will be converted to a 2D representation of the 3D graph. We have however also implemented the import of data such that the original position of nodes in the GraphML are imported and may be used in the program's 2D mode. Open the *Properties* window (Shift+P) and under the *General* tab click the text box *yEd-style rendering of GraphML imported files* (second row down under sub-heading *General 2D Graph Settings*). The graph will then be rendered using the node locations encoded in the GraphML file (Figure 6c). In this pathway diagram, the nodes in the graph represent different biological entities and a variety of relationships between them. The various classes of nodes may be distinguished using a combination of node properties, namely node size, shape and colour. The following section describes how node properties may be manipulated.

43. There are a number of ways to select nodes for the editing of their properties. Nodes of a similar type may be searched for by name or part of their name. For example there are multiple nodes in the example graph with the same name that represent processes e.g. activation (A), inhibition (I), translocation (T) or logic nodes e.g. AND, OR. To search for a given class of nodes select *Search* from the top menu bar and select *Find By Name* (Ctrl+F). As an example type A. This will select all nodes representing protein/gene activation.

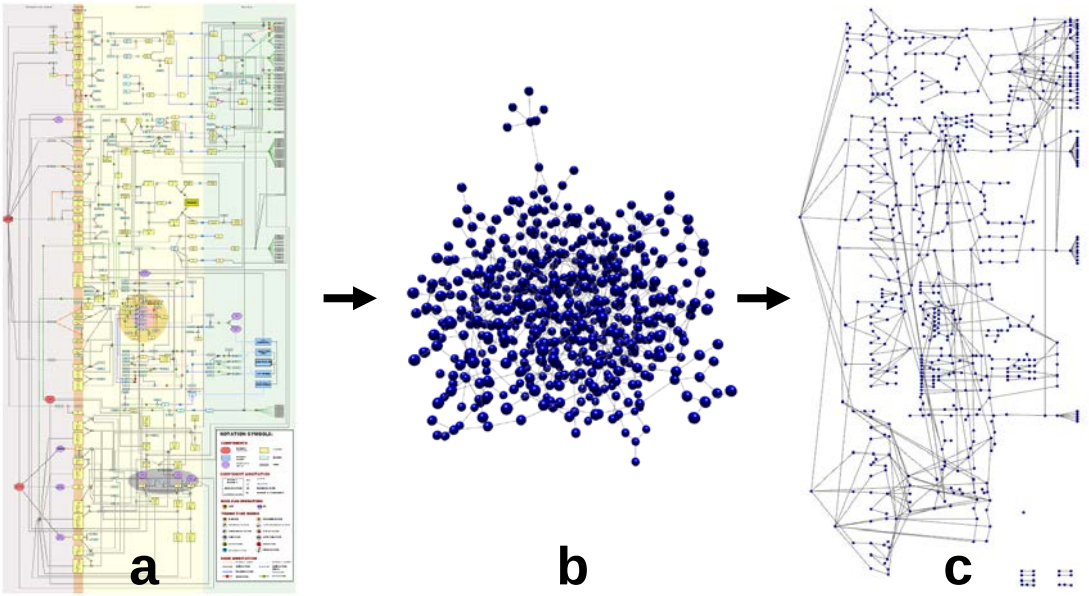


*Figure 6: Rendering of **GraphML** files in BioLayout Express*[3D]

44. To change the visual properties of the selected nodes, open the *Properties Menu* (Shift+P) and the window will open on the *Nodes* tab. This menu provides the opportunity to change the nodes shape (in 2D as well as 3D), colour, transparency, as well as size. Try changing these settings and then click *Apply*. A dialog window will appear asking '*Are you sure you want to Override Class Colour?'.* Click *Yes* and the changes you have made will take effect.

45. In this way, the visual characteristics of the graph may be changed such that nodes belonging to different classes may be assigned distinct visual properties. Saving the graph *File -> Save Graph As…* (Ctrl+S) will create a layout file where the changes to the node characteristics will be saved.

46. In a similar way, nodes can also be assigned to classes such that we can select all nodes belonging to given class of protein. Using again the example of the Raza *et al.*, macrophage activation pathway select all the process nodes representing activation by selecting *Search* from the top menu bar and select *Find By*, type 'A'. Open the *Properties* window and go to the *Classes* tab. Go to *Create Class Set* and type in 'Process nodes' and in the *Create Class* box type 'Activation'. Then click *Apply*. Moving to the *Nodes* tab you may change the selected nodes shape, size and colour in 2D and 3D, and finally confirm their class membership by going down to the *Node Class* section of the *Nodes* tab and click on the *Containing Class* drop-down menu and select the newly defined class. When you click *Apply* a dialog window will appear asking '*Are you sure you want to Override Class Colour?'.* Click *Yes* and the changes you have made will take effect. This procedure may be repeated adding nodes to classes and changing their properties. A version of this pathway diagram where all the nodes in this pathway have been classified under node type and cellular location is also available on the BioLayout *Express*[3D] website just below the GraphML file in the download section (Figure 7). It is also worth noting that nodes can be assigned to classes by directly editing the input file. See *Creation of Classes* in Box 1.

44. In 2D mode the directionality of edges may be shown by opening the *Properties* dialog box and under the *General* tab selecting *Directional Edges* under *General 2D Graph Options* heading. The arrowhead size can be adjusted under the *Edges* tab.

45. A useful feature when following the connectivity of components in a pathway diagram is the selection of parent (upstream) or children (downstream) nodes. To do this select a node in the diagram e.g. a pathway input such as IFNG and go to *Edit -> Selection -> Select Children* (Ctrl+Alt+C). The node(s) immediately downstream i.e. output edges will be selected. Repeating this action will allow you follow the flow of connectivity. *Edit -> Selection -> Select All Children* (Ctrl+Alt+Shift+C) will select all nodes downstream of a node selection. Alternatively, nodes upstream (inputs) of nodes may be selected using the commands *Edit -> Selection -> Select Parents* (Ctrl+Alt+P) or *Select All Parents* (Ctrl+Alt+Shift+P).

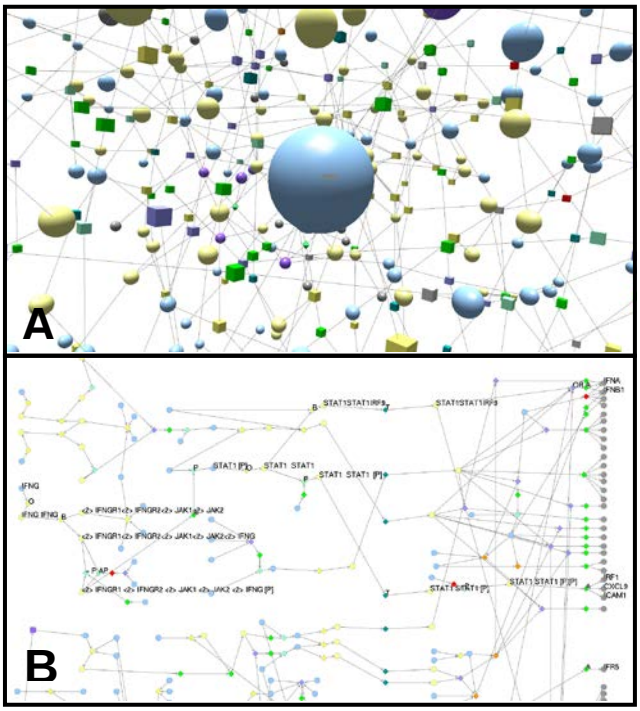


*Figure 7: Rendering of pathways in BioLayout Express^{3D}. **A) 3D mode** and **B) 2D mode** using node shape, size and colour to distinguish between classes of nodes. In the bottom panel labels have been placed on nodes downstream of interferon-gamma (IFNG).*

Summary

BioLayout *Express*^{3D} provides a powerful environment for the visualisation, analysis and integration of large network graphs. The tool has been designed primarily for rendering of data derived from biological pathways, protein-protein and gene expression data, but also supports the analysis of networks derived from data from any source. This article provides protocols for the use of BioLayout *Express*^{3D} in the visualisation and analysis of gene expression data and biological pathways imported as GraphML files. In doing so it provides a workflow for dealing with graphs and describes the basic functionality of the tool that may be used for other applications. Development of the program is ongoing and many new features will be added in the near future as we refine the tools for existing applications, develop new applications and

optimise its implementation to make best use of the rapid advances in Java/OpenGL and improvements in hardware, particularly graphics cards.

References

1. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-7 (2004).
2. Freeman, T.C. et al. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol* **3**, 2032-42 (2007).
3. Brohee, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 488 (2006).
4. van Dongen, S. Graph Clustering by Flow Simulation. *PhD Thesis, University of Utrecht* (2000).
5. Dennis G Jr, S.B., Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, P3 (2003).
6. Subramanian A, T.P., Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* **102**, 15545-50 (2005).
7. Raza, S. et al. A logic-based diagram of signalling pathways central to macrophage activation. *BMC Syst Biol* **2**, 36 (2008). Raza S., McDerment N., Lacaze P.A., Robertson K., Watterson S., Chen Y., Chisholm M., Eleftheriadis G., Monk S., O'Sullivan M., Turnbull A., Roy D., Theocharidis A., Ghazal P. and *Freeman T.C.* Construction of a Large Scale Integrated Map of Macrophage Pathogen Recognition and Effector Systems. BMC *Systems Biology* **4**:63, 2010.

3. Assembly of Logic-Based Diagrams of Biological Pathways

Introduction

Complete genome sequencing of hundreds of pathogenic and model organisms over the last decade has provided us with the parts list of life (Janssen et al, 2003). At the same time enormous amounts of data pertaining to the nature of genes and proteins and their potential cellular interactions have now been generated using new analytical platforms including, but not limited to: gene coexpression analysis, yeast two-hybrid assays, mass spectrometry, and RNA interference (Reed et al, 2006). With the advent of next generation sequencing technologies and advances in other fields, this deluge of data on biological systems only looks set to continue and increase. Whilst the data from these 'omics platforms can be overwhelming these analyses finally allow us to open a window on to the complex cellular and molecular networks that underpin life (Kitano, 2002; Nurse, 2003). The main problem we now face is how to interpret all this data and use it to better understand the structure and function of biological pathways in health and disease (Cassman, 2005).

Our existing knowledge of biological pathways and systems is still largely based on the painstaking efforts of countless investigators whose work has and continues to be, focused on a specific cell type and the function of one or a small number of proteins within that cell. Their studies have produced our current framework of understanding of how proteins and genes interact with each other to form the metabolic, signalling and effector systems that together regulate biological form and function. Much of this work however remains locked inside the literature where specific insights into the functional role of cellular components are subject to the semantic irregularities that come with their description by different authors. As a result, the details of a given pathway have traditionally been known only to a few experts in the field whose research is often focused on a single protein and its immediate interaction partners within that pathway. These pathways are understood more generally by their description in reviews and diagrams produced on an *ad hoc* basis.

To a certain degree the concept of a biological pathway is an artificial construct and in reality there is only one big integrated network of molecular interactions operating within a cell. However, it is still useful to think in terms of pathways as being connected modules of this network. As such a pathway may be considered to consist of a specific biological input or event that initiates a series of directional interactions between the components of a system leading to an appropriate shift in cellular activity. In other words a biological pathway might be viewed as starting from the engagement of a ligand with its receptor to all the downstream consequences of that interaction. This not to say that the cellular components utilised for such a pathway will be necessarily unique to it, only that they are connected in this context. As we begin to appreciate the complexity of these molecular networks, their topology and interconnectivity, there is increasing interest in moving away from the tradition gene-centric view of life to a systems or pathway level appreciation of biological function. To do this we need to create models of these pathways.

Pathway diagrams act as a visual representation of known networks of interaction between cellular components and modelling them is fundamental to our understanding of them. At their best formalised diagrams of biological pathways act as a clear and concise visual representation of the known interactions between cellular components. However, the task of assimilating the large amounts of available data on a particular pathway and representing this information in an

intuitive manner remains an ongoing challenge. Indeed, there are numerous different ways that one can represent a pathway and pathway diagrams are currently available in a plethora of different forms. Using the term in the broadest sense, they can be a picture that accompanies a review article, wall charts distributed by a journals and companies, small schematic diagrams used to support mathematical modelling efforts or network graphs reflecting all known protein interactions based on the results of large scale interaction studies or literature mining. As such, pathway models are an invaluable resource for interpreting the results of genomics studies (Antonov et al, 2008; Arakawa et al, 2005; Babur et al, 2008; Cavalieri & De Filippo, 2005; Dahlquist et al, 2002; Ekins et al, 2007; Pandey et al, 2004), for performing computational modelling of biological processes (Eungdamrong & Iyengar, 2004; Kwiatkowska & Heath, 2009; Ruths et al, 2008; van Riel, 2006; Watterson et al, 2008) and fundamentally important in defining the limits of our existing knowledge. To support these efforts there are also a growing number of databases that serve up a wide range of pathways which are either curated centrally (http://www.biopax.org/; http://www.ingenuity.com/; Kanehisa & Goto, 2000; Thomas et al, 2003) or increasingly by the community (Joshi-Tope et al, 2005; Pico et al, 2008; Schaefer et al, 2009; Vastrik et al, 2007). These offer searchable access to pathway diagrams and interaction data derived from a combination of manual and automated (text mining) extraction of primary literature, reviews and large-scale molecular interaction studies. The sheer range of resources available (Bader et al, 2006) reflects the current interest in pathway science. Whilst invaluable and in many ways the best we have, a major problem with these efforts is that the information content of these diagrams is frequently limited, generic and visualisations of these systems are of variable and often poor quality; Pathways are drawn using informal and idiosyncratic notation systems using a variety of shapes (glyphs) to illustrate component ‘type’. There are variable degrees of accuracy and specificity in defining what pathway components are being depicted and the relationships between them. Resources are often fragmented with some proteins or metabolites being members of numerous pathways; the concept of pathway membership being a highly subjective division. The pathways themselves are rarely available as a cohesive network and there are numerous pathway exchange formats in current use (Hucka et al, 2003; Lloyd et al, 2004; Luciano, 2005). Finally, pathway diagrams are generally highly subjective reflecting the curator's bias, such that two diagrams depicting the ‘same’ pathway may share little in common. Together these factors commonly result in uncertainty as to what exactly is being shown. All in all, despite the huge efforts in time and resources that has been poured into pathway science the state of the art leaves a lot to be desired.

As our appreciation of systems-level biology increases rapidly, there has been an increasing realisation of the need for comprehensive well constructed maps of known pathways. Over the past ten years a number of groups have suggested formalised notation schemes and syntactical rules for drawing ‘wiring diagrams’ of cellular pathways (Cook et al, 2001; Kitano et al, 2005; Kohn, 1999; Moodie et al, 2006; Pirson et al, 2000). These have been used to construct a number of large pathway diagrams (Calzone et al, 2008; Oda et al, 2004; Oda & Kitano, 2006). These pioneering efforts have all contributed to the field and more recently the Systems Biology Graphical Notation (SBGN) group have proposed a series of formalised pathway notation schemes to be adopted by all (Novere et al, 2009). Of course in principle this is an excellent idea but it remains to be seen whether the SBGN schemes are going to be widely taken up or indeed whether they are flexible enough to suit all purposes.

Our own efforts on pathway modelling stem from our interest in macrophage biology and in understanding pathways known to be activated in these cells during infectious and inflammatory disease. Therefore last few years we have been constructing large graphical

models of macrophage-related pathways as a way of recording what is known about the signalling events controlling this cell's immune biology (Raza et al, 2008, 2010). In so doing our main objectives have been to create models that:

1. support the detailed representation of diverse range of biological entities, interactions and pathway concepts
2. represent a consensus view of pathway knowledge in a semantically and visually unambiguous manner
3. are easy to assemble and understandable by a biologist
4. are useful in the interpretation of 'omics data
5. are sufficiently well defined that software tools can convert these graphical models into formal models, suitable for analysis and simulation

In attempting to achieve these goals we have faced one of the central challenges in pathway biology: How exactly does one construct clear concise pathway diagrams of the known interactions between cellular components that can be understood by and useful to a biologist? In the beginning our efforts were largely based on the principles of the process diagram notation (PDN) (Kitano et al, 2005) and the original Edinburgh Pathway Notation (EPN) scheme (Moodie et al, 2006). However during the course of working with these notation schemes it became apparent that the available diagrams drawn using these systems were not always easy to interpret and the schemes were a challenge to implement. Furthermore, we found that these notation schemes did not support all of the concepts that we wished to represent in order to reflect the full diversity of pathway components and the relationships between them. As a result of our efforts we have significantly modified these existing schemes and created what has now been named the 'modified Edinburgh Pathway Notation' (mEPN) scheme (Freeman et al, 2010). Below I describe the basic principles behind the mEPN scheme and illustrate how it can be used to depict a wide variety of biological pathways.

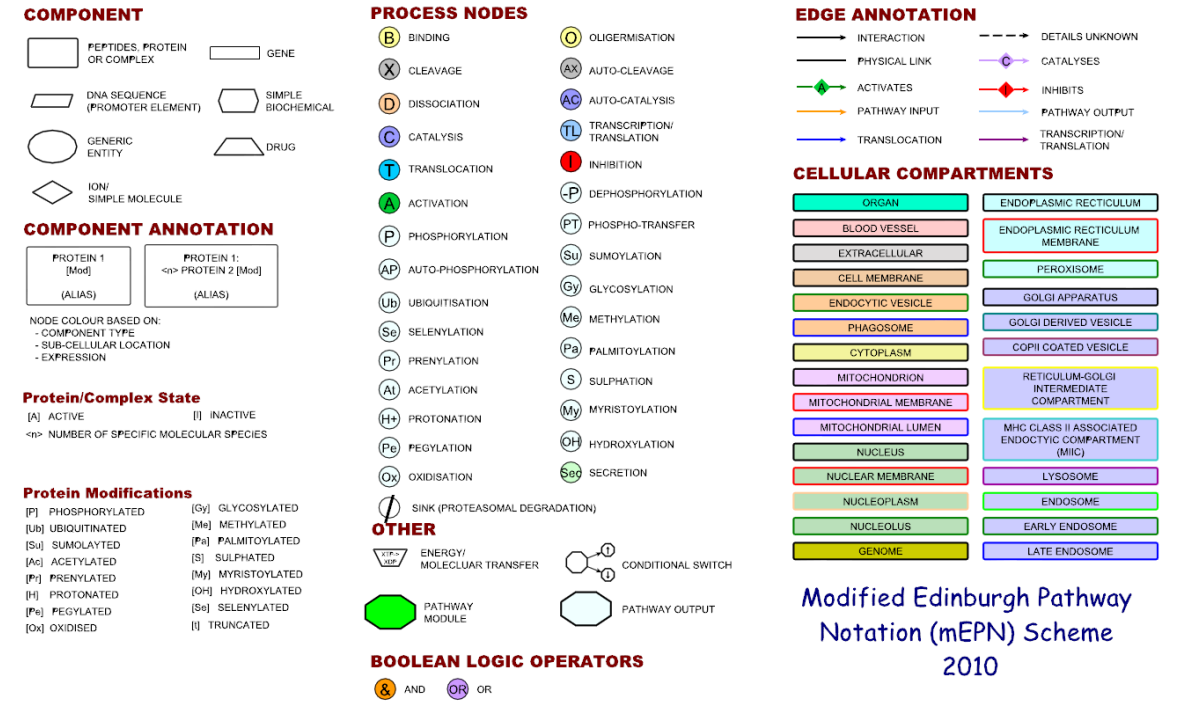
1. Definition of the modified Edinburgh Pathway Notation (mEPN) Scheme

A *pathway* may be considered to be a directional network of molecular interactions between components of a biological system that act together to regulate a cellular event or process. In this context a *component* is any physical entity involved in a pathway that contributes or influences its activity e.g. a protein, protein complex, nucleic acid (DNA, RNA), molecule, etc. The *mEPN scheme* is a collection of formalised symbols that form the constituent parts of a graphical system for depicting the components of a biological pathway and the interactions between them. The mEPN scheme is based on the *node* and *edge* principles of depicting networks. This which allows one to use ideas and tools previously developed in graph theory and applied more recently to computational systems biology. Cellular components are represented as nodes (vertices) in the network and specific *glyphs* (stylised graphical symbols) are used that impart information nonverbally on the class of biological entity portrayed e.g. protein, gene, biochemical etc. The processes that connect components are also represented by nodes using different glyphs and the connectivity between them is defined by *edges* (lines/arcs). Edges represent interactions or relationships between one component and another usually where one component influences the activity of another e.g. through its binding to, inhibition of, catalytic conversion of, etc. The network of interactions between cellular components and processes thereby defines a pathway.

2.1 Depiction of Pathway Components

When drawing pathways one has to decide about the level of biological detail that you wish to depict. It is not uncommon in pathway depiction to use component glyphs that infer structural or functional characteristics of the entities depicted. For instance, receptors may be shown using a glyph with a specific ligand binding site or possibly as a protein containing membrane spanning domains. Whilst on one level this approach is appealing to the eye and imparts visually information on the nature of the molecular species depicted, it can lead to complications. After all both depictions described above may be appropriate for any receptor and a protein may also have other functional domains which could be graphically depicted. If one tries to impart all this information visually it leads to a notation system that is difficult to implement and to remember. Such a system also requires the development of specific pathway editing tools that support it. In contrast we have used a set of standard shapes to represent different classes of components (molecular species) and in so doing created a notation scheme that is supported by generic network editing/visualisation tools, in particular the tool of choice for all our work has been the freely available yEd (yFiles, Tubingen). There is however a variety of other pathway and network editing tools available (Pavlopoulos, 2008). It is worth remembering that the ability graphically depict a wide variety of pathway concepts depends not only on the tool used to construct and display them, but also the pathway notation scheme employed.

The mEPN scheme as described here is based on the concepts first described for the process diagram notation (PDN) scheme (Kitano et al, 2005). However, our experience in building large-scale pathway models of a variety of biological systems has required us to depict concepts that were not supported by the original PDN scheme. Furthermore, a lack of available pathway editing tools when we began this work, as well as the scale our diagrams, have both played their part in determining our approach to pathway depiction. As a result there are a number of important differences that exist between the mEPN scheme described here and other PDN schemes. Firstly, in common PDN, the mEPN uses simple shapes to define the class of a component but only a labelling system to define the exact identity of components (nodes). Other schemes use circles overlaid on nodes to depict protein modifications. We have found this a considerable overhead to implement and can interfere the clarity of what is depicted rather than enhancing it. Furthermore, the PDN scheme is not supported by many of the general purpose network visualisation tools e.g. yEd, Cytoscape, Biolayout *Express*^{3D} (Freeman et al, 2007; http://www.yworks.com; Yeung et al, 2008), requiring instead the use of dedicated pathway editing software e.g. CellDesigner (Funahashi et al, 2008). Secondly, we have avoided the use of different styles of arrowheads to depict the nature of interactions (edges) which limits the vocabulary of edges and is a system that can be challenging to remember. Instead where appropriate, we have chosen to use inline annotation nodes to depict the meaning of edges; these carry a visual clue (a letter symbolising the meaning of the edge e.g. A for activation, I for inhibition) and can potentially support a wider range of edge meanings. Again the use of a wide variety of arrowheads is not supported by many pathway/network editing software packages. Finally, we explicitly state the nature of interactions by the use of labelled process nodes. Under other PDN-based schemes process nodes are used but generally not as a means to convey the nature of interactions except in the case of protein binding (association) and dissociation. When pathways are large and the distance between interacting species may be great, having a visual clue as to the nature of interactions is very important in our experience.



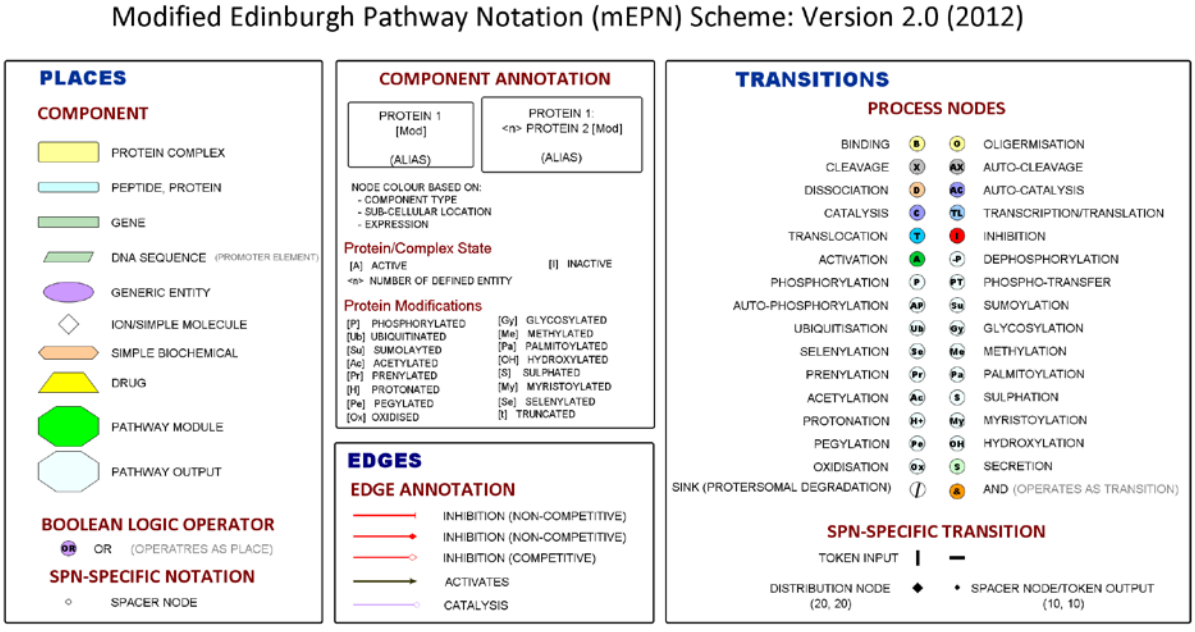


Figure 1. List of the Glyphs used by the modified Edinburgh Pathway Notation (mEPN) scheme. Unique shapes and identifiers are used to distinguish between each element of the notation scheme. The notation scheme essentially consists of the following categories of nodes representing; cellular components, processes and Boolean logic operators. Edges are used to denote the interactions between components, the nature of the relationship between them being described using process nodes and Boolean operators and edge annotations. The cellular compartment in which these components reside being depicted by their spatial localisation in the network and background colour.

The full set of glyphs employed in the mEPN scheme is shown in figure 1. Under the scheme *peptides, proteins* and *protein complexes* are all represented by a rounded rectangle and *genes* depicted using a rectangle. Parallelograms may be used to show a specific *DNA sequence* known

to play a specific functional role e.g. promoter sequence. This may be shown on its own or associated with a gene or other genomic feature. *Simple biochemicals* e.g. sugars, amino acids, nucleic acids, metabolites are represented using a hexagon. It is often the case that an interacting component of a pathway is not an exact molecular entity but rather molecular class or complex entity such as a virus or other pathogen. In this case we use a flattened circle (ellipse) to depict any *generic entity*. A small molecule or biologic known to affect a biological system is show using a trapezoid. These may be licensed as a *drug* or used for experimental manipulation of biological components e.g. enzyme inhibitor, siRNA etc. Finally, a *ions* e.g. Ca^{2+}, Na^{+}, Cl^{-},or other *simple molecules* H_2O, NO, O_2, CO_2 are represented using a diamond shaped glyph.

2.2 Component annotation

Multiple component names are often available to describe any given component. For example, the same protein may be called several different names in the literature. In other cases the same name has been used to describe different proteins and some protein names are quite different from the gene name. Other names sometimes used for labelling components in pathways do not represent any specific entity at all e.g. NF-ĸB. Therefore when non-standard nomenclature is used to name pathway components it frequently leads to ambiguity as to the exact identity of what is being depicted. Use of standard nomenclature to denote a components identity removes this uncertainty and also assists in the comparison and overlay of experimental data with pathway models. Under mEPN we recommend the use of standard gene nomenclature systems e.g. human genome nomenclature committee (HGNC) or mouse genome database (MGD) systems to name human or mouse genes/proteins, respectively. These nomenclature systems now provide a near complete annotation of all human and mouse genes. Their use in the naming of proteins as well as genes provides a direct between the two. Therefore when a protein or gene is discussed within a paper almost the first act is to search the databases in order to record the identity of the component according to standard nomenclature. Where other names (alias') are in common use these name(s) may be shown as an addition to the label on the glyph representing the protein and included after the official gene symbol in rounded ( ) brackets. Protein complexes are named as a concatenation of the proteins belonging to the complex separated by a colon. Again if the complex is commonly referred to by a generic name this may be shown. There are no strict rules as to the order in which the protein names are shown in the complex and are often shown in the order in which the proteins join the complex, in the position they are likely to hold relative to other members of the complex (where known) or position relative to cellular compartments e.g. with receptor proteins in a membrane bound protein complex protruding into the extra-cellular space. Where a specific protein is present multiple times within a complex, this may be represented by placing the number of times a protein is present within the complex in angle brackets < >. If the number of proteins in the complex is unknown this may be represented by <n>. The particular 'state' of an individual protein or a protein within a complex may be altered as a consequence of a particular process. This change in the component's state is marked using square [ ] brackets following the component's name; each modification being placed in separate brackets. This notation may be used to describe the whole range of protein modifications from phosphorylation [P], truncation [t], ubquitinisation [Ub] etc. Where details of the site of modification are known this may be represented e.g. [P-L232] = phosphorylation at leucine 232. Alternatively the details of a particular modification may be placed as a note on the node visible only during 'mouse-over' or when viewing a node's properties. Where multiple sites are modified this may be shown using multiple brackets, each modification (state) being shown in separate brackets. Unfortunately, there appears to be no universally recognised nomenclature

system for many of other classes of biologically active molecules e.g. lipids, metabolites, drugs, and therefore when included in a pathway we have generally used names commonly recognised by biologists.

Colour may also be added to the diagrams to assist in their interpretation. Components may be coloured to impart information on components type, location or state e.g. to visually differentiate between a protein and a complex, to denote cellular location or denote a component's expression level. In addition process nodes, Boolean operators, compartments and edge annotations are generally coloured to improve the visual impact of the diagram. However it must be stated that the exact choice of colours is down to individual taste and colour recognition capabilities and the mEPN scheme has been designed to work even in the absence of colour.

2.3 Depiction of Biological Processes

A process node in the context of this notation system can be defined as a specific action, transformation, transition or process occurring between components or to a component and is represented by a process node. Process nodes impart information on the type of process that is associated with transformation of a component from one state to another or movement in cellular location. They also act as junctions between components and as such may have multiple inputs or outputs to or from components. All process nodes are represented by a small circular glyph and the process they represent is indicated by a one-to-three letter code. Colour has been used as a visual clue to group processes into 'type' but is not necessary for inferring meaning. There are currently 32 process nodes recorded under the mEPN. Different process nodes generally have different network connectivity. For instance a process node depicting a component's translocation from one compartment to another will generally only have one input and output edge (Figure 2a). In contrast a 'binding' node will have multiple inputs and one output (Figure 2b), the opposite is true for a dissociation node (Figure 2c). Process nodes also act as way of collating information about a given event; for example protein X may be converted from one state to another by a process activated by protein Y (Figure 2d). However, this process may also be inhibited by such a protein (Figure 2e).

2.4 Boolean Logic Operators

Components in a pathway are dependent on each other. For example if a process requires X and Y to be present for it to proceed, perhaps because they are independently acting cofactors in a given reaction then the process will not proceed unless both present. Alternatively, if a given process can be catalysed by either X or Y, then the process will proceed if either component is present. Such dependencies can be captured using Boolean logic operators which are used to define the relationships between multiple inputs into a process. An 'AND' operator is used when two or more components are required to bring about a process i.e. an event is dependent on more than one factor being present (Figure 2f). In modelling of flow through networks these act in a similar manner to 'bind' process nodes i.e. all inputs must be present before a product is formed or reaction proceeds. In contrast an 'OR' operator is used when one component or another may orchestrate the same change in another component (Figure 2g). For instance multiple kinases e.g. MAP2K3, MAP2K6, MAP2K7 may catalyse the phosphorylation of p38 (MAPK14) and are therefore shown connecting with p38 via an OR operator. OR operators have also occasionally been used to infer that a component(s) can potentially lead to multiple outcomes.

2.5 Depiction of Other Concepts

There are a number of glyphs that represent concepts that do not sit neatly under the headings of being a component, a process or logic operator. These include:

Energy/molecular transfer nodes are used to represent simple co-reactions associated with or required to drive certain processes (e.g. ATP→ADP, GTP→GDP, NADPH→$NADP^+$). They are linked directly to the node representing the process in which they take part (Figure 2h).

Conditional gates are used where there are potentially multiple fates of a component and the output is dependant on other factors such as the components concentration, time or is associated with a cellular state (Figure 2i). These have been used to depict events such as the check point controls in the cell cycle where the decision to go on to the next phase cell replication is under the control of a number of factors and two or more outcomes are possible. Another example is where cholesterol, depending on its intracellular concentration, may be either exported out of the cell or trigger the cholesterol biosynthetic pathway.

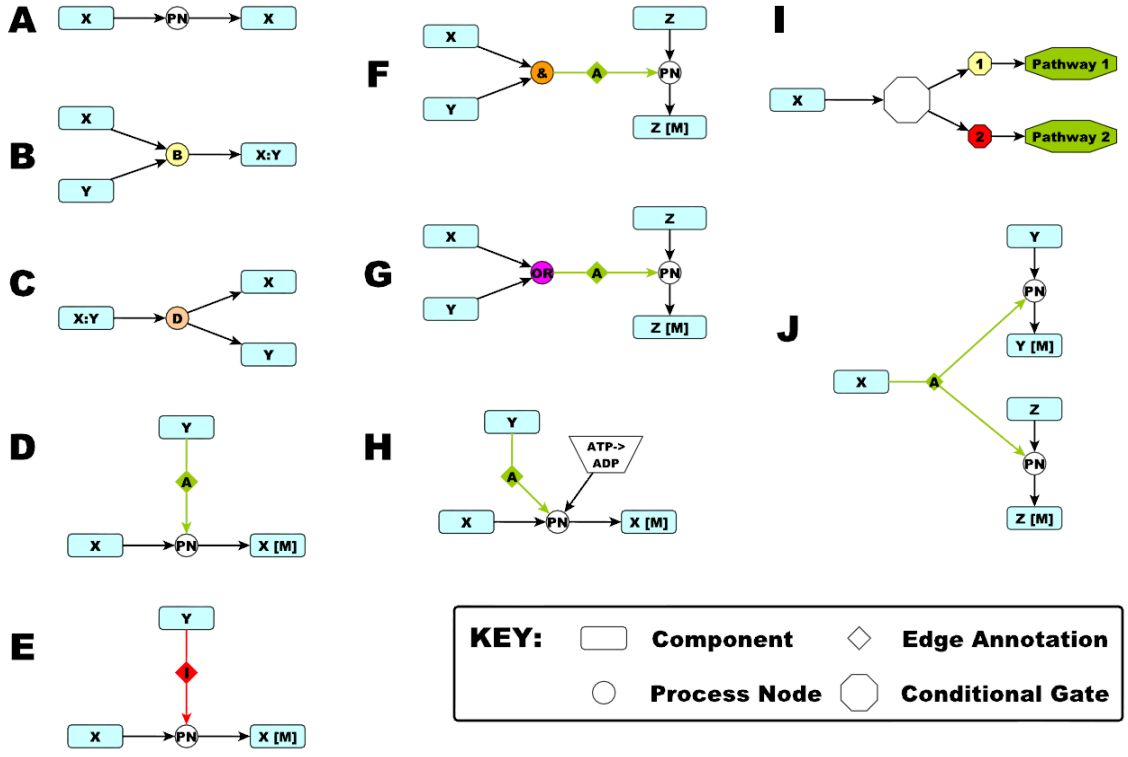


Figure 2. Depiction of Basic Concepts in Pathway Biology using the mEPN scheme. (A) Depiction of the transition of a component from one location or state to another e.g. the translocation of a protein from the cytoplasm to the nucleus or transcription/translation of a gene to protein. (B) Binding (association) of two proteins to form a complex. (C) Dissociation of a complex into its constituent parts. (D) Activation of the transformation of one component by another. (E) Inhibition of the transformation of one component by another. (F) Activation of the transformation of two components by another. (G) Absolute requirement (co-dependency) of two components for the activation of a process. (H) Requirement of either of two components for the activation of a process. (I) Activation of the transformation of one component by another that requires ATP. (J) Depiction of a 'conditional gate' that indicates the start of potentially multiple alternative pathway outcomes which are dependent on other factors. The main

octagon is labelled with the process name e.g. G1 to S phase checkpoint and the other smaller octagons used to denote the factors that influence progression down one pathway or another.

Pathway modules define complicated processes or events that are not otherwise fully described. Examples include signalling cascades, endocytosis, compartment fusion etc. They are a short-hand way of representing molecular events that are not known, not recorded or not shown.

Pathway outputs detail the cumulative output of series of interactions or function of an individual component at the end of a pathway. Pathway outputs are shown in order to describe the significance of those interactions in the context of a biological process or with respect to the cell. The input lines leading into a pathway output node have been coloured light blue to emphasise the end of the pathway description.

2.5 Depiction of Interactions between Components and the Use of Edges

Interactions are depicted by edges, sometimes referred to as lines or arcs (a directional edge). They signify a relationship between components/processes in a pathway and convey the directionality of that interaction. The nature of an interaction is inferred through the use not only of process nodes and Boolean logic operators, but also edge annotation nodes. An edge annotation node is characterised as having only one input (with no arrowhead) and one output and functions to describe the type of activity implied by the line e.g. activation, inhibition, catalysis (Figure 2). A number of notation schemes use of different arrowheads to indicate the 'type' of interaction but their use has been avoided in the mEPN scheme for several reasons; firstly, there is a limit to the number of differing types of arrowheads which potentially falls below the possible number biological concepts one may need to depict. Secondly, differentiating between different arrowheads is sometimes difficult when viewed at a distance. Thirdly, few arrowheads are symbolic or indicative of the action they are designed to describe requiring them to be committed to memory. Finally, multiple arrowhead types are not always supported by different network-editing/visualisation software. Interaction edges may be coloured for visual emphasis but as with nodes, the definition of meaning is not reliant on colour. However, in certain instances they can be used as distribution nodes e.g. where one component activates many others such as with transcriptional activation of a number of genes by a transcription factor it can reduce the number of edges emanating from the transcription factor and therefore simply the representation (Figure 2j and 3). Where separate depiction of modules belonging to the same component is desirable an undirected edge (no arrowhead) is used to denote a physical connection (bond) between two or more components.

2.6 Cellular Compartments

Pathway components exist in different cellular compartments. A cellular compartment can be a region of the cell, an organelle or cellular structure, dedicated to particular processes and/or hosting certain sub-sets of components e.g. genes are found only in the nuclear compartment. In principle a sub-cellular compartment can be any size or shape. Compartments are defined by a labelled background to the pathway and arranged with spatial reference to cell structure. Compartments are coloured differently for emphasis. Similar or related compartments are shown to share the same fill colour but different coloured perimeters. This has been used to differentiate between different but related compartments e.g. different classes of vesicles derived from the endoplasmic reticulum or plasma membrane.

2. Collation of Information and Pathway Assembly

The assembly of a pathway diagram is an extraordinarily interesting and informative exercise. The act of converting text-based information into a visual resource forces one to understand the information that is being presented to a level that the mere reading of an article never requires. When presented with a long textual description of a process involving numerous components all interacting through a complex series of events, it is easy to read about them but far more difficult to construct an accurate picture of them in the mind's eye. Furthermore, the semantics of the written word does not always make sense when drawn, at least not when done in a logical fashion. The art of pathway construction therefore relies on the ability to convert numerous textual descriptions where different words may be used to describe the same or similar processes between multiple components which in turn may or may not be designated the same name, into a concise and unambiguous model of events.

When embarking on the construction of pathway diagram there is a need to define the specific areas that are of interest to you. This sounds obvious but in reading the literature on one system, it is common to find that other systems are discussed (the one big network scenario) and it is easy to stray from the area of original interest. This in itself is not a problem and indeed part of the learning exercise, as long as the area covered has been documented correctly before moving on. The danger is that after a mapping exercise has been 'completed' what results is a sketch covering many components in related systems, where the relationships between them have not been documented to a sufficient level of detail to render the diagram truly useful or informative. It is therefore better to aim for quality over quantity when engaging in this activity. It is also true that what makes sense to the pathway curator does not necessarily make sense to another individual. Great emphasis has should therefore be placed on the need to discuss and justify the information represented to others. If the knowledge gained by the curator can not be communicated clearly and effectively, then they have not done their job properly. Pathway content, adherence to the notation system and layout should always be assessed by others to ensure that the graphical depiction of pathway/interactions is intelligible and unambiguous to another individual familiar with the notation scheme. Ideally the work should also be inspected by those intimately familiar with the field of research that one is attempting to depict, this is always a good test of the accuracy and completeness of the information.

The best source of information about pathways is in buried in the primary literature. However the amount of pathway information that can be gleaned from any one paper is generally limited as a given piece of work will tend to focus only on one or a small number of components and their interacting partners. It is therefore advisable to spend some time gaining a high level view of any given pathway or system of interest. Internet searches for images of the pathway or specific complexes within it provide a framework understanding the pathway of interest. Pathway databases such as Reactome or Kegg (Kanehisa & Goto, 2000; Vastrik et al, 2007) can be used to gain a high level view of the pathway. Interaction databases e.g. String, IntAct, Ingenuity, HPRD or Bind (Alfarano et al, 2005; Hermjakob et al, 2004; Jensen et al, 2009; Mishra et al, 2006) might also be used to gain a view of molecular interactions of a given component. Our experience however has been that such resources present such a generic network view of pathways and often capture seemingly erroneous interactions thereby limiting their utility for this purpose. One of the best starting points is literature reviews. Whilst they frequently discuss information of limited use to pathway construction e.g. concerning protein structure, evolution of protein families, high level concepts etc., they frequently provide graphical depictions of subsystems and are an excellent portal into the primary literature. The point is not to get too

involved too early but to take snapshot of the current understanding of the system and construct a framework of understanding and sources of available information prior going into detail.

During the course of pathway mapping exercise many papers will be read and snippets of information will be mentally recorded concerning all aspects of pathway biology. It is important to have mechanisms in place that allow the curator to record this information and its source otherwise all this information will be lost. Evidence to support an interaction derived from the primary literature (and reviews) must be recorded in an interaction table. This must include the identity of interacting partners, the direction of the interaction e.g. HGNC1 -> HGNC2, the type of interaction (phosphorylation, cleavage), method by which the interaction was determined, PubMed ID of the paper reporting the interaction and site of specific change of state e.g. phosphorylation of Serine 123. Of course more than one paper may be used to support the same interaction and arguably two or more references are preferable to a single work reporting an interaction. Indeed no interaction should be included within the pathway without published evidence to back it up. An example of a pathway interaction table is shown in reference Table 1. Additional notes and hyperlinks to external databases are also useful in linking additional information on the biology depicted. Graphml files support this activity and pathway diagrams may include URL-links to Entrez Gene (or other database of choice) for each protein or gene component in the pathway. Furthermore, component descriptions obtained from databases, PubMed IDs and textual descriptions can be included and stored on appropriate edges or nodes. These can be accessed under the properties-description tab for nodes or edges, or appear when hovering over a node or edge thereby supplementing what is shown graphically.

	Interaction No.	1	2	3	4	5	6
Interacting Partner 1	Official Gene Symbol	ATM	ATM:IKBKG	ATM:IKBKG	BCL2	CDC37	CHUK
	Gene ID	472	472:4214	472:4214	596	11140	1147
	Interactant Type	Protein	Complex	Complex	Gene	Protein	Protein
	Interactant as on map	ATM[P]	ATM[P]:IKBKG[P][Ub]	ATM[P]:IKBKG[P][Ub]	BCL2	CDC37	CHUK
Interacting Partner 2	Official Gene Symbol	IKBKG	CHUK	ERC1	NFKB1(p50):NFKB1(p50)	HSP90AA1	CHUK
	Gene ID	4214	1147	23085	N/A	3320	1147
	Interactant Type	Protein	Protein	Protein	Complex	Protein	Protein
	Interactant as on map	IKBKG[SU]	CHUK	ERC1(ELKS)	NFKB1(p50):NFKB1(p50)	HSP90AA1	CHUK
Interaction Type		Binding	Binding	Binding	Activation	Binding	Binding
Interaction Location		nucleus	cytoplasm	cytoplasm	nucleus	cytoplasm	cytoplasm
NCBI-PubMed ID		16497931	16497931	16497931	14668329	15371334	15145317

Table 1. Example of the information that should be stored when recording an interaction associated with the construction of a pathway.

As a final note on pathway construction, it should be emphasised that the visualisation of specific events as well as overall layout of a diagram is everything in ensuring the pathways usability. Under the mEPN system each step in a given process is explicitly depicted. For example if the activation of given signalling pathway requires the receptor complex to go through a series of changes e.g. binding, phosphorylation or dissociation events following ligand binding,

then each intermediate stage should ideally be shown (see Figure 3). Whilst this can make the depiction of events long winded it accurately reflects what is known and may ultimately be important in understanding the pathways regulation. Another important rule is that although a given pathway component can play a role in numerous different processes it may only be represented once in any given cellular compartment. Whilst this rule can potentially lead to a tangle of edges due to certain components possessing numerous connections to other components spread across the pathway, the benefits of the rule outweigh the issues in adhering to it. The number of edges leaving each node gives the reader an exact indication of a component's interactions with other components and hence potential activity, without the need for scanning the entire diagram to find other instances where the component is described. A component may however be shown more than once in a given cellular compartment if it changes from one state to another e.g. from an inactive form to an active form, in which case both forms are represented as separate components.

As a general rule nodes (components, processes, operators) and edges (interactions) should be drawn in such a way as to make the diagram compact with a minimum about of crossing over, changes in direction of edges and length i.e. edges should be easy to follow. Hierarchical relationships between components should be shown in the layout of interactions. In order to do this an orientation of pathway flow is chosen e.g. left to right or top to bottom and where possible should be maintained throughout the diagram. Ideally the direction of interactions should follow the 'flow' through the pathway, although it is appreciated this becomes more difficult in larger diagrams. A certain degree of consistency should also be aimed for when depicting components and their interactions e.g. components should be depicted using nodes of a similar size, similar pathway relationships should be drawn in a consistent manner. Visual clarity relies on a 'clean' layout of pathways and whilst there are a number of automated algorithms available for network layout, they are currently no substitute for a curator with an attention to detail and an artistic eye.

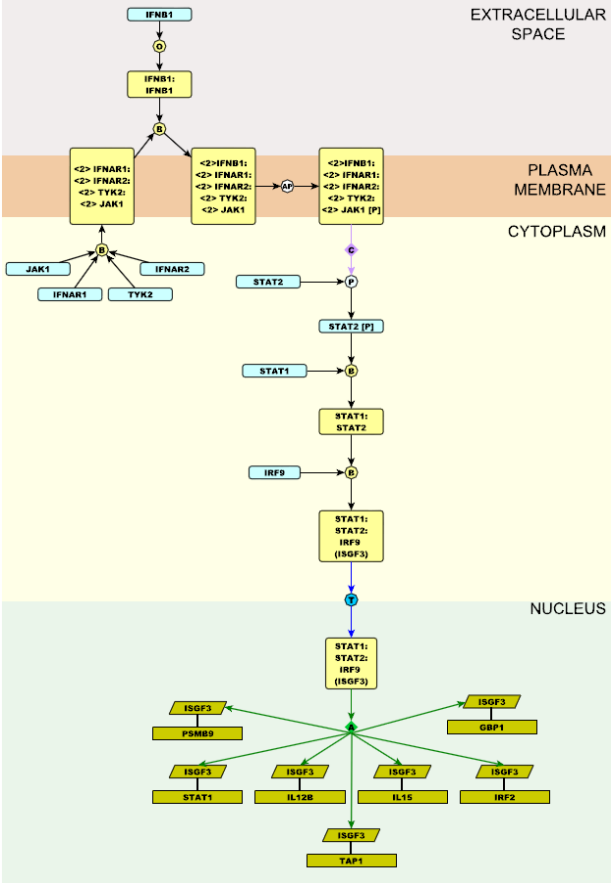


Figure 3. Example of a small pathway depicted using the mEPN scheme. Interferon B (IFNB) is a cytokine released from many cell types in response to immune stimulation. It homodimerises and binds to a cell surface receptor complex composed of the receptor proteins IFNAR1 and IFNAR2 and the intracellular kinases TYK2 and JAK2. The complex is composed of 2 of each of these proteins. Binding causes a conformation change in the complex resulting in the autophosphorylation of JAK1. Once activated the complex catalyses the phosphorylation of STAT2 to which forms a heterodimer with STAT1. This complex then binds interferon regulatory factor 9 (IRF9) forming the complex often referred to as ISGF3 and translocates to the nucleus. Here it binds to the ISGF3 element in the promotor of a number of genes including IRF2, IL12B, STAT1, IL15, TAP1, GBP1, PSMB9 initiating their transcription. For a more detailed view of this and other immune-related pathways, see Raza et al 2008, 2010.

3. Summary

The networks of molecular interactions that underpin cellular function are highly complex and dynamic. The topology, behaviour and logic of these systems, even on a relatively small scale, are far too complicated to understand intuitively. Formalised models provide a possible solution to the problem. However the challenge of creating models that reflect our current understanding of these systems and display this information in an intuitive and logical manner is not trivial. The task of constructing pathway diagrams is time consuming and laborious involving many hours of work. On the other hand, it summarises the results of investigations that may have taken many 1000's of hours of time to perform and it is difficult to envisage how one could précis such a body of work in any other meaningful way. The act of creating a pathway model forces you to formalise what you know about a system and justify it using appropriate sources. It allows you to explore the nature of relationships that might have existed as mental picture but the need to graphically depict them in a formalised way is in itself highly informative. As well as defining what you do know about a system, it is equally useful in defining what you don't.

The mEPN scheme described here provides a system where pathways can be represented in a logical, unambiguous and biologist-friendly fashion, whatever the system of interest. What we would like to see and believe is essential, is the support of the wider community in assembling and editing such diagrams. Such efforts are underway (Pico et al, 2008; Schaefer et al, 2009; Vastrik et al, 2007) and are already providing a vital forum for debate on the known details of pathways in different cell systems. Ideally they these efforts will result in detailed models of biological systems that can be shared and assimilated. However, in order to achieve this end pathway models clearly need to be assembled using standard rules and graphical languages. We

therefore hope our work will contribute to the ongoing community effort to develop such standards (Le Novère, 2009).

To gain a systems level view of these pathways is to gain an insight into the molecular networks that regulate normal function and whose malfunction underpins disease pathology. Greater understanding of the overall architecture of the pathways and their susceptibility to deregulation by disease causing agents, should ultimately lead to new strategies and targets for therapeutic intervention. For my group the creation of pathway models has provided a resource for training, computational pathway modelling, literature/data interpretation, computational pathway modelling and hypothesis generation. As such the approach is now central to our ongoing investigations of macrophage biology and has transformed the way we think about these cells and our interpretation of results of investigations into their immune biology.

References

1. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, Buzadzija K, Cavero R, D'Abreo C, Donaldson I, Dorairajoo D, Dumontier MJ, Dumontier MR, Earles V, Farrall R, Feldman H et al (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33:** D418-424

2. Antonov AV, Dietmann S, Mewes HW (2008) KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol* **9:** R179

3. Arakawa K, Kono N, Yamada Y, Mori H, Tomita M (2005) KEGG-based pathway visualization tool for complex omics data. *In Silico Biol* **5:** 419-423

4. Babur O, Colak R, Demir E, Dogrusoz U (2008) PATIKAmad: putting microarray data into pathway context. *Proteomics* **8:** 2196-2198

5. Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic Acids Res* **34:** D504-506

6. Calzone L, Gelay A, Zinovyev A, Radvanyi F, Barillot E (2008) A comprehensive modular map of molecular interactions in RB/E2F pathway. *Mol Syst Biol* **4:** 173

7. Cassman M (2005) Barriers to progress in systems biology. *Nature* **438:** 1079

8. Cavalieri D, De Filippo C (2005) Bioinformatic methods for integrating whole-genome expression results into cellular networks. *Drug Discov Today* **10:** 727-734

9. Cook DL, Farley JF, Tapscott SJ (2001) A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biol* **2:** RESEARCH0012

10. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* **31:** 19-20

11. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T (2007) Pathway mapping tools for analysis of high content data. *Methods Mol Biol* **356:** 319-350

12. Eungdamrong NJ, Iyengar R (2004) Modeling cell signaling networks. *Biol Cell* **96:** 355-362

13. Freeman TC, Goldovsky L, Brosch M, van Dongen S, Maziere P, Grocock RJ, Freilich S, Thornton J, Enright AJ (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol* **3:** 2032-2042

14. *Freeman T.C.*, Raza S., Theocharidis A. and Ghazal P. The mEPN Scheme: An Intuitive and Flexible Graphical System for Rendering Biological Pathways. *BMC BMC Systems Biology* **4**:65, 2010

15. Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H (2008) CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings of the IEEE* **96:** 1254-1265

16. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, Margalit H, Armstrong J, Bairoch A, Cesareni G, Sherman D, Apweiler R (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32:** D452-455

17. http://www.biopax.org/. Biological Pathways Exchange

18. http://www.ingenuity.com/. Ingenuity Pathway Analysis.

19. http://www.yworks.com. yEd Graph Editior - yWorks the diagramming company.

20. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin, II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19:** 524-531

21. Janssen P, Audit B, Cases I, Darzentas N, Goldovsky L, Kunin V, Lopez-Bigas N, Peregrin-Alvarez JM, Pereira-Leal JB, Tsoka S, Ouzounis CA (2003) Beyond 100 genomes. *Genome Biol* **4:** 402

22. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37:** D412-416

23. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33:** D428-432

24. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27-30

25. Kitano H (2002) Computational systems biology. *Nature* **420:** 206-210

26. Kitano H, Funahashi A, Matsuoka Y, Oda K (2005) Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol* **23:** 961-966

27. Kohn KW (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol Biol Cell* **10:** 2703-2734

28. Kwiatkowska MZ, Heath JK (2009) Biological pathways as communicating computer systems. *J Cell Sci* **122:** 2793-2800

29. Le Novère N, Hucka M, Mi H, Moodie S, Shreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H. (2009) The Systems Biology Graphical Notation. *Nature Biotechnology* **27:** 735-741

30. Lloyd CM, Halstead MD, Nielsen PF (2004) CellML: its future, present and past. *Prog Biophys Mol Biol* **85:** 433-450

31. Luciano JS (2005) PAX of mind for pathway researchers. *Drug Discov Today* **10:** 937-942

32. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS et al (2006) Human protein reference database--2006 update. *Nucleic Acids Res* **34:** D411-414

33. Moodie SL, Sorokin A, Goryanin I, Ghazal P (2006) A Graphical Notation to Describe the Logical Interactions of Biological Pathways. *Journal of Integrative Bioinformatics* **3:** 11

34. Novere NL, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villeger A, Boyd SE, Calzone L, Courtot M et al (2009) The systems biology graphical notation. *Nat Biotechnol* **27:** 735-741

35. Nurse P (2003) Systems biology: understanding cells. *Nature* **424:** 883

36. Oda K, Kimura T, Matsuoka Y, Funahashi A, M. M, Kitano H. (2004) Molecular Interaction Map of a Macrophage. *The Alliance for Cellular Signaling (AfCS) Research Reports*, Vol. 2.

37. Oda K, Kitano H (2006) A comprehensive map of the toll-like receptor signaling network. *Mol Syst Biol* **2:** 2006 0015

38. Pandey R, Guru RK, Mount DW (2004) Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics* **20:** 2156-2158

39. Pavlopoulos GA, Wegener, A-L., Schneider, R. (2008) A survey of visualization tools for biological network analysis
40. *BioData Mining* **1:** 12

41. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C (2008) WikiPathways: pathway editing for the people. *PLoS Biol* **6:** e184

42. Pirson I, Fortemaison N, Jacobs C, Dremier S, Dumont JE, Maenhaut C (2000) The visual display of regulatory information and networks. *Trends Cell Biol* **10:** 404-408

43. Raza S, Robertson KA, Lacaze PA, Page D, Enright AJ, Ghazal P, Freeman TC (2008) A logic-based diagram of signalling pathways central to macrophage activation. *BMC Syst Biol* **2:** 36

44. Raza S., McDerment N., Lacaze P.A., Robertson K., Watterson S., Chen Y., Chisholm M., Eleftheriadis G., Monk S., O'Sullivan M., Turnbull A., Roy D., Theocharidis A., Ghazal P. and *Freeman T.C.* Construction of a Large Scale Integrated Map of Macrophage Pathogen Recognition and Effector Systems. BMC *Systems Biology* **4**:63, 2010.

45. Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. *Nat Rev Genet* **7:** 130-141

46. Ruths D, Muller M, Tseng JT, Nakhleh L, Ram PT (2008) The signaling petri net-based simulator: a non-parametric strategy for characterizing the dynamics of cell-specific signaling networks. *PLoS Comput Biol* **4:** e1000005

47. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res* **37:** D674-679

48. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky-Lazareva B, Muruganujan A, Rabkin S, Vandergriff JA, Doremieux O (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res* **31:** 334-341

49. van Riel NA (2006) Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief Bioinform* **7:** 364-374

50. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* **8:** R39

51. Watterson S, Marshall S, Ghazal P (2008) Logic models of pathway biology. *Drug Discov Today* **13:** 447-456

52.Yeung N, Cline MS, Kuchinsky A, Smoot ME, Bader GD (2008) Exploring biological networks with Cytoscape software. *Curr Protoc Bioinformatics* **Chapter 8:** Unit 8 13

7. RNAi

RNAi Practical Session

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves should be worn during all practical sessions

M9 Buffer –irritant
2 x TY + ampicillin - harmful

7 RNAi

Introduction

RNAi in outline

RNA-mediated interference (RNAi) is one of the easiest and most direct ways to investigate the *in vivo* functions of a novel gene. In outline, RNAi allows one to knock down the expression of any target gene and ask what happens. Any gene of known sequence can be targeted by RNAi — dsRNA of sequence identical to the target gene is introduced into the cells or organism and this ultimately causes the specific degradation of the target gene mRNA by the RISC complex. RNAi works in all animal model organisms as well as mammalian cells in culture and thus is widely applicable. The technical details of how to do RNAi varies greatly between different organisms — some of these will be covered in this course.

As well as being a great tool to look at the functions of single genes, we can use RNAi to screen many genes to find which ones affect a particular process. This is basically like doing a classical genetic screen but without the need for positional cloning — a big advantage! Genome-wide RNAi screens have been carried out in worms and flies to identify genes involved in processes as diverse as development, fat metabolism, aging, cell morphology, and DNA repair. Reagents are also available for targeting over a third of predicted human genes for RNAi, and RNAi screens at last bring the power of the genetic screen to mammalian cells in culture. Large-scale RNAi screens allow us to take an unbiased look at gene function, and will certainly be an excellent way of exploring gene functions at the scale of entire genomes.

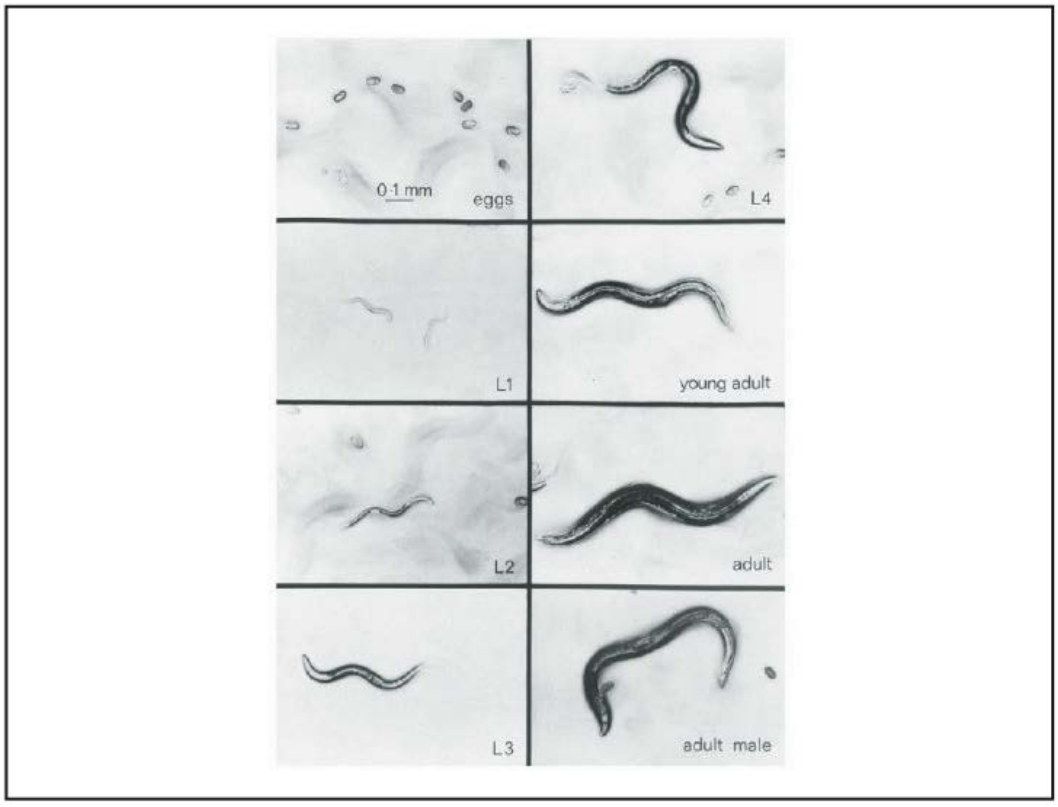
In this course, we will be doing RNAi screens in *C. elegans*. RNAi screens in the worm are fairly straightforward, sort-of foolproof, and should illustrate how rapidly RNAi screens can shed light on the molecular biology underlying complex phenotypes and processes.

Worm biology — a little aside

First off — don't worry! Worms are easy — all you need to do to be able to grow and use worms is the ability to inoculate a bug culture, and they will do all the rest for you. *C. elegans* is a self-fertilising hermaphrodite with a three day life-cycle — in practical terms, this means if you leave a single adult worm alone on a plate, three days later you will have 200-300 new adult worms. They are very robust — if they starve they will still be

fine left alone for a couple of months; if they get contaminated, you can clean the worms in a few minutes using bleach; and if you get so bored of them that you don't want to see them for a while, you can freeze them as easily as bugs or yeast. This is all good! In addition, every worm is identical, so once you've seen one you've seen them all. Learning worm biology is thus much easier and you do not really have to worry about too much variation.

We generally grow worms at ~20°C on NGM agar plates (NGM is essentially a buffered salt medium) which have a lawn of OP50 bacteria (a slow-growing strain of *E. coli*) as food. They can also be grown in liquid culture, and we will be doing both. Embryogenesis lasts ~24hrs after which the worms proceed through 4 larval stages (L1 to L4) before they become adult. This entire cycle lasts ~72hrs at ~20°C. Images of worms at different stages are shown below to give a rough idea of sizes and morphologies.



The different stages in the worm life cycle.

If embryos are harvested on Day 1, you will see L1-L3 larvae by the end of Day 2 and L4 and young adults by the end of Day 3. The germline starts to develop at the L3 stage

and the animal is fully fertile only at the adult stage. An average adult lays ~300 embryos; this number is limited by the number of sperm, rather than oocytes.

RNAi in C. elegans

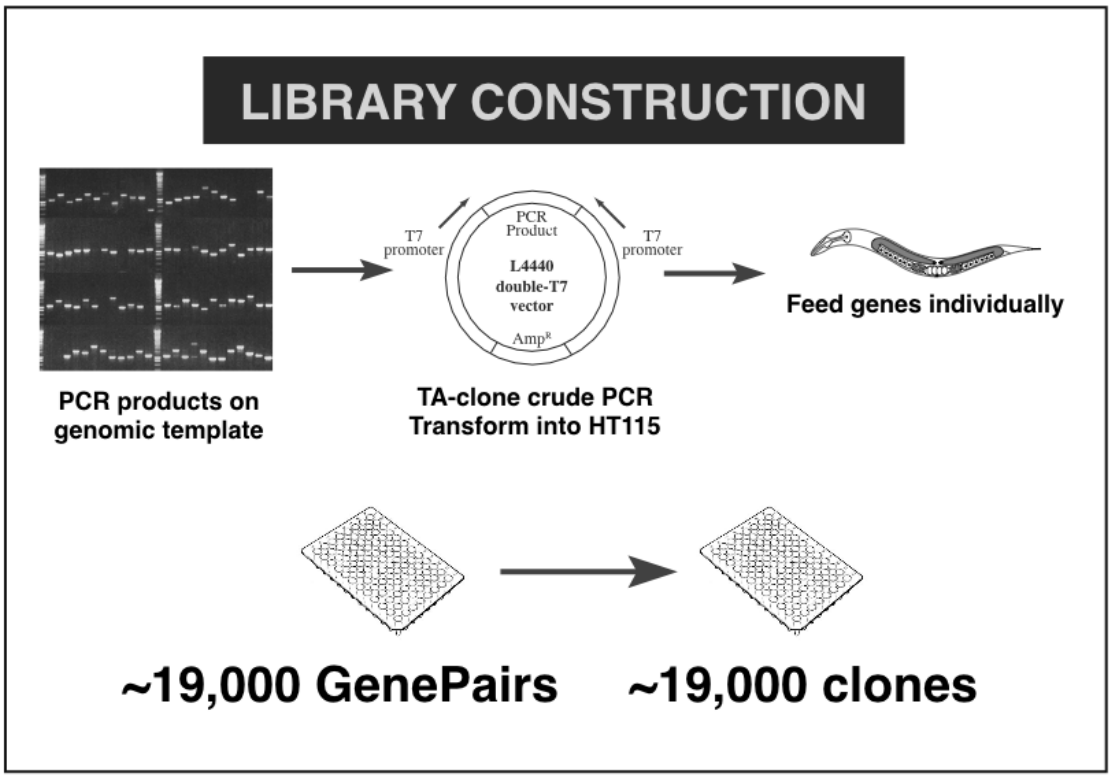
An outline : injection, soaking and feeding techniques

Much of the pioneering work on RNAi has been done in *C. elegans* including the first genome-wide RNAi screens. In part, this is because RNAi is very easy in the worm: almost any way of getting dsRNA into the animal will generate a potent and specific knock down of the target gene. dsRNA can be made *in vitro* and injected into the gonad adult worm — this causes a very robust knock down in the progeny of the injected animal, but is obviously labour intensive. Alternatively, adult worms can be soaked in dsRNA containing solution — again this is very effective in generating knock downs in the progeny, but this becomes expensive if you want to screen many genes since *in vitro* RNA synthesis is still relatively costly. Lastly, and most easily, if worms are fed on a diet of bugs expressing dsRNA identical to a target gene, this also generates a potent RNAi effect. Precisely how the dsRNA is transmitted from the gut (where it ends after the worms eat the bugs) into the gonad (where it must be to affect the next generation) is mysterious although some components of the machinery involved are known. The key point, though, is that feeding bugs to worms is very easy — RNAi by feeding is thus a very cheap and cheerful way of doing RNAi in worms and a researcher can analyse the RNAi phenotypes of hundreds of genes a day using this technique. This is how you will doing RNAi on the course.

The RNAi library

RNAi in the worm is made still easier since there is a publicly available library of ~17,000 dsRNA-expressing bacterial strains that covers ~85% of all predicted worm genes. Each bacterial strain targets a single predicted gene and thus this library allows the researcher to examine the phenotypes of almost all *C. elegans* genes just by feeding each bugs strain to the worms. We will be using strains from this library for all the RNAi experiments.

Technically, each bacterial strain contains a plasmid in which a fragment of *C. elegans* genomic DNA spanning one predicted gene has been cloned in the vector shown in over page.



Construction and design of the *C. elegans* RNAi library
Primers were designed to amplify fragments from *C. elegans* genomic DNA that overlapped with the predicted coding region of each predicted gene. The fragments were cloned into L4440; this contains a T7 promoter to drive transcription of both sense and antisense strands. These constructs were transformed into the HT115 bacterial strain and individual clones were fed to worms.

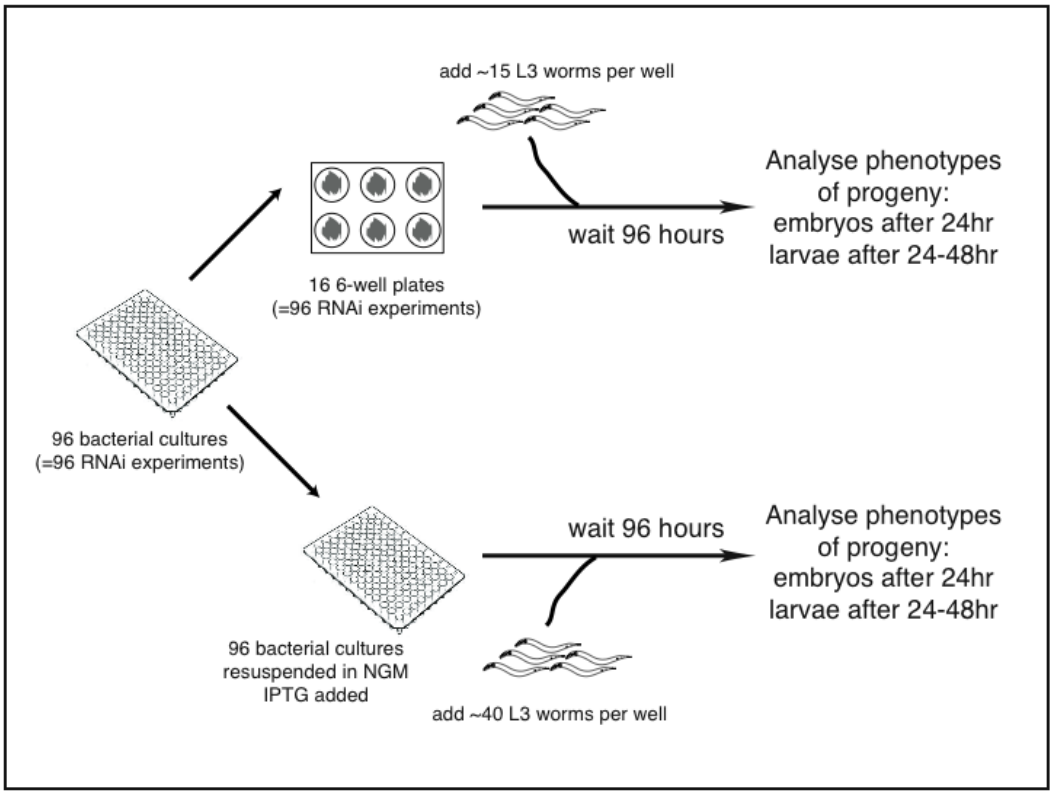
These fragments range from 300bp – 3kb with a mean size of 1.3kb. The bacterial strain expresses T7 RNA polymerase following IPTG induction — this transcribes both sense and antisense strands of the inserted *C. elegans* genomic fragment, and these anneal in the bacteria. The bacterial strain has also been engineered so it lacks RNAse III — the dsRNA expressed in the bacteria is therefore not degraded. RNAi using the library is therefore very simple. All you need to do is grow bacteria, add IPTG to induce dsRNA expression and feed the bugs to the worms.

RNAi by feeding: plates versus liquid

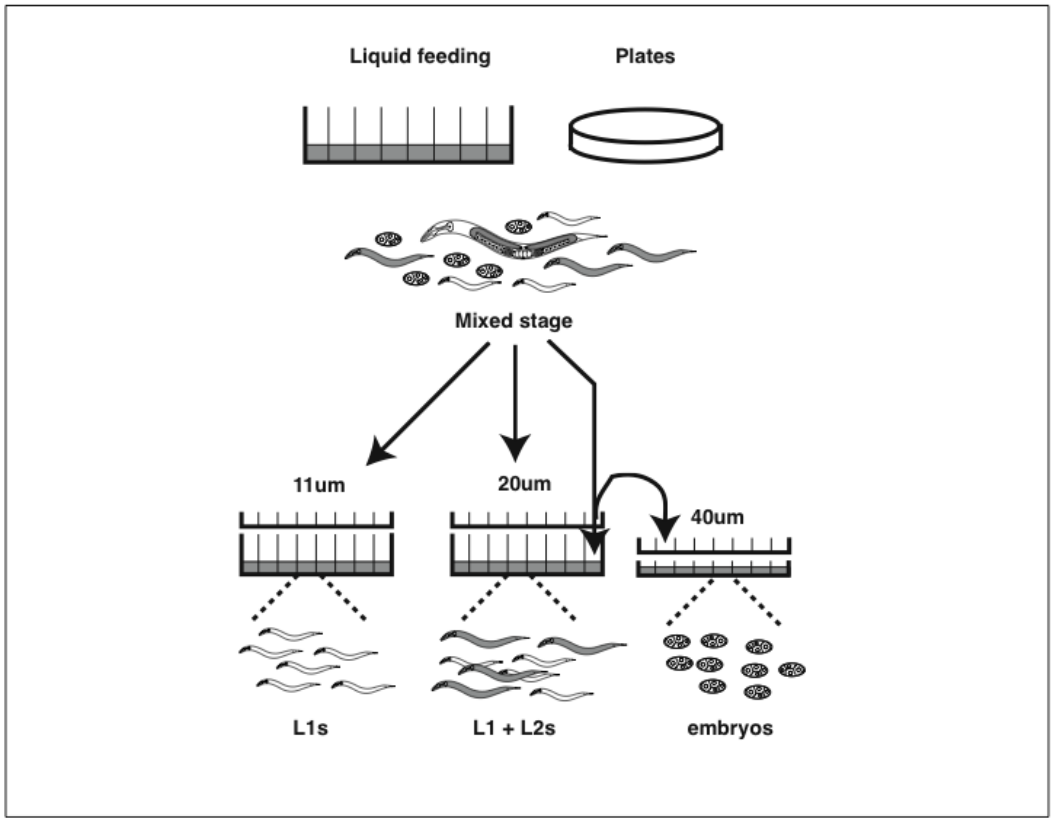
The most common way to do RNAi by feeding is on plates. Each plate contains Amp (to maintain the plasmid carrying the genomic fragment that will be transcribed into dsRNA)

and IPTG (to induce dsRNA expression). The bacterial strains to be fed are grown up to saturation in liquid and spotted on these plates to make a bacterial lawn on which the worms will feed. In general a single well of a 12-well or a 6-well plate will hold enough bugs to feed several adult worms and all their progeny through to the next generation; each well is thus a different RNAi experiment. Worms are typically placed on the bacterial lawns when they are at the L3-L4 stage, then left to feed for 96hours. By this time, they are adults and have laid embryos — it is these embryos that we study to look for effects, both pre- and post-hatching. We look at the phenotypes in the progeny of the fed worms rather than in the worms themselves since in this way we can monitor the effect on development, and also since it allows us to observe the maximal RNAi effect.

Feeding on plates generates robust phenotypes and has the advantage that worms are most easily observed on plates, so it is straightforward to analyse phenotypes this way. However, feeding on plates cannot be done in 96-well format which is a pain. For this, liquid feeding is far easier. Bugs are grown up in 96-well format (1 well per gene to be targeted), spun down and resuspended in NGM buffer (so the worms are happy during feeding) and IPTG is added to induce dsRNA. ~40 worms are then added per well, and are allowed to eat the bugs for 96 hours at 16°C. At this point, the phenotypes of the progeny can be analysed. The downside of liquid feeding is that worms cannot be analysed directly in the bacterial suspension, but need to be purified in some manner. We use a system of meshes (see figure below) to purify worms of different stages from liquid cultures; they can then be observed either directly or by placing on plates.



Comparison of RNAi feeding protocols using plates and liquid.
In both methods, worms are exposed to dsRNA-expressing bugs at the L3 stage and allowed to feed on the bugs for 96hrs while they develop into adults and begin laying embryos. The embryos laid after 72-96 hours are most strongly affected and it is these that we analyse.



Using meshes to isolate purified populations of worms from mixed stage cultures.
L1s can be purified by spinning through an 11um mesh, L1 and L2s through a 20um mesh. Embryos are larger than either L1 or L2s and can be purified by first filtering with a 20um mesh — this retains embryos, L3 and L4 larvae and adults, but discards L1 and L2s. Embryos then pass through a 40um mesh whereas paralysed worms cannot.

RNAi screens

You will be carrying out RNAi screens for three different phenotypes, all of which are non-viable: embryonic lethality, sterility and growth defects. Each person will screen 100 genes in the lethality/sterility assay. These genes are all chosen since they are thought to play roles in signal transduction or transcriptional regulation based on their GO annotation or on their Pfam domains (e.g. tyrosine kinases, bHLH transcription factors etc).

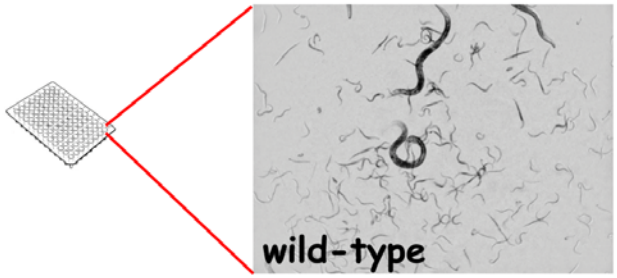
Embryonic lethality and adult sterility

These are the easiest assays of all — they cannot fail! We will be feeding the same genes in several different worm strains: the first is a wildtype, while the other carfry mutations in different genes. The object of this experiment is identify genes that have a different degree of lethality and/or sterility in the mutant backgrounds compared with wild-type animals. If there are differences, we can infer some redundancy or functional connection between the gene mutated in the mutant strain and the target gene. This kind of genetic interaction provides valuable pathway information and lets us probe redundancy.

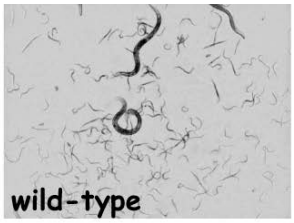
The screens are easy to set up. All you need is to grow bugs overnight, then add the cultures to synchronised populations of L1 larvae in 96 well format. We carry out each experiment in duplicate to minimise variation and noise.

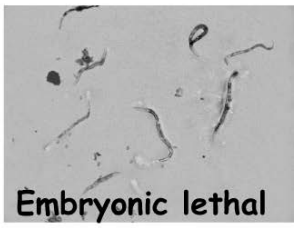
In outline, we add ~15 L1 larvae per well, let these grow to become adults while eating the bugs, then allow them to lay the next generation and for these to develop for a day or so. At this point, we have a slightly mixed stage population that has eaten almost all the bugs so we can see the worms easily (see fig below). In the fig at the top, you can see a wildype population; in the fig below that you can see various mutant phenotypes: embryonic lethal, sterile and growth defects. This is what you will be looking for. We will set up duplicates of wildtype worms and mutants, so for each gene screened by RNAi you will analyse 4 wells, looking for a consistent phenotype, and (if there is one) a consistent difference between the strains…

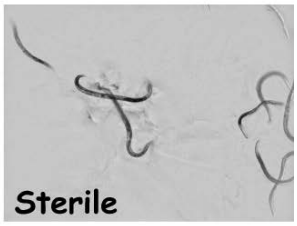
High throughput RNAi screening in worms



~2500 genes per person per day
Detects >85% of previously described non-viable/growth defects
>90% reproducibility







Protocols

The obvious

You are generating RNAi phenotypes by feeding bugs to worms. The more of your bugs they eat, the better your phenotype — the more contamination they eat, the worse it will be. Most frequent source of contamination is usually the worms, so be careful when manipulating these. Also, the only way you know which gene you fed is via its position on a 96-well plate. Aligning A1 with A1 is good; A1 with H12 is bad.

Purifying L1 worms

Wash worms off plates with M9 buffer into a 15ml falcon.
Spin 1000rpm, 1 min.
Take off supernatant. (NB worms will not be in a hard pellet! Leave the last ~1ml when you aspirate…and do NOT try to pour off the snt — this is unwise!)
Resuspend in 3.5 ml.
Place a 96-well 10um mesh onto a 2.2ml 96-well plate – this should fit on nicely.
Pipette 0.3ml of worm suspension per well across one row. Spin for 30s, 1000rpm.
Pool flow-through into a 15ml — here are your worms. Top up to 15ml and give them one more 1000rpm, 1 min spin. Discard snt.
Count the number of L1s (take 2 10ul samples). (NB. Worms are big. They settle pretty quickly. Agitate the worm suspension pretty often otherwise your numbers will go nuts.)

Feeding on plates

Bug cultures

Grow up 800ul cultures in 2xTY + 100ug/ml Amp in 2.2 ml 96-well plates for 8hrs at 37°C with vigorous shaking. Inoculate the cultures from the bacterial spots using a multi-channel pipette – take as much of the spot as you can, since your cultures will grow faster. Seal the 2.2ml plates with the breathable membrane.

Spotting plates

You will need to spot the cultures onto the 6 well plates. These contain Amp + 1mM IPTG for induction of dsRNA expression. Each well needs to be inoculated with a single culture — you do NOT want to flood the well when you do this since it will not dry! About 4-5 drops from a P1000 blue tip is plenty. Leave these to dry overnight by leaving the lids skew on the plates. Label the plates, NOT the lids to avoid future misery and pain.

Adding worms

Add ~40 L1 animals per well, after purification with the mesh. (NB. Worms are big. They settle pretty quickly. Agitate the worm suspension pretty often otherwise your numbers will go nuts.)

Removing worms

After 96hrs, you have adults + larvae + embryos on the plates. To help you score phenotypes, it is easiest when you have a synchronous population. We therefore remove all the worms and leave the embryos. You could do this in many ways (one-by-

one by hand, anyone?) but we exploit the fact that embryos stick to agar and bugs, whereas worms swim.
Gently pour ‘a bit’ of M9 into each well — this should be enough to cover the bottom easily. Give it a gentle rock and flip off the liquid. If you do this right, the bugs and embryos will all be there, and the worms will not.

Feeding in liquid

Bug cultures

Grow up 800ul cultures in 2xTY + 100ug/ml Amp in 2.2 ml 96-well plates overnight at 37C with vigorous shaking. Inoculate the cultures from the bacterial spots using a multi-channel pipette – take as much of the spot as you can, since your cultures will grow faster. Seal the 2.2ml plates with the breathable membrane.

The next day, you have plenty bugs and need to induce dsRNA expression. Add IPTG to 4mM and leave at RT for 90mins. You are now ready to add these to the worms.

Adding worms

Add ~15 L1 animals per well, after purification with the mesh. (NB. Worms are big. They settle pretty quickly. Agitate the worm suspension pretty often otherwise your numbers will go nuts.)

Recipes

NGM Agar

NaCl 3 g
agar 17 g
peptone 2.5 g
H2O 975 ml
Autoclave; then, using sterile technique, add the following and mix after each addition.
cholesterol (5 mg/ml in EtOH) 1 ml
CaCl2 1M 1 ml
MgSO4 1M 1 ml
potassium phosphate 1M pH6 25 ml

NGM liquid

NaCl 3 g
peptone 2.5 g
H2O 975 ml
Autoclave; then, using sterile technique, add the following and mix after each addition.
cholesterol (5 mg/ml in EtOH) 1 ml
CaCl2 1M 1 ml
MgSO4 1M 1 ml
potassium phosphate 1M pH6 25 ml

M9 Buffer

KH2PO4 3 g
Na2HPO4 6 g
NaCl 5 g
MgSO4 1M 1 ml
H2O 1 liter

References

The basic stuff

Technology

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998 Feb 19;391(6669):806-11.

Bernstein E, Caudy AA, Hammond SM, Hannon GJ. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature. 2001 Jan 18;409(6818):363-6.

Hammond SM, Bernstein E, Beach D, Hannon GJ. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature. 2000 Mar 16;404(6775):293-6.

Tuschl T, Zamore PD, Lehmann R, Bartel DP, Sharp PA. Targeted mRNA degradation by double-stranded RNA in vitro. Genes Dev. 1999 Dec 15;13(24):3191-7.

Hammond SM, Bernstein E, Beach D, Hannon GJ. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature. 2000 Mar 16;404(6775):293-6.

Zamore PD, Tuschl T, Sharp PA, Bartel DP. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell. 2000 Mar 31;101(1):25-33.

Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature. 2001 May 24;411(6836):494-8.

Paddison PJ, Caudy AA, Bernstein E, Hannon GJ, Conklin DS. Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev. 2002 Apr 15;16(8):948-58.

Kunath T, Gish G, Lickert H, Jones N, Pawson T, Rossant J. Transgenic RNA interference in ES cell-derived embryos recapitulates a genetic null phenotype. Nat Biotechnol. 2003 May;21(5):559-61.

Hemann MT, Fridman JS, Zilfou JT, Hernando E, Paddison PJ, Cordon-Cardo C, Hannon GJ, Lowe SW.
An epi-allelic series of p53 hypomorphs created by stable RNAi produces distinct tumor phenotypes in vivo. Nat Genet. 2003 Mar;33(3):396-400.

Capodici J, Kariko K, Weissman D. Inhibition of HIV-1 infection by small interfering RNA-mediated RNA interference. J Immunol. 2002 Nov 1;169(9):5196-201.

Coburn GA, Cullen BR. Potent and specific inhibition of human immunodeficiency virus type 1 replication by RNA interference. J Virol. 2002 Sep;76(18):9225-31.

Jacque JM, Triques K, Stevenson M. Modulation of HIV-1 replication by RNA interference. Nature. 2002 Jul 25;418(6896):435-8.

Rubinson DA, Dillon CP, Kwiatkowski AV, Sievers C, Yang L, Kopinja J, Rooney DL, Ihrig MM, McManus MT, Gertler FB, Scott ML, Van Parijs L. A lentivirus-based system to functionally silence genes in primary mammalian cells, stem cells and transgenic mice by RNA interference. Nat Genet. 2003 Mar;33(3):401-6. Epub 2003 Feb 18.

Screens

First mammalian

Aza-Blanc P, Cooper CL, Wagner K, Batalov S, Deveraux QL, Cooke MP. Identification of modulators of TRAIL-induced apoptosis via RNAi-based phenotypic screening. Mol Cell. 2003 Sep;12(3):627-37.

First big human

Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, Agami R, Ge W, Cavet G, Linsley PS, Beijersbergen RL, Bernards R. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature. 2004 Mar 25;428(6981):431-7.
and
Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, Balija V, O'Shaughnessy A, Gnoj L, Scobie K, Chang K, Westbrook T, Cleary M, Sachidanandam R, McCombie WR, Elledge SJ, Hannon GJ. A resource for large-scale RNA-interference-based screens in mammals. Nature. 2004 Mar 25;428(6981):427-31.

First barcode

Brummelkamp TR, Fabius AW, Mullenders J, Madiredjo M, Velds A, Kerkhoven RM, Bernards R, Beijersbergen RL. An shRNA barcode screen provides insight into cancer cell vulnerability to MDM2 inhibitors. Nat Chem Biol. 2006 Apr;2(4):202-6. Epub 2006 Feb 13.

Other Organisms

C. elegans
First big
Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J.
Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. Nature. 2000 Nov 16;408(6810):325-30.
and
Gonczy P, Echeverri C, Oegema K, Coulson A, Jones SJ, Copley RR, Duperon J, Oegema J, Brehm M, Cassin E, Hannak E, Kirkham M, Pichler S, Flohrs K, Goessen A, Leidel S, Alleaume AM, Martin C, Ozlu N, Bork P, Hyman AA.
Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III. Nature. 2000 Nov 16;408(6810):331-6.

First genome-wide:
Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature. 2003 Jan 16;421(6920):231-7.

First high-resolution:
Sonnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, Brehm M, Alleaume AM, Artelt J, Bettencourt P, Cassin E, Hewitson M, Holz C, Khan M, Lazik S, Martin C, Nitzsche B, Ruer M, Stamford J, Winzi M, Heinkel R, Roder M, Finell J, Hantsch H, Jones SJ, Jones M, Piano F, Gunsalus KC, Oegema K, Gonczy P, Coulson A, Hyman AA, Echeverri CJ. Related Articles, Links
Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. Nature. 2005 Mar 24;434(7032):462-9.

Drosophila
First
Kiger AA, Baum B, Jones S, Jones MR, Coulson A, Echeverri C, Perrimon N.
A functional genomic analysis of cell morphology using RNA interference. J Biol. 2003;2(4):27.

New library
Bjorklund M, Taipale M, Varjosalo M, Saharinen J, Lahdenpera J, Taipale J.
Related Articles, Links
Identification of pathways regulating cell size and cell-cycle progression by RNAi. Nature. 2006 Feb 23;439(7079):1009-13.

in vivo by injection
Kim YO, Park SJ, Balaban RS, Nirenberg M, Kim Y. A functional genomic screen for cardiogenic genes using RNA interference in developing Drosophila embryos. Proc Natl Acad Sci U S A. 2004 Jan 6;101(1):159-64. Epub 2003 Dec 18.

Reddien PW, Bermange AL, Murfitt KJ, Jennings JR, Sanchez Alvarado A. Identification of genes needed for regeneration, stem cell function, and tissue homeostasis by systematic gene perturbation in planaria. Dev Cell. 2005 May;8(5):635-49.

Human and Mouse libraries

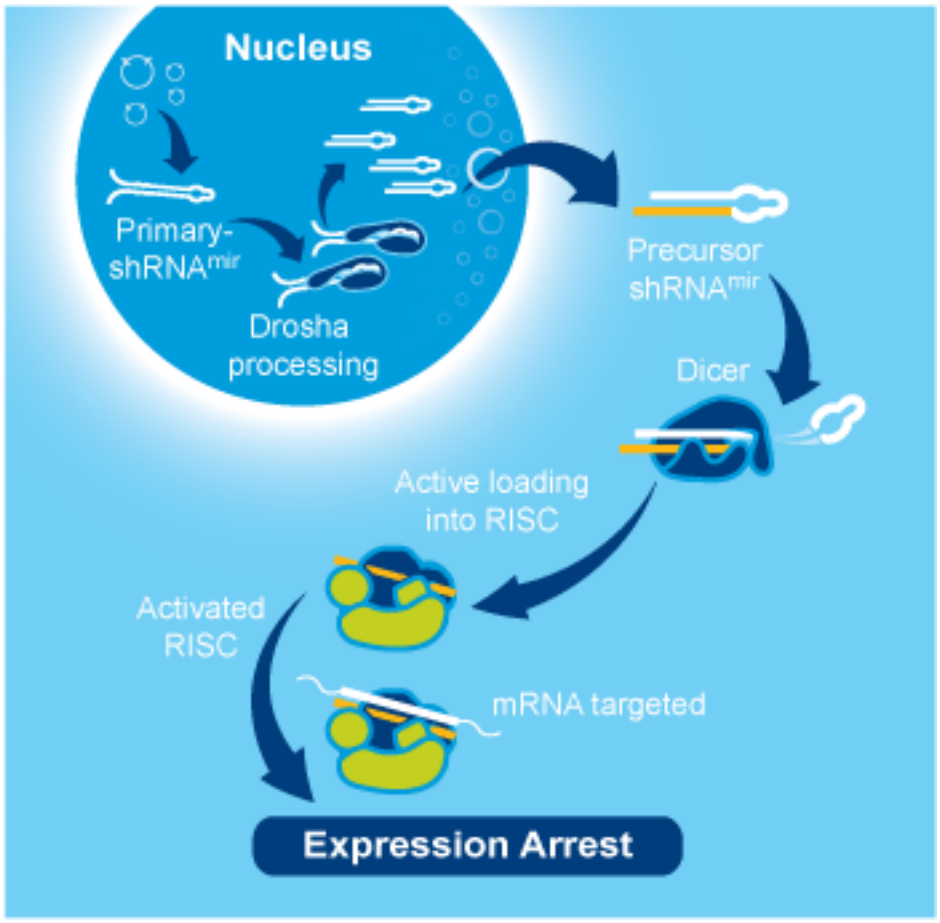
siRNA libraries
Dharmacon
Ambion

Both: genome-wide, good rules, cell-ready

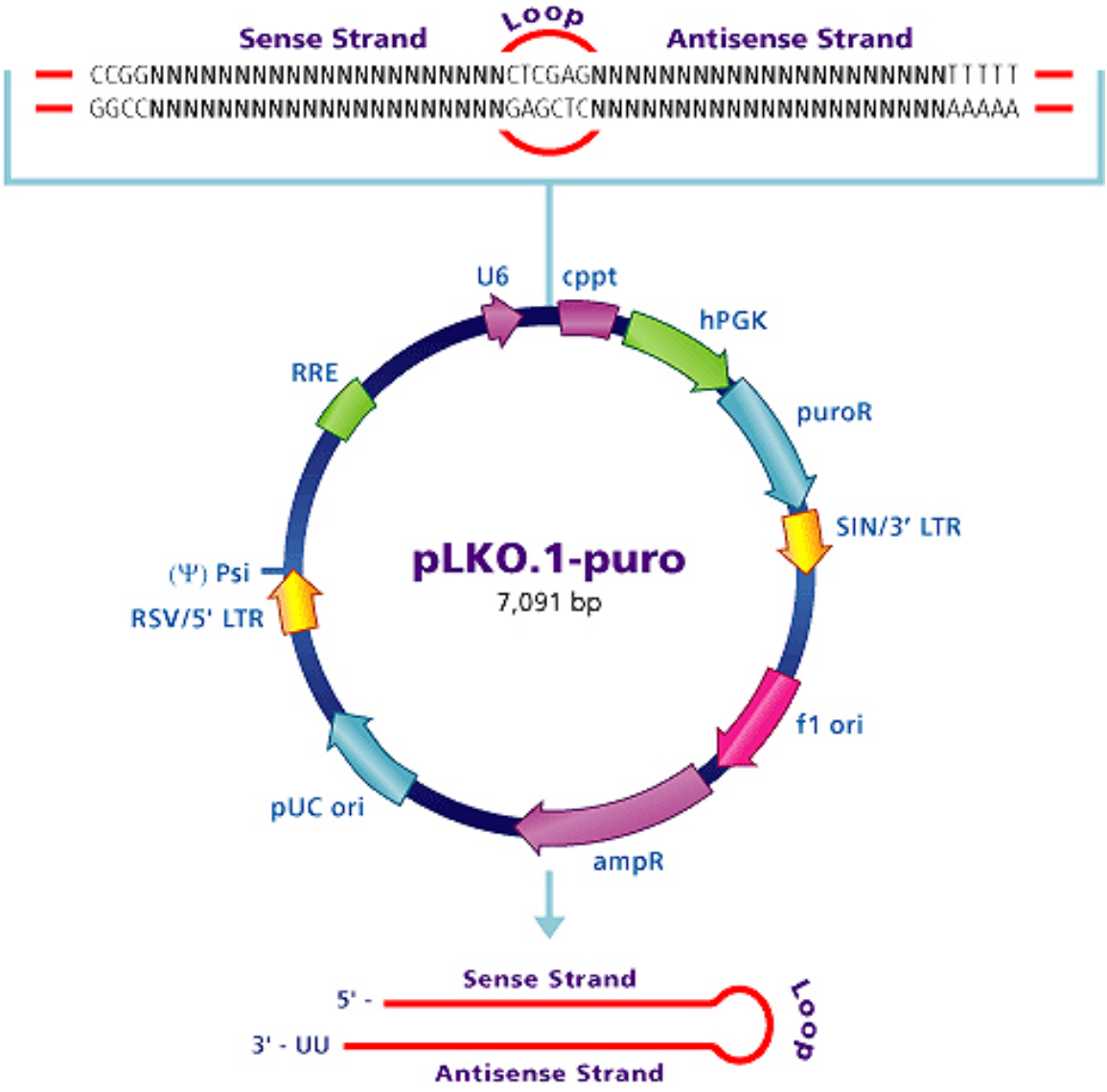
http://www.openbiosystems.com/RNAi/
Distributes Hannon and Broad libraries

Hannon
retroviral and lentiviral
now 5K, genome mouse and human by ‘mid-2006’

Silva JM, Li MZ, Chang K, Ge W, Golding MC, Rickles RJ, Siolas D, Hu G, Paddison PJ, Schlabach MR, Sheth N, Bradshaw J, Burchard J, Kulkarni A, Cavet G, Sachidanandam R, McCombie WR, Cleary MA, Elledge SJ, Hannon GJ. Second-generation shRNA libraries covering the mouse and human genomes.
Nat Genet. 2005 Nov;37(11):1281-8. Epub 2005 Oct 2

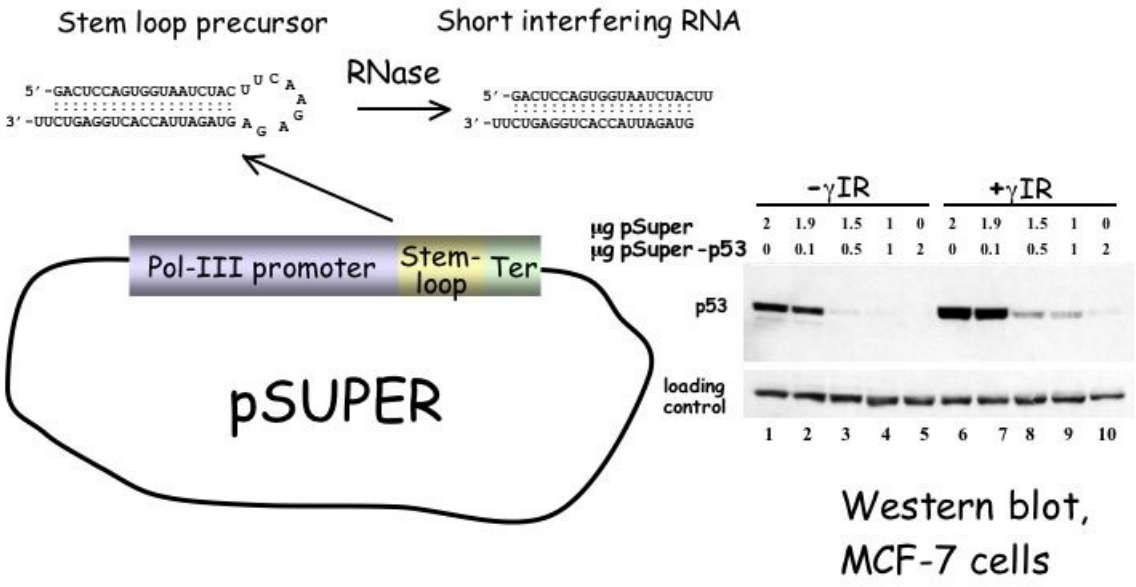


http://www.broad.mit.edu/genome_bio/trc/rnai.html
50,000 clones targeting 7,100 human and 2,900 mouse genes



Rene Bernards (NKI)
23,742 RNAi retroviral pSUPER vector that together target 7,914 human genes

A vector that mediates persistent RNA interference



Western blot, MCF-7 cells

Reviews
Fuchs F, Boutros M. Cellular phenotyping by RNAi.
Brief Funct Genomic Proteomic. 2006 Mar;5(1):52-6. Epub 2006 Feb 23.

Dykxhoorn DM, Novina CD, Sharp PA. Killing the messenger: short RNAs that silence gene expression. Nat Rev Mol Cell Biol. 2003 Jun;4(6):457-67.

Sharp PA. RNAi and double-strand RNA. Genes Dev. 1999 Jan 15;13(2):13

8. Expression Profiling Using Next Generation Sequencing

DIRECTIONAL SEQUENCING OF rRNA-DEPLETED RNA FRACTION / POLYADENYLATED RNA FRACTION

The whole set of transcribed elements of the genome would constitute what is called the transcriptome. It will contain mRNA transcripts, rRNA and other non-mRNA transcripts. Since large rRNA account for 90-95% of the total RNA sample, it is important to try to avoid this fraction when sequencing in order to increase the coverage of other RNA species. Some methods are conceived to purify mRNA transcripts and they will use a positive selection of polyadenylated molecules. Others are aimed to do a negative selection, depleting rRNA. Therefore the resulting samples obtained from these last methods would contain both mRNA transcripts and non-mRNA-non-rRNA transcripts.

We will apply these two methods to the same RNA samples (obtained from cancer cell lines). The polyadenylated RNAs will be selected in the first method by using oligo(dT) beads (Dynabeads, Invitrogen, Life Technologies). In the second method, rRNA will be removed using a hybridization/bead capture procedure that selectively binds rRNA species using biotinylated capture probes (RiboZero rRNA removal beads, Epicentre). The purified RNA samples obtained from both methods will then be used to prepare libraries for subsequent sequencing on the Illumina Cluster Station and Genome Analyser, with the particularity of being able to determine the polarity of the RNA transcript. Knowing directionality brings some important advantages such as the ability to resolve overlapping genetic features, to detect antisense transcription and to assign the sense strand for non-coding RNAs.
Finally conventional cloning will be performed in order to have a “flavour” of the libraries prepared and compare both RNA purification methods.

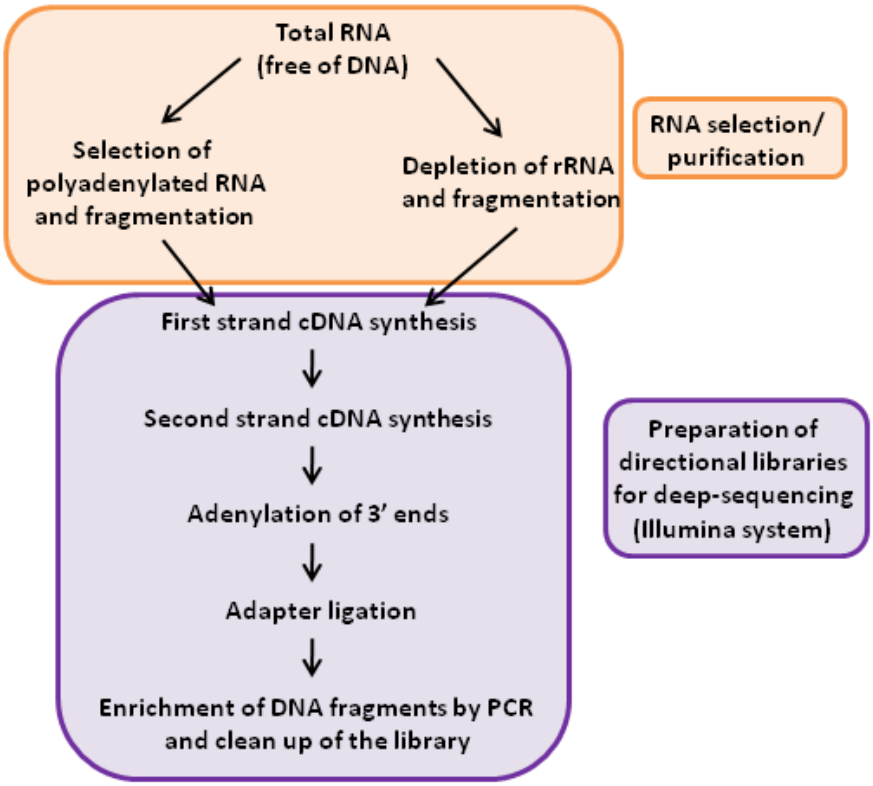


Figure 1. Overview of the procedure for preparation of directional RNA libraries.

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all time during this practical

Bead Binding Buffer - Irritant
Bead Binding Wash Buffer - Irritant
RNA Purification Beads - Irritant
Stop Ligation Buffer – Irritant/Corrosive
rRNA Binding Buffer - Irritant
Binding Buffer - Irritant
70% Ethanol – Highly Flammable
80% Ethanol– Highly Flammable
Ampure XP Beads – Toxic/Irritant
Agilent high Sensitivity Kit –Dye concentrate contains dimethyl sulphoxide – avoid contact with skin and eyes

DAY 1

It is important to use RNA that is DNA free. Therefore it is advised to do a DNAse treatment of your RNA samples before starting. Here we use RNA from a commercial source that is certified to be free of DNA.

A) DEPLETION OF RIBOSOMAL RNA OR SELECTION OF POLYADENYLATED TRANSCRIPTS FROM TOTAL RNA SAMPLES AND SUBSEQUENT FRAGMENTATION OF THE SPECIFIC RNA FRACTIONS OBTAINED ON THESE PROCESSES.

OPTION 1. DEPLETION OF RIBOSOMAL RNA FROM TOTAL RNA SAMPLES

You will receive a sample containing 1µg of total RNA in a 10µl volume (referred here as tube RD-A).

Depletion of ribosomal RNA from total RNA and subsequent fragmentation of the RNA fractions retained during this process.

1) Add 5µl of rRNA Binding Buffer to the total RNA sample provided (tube RD-A).
2) Add 5µl of Ribo-Zero rRNA Removal Mix – Gold to tube RD-A.
3) Gently pipette the entire volume up and down 6 times to mix thoroughly
4) Place the tube RD-A in a thermal cycler with the lid heated at 100˚C. Close the lid and incubate the sample at 68˚C for 5 minutes, then transfer it to a rack and incubate at room temperature for 1 minute.
5) Vortex the rRNA Removal Bead tube vigorously to completely resuspend the beads (it is important that the beads are at room temperature).
6) Add 35µl of rRNA Removal Beads to a new 1.5mL low-binding tube (tube RD-B) and then transfer all the contents (20µl) from your sample tube (tube RD-A, from step 4) into tube RD-B in order to mix it with the beads. It is important to do it in this order to ensure optimal performance.

7) Adjust the pipette to 45µl, then with the tip of the pipette at the bottom of the tube RD-B, pipette quickly up and down 20 times to mix thoroughly. It is important to mix well to maximise rRNA depletion as well as to avoid foaming as much as possible.
8) Incubate the tube containing RNA and beads (tube RD-B) at room temperature for 1 minute.
9) Place tube RD-B on a magnetic stand at room temperature for 1 minute.
10) Transfer all of the supernatant from tube RD-B to a new 1.5mL low-binding tube (tube RD-C).
11) Place tube RD-C on a magnetic stand at room temperature for 1 minute.
12) Transfer all of the supernatant from tube RD-C to a new 1.5mL low-binding tube (tube RD-D). Please be sure there are no beads in tube RD-D, otherwise place the tube again on the magnetic stand for 1 minute and transfer the supernatant to a new tube. Repeat as necessary until there are no beads remaining. The last tube containing bead-free supernatant will be a 1.5mL low-binding tube referred as tube RD-D. This contains the ribosomal depleted RNA.
13) Vortex the RNAClean XP beads (set at room temperature) until they are well dispersed (typically vortex the beads at full speed for 1-2 minute), then add 99µl of beads to tube RD-D. Gently pipette the entire volume up and down 10 times to mix thoroughly.
14) Incubate tube RD-D at room temperature for 15 minutes
15) Place tube RD-D on a magnetic stand at room temperature, for 5 minutes to make sure that all of the beads are bound to the side of the tube.
16) Remove and discard all of the supernatant from tube RD-D.
17) With tube RD-D still remaining on the magnetic stand, add 200µl of freshly prepared 70% ethanol without disturbing the beads.
18) Incubate at room temperature for 30 seconds, then remove and discard all of the supernatant from tube RD-D.
19) Let tube RD-D air dry on the magnet for 15 minutes (with the lid open) and then remove from the magnetic stand.
20) Centrifuge the Elution Buffer (at room temperature) to 600 xg for 5 seconds.
21) Add 16µl of Elution Buffer to tube RD-D. Gently pipette the entire volume up and down 10 times to mix thoroughly.
22) Incubate tube RD-D at room temperature for 2 minutes.
23) Place tube RD-D on a magnetic stand at room temperature for 5 minutes.
24) Transfer 12µl of the supernatant from tube RD-D to a new 1.5mL low-binding tube (tube RD-E). Be careful not to transfer beads. In case this happens, return all the liquid and beads to tube RD-D and incubate on the magnetic rack for 5 minutes or until the beads are bound to the side of the tube.
25) Transfer 8.5µl from tube RD-E to a new 0.2mL PCR tube (label this tube as RD-F plus your group number) and **keep tube E at -20°C for further analysis**. We will use the contents of tube RD-E to measure the amount and size distribution of the ribosomal depleted RNA obtained with this process.
26) Continue the protocol with the 0.2mL PCR tube containing 8.5µl of ribosomal depleted RNA (RD-F, from last step). In order to fragment the RNA, add 8.5µl of “Elute, Prime, Fragment High Mix”. Gently pipette the entire volume up and down 10 times to mix thoroughly.
27) Place the tube on a thermal cycler with the lid heated at 100°C. Close the lid and incubate the tube on the following conditions:

Temperature	Time
94°C	8 minutes
4°C	hold

28) Remove the tube from the thermal cycler when it reaches 4°C and centrifuge briefly.
29) Proceed immediately to section B (First strand cDNA synthesis).

OPTION 2. SELECTION OF POLYADENYLATED TRANSCRIPTS FROM DNASE TREATED TOTAL RNA SAMPLES- OPTIONAL, PROTOCOL FOR REFERENCE ONLY:

Start with 1µg of total RNA in a 50µl volume (referred here as tube PL-A).

Selection of polyadenylated RNA from total RNA and subsequent fragmentation of the RNA fractions retained during this process.

1) Vortex the RNA purification Beads tube (at room temperature) vigorously to completely resuspend the oligo-dT beads (typically vortex the beads at full speed for 1-2 minutes).
2) Add 50 µl of RNA purification beads to the RNA sample provided (PL-A) in order to bind the polyadenylated RNA to the oligo dT magnetic beads. Gently pipette the entire volume up and down 6 times to mix thoroughly.
3) Place the tube on a thermal cycler with the lid heated at 100°C. Close the lid and incubate on the following conditions:

Temperature	Time
65°C	5 minutes
4°C	hold

4) Remove the tube from the thermal cycler when it reaches 4°C.
5) Place the tube on the bench and incubate at room temperature for 5 minutes to allow the RNA to bind to the beads.
6) Transfer all the contents from tube PL-A to a new 1.5mL low-binding tube and label it as PL-B. Place PL-B tube on a magnetic stand at room temperature for 5 minutes to separate the poly-A RNA bound beads from the solution.
7) Remove and discard all of the supernatant from PL-B tube.
8) Remove PL-B tube from the magnetic stand.
9) Wash the beads by adding 200 µl of Bead Washing Buffer to PL-B tube. Gently pipette the entire volume up and down 6 times to mix thoroughly.
10) Place PL-B tube on a magnetic stand at room temperature for 5 minutes or until the solution is clear.
11) Centrifuge the thawed Elution Buffer to 600 xg for 5 seconds.
12) Remove and discard all of the supernatant from PL-B tube while staying on the magnetic rack. The supernatant contains the majority of the ribosomal and other non-messenger RNA.
13) Remove the PL-B tube from the magnetic stand.

14) Add 50 µl of Elution Buffer and gently pipette the entire volume up and down 6 times to mix thoroughly. Transfer the entire volume to a 0.2mL PCR tube and label it as PL-C.
15) Place the PL-C tube on a thermal cycler with the lead heated at 100°C. Close the lid and incubate on the following conditions:

Temperature	Time
80°C	2 minutes
25°C	hold

This will release all the RNA (mostly polyadenylated RNA) that has bound to the beads.

16) Remove the PL-C tube from the thermal cycler when it reaches 25°C.
17) Place the PL-C tube on the bench at room temperature and transfer all the contents to a new 1.5mL low-binding tube (label it as PL-D).
SPECIALLY ADDED STEP: the following step is not necessary and the purpose to do it is to have an aliquot of RNA after one round of selection in order to be able to compare to ribo-depletion (option 1). For this, place tube PL-D on a magnetic stand for 5 minutes or until the supernatant is clear. Take 4 µl of the supernatant from PL-D to a new 1.5 mL low-binding tube (label it as PL and your group number). Keep it at -20°C for further analysis. We will use the contents of this tube to measure the amount and size distribution of the polyadenylated RNA obtained with this first round of selection. Continue the protocol with PL-D tube.

18) Vortex tube PL-D in order to mix again its contents (supernatant and beads)..
19) Centrifuge the thawed Bead Binding Buffer to 600 xg for 5 seconds. Add 50 µl of Bead Binding Buffer to the PL-Dtube. This allows the polyadenylated RNA to specifically rebind the beads, while reducing the amount of non polyadenylated RNA (mostly rRNA) that binds non-specifically to the beads. Gently pipette the entire volume up and down 6 times to mix thoroughly.
20) Incubate at room temperature for 5 minutes.
21) Place the PL-D tube on a magnetic stand at room temperature for 5 minutes.
22) Remove and discard all of the supernatant from the PL-D tube.
23) Remove the PL-D tube from the magnetic stand.
24) Wash the beads by adding 200 µl of Bead Washing Buffer to the PL-D tube. Gently pipette the entire volume up and down 6 times to mix thoroughly.
25) Place the PL-D tube on the magnetic stand at room temperature for 5 minutes.
26) Remove and discard all of the supernatant from the PL-D tube while staying on the magnetic rack. The supernatant contains residual rRNA and other contaminants that were released in the first elution and did not rebind the beads.
27) Remove the PL-D tube from the magnetic stand.
28) Add 19.5 µl of Fragment, Prime, Finish Mix to the tube. Gently pipette the entire volume up and down 6 times to mix thoroughly and transfer all the contents to a 0.2mL PCR tube (labelled as PL-E).
29) Place the PL-E tube on a thermal cycler with the lid heated at 100°C. Close the lid and incubate the tube on the following conditions:

Temperature	Time
94°C	8 minutes
4°C	hold

30) Remove the PL-E tube from the thermal cycler when it reaches 4°C and centrifuge briefly.
31) Transfer all the contents from PL-E tube to a new 1.5 mL low-binding tube (label it as PL-F). Place tube PL-F on a magnetic stand at room temperature for 5 minutes or until the solution is clear.
32) With tube PL-F still on the magnet, transfer 17µl of supernatant to a new 0.2 mL PCR tube (label it as PL-G plus your group number). Proceed immediately to section B (First Strand cDNA synthesis).

B) FIRST STRAND cDNA SYNTHESIS

This process reverse transcribes the cleaved RNA fragments that were obtained in previous section. We have two different samples: RD-F (17 µl), which corresponds to total RNA depleted of ribosomal RNA fraction; and PL-G (17ul), which is enriched in polyadenylated RNA fraction.

1) Prepare the first strand synthesis mix by combining the following reagents in a tube. You can prepare a master mix for all the reactions that need to be done:

Reagent	**Amount for 1 reaction**
First strand synthesis Act D Mix	9µl
SuperScript II	1µl

Mix gently, but thoroughly, and centrifuge briefly.

2) Add 8µl of the First strand synthesis mixture with Superscript II to each sample (RD-F and PL-G). Gently pipette the entire volume up and down 6 times to mix thoroughly and spin down briefly.

3) Place the tubes on a thermal cycler with the lid heated at 100°C. Close the lid and incubate on the following conditions:

Temperature	Time
25°C	10 minutes
42°C	15 minutes
70°C	15 minutes
4°C	hold

4) When the thermal cycler reaches 4°C, remove the tubes and proceed immediately to section C (Second strand cDNA synthesis).

C) SECOND STRAND cDNA SYNTHESIS

This process removes the RNA template and synthesises a replacement strand, incorporating dUTP in place of dTTP to generate ds cDNA. This allows the distinction between both strands.

1) Add 5µl of Resuspension buffer to the samples (RD-F and PL-G) after first strand cDNA synthesis (previous section, step 4).
2) Centrifuge the thawed Second Strand Marking Master Mix to 600xg for 5 seconds.
3) Add 20µl of thawed Second Strand Marking Master Mix to each sample (RD-F and PL-G). Gently pipette the entire volume up and down 6 times to mix thoroughly.
4) Place the samples on a thermal cycler pre-heated at 16°C. Incubate at 16°C for 1hour (close the lid if its temperature can be set up at 30°C. Otherwise, leave the lid open).
5) Immediately after 1h incubation at 16°C, remove the samples from the thermal cycler and let stand to bring them to room temperature. Transfer the entire volume of each sample to a new 1.5mL low binding tube (label the new tubes as RD-G and PL-H).
6) Vortex the AMPure XP beads (at room temperature) until they are well dispersed, then add 90µl to each sample obtained in previous step (RD-G and PL-H). Gently pipette the entire volume up and down 10 times to mix thoroughly.
7) Incubate the samples at room temperature for 15 minutes.
8) Place the tubes with your samples and beads on a magnetic stand at room temperature, for 5 minutes to make sure that all of the beads are bound to the side of the tubes.
9) Remove and discard 135µl of the supernatant from each sample.
10) With the tubes remaining on the magnetic stand, add 200µl of freshly prepared 80% ethanol without disturbing the beads.
11) Incubate at room temperature for 30 seconds, then remove and discard all of the supernatant from each tube.
12) Repeat steps 10 and 11 once for a total of two 80% ethanol washes.
13) Let the tubes stand at room temperature for 15 minutes to dry (with the lid open) and then remove them from the magnetic stand.
14) Centrifuge the Resuspension buffer (thawed and at room temperature) to 600xg for 5 seconds.
15) Add 17.5 µl of Resuspension Buffer to each sample. Gently pipette the entire volume up and down 10 times to mix thoroughly.
16) Incubate at room temperature for 2 minutes.
17) Place the samples on a magnetic stand at room temperature for 5 minutes.
18) Transfer 15 µl of the supernatant (containing ds cDNA) from each sample to a new 0.2mL PCR tube (label the new tubes as RD-H and PL-I followed by your group number) and store them at -20°C or proceed to section D.

DAY 2

D) ADENYLATION OF 3' ENDS

A single "A" nucleotide is added to the 3' ends of the blunt fragments to prevent them from ligating to one another during the adapter ligation reaction.

1) Add 2.5 µl of Resuspension Buffer to the ds cDNA samples obtained from last section (RD-H and PL-I, step18).
2) Add 12.5 µl of thawed A-tailing mix to each sample. Gently pipette the entire volume up and down 10 times to mix thoroughly.
3) Place the samples on a thermal cycler with the lid heated at 100°C. Close the lid and incubate on the following conditions:

Temperature	Time
37°C	30 minutes
70°C	5 minutes
4°C	hold

4) When the thermal cycler temperature is 4°C, remove the samples from the thermal cycler, then proceed immediately to section E (Adapter Ligation).

E) ADAPTER LIGATION

The adaptors have a single "T" nucleotide on the 3' end that provides a complementary overhang for ligating the adapter to the adenylated ds cDNA products.

1) Add 2.5 µl of resuspension Buffer to the samples obtained from last section (RD-H and PL-I, step 4).
2) Add 2.5 µl of Ligation Mix to each sample.
3) Add 2.5 µl of the desired thawed RNA adapter index to each sample (important if we were about to multiplex this sample with others for sequencing on the same lane). Gently pipette the entire volume up and down 10 times to mix thoroughly.
4) Centrifuge the samples at 280 xg for 1 minute.
5) Place the samples on a pre-heated thermal cycler (30°C) with the lid heated at 100°C. Close the lid and incubate samples at 30°C for 10 minutes.
6) Remove the samples from the thermal cycler and add 5 µl of Stop Ligation Buffer to inactivate the ligation. Gently pipette the entire volume up and down 10 times to mix thoroughly. Transfer the whole content of each sample to a new 1.5mL low-binding tube (label them as RD-I and PL-J).
7) Vortex the AMPure XP Beads (at room temperature) until they are well dispersed, and then add 42 µl of beads to the RD-I and PL-J samples. Gently pipette the entire volume up and down 10 times to mix thoroughly.
8) Incubate the samples at room temperature for 15 minutes.
9) Place the RD-I and PL-J samples on a magnetic stand at room temperature for 5 minutes or until the liquid appears clear.
10) Remove and discard 79.5 µl of the supernatant of each sample.

11) With the RD-I and PL-J tubes remaining on the magnetic stand, add 200 µl of freshly prepared 80% ethanol to each tube without disturbing the beads.
12) Incubate the RD-I and PL-J tubes on the magnet at room temperature for 30 seconds, then remove and discard all of the supernatant from the tube.
13) Repeat steps 11 and 12 once for a total of two 80% ethanol washes.
14) While keeping RD-I and PL-J tubes on the magnetic stand, let the samples air dry at room temperature for 15 minutes (leave the tubes open) and then remove the tubes from the magnetic stand.
15) Resuspend the dried pellets on each tube with 52.5 µl of Resuspension Buffer. Gently pipette the entire volume up and down 10 times to mix thoroughly.
16) Incubate the RD-I and PL-J tubes at room temperature for 2 minutes.
17) Place the RD-I and PL-J tubes on a magnetic stand at room temperature for 5 minutes or until the liquid appears clear.
18) For each sample, transfer 50 µl of the clear supernatant to a new 1.5mL low-binding tube (label them as RD-J and PL-K).
19) Vortex the AMPure XP beads until they are well dispersed, then add 50 µl of the beads to the RD-J and PL-K tubes for a second clean up. Gently pipette the entire volume up and down 10 times to mix thoroughly.
20) Incubate the RD-J and PL-K tubes at room temperature for 15 minutes.
21) Place the RD-J and PL-K tubes on the magnetic stand at room temperature for 5 minutes or until the liquid appears clear.
22) Remove and discard 95 µl of the supernatant from RD-J and PL-K tubes.
23) With the RD-J and PL-K tubes remaining on the magnetic stand, add 200 µl of freshly prepared 80% ethanol without disturbing the beads.
24) Incubate the RD-J and PL-K tubes on the magnet for 30 seconds, then remove and discard all of the supernatant from the tube.
25) Repeat steps 23 and 24 once for a total of two 80% ethanol washes.
26) While keeping the RD-J and PL-K tubes on the magnetic stand, let the samples air dry at room temperature for 15 minutes (leave the tubes open) and then remove the plate from the magnetic stand.
27) Resuspend the dried pellets from each sample with 22.5 µl Resuspension Buffer. Gently pipette the entire volume up and down 10 times to mix thoroughly.
28) Incubate the RD-J and PL-K tubes at room temperature for 2 minutes.
29) Place the RD-J and PL-K tubes on a magnetic stand at room temperature for 5 minutes or until the liquid appears clear.
30) Transfer 20 µl of the clear supernatant from each sample to a new 0.2mL PCR tube (label them as RD-K and PL-L, followed by your group number). Proceed immediately to section F (Enrichment of DNA fragments by PCR) or store the tube at -20°C.

DAY 3

F) ENRICHMENT OF DNA FRAGMENTS BY PCR

The PCR is performed with a PCR primer cocktail that anneals to the ends of the adapters in order to selectively enrich and amplify those DNA fragments that have adapter molecules on both ends.

1) Add 5 µl of thawed PCR Primer Cocktail to each of the sample obtained in last section (RD-K and PL-L, step 30).
2) Add 25 µl of thawed PCR Master Mix to each sample. Gently pipette the entire volume up and down 10 times to mix thoroughly.
3) Place the tubes on a thermal cycler with the lid heated to 100°C. Close the lid and amplify with the following cycling conditions:

Temperature	Time	Number of cycles
98°C	30 seconds	1 cycle
98°C	10 seconds	15 cycles
60°C	30 seconds	
72°C	30 seconds	
72°C	5 minutes	1 cycle
4°C		hold

4) Once the PCR is finished, transfer each product into a new 1.5mL low-binding tube (label them as RD-L and PL-M).
5) Vortex the AMPure XP Beads until they are well dispersed, then add 50 µl to each of the 1.5mL tubes containing the PCR amplified library (RD-L and PL-M). Gently pipette the entire volume up and down 10 times to mix thoroughly.
6) Incubate the samples at room temperature for 15 minutes.
7) Place the samples on a magnetic stand at room temperature for 5 minutes or until the liquid appears clear.
8) Remove and discard 95 µl of the supernatant from each sample.
9) With the tubes remaining on the magnetic stand, add 200 µl of freshly prepared 80% ethanol to each sample without disturbing the beads.
10) Incubate the tubes on the magnet for 30 seconds, then remove and discard all of the supernatant from the samples.
11) Repeat steps 9 and 10 once for a total of two 80% ethanol washes.
12) While keeping the samples on the magnetic stand, let them air dry at room temperature for 15 minutes (leave the tubes open) and then remove the tubes from the magnetic stand.
13) Resuspend the dried pellets of each sample with 32.5 µl Resuspension Buffer. Gently pipette the entire volume up and down 10 times to mix thoroughly.
14) Incubate the tubes at room temperature for 2 minutes.
15) Place the tubes on a magnetic stand at room temperature for 5 minutes or until the liquid appears clear.
16) Transfer 30 µl of the clear supernatant of each sample to a new 1.5mL low-binding tube (label them as RD Lib and PL Lib, followed by your group number). These are your final sequencing libraries prepared from ribo-depleted RNA (RD Lib) and polyadenylated RNA (PL Lib).
17) We will measure the concentration of the libraries in the nanodrop and check their size distribution in the bioanalyser by using Agilent DNA-1000 chips. The final product for both RD Lib and PL Lib should be a band at approximately 260bp.

PREPARING SAMPLES FOR SEQUENCING OF SMALL RNA

Small RNAs have less than 200nt and are contained within extracted total RNA, provided there was no size based purification. Several classes of non-coding small RNAs have been described including microRNAs (miRNAs). MicroRNAs are post-transcriptional regulators that bind to complementary sequences in the three prime untranslated regions (3' UTRs) of target mRNA transcripts, usually resulting in the impairment of protein translation and/or mRNA degradation. They have been involved in several biological pathways and its aberrant expression has been linked to several pathological processes including cancer.

This protocol explains how to prepare libraries of small RNA for subsequent clustering and cDNA sequencing on the Illumina platform. It is a modification of the Illumina-provided protocol for preparation of libraries of small RNAs. In the original protocol, total RNA is used directly for ligation of the adapters necessary for use during cluster reaction. In the protocol we propose, total RNA is migrated in a denaturing gel and the desired small RNA fraction is purified by removing a band that corresponds to the nucleotide length of interest. The adapters are then ligated to the purified fraction and the resulting product is reverse-transcribed and amplified, obtaining DNA constructs that are suitable as final template for sequencing on an Illumina platform.

Most microRNAs and some other small RNAs have a 5’ phosphate group and a 3’ hydroxyl group as a result of the enzymatic cleavage performed by DICER during their generation process. The 5' small RNA adapter will only bind to the 5’ end and is necessary for amplification of the small RNA fragment. This adapter also contains the DNA sequencing primer binding site. The 3' small RNA adapter has been modified to bind specifically to the 3’ end of molecules containing a 3’ hydroxyl group. It is necessary for reverse transcription and corresponds to the surface bound amplification primer on the flow cell used on the Cluster Station.

In order to sequence the microRNA fraction, we will select the 20-30nt fraction from total RNA. This would help to increase specificity and therefore to reduce the artefacts in the final library product. The starting material, total RNA, has to be isolated using a method that retains small RNA.

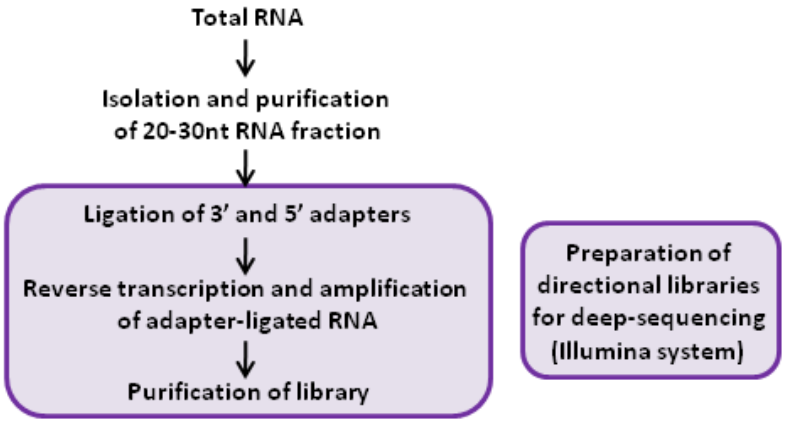


Figure 1. Overview of the procedure for preparation of small RNA libraries.

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all time during this practical

1x TBE – Harmful
Sybr Gold (1x) – Irritant
75% Ethanol – Highly Flammable
100% Ethanol – Highly Flammable
Agilent high Sensitivity Kit –Dye concentrate contains dimethyl sulphoxide – avoid contact with skin and eyes
Transilluminator – UV light source wear protective UV face shield, lab coat and gloves

DAY 1 - OPTIONAL

A) ISOLATE SMALL RNA BY DENATURING PAGE

1) Determine the volume of 1X TBE buffer needed. Dilute the 5X TBE buffer to 1X for use in electrophoresis. Assemble the gel electrophoresis apparatus per the manufacturer's instructions.
2) Pre-run the 15% TBE-urea gel for 15–30 minutes at 200 V and wash the wells using 1X TBE.
3) While the gel is pre-running, mix 2 µl of 10bp ladder with 2 µl of RNA gel loading dye in a sterile, RNase-free, 200 µl PCR tube.
4) Mix 5 µl (1µg) of total RNA (provided to you already in a tube) with 5 µL of RNA gel loading dye in a sterile, RNase-free 200 µl PCR tube.
5) Heat the sample and ladder tubes at 65°C for 5 minutes in a thermal cycler.
6) Place the tubes on ice.
7) Centrifuge the tubes to collect the entire column of the tube.
8) Load both the entire 10bp ladder and sample RNA on the same gel with several lanes between them.
9) Run the gel at 200 V for 1 hour.
10) During this time, puncture the bottom of a sterile, nuclease-free, 0.5 ml microcentrifuge tube 4–5 times with a 21-gauge needle.
11) Remove the gel from the apparatus.
12) Pry apart the cassette and stain the gel with SYBR Gold (1X) in a clean container for 10 minutes.
13) View the gel on a Dark Reader transilluminator or a UV transilluminator. The 10bp ladder is from 10–100 bases in 10 base increments.
14) Using a clean scalpel, cut out a band of gel corresponding to the 18–30 nucleotide bands in the marker lane.
15) Place the gel slice into the 0.5 ml punctured microcentrifuge tube you have prepared previously. Place the 0.5 ml punctured microcentrifuge tube into a sterile, round-bottom, nuclease-free, 2 ml microcentrifuge tube.
16) Centrifuge the stacked tubes at 14000 rpm in a microcentrifuge for 2 minutes at room temperature to move the gel through the holes into the 2 ml tube.
17) Add 300 µl of 0.3 M NaCl to the gel debris in the 2 ml tube and elute the RNA by rotating the tube gently at room temperature for 4 hours.

18) Transfer the eluate and the gel debris to the top of a Spin X cellulose acetate filter.
19) Centrifuge the filter in the microcentrifuge for 2 minutes at 14000 rpm.
20) Add 1 µl of glycogen (20mg/ml) and 750 µl of room temperature 100% ethanol.
21) Incubate at -80°C for at least 30 minutes.
22) Immediately centrifuge to 14000 rpm for 25 minutes on a 4°C microcentrifuge. Remove the supernatant and discard it.
23) Wash the pellet with 750 µl of room temperature 75% ethanol. Centrifuge to 14000 rpm for 25 minutes on a 4°C microcentrifuge. Remove the supernatant and discard it.
24) Allow the RNA pellet to air dry.
25) Resuspend the RNA pellet in 7 µl of ultra pure water. We will use 1 µl to measure the concentration in the nanodrop and 1 µl for checking the size distribution of the RNA molecules in the bioanalyser.

DAY 2

B) LIGATION OF 3' AND 5' ADAPTERS

We will use the Illumina supplied RNA 5' adaptor and the RNA 3' adaptor from TruSeq Small RNA sample preparation kit. We can try to coordinate all groups together in order to prepare common master mix for all reactions. All amounts stated in the protocol are for one reaction.

1) Mix the following in a 200µl PCR tube:
   RNA sample obtained in previous section (step 25) (5µl) or up to 1ug of total RNA in 5ul
   RNA 3' adaptor (RA3) (1µl)
2) Incubate at 70˚C for 2 minutes in a thermal cycler (with lid at 100˚C) and place immediately on ice.
3) Pre-heat the thermal cycler to 28˚C with the lid at 100˚C.
4) Prepare the following mix in a separate tube on ice. Multiply each reagent volume by the number of samples being prepared (Make 10% extra reagent if you are preparing multiple samples):

Reagent	**Amount for 1 reaction**
Ligation Buffer (HML)	2.0µl
RNAse Inhibitor	1.0µl
T4 RNA Ligase2, Deletion mutant	1.0µl
TOTAL	**4.0µl**

5) Mix well and spin down. Add 4 µl of the mix to the reaction tube from step 1 (denaturated RNA sample and RNA 3' adaptor) and gently pipette the entire volume up and down 6–8 times to mix thoroughly. The total volume of the reaction should be 10 µl.
6) Incubate at 28˚C for 1 hour in a thermal cycler with the lid closed.
7) With the reaction tube remaining on the thermal cycler, add 1 µl Stop Solution (STP) and gently pipette the entire volume up and down 6–8 times to mix

thoroughly. Continue to incubate the reaction tube on the thermal cycler at 28°C for 15 minutes, and then place the tube on ice.

8) Determine the required amount of the 5' adaptor from the stock (1.1µl per reaction). Aliquot 1.1 x N µl of the RNA 5' Adapter (RA5) into a separate, nuclease-free 200 µl PCR tube, with N equal to the number of samples being processed for the current experiment.
9) Incubate the adapter on the pre-heated thermal cycler at 70°C for 2 minutes and then immediately place the tube on ice.
10) Pre-heat the thermal cycler to 28˚C with the lid at 100˚C.
11) Add 1.1 X N µl of 10mM ATP to the aliquoted RNA 5' Adapter tube, with N equal to the number of samples being processed for the current experiment. Gently pipette the entire volume up and down 6–8 times to mix thoroughly.
12) Add 1.1 X N µl of T4 RNA Ligase to the aliquoted RNA 5'Adapter tube, with N equal to the number of samples being processed for the current experiment. Gently pipette the entire volume up and down 6–8 times to mix thoroughly.
13) Add 3 µl of the mix from the aliquoted RNA 5' Adapter tube to the reaction from step 7 of Ligate 3' Adapter. Gently pipette the entire volume up and down 6–8 times to mix thoroughly.
    The total volume of the reaction should now be 14 µl.
14) Incubate the reaction tube on the pre-heated thermal cycler at 28°C for 1 hour with the lid closed and then place the tube on ice.

C) REVERSE TRANSCRIPTION AND AMPLIFICATION OF THE ADAPTER-LIGATED RNA

Reverse transcription followed by PCR is used to create cDNA constructs based on the small RNA ligated with 3' and 5' adapters. This process selectively enriches those fragments that have adapter molecules on both ends. PCR is performed with two primers that anneal to the ends of the adapters.

1) Combine the following in a 200µl PCR tube:
   5' and 3' Adapter-ligated RNA from previous section (6µl)
   RNA RT primer (RTP) (1µl)
2) Gently pipette the entire volume up and down 6–8 times to mix thoroughly, then centrifuge briefly.
3) Heat the mixture at 70˚C for 2 minutes in a thermal cycler and then immediately place the tube on ice. Briefly spin in order to collect all the contents.
4) Pre-heat the thermal cycler to 50˚C with the lid at 100˚C.
5) Prepare the following mix in a separate, sterile, nuclease-free, 200 µl PCR tube placed on ice. Multiply each reagent volume by the number of samples being prepared. Make 10% extra reagent if you are preparing multiple samples.

Reagent	**Amount for 1 reaction**
5X first strand buffer	2.0µl
12.5mM dNTPs	0.5µl
100mM DTT	1.0µl
RNAse Inhibitor	1.0µl
Superscript II Reverse Transcriptase	1.0µl
TOTAL	**5.5µl**

6) Gently pipette the entire volume up and down 6–8 times to mix thoroughly, then centrifuge briefly.
7) Add 5.5 µl of the mix to the reaction tube from step 3. Gently pipette the entire volume up and down 6–8 times to mix thoroughly, then centrifuge briefly. The total volume should now be 12.5 µl.
8) Incubate the tube in the thermal cycler at 50°C for 1 hour with the lid closed and then place the tube on ice.
9) Prepare the PCR master mix for each index used, in the following order, in a separate, sterile, nuclease-free, 200 µl PCR tube placed on ice. In this case, each group has to prepare its own mix for one reaction since each group will use a different index to identify the sample:

Reagent	**Amount for 1 reaction**
Ultra pure water	8.5µl
PCR mix (PML)	25.0µl
RNA PCR Primer (RP1)	2.0µl
RNA PCR Primer Index (RPIX)	2.0µl
TOTAL	**37.5µl**

10) Gently pipette the entire volume up and down 6–8 times to mix thoroughly, then centrifuge briefly, then place the tube on ice.
11) Add 37.5 µl of PCR master mix to the reaction tube from step 8 (reverse transcription product).
12) Gently pipette the entire volume up and down 6–8 times to mix thoroughly, then centrifuge briefly and place the tube on ice.
The total volume should now be 50 µl.
13) Amplify the tube in the thermal cycler using the following PCR cycling conditions:

Temperature	Time	Number of cycles
98°C	30 seconds	1 cycle
98°C	10 seconds	11 cycles
60°C	30 seconds	
72°C	15 seconds	
72°C	10 minutes	1 cycle
4°C		hold

DAY 3

D) PURIFICATION OF THE AMPLIFIED cDNA CONSTRUCT

This protocol gel purifies the amplified cDNA construct in preparation for loading on the Illumina Cluster Station.

1) Prepare the Gel Electrophoresis Reagents and Apparatus. Determine the volume of 1X TBE buffer needed. Dilute the 5X TBE buffer to 1X for use in electrophoresis. Assemble the gel electrophoresis apparatus per the manufacturer's instructions.
2) Three ladders will be available:
Ladder 1: hypperladder V from Bioline (already ready to load)

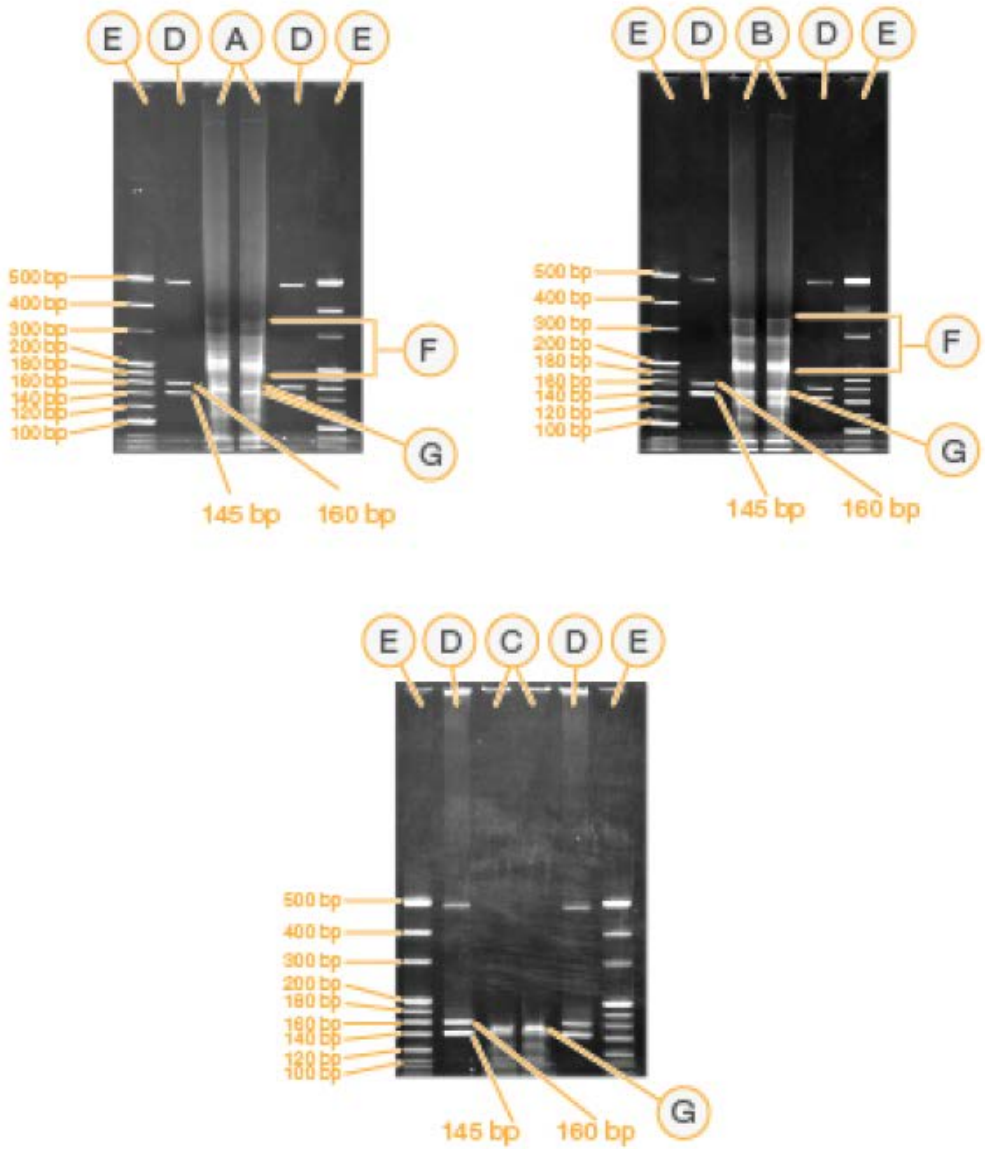
Ladder 2: Illumina Custom ladder (prepare it by mixing 1 µl of Custom Ladder with 1 µl of DNA Loading Dye per lane needed).
Ladder 3: Illumina High Resolution Ladder (prepare it by mixing 1 µl of High Resolution Ladder with 1 µl of DNA Loading Dye per lane needed).

3) Mix 50 µl of amplified cDNA construct with 12.5µl of 5X DNA loading dye.
4) Load the ladders in three different wells on the 6% PAGE gel (for each ladder, load 2ul/well).
5) Load two wells with 25µl each of amplified construct and loading dye mix on the 6% PAGE gel.
6) Run the gel for 60 minutes at 145 V (the gel has to be monitored during the last minutes of the run to be sure we stop just after the blue front dye exits the gel).
7) While the gel is running, puncture the bottom of a sterile, nuclease-free, 0.5 ml microcentrifuge tube 4–5 times with a 21-gauge needle.
8) Dilute the 10X gel elution buffer into a fresh tube as indicated below. Multiply each volume by the number of samples being prepared. Prepare 10% extra reagent mix if you are preparing multiple samples.

Reagents	**Volume (µl)**
Ultra pure water	270
10X gel elution buffer	30
total volume	**300**

9) Remove the gel from the apparatus.
10) Pry apart the cassette and stain the gel with SYBR gold (1X) in a clean container for 5 minutes.
11) View the gel on a Dark Reader transilluminator or a UV transilluminator. The ladder 1 (Hyperladder V from Bioline) has DNA fragments between 25bp and 500bp in 25bp increments with brighter bands at 100bp and 200bp. The ladder 2 (Illumina Custom Ladder) consists of three dsDNA fragments 145 bp, 160 bp, and 500 bp (see figure below, lane D). Ladder 3 (High Resolution ladder) has DNA fragments between 100 and 500bp as shown in figure below (lane E). For each sample we expect to see something similar to the bottom gel shown in the figure below. The 120nt band would correspond to adaptor-adaptor products. The 147 nt band primarily contains mature microRNA generated from approximately 22 nt small RNA fragments. The 157 nt band is generated from approximately 30 nt small RNA fragments that include piwi-interacting RNAs as well as some microRNAs and other regulatory small RNA molecules. Sometimes is not possible to discriminate the 2 latest bands and they appear as a unique band of size between 140-160 bp.
12) Using a clean scalpel, cut out the bands corresponding to approximately the adapter-ligated constructs derived from 22nt and 30nt small RNA fragments (so 147bp and 157bp bands). Typically this is the area comprised between the 160bp and 145bp band of the Illumina custom ladder. Sequencing can be conducted on individual bands or from pooled bands.
NOTE: Do not cut the 120bp band since it contains adapter dimmers.

Figure 12 Small RNA Library from Total RNA Samples



A Human Brain Total RNA
B Mouse Brain Total RNA
C Pre-purified microRNA
D Custom Ladder
E High Resolution Ladder
F Small Non-coding RNAs
G MicroRNA

Small RNA libraries prepared from total RNA (top gels) and from a purified small RNA fraction of total RNA (bottom gel) (figure obtained from Illumina).

13) Place the band into the 0.5 ml microcentrifuge tube prepared previously. Place the 0.5 ml microcentrifuge tube into a sterile, round-bottom, nuclease-free, 2 ml microcentrifuge tube.

14) Centrifuge the stacked tubes at 14000 rpm in a microcentrifuge for 2 minutes at room temperature to move the gel through the holes into the 2 ml tube.
15) Add 300µl of ultra pure water to the gel debris in the 2ml tube.
16) Elute the DNA by rotating the tube gently at room temperature for 2 hours (can be done overnight).
17) Transfer the eluate and the gel debris to the top of a Spin-X filter.
18) Centrifuge the filter for 2 minutes at 14000 rpm.
19) Add 2µl of glycogen, 30µl of 3M NaOAc, 2 µl of 0.1X pellet paint (optional) and 975µl of pre-chilled -20°C 100% ethanol.
20) Immediately centrifuge to 14000 rpm for 20 minutes in a benchtop microcentrifuge. Alternatively it can be left precipitating for 30 minutes at -80°C or o/n at -20°C or -80°C.
21) Remove and discard the supernatant, leaving the pellet intact.
22) Wash the pellet with 500µl of room temperature 70% ethanol.
23) Remove and discard the supernatant, leaving the pellet intact.
24) Dry the pellet by placing the tube, lid open, in a 37°C heat block for 5–10 minutes or until dry.
25) Resuspend the pellet in 10µl of 10mM Tris-HCl, pH=8.5) or ultra pure water.
26) We will measure the concentration of the library in the nanodrop and check its size distribution in the bioanalyser by using Agilent high sensitivity DNA chips.

CONVENTIONAL CLONING AND SEQUENCING OF DEEP-SEQUENCING LIBRARIES

The aim of this experiment is to check for the quality RNA libraries created for Illumina/Solexa digital sequencing (suitable for the three types of libraries created during this course). We will clone a certain amount of library in Invitrogen TOPO vectors and sequence using conventional technology. If the experiment has worked correctly, the sequences obtained should match the human genome. In the case of the polyadenylated RNA libraries, we expect to get sequences that match mostly to mRNAs. In contrast, rRNA-depleted RNA libraries can give rise to sequences matching mRNA but also non-coding RNAs. However it is very likely for both types of libraries that a certain percentage of sequences correspond to rRNA. We will try to determine which method was more efficient on purifying total RNA. Finally the sequences obtained from the small RNA library should mostly correspond to miRNAs.

Zero Blunt PCR Cloning kit (Invitrogen).

The Zero Blunt® PCR Cloning Kit is designed to clone blunt PCR fragments (or any blunt DNA fragment) with a low background of non-recombinants. The pCR®-Blunt vector contains the lethal E. coli ccdB gene fused to the C-terminus of LacZα. Ligation of a blunt PCR fragment disrupts expression of the lacZα-ccdB gene fusion permitting growth of only positive recombinants upon transformation. Cells that contain non-recombinant vector are killed when the transformation mixture is plated.

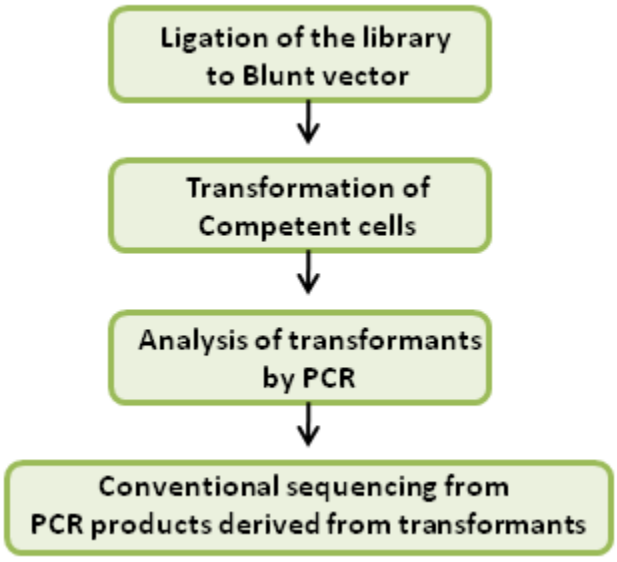


Figure 1. Overview of the procedure for conventional cloning and sequencing of deep-sequencing libraries.

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all time during this practical

Magnesium Chloride – Irritant
3M Sodium Acetate – Irritant
75% Ethanol – Highly Flammable
100% Ethanol – Highly Flammable
Kanamycin Monosulphate – Respiratory Sensitizer/ Reproductive toxin, may harm unborn child. Not to be handled by new or expectant mothers.

DAY 1

A) LIGATION OF THE RNA LIBRARY TO THE BLUNT VECTOR

1) Set up the following 10µl ligation reaction:

	RD Lib and PL Lib	**Small RNA library**
Reagent	**Amount per reaction**	**Amount per reaction**
RNA library (PCR product)	**2µl**	**1µl**
pCR®-Blunt (25 ng)	1µl	1µl
10X Ligation Buffer (with ATP)	1µl	1µl
Sterile water	**5µl**	**6µl**
T4 DNA Ligase (4 U/µl)	1µl	1µl
Total Volume	**10µl**	**10µl**

2) Incubate the ligation reaction 16°C for 1 hour. Optimal ligation occurs at 16°C.
3) Store the ligation reaction at -20°C or proceed with transformation.

B) TRANSFORMATION OF COMPETENT CELLS

1) Thaw on ice one 50µl vial of One Shot® TOP10 cells for each transformation.
2) Pipette 2µl of each ligation reaction directly into the vial of competent cells and mix by stirring gently with a pipette tip. **Do not mix cells by pipetting.**
3) Incubate the vials on ice for 30 minutes. Store the remaining ligation mixtures at -20°C.
4) Heat shock the cells for 45 seconds at 42°C without shaking. Immediately place the vials on ice for 2 minutes.
5) Add 250µl of room temperature S.O.C. medium to each vial.
6) Shake the vials horizontally at 37°C for 1 hour at 225 rpm in a shaking incubator.
7) Spread 60µl and 80µl from each transformation vial on prewarmed LB plates containing 50µg/ml kanamycin (in order to ensure well-spaced colonies).
8) Incubate plates overnight at 37°C.

DAY 2

C) ANALYSING TRANSFORMANTS

1) Pick at least 10 transformants for analysis. Pick the colony with a tip and put it in a tube containing 15ul of water. Keep the tip inside of the tube.
2) We will do PCR in order to verify if the colony has the vector+insert. We will use M13F and M13R primers (both priming the blunt vector). Alternatively, a combination between a M13 primer and a primer specific for the insert can be used.
3) Prepare the PCR reaction mixture for the number of reactions needed (as many as colonies have been picked):

Reagent	**Amount for 1 reaction**
10X AmpliTaq buffer	2.5µl
$MgCl_2$ (25mM)	0.5µl
Ford. Primer (10uM)	1µl
Rev. primer (10uM)	1µl
dNTP (25mM)	0.5µl
AmpliTaq polymerase	0.5µl
Sterile water	17µl
Total volume	**23µl**

Add 2ul of the water where colony has been disrupted (for each colony picked).

4) PCR cycling conditions:

95°C	5 min	Hold
95°C	30 sec	30 cycles
56°C	30 sec	
72°C	45 sec	
15°C		hold

5) In order to visualise the PCR result, add 1.25µl of 5X DNA loading buffer to 5µl of the PCR product and run it in a 2% agarose gel. Don't forget to load at least one well with 2 µl of Hypperladder V for resolving the size of the bands. If the colony contains an empty vector (with no library product ligated), the PCR product size should be ~238bp. If the colony has incorporated a vector containing a library product, then the PCR product size should be ~390bp and ~500bp (or longer) for small RNA libraries and stranded RNA libraries (RD and PL), respectively.
6) We will use the remaining PCR product for sequencing. Please keep it at -20˚C until use.

Alternatively, sequencing can be performed using plasmid DNA as a template (this can give much cleaner sequencing results). In order to proceed with this option, positive colonies (containing a vector with a library product) must be grown in suspension cultures. These can be set up by disposing 3ml of LB

containing 50µg/ml Kanamycin in round bottom culture tubs. Then add the tip you have used to originally to pick the colony and the remaining water where the colony was disposed. Incubate the tubes at 37˚C overnight at 225rpm. Next morning, centrifuge the cultures at 3000rpm for 5 minutes. Discard the supernatant. Isolate plasmid DNA from culture pellets by using Qiagen kit (as indicated in the manual provided by the kit).

D) SEQUENCING OF PCR PRODUCTS DERIVED FROM POSITIVE COLONIES (CONTAINING A VECTOR+INSERT)

Spiky plates (Greiner Bio-one 96 well half skirted plate -652290) must be used for these reactions as these are compatible with ABI 3730 sequencers.

1) Prepare 1:1 mix of Exonuclease I and Shrimp alkaline phosphatase
2) In a 96 well plate, dispose in each well 1µl of Exonuclease I/ Shrimp alkaline phosphatise mixture and 1µ of PCR product. Centrifuge briefly.
3) Incubate at 37˚C for 30 minutes then deactivate enzymes by incubation at 80˚C for 15 minutes. Centrifuge briefly.
4) Prepare the following sequencing master mix:

Reagent	Amount for 1 reaction
Big Dye version 3.1	0.25µl
Big dye buffer	1.875µl
Primer	2 µl (at 2.5 pmol/µl)
water	3.875µl
Total volume	8µl

5) Add 8µl of sequencing mix to each well of the plate containing the treated PCR products. Centrifuge briefly.

6) Amplify DNA with the following cycling conditions:

96°C	10 sec	25 cycles
50°C	5 sec	
60°C	4 min	
4°C		hold

E) CLEAN UP PROCEDURE

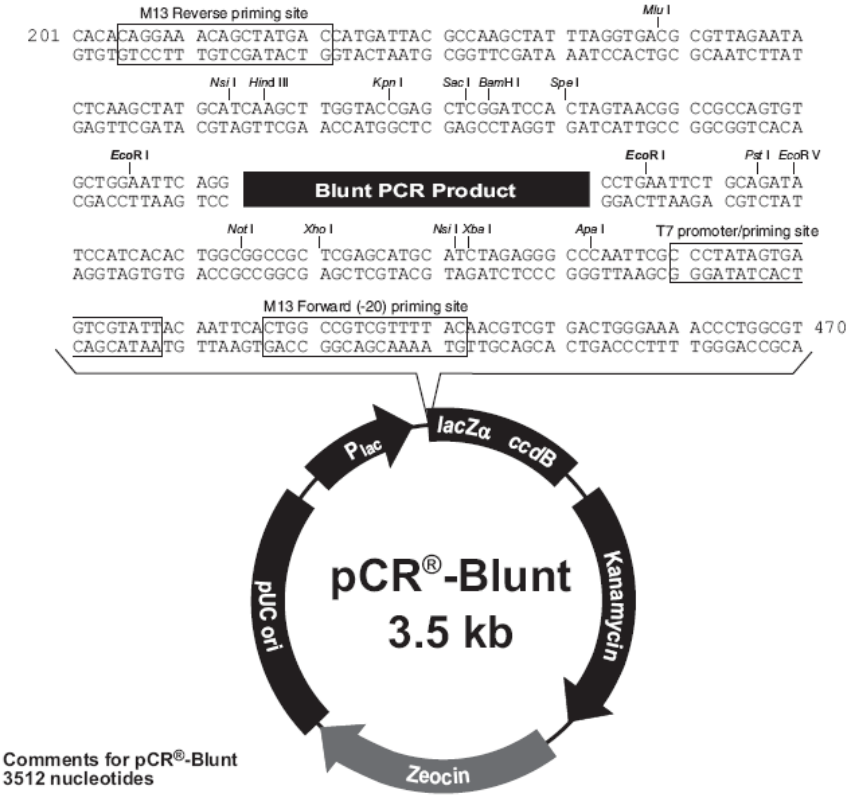
1) Add 2µl of 125mM EDTA to each well.
2) Add 2µl of 3M sodium acetate to each well.
3) Add 50µl of 100% ethanol to each well.
4) Use good aluminium tape and vortex briefly or invert 4 times
5) Leave at room temperature for 15 minutes (do not exceed 15 minutes)
6) Spin at 3000x g at 4˚C for 30 minutes (proceed to next step immediately or spin for 2 more minutes before proceeding)
7) Remove the foil gently. Invert the plate and spin up to 185x g for 10-15 seconds.
8) Add 70µl of 70% ethanol to each well.
9) Spin at 1650x g at 4˚C for 15 minutes

10) Invert the plate and spin up to 185x g for 1 minute. Start timing when the rotor starts moving.
11) Air dry the pellets by leaving the plate 15 minutes at room temperature. Protect the plate from the light.
12) Store at -20˚C.

Map and Features of pCR®-Blunt

Map of pCR®-Blunt

The figure below summarizes the features of the pCR®-Blunt vector. Restriction sites that are only found in the polylinker are shown. **The complete sequence is available for downloading from our Web site (www.invitrogen.com) or by contacting Technical Service (see page 19).**



Comments for pCR®-Blunt 3512 nucleotides

Lac promoter/operator region: bases 95-216
M13 Reverse priming site: bases 205-221
*Lac*Z-alpha ORF: bases 217-570
T7 promoter priming site: bases 400-419
M13 Forward (-20) priming site: bases 427-442
Fusion joint: bases 571-579
*ccd*B lethal gene ORF: bases 580-882
Kanamycin resistance ORF: bases 1231-2025
Zeocin resistance ORF: bases 2231-2605
pUC origin: bases 2673-3386

Sequencing The Transcriptome Of Single Cells

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves should be worn during all practical sessions

C1 single cell auto prep kit box 1 (Fluidigm) – Irritant

NEXTRA XT DNA Sample Preparation Kit (Cat# FC131-1096)
GS#-AM1 (part of Nextra XT DNA sample preparation kit (96 samples) -contains citric acid, formamide, sodium chloride – teratogen and irritant – not to be handled by new or expectant mothers
LP#-LNA1 (part of Nextra XT DNA sample preparation kit (96 samples) - contains formamide, sodium chloride - teratogen and irritant – not to be handled by new or expectant mothers
NX#- TD Tagment DNA Buffer (part of Nextra XT DNA sample preparation kit (96 samples) - contains dimethyl formamide – highly flammable, teratogen and irritant – not to be handled by new or expectant mothers

RNeay plus mini kit((Cat#74035)
Buffer RLT plus contains guanidine thiocyanate and t-Octylphenoxypolyethoxyethanol (part of RNeasy plus Minikit) – Harmful
Buffer RW1 plus contains guanidine thiocyanate and ethanol (part of RNeasy plus Minikit) – flammable and harmful

RNase free DNAse Kit Cat# 79254)
DNase I - contains Deoxyribonuclease (Part of RNase free DNase Kit) –harmful and skin sensitizer

Live /Dead Viability/Cytotoxicity Kit
Calcein AM 4mM solution in DMSO – Toxic/Corrosive/Irritant

Qubit RNA reagent (Q32852 Componant A) contains dimethyl suphoxide (60-100%) – harmful
70% Ethanol – highly flammable
Ampure Beads - irritant

High Sensitivity RNA screen tape sample buffer - Irritant

Purpose: Prepare Illumina Sequence Ready mRNA-to-cDNA Libraries from single cells, bulk cells and purified total RNA

Process Overview : Cells are stained with Calcein-AM and Ethidium homodimer-1, which fluoresces green and red indicating live or dead cells respectively. Stained cells are mixed with 10um beads and pipetted onto a hemocytometer for viewing under a fluorescence microscope (Fig 1). The viability, number and size of the cells are then determined. Cell suspension are then loaded onto an Integrated Fluidic Chip (IFC), which segregates up to 96 single cells inside a Fluidigm C1 machine. Every single cell captured on the IFC are imaged under a fluorescence microscope and its viability status recorded. Next, the C1 machine continues with cell lysis, reverse transcription of mRNA to cDNA and PCR amplification of the cDNA library. Since every cell gets its own set of reaction chambers, distinct gene expression profiles are maintained. Selecting only cDNA libraries from healthy cells for downstream processing is possible, by referring back to the viability status recorded earlier. In parallel, several cDNA libraries called tube controls (TC) are made from 1) bulk cells, 2) purified RNA from bulk cells and 3) cell buffer (No template control). TC data are necessary for quality control purposes. Finally, these cDNA libraries are converted into sequencing ready libraries using the Illumina Nextera XT Kit.

Work Plan

		Procedure	Hands on time	Incubation time	Total time	Cumulative Time
First Day	Pre-lab setup	Aliquot ERCC spike-in mix	10 min		10 min	
		Organize reagents	10 min		10 min	
	Step1: Cell conc , size and viability	Pellet and wash cells		30 min	30 min	
		LIVE/DEAD Stain	5 min	10 min	15 min	1 hr
		Fluorescent microscopy	10 min		10 min	
	Step 2. RNA integrity Check	Purify total RNA from bulk cells	5 min	15 min	20 min	
		Check RNA integrity	5 min	10 min	15 min	
	Step 3-5. Single cell capture and cDNA synthesis	Prime IFC	5 min	10 min	15 min	2 hrs
		Record single cell status	10 min	30 min	40 min	
		Prepare Lysis reaction	5 min	15 min	20 min	3 hrs
		Prepare RT reaction	5 min	2 hr 40 min	2 hr 45 min	6 hrs
		Prepare PCR reaction	5 min	4 hrs	4 hrs	Overnight
		cDNA profile and quantification	30 min	1 hr 30 min	2 hrs	
Second Day	Nextera Library Prep	Dilute cDNA	30 min		30 min	
		Tagmentation	5 min	30 min	35 min	3 hrs
		Stop Tagmentation	5 min	5 min	10 min	
		library indexing and PCR	5 min	30 min	35 min	
		Library quality and quantification	5 min	1 hr	1 hr	5 hrs
		Library pooling and quantification	30 min	1 hr	1 hr 30 min	6.5 hrs

Equipment and Material

- PCR hood
- Eppendorf Centrifuge 5424R or 5424 (or equivalent)
- Micro-Centrifuge for 0.2mL PCR tube
- Life Techonologies Qubit 2.0 Fluorometer
- Qubit RNA HS Assay Kit (Q32852)
- Qubit Assay Tubes (Q32856)
- Agilent 2200 TapeStation
- Agilent High Sensitivity RNA ScreenTape (5067-5579)
- Agilent High Sensitivity RNA ScreenTape Ladder (5067-5581)
- Agilent High Sensitivity RNA ScreenTape Sample Buffer (5067-5580)
- Agilent High Sensitivity DNA ScreenTape (5067- 5584)
- Agilent High Sensitivity DNA Reagents (5067- 5585)
- BioRad Thermo Cycler T100
- Qiagen RNeasy Plus Mini Kit (74035)
    - RNeasy Mini Spin Columns (pink)
    - RNeasy Collection Tubes
    - RNeasy Buffer RLT
    - RNeasy Buffer RW1
    - RNeasy Buffer RPE
- Qiagen RNase-Free DNase Set (79254)
    - DNase I
    - Buffer RDD
- LifeTechnologies LIVE/DEAD Viability/Cytotoxicity Kit (L03224)
- Fluidigm C1 Single-Cell Auto Prep Reagent Kit for mRNA Seq (100-6201)
- Nextera XT DNA Sample Preparation Kit (96 Samples) FC-131-1096
- Nextera XT Index Kit (96 Indices, 384 Samples) FC-131-1002
- Agencourt Ampure XP Beads (A63880)
- LifeTechnologies ERCC RNA Spike-In Mix (4456740)
- Ambion RNA Storage Solution (AM7000)
- Eppendorf Twin.Tec 384 well plate (951020702)
- Eppendorf DNA LoBind Tubes 1.5mL (022431021)
- VWR PCR 8-Tube Strip 0.2mL (53509-304)
- iNCyto C-Chip Disposable Hemocytometer (DHCN012)
- Ethanol 100%
- DNase/RNase free Water

Pre-Lab setup

Aliquot ERCC RNA Spike-In Mix (LifeTechnologies 4456740)

1. Thaw ERCC RNA Spike-In Mix.
2. Dilute 1ul of ERCC RNA Spike-In Mix into 9uL RNA Storage Solution (Ambion AM7000). This is "**ERCC-1/10**".
3. Aliquot "**ERCC-1/10**" into 0.5µL volumes and store at -80°C until use. Use one aliquot for every C1 run.

Organize Reagents

Steps	Reagent	Preparation	Source
Cell Mix	Cells	Keep on ice	-
	Ethidium homodimer-1	−20°C to room temp, avoid light	LIVE/DEAD Kit, LifeTech
	Calcein AM	−20°C to room temp, avoid light	LIVE/DEAD Kit, LifeTech
	Cell Wash Buffer	4°C to room temp	Module 1, Fluidigm
	Cell Suspension Reagent	4°C to room temp, vortex well	Module 1, Fluidigm
Prime IFC	C1 Preloading Reagent	−20°C to room temp	Module 2, Fluidigm
	C1 Harvest Reagent	−20°C to room temp	Module 2, Fluidigm
	C1 Blocking Reagent	4°C to room temp	Module 1, Fluidigm
Lysis Mix	0.5ul Aliquot of ERCC 1/10	−80°C to thaw on ice	-
	Loading Reagent	−20°C to room temp	Module 2, Fluidigm
	RNase Inhibitor	−20°C to thaw on ice	SMARTer Kit, Clontech
	3'SMART CDS Primer IIA	−20°C to thaw on ice	SMARTer Kit, Clontech
	SMARTer Dilution Buffer	−20°C to room temp	SMARTer Kit, Clontech
RT Mix	5X First-Strand Buffer	−20°C to thaw on ice	SMARTer Kit, Clontech
	Dithiothreitol (DTT)	−20°C to thaw on ice	SMARTer Kit, Clontech
	dNTP Mix (dATP, dCTP, dGTP, dTTP each at 10 mM)	−20°C to thaw on ice	SMARTer Kit, Clontech
	SMARTer IIA Oligonucleotide	−80°C to thaw on ice	SMARTer Kit, Clontech
	SMARTScribe Reverse Transcriptase	−20°C to thaw on ice	SMARTer Kit, Clontech
PCR Mix	PCR-Grade Water	Keep at room temp	Advantage2 PCR Kit, Clontech
	10X Advantage2 PCR Buffer	−20°C to thaw on ice	Advantage2 PCR Kit, Clontech
	50X dNTP Mix	−20°C to thaw on ice	Advantage2 PCR Kit, Clontech
	IS PCR primer	−20°C to thaw on ice	SMARTer Kit, Clontech
	50X Advantage2 Polymerase Mix	−20°C to thaw on ice	Advantage2 PCR Kit, Clontech
Purify Total RNA	DNase I	−20°C to thaw on ice	Qiagen RNase-Free DNase Set
	Buffer RDD	−20°C to thaw on ice	Qiagen RNase-Free DNase Set
Harvest cDNA	C1 DNA Dilution Reagent	−20°C to room temp	Module 2, Fluidigm

Day1 Step1: Determine the cell concentration, size and viability

Note: (Keep cells on ice unless specified otherwise).

1. Pipette 10uL of cells onto an iNCyto Disposable Hemocytometer (DHCN012).
2. Pellet rest of the cells by centrifugation for 10 min at 300 x g. During centrifugation
   - **a.** count the cells on the hemocytometer.
   - **b.** make LIVE/DEAD Stain by adding 1uL Calcein-AM and 4uL Ethidium Homodimer-1 to 500uL of Cell Wash Buffer.
3. Aspirate the liquid from the cell pellet, and re-suspend pellet in 500uL of LIVE/DEAD Stain. Incubate at room temperature for 10 min.
4. Pellet, aspirate and re-suspend cells in 500uL of Cell Wash Buffer. Do this twice but re-suspend the final pellet in appropriate amount of Cell Wash Buffer to get a cell suspension of ~1000-3000 cells/uL. Use the cell count obtained from step **2a**.
5. Make 50uL of cell suspension with 200-300 cells/uL.
6. Mix 8uL of cell suspension (200-300 cells/uL) with 2uL of Countess 10um beads then pipette the 10uL mix onto a hemocytometer.
7. View hemocytometer with fluorescent microscope to get cell **concentration, size and viability** (Fig 1).
   a. Re-wash the cell stock if necessary to eliminate large debris which may clog the microfluidics inside the IFC.
   b. Estimate the cell size relative to the 10um beads (Fig1C). Use this cell size to determine the appropriate IFC to use. IFC have capture sites for cells with diameter of 5-10um, 10-17um or 17-25um.
   c. Find the percentage of live cells to estimate the number of live single cells that can be captured.

   Live Stain (Green) = Calcein, Ex/Em=494/517nm
   Dead Stain (Red) = Ethidium homodimer-1, Ex/Em=528/617nm

8. Reserve 50uL of the stock cell (1000-3000 cells/uL). Pellet rest of the cells and proceed to Day1 Step2 - Total RNA extraction.

A. **B.** **C.**

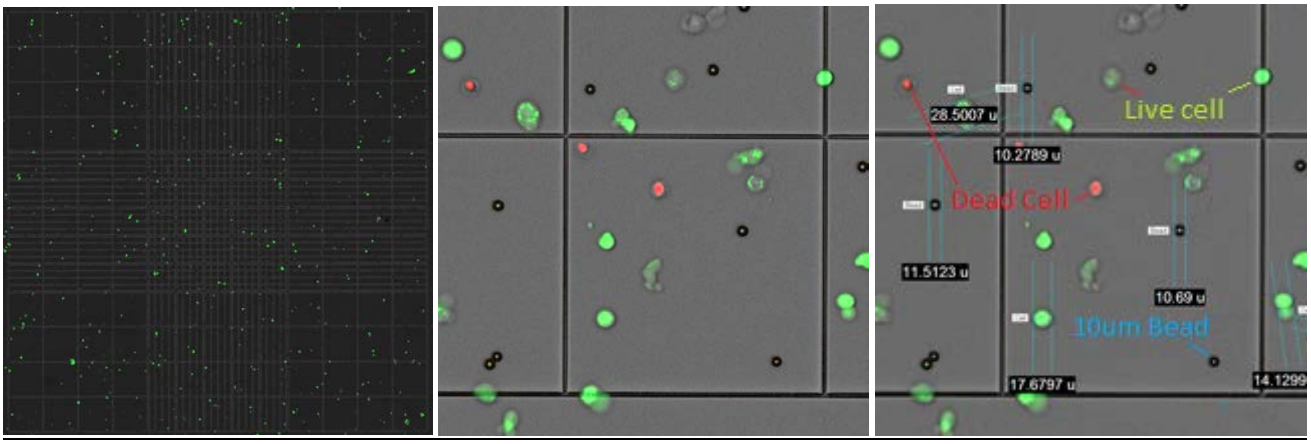


Fig1. Image of live dead stained cells on a hemocytometer

A. 10X image of a heterogeneous cell population stained with Live/Dead stain.
B. Zoomed-in image of live dead stained cells on a hemocytometer
C. Cell size ranged from 14-17um thus IFC with 10-17um capture sites will be appropriate.

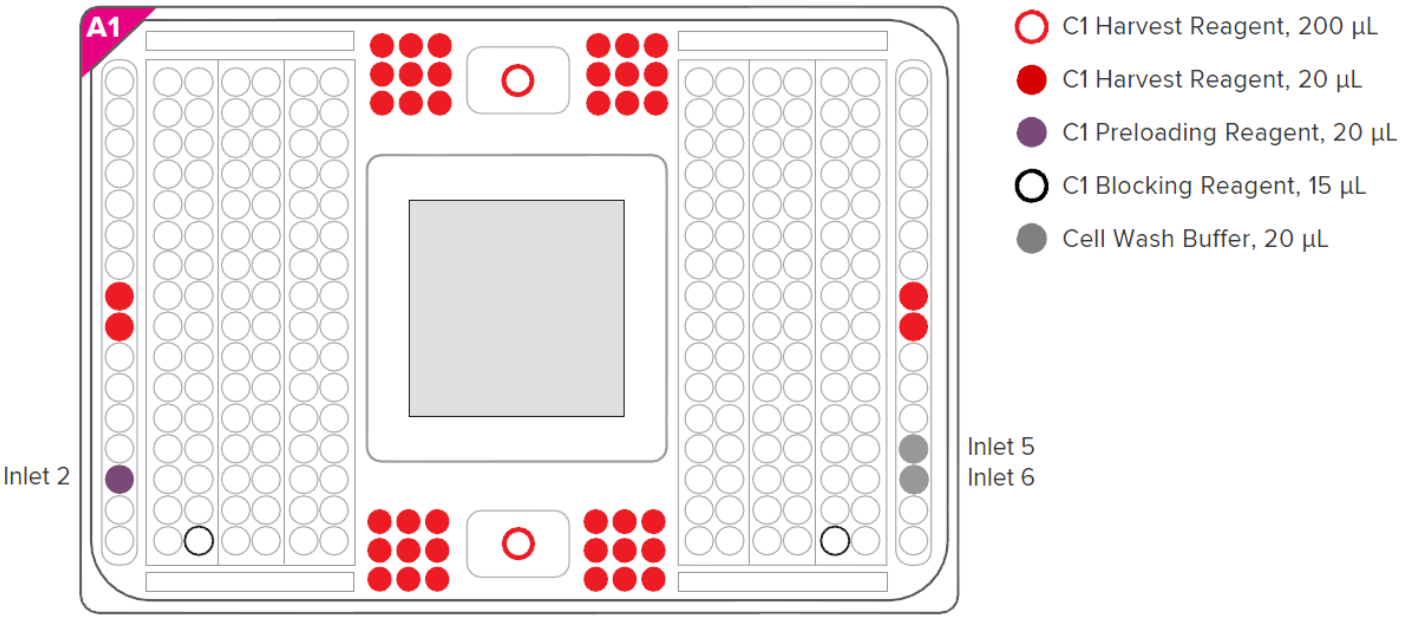
Day1 Step2: Purify total RNA (using RNeasy Mini Kit) and determine its integrity (using RNA ScreenTape)

Note: When using RNeasy for the first time, add 4 volumes of ethanol (96–100%) to RNeasy Buffer RPE as indicated on the bottle to obtain a working solution.

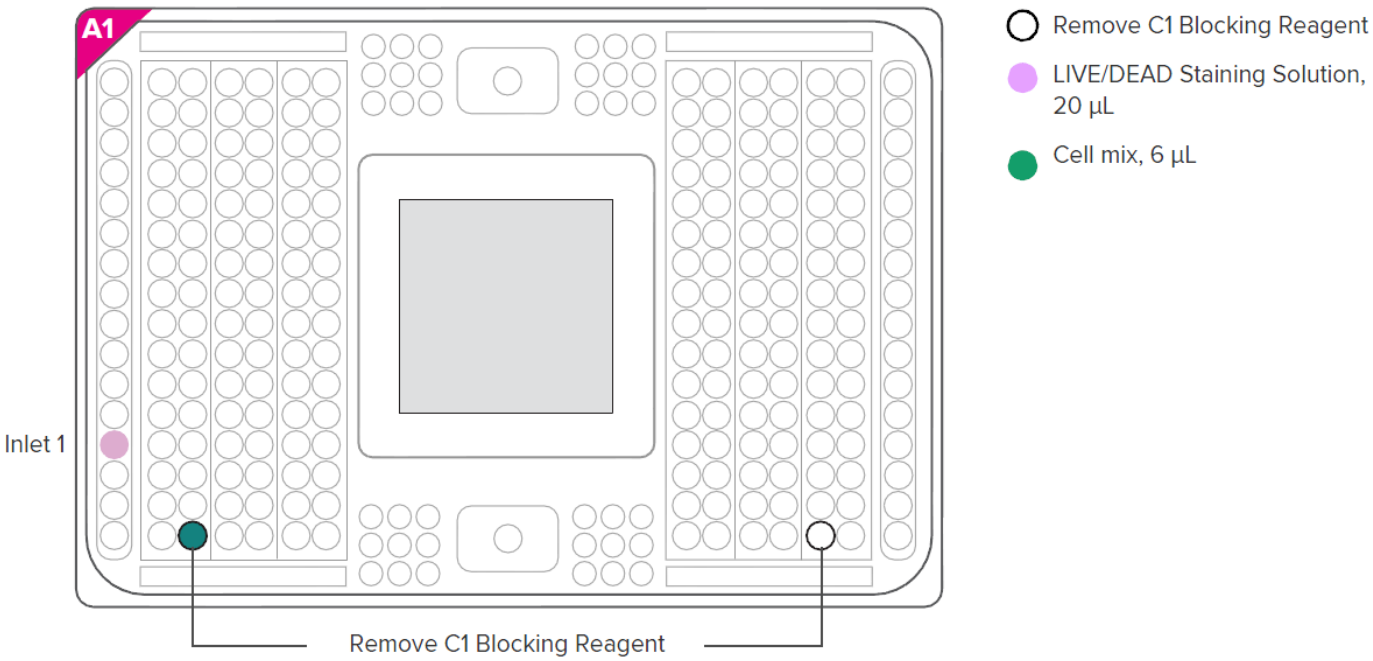
1. Top up cell suspension to 350uL with RNeasy Buffer RLT.
2. Invert and mix lysate well to make sure cells are lysed.
3. Add 1 volume of 70% ethanol, invert to mix well and spin down briefly.
4. Transfer up to 700uL of lysate to a RNeasy Mini Spin Column (pink color).
5. Centrifuge for 15s at >= 8000 x g and discard flow through.
6. Add 350µL RNeasy Buffer RW1 to the RNeasy spin column.
7. Centrifuge for 15s at >= 8000 x g and discard flow through.
8. Mix 10µL DNase I stock solution with 70uL Buffer RDD and add to spin column membrane.
9. Incubate at room temperature for 15 min.
10. Add 350µl RNeasy Buffer RW1 to the RNeasy spin column.
11. Centrifuge for 15s at >= 8000 x g and discard flow through.
12. Add 500µl Buffer RPE.
13. Centrifuge for 15s at >= 8000 x g and discard flow through.
14. Add 500µl Buffer RPE.
15. Centrifuge for 2 min at >= 8000 x g and place the RNeasy spin column into a new DNA/RNA low binding tube.
16. In a PCR hood, open the lid of the RNeasy spin column and allow the residual ethanol to evaporate (~10 min).
17. Pipette 25uL of DNase/RNase free water onto the membrane inside the spin column. Elute the total RNA by centrifuge for 30 sec at >= 8000 x g. Repeat one more time to get 50uL of total RNA.
18. Quantify the total RNA with Qubit, then load 2uL of total RNA (5pg-5ng/uL) onto a RNA ScreenTape to determine its integrity.
19. Continue to C1 if RNA integrity number (RIN) is 7.5 or higher.

Day1 Step3 : Prime the IFC and single cell capture

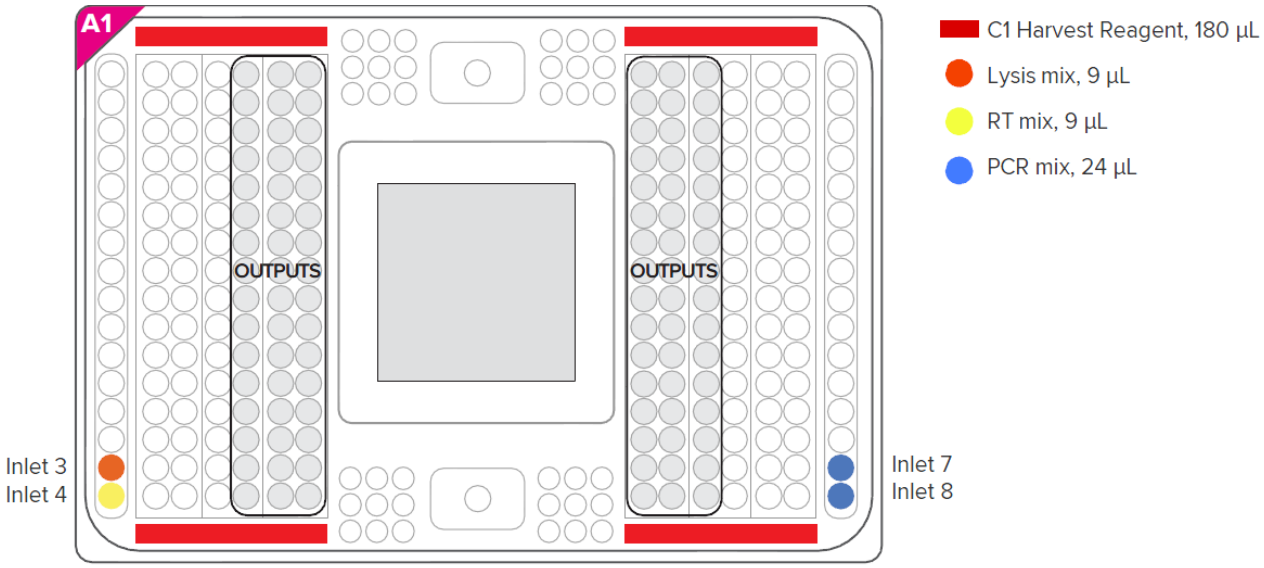
1. Base on the cell size pick the appropriate IFC (5-10µm, 10-17µm or 17-25µm).
2. Prime the IFC (**10 minutes**):
    a. Load reagents onto the IFC according to diagram below
    b. Load IFC onto C1 and run “mRNA Seq: Prime” script



3. Capture single cells with IFC **(15min. for 5–10µm cells, 30min. for larger cells)**
    a. Take out the IFC and remove blocking reagent
    b. Load reagents onto the IFC according to diagram below
    c. Mix 2uL of Cell Suspension Reagent with 3uL of cell suspension (200-300 cells/uL). Load the 5uL "Cell mix onto the IFC.
    d. Load IFC onto C1 and run "mRNA Seq: Cell Load" script. During "Loading" goto **Day1 Step4: Prepare Clontech Master Mixes.**
    e. Eject IFC and record viability status of captured single cells.



4. Run Lysis, Reverse Transcription and PCR program on the C1 System **(8 hours)**
    a. Load reagents onto the IFC according to the diagram below.
    b. Load IFC onto C1 and run "mRNA Seq: RT & Amp" script.
    c. Set a convenient time on the C1 to harvest the cDNA.



Day1 Step4: Prepare Clontech Master Mixes

1. Label 3 tubes **LYS, RT, PCR** for the 3 master mixes,
2. Dilute a 0.5uL aliquot of "ERCC1/10" by adding 24.5uL of C1 Loading Reagent (Fluidigm). This makes the "ERCC1/500".
3. Make the following master mixes

	Lysis Mix	vol (uL)
1	ERCC 1/500	1.50
2	RNase Inhibitor (Clontech, 40 U/uL)	0.75
3	3' SMART CDS Primer IIA (Clontech, 12uM)	10.50
4	Clontech Dilution Buffer (Do not vortex)	17.25
	Total =	30.00

	RT Mix	vol (uL)
1	C1 Loading Reagent (Fluidigm)	1.50
2	5X First-Strand Buffer (Clontech)	14.00
3	Dithiothreitol (Clontech, 100 mM)	1.75
4	dNTP Mix (Clontech , dATP, dCTP, dGTP, and dTTP, each at 10 mM)	7.00
5	SMARTer IIA Oligonucleotide (Clontech, stored at −80 ºC)	7.00
6	RNase Inhibitor (Clontech, 40 U/uL)	1.75
7	SMARTScribe Reverse Transcriptase (Clontech, 100U/uL)	7.00
	Total =	40.00

	PCR Mix	vol (uL)
1	PCR-Grade Water	88.90
2	10X Advantage 2 PCR Buffer (not SA, short amplicon) (Advantage 2 Kit)	14.00
3	50X dNTP Mix (Advantage 2 PCR Kit)	5.60
4	IS PCR primer (Clontech SMARTer Kit)	5.60
5	50X Advantage 2 Polymerase Mix (Advantage 2 PCR Kit)	5.60
6	C1 Loading Reagent (Fluidigm)	6.30
	Total =	126.00

Day1 Step5: Prepare Tube Control reactions (Lysis > RT > PCR)

1. Pipette 2uL of **Lysis Mix** into 7 PCR tubes for the following TC samples…
    - **a.** TC_200cells_1of2 SMARTer RT_cDNA
    - **b.** TC_200cells_2of2 SMARTer RT_cDNA
    - **c.** TC_1000cells_1of2 SMARTer RT_cDNA
    - **d.** TC_1000cells_2of2 SMARTer RT_cDNA
    - **e.** TC_RNeasy_1of2 SMARTer RT_cDNA
    - **f.** TC_RNeasy_1of2 SMARTer RT_cDNA
    - **g.** SMARTer_NTC (No Template Control) RT_cDNA
2. Pipette 1uL of cell suspension (200-300 cells/uL) into TC_200 tubes.
3. Pipette 1uL of cell suspension (1000 cells/uL) into TC_1000 tubes.
4. Pipette 1uL of RNeasy total RNA (up to 7ng/uL) into TC_RNeasy tubes.
5. Pipette 1uL of Clontech Wash Buffer into the Smarter_NTC tube.
6. Mix well, spin down briefly and run the lysis protocol. Each lysis reaction volume is 3ul.

Lysis protocol

Temperature	Time	Purpose
72°C	3 min	Cell lysis, unfolding of RNA secondary structures, Poly-T primer binding
4°C	10 min	Poly-T primer binds
25°C	1 min	Poly-T primer binds more specifically
4°C	Hold	

7. Pipette 4uL of RT mix to each lysis reaction.
8. Vortex to mix well, spin down briefly and run the RT protocol. Each RT reaction volume is 7uL.

RT protocol

Temperature	Time	Cycle	Purpose
42°C	90 min	1	RT and template-switching
70°C	10 min	1	Enzyme inactivation
4°C	Hold	1	

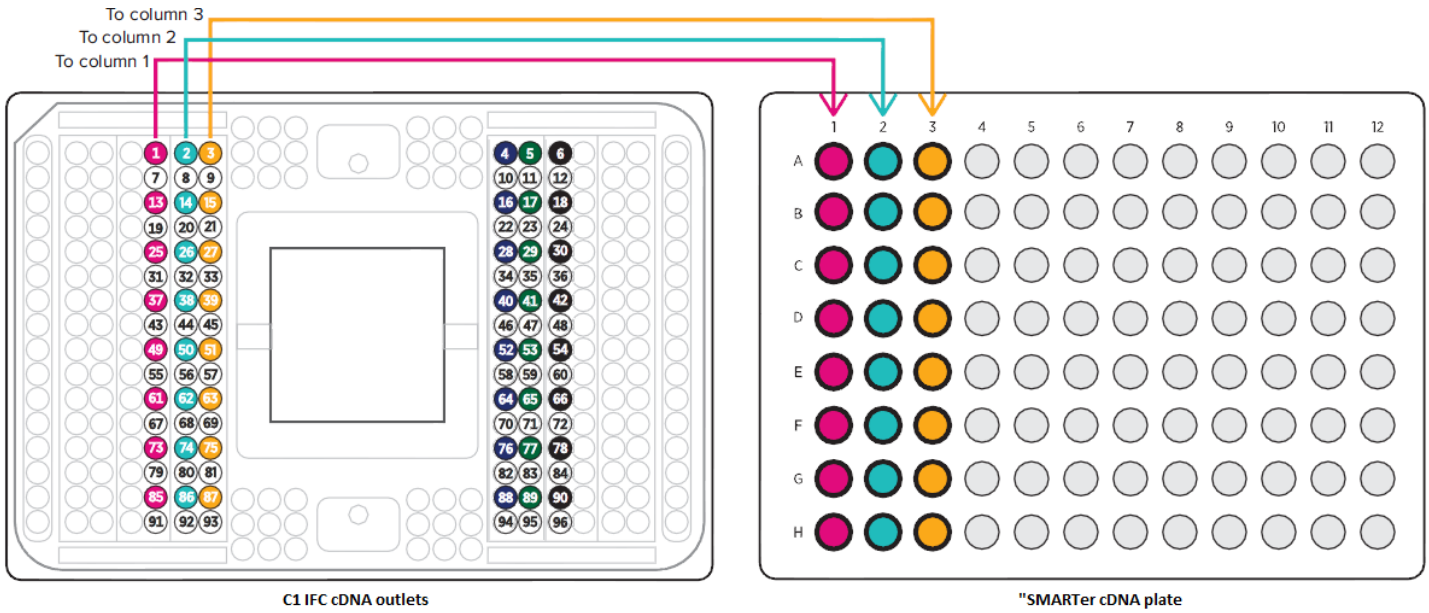
9. Pipette 9uL of PCR Mix into 7 new PCR tubes for the following…
    - **a.** TC_200cells_1of2 Smarter PCR_cDNA
    - **b.** TC_200cells_2of2 Smarter PCR_cDNA
    - **c.** TC_1000cells_1of2 Smarter PCR_cDNA
    - **d.** TC_1000cells_2of2 Smarter PCR_cDNA
    - **e.** TC_RNeasy_1of2 Smarter PCR_cDNA
    - **f.** TC_RNeasy_1of2 Smarter PCR_cDNA
    - **g.** Smarter_NTC (No Template Control) PCR_cDNA
10. Pipette 1uL of RT_cDNA to the corresponding PCR cDNA tubes.
11. Vortex to mix well, spin down briefly and run PCR protocol. Each PCR reaction volume is 10uL.

PCR protocol

Temperature	Time	Cycle
95°C	1 min	1
95°C	20 sec	5
58°C	4 min	
68°C	6 min	
95°C	20 sec	9
64°C	30 sec	
68°C	6 min	
95°C	30 sec	7
64°C	30 sec	
68°C	7 min	
72°C	10 min	1
4°C	Hold	1

Day2 Step1: Harvest, analyze and quantify each single cell cDNA libraries

1. Label a 96 well plate "**SMARTer cDNA**" and pipette 20uL of C1 DNA Dilution Reagent (Fluidigm) to each well.
2. Transfer ~3uL of single cell cDNA libraries from IFC outlets into "**SMARTer cDNA"** plate.



C1 IFC cDNA outlets "SMARTer cDNA plate

3. Get the cDNA profile of all single cell and tube controls using High Sensitivity LabChip (Fig 2).

    a. Wash the High Sensitivity LabChip.
    b. Prime the High Sensitivity LabChip.
    c. Pipette 15uL of Single cell cDNA onto a 384 well plate.
    d. Pipette 14uL water onto the 384 well plate then add 1uL of the TC cDNA libraries.
    e. Pipette 15uL of water onto an empty well in 384 well plate. This will be the background trace.
    f. Vortex to mix well and centrifuge at max to get rid of bubbles.
    g. Run the High Sensitivity LabChip.

4. Make a list of cDNA libraries to be sequenced based on the single cell status recorded earlier and their cDNA profile.

5. Label a 96 well plate "**SMARTer cDNA 0.3 ng/uL**" and pipette a calculated amount of DNA Dilution Buffer to add, so that when 2uL of cDNA library is added, it will have a cDNA concentration of 0.1-0.3ng/uL.
6. Label a 96 well plate "**Tag cDNA**" and pipette 1.25uL of cDNA from "**SMARTer cDNA 0.3 ng/uL**" to "**Tag cDNA**" plate.

Day2 Step2: Nextera XT DNA Library Preparation

1. Based on the number of cDNA libraries to be sequenced, make the Tagmentation Master Mix.

	Tagmentation Master Mix per cDNA library	vol (uL)
1	Tagment DNA Buffer	2.5
2	Ampicon Tagment Mix	1.25
	Total =	3.75

2. Pipette 3.75 uL of Tagmentation Master Mix to the 1.25uL of cDNA libraries on the "**Tag cDNA**" plate.
3. Vortex to mix well, spin down briefly and run Nextera Tagmentation protocol. Each Tagmentation reaction is 5u**L.**

Nextera Tagmentation protocol

Temperature	Time
55°C	10 min
10°C	Hold

4. Immediately neutralize the Tagmentation reaction by adding 1.25 uL of NT Buffer to each sample.
5. Spin down briefly and vortex at medium speed for 5 min. Each reaction is now 6.25u**L.**
6. PCR amplification of the Tagmented fragments.

    a. Pipette 3.75uL NPM (PCR mix) to each sample.
    b. Plan a unique combination of Index1+Index2 for each sample.
    c. Pipette 1.25uL of Index1 N7XX (Forward primer) to each sample.
    d. Pipette 1.25uL of Index2 S5XX (Reverse primer) to each sample.
    e. Vortex to mix well, spin down briefly and run the Nextera PCR protocol. Each PCR reaction is 12.5uL.

Nextera PCR protocol

Temperature	Time	Cycle
72°C	3 min	1
95°C	30 sec	1
95°C	10 sec	12
55°C	30 sec	
72°C	60 sec	
72°C	10 min	1
10°C	hold	

Day2 Step3: Library Pooling and Clean up

1. Get the Tagmented cDNA profiles using High Sensitivity LabChip (Fig 3).

    a. Wash the High Sensitivity LabChip.
    b. Prime the High Sensitivity LabChip.
    c. Pipette 14uL water onto the 384 well plate then add 1uL of the Tagmented cDNA libraries.

d. Pipette 15uL of water onto an empty well in 384 well plate. This will be the background trace.
e. Vortex to mix well and centrifuge at max to get rid of bubbles.
f. Run the High Sensitivity LabChip.

2. Based on the concentration of the Tagmented cDNA in the 200bp-500bp range, pool equal amount of Tagmented cDNA library from all samples into a 1.5ml DNA low bind tube labeled "**Pooled library**".
3. Add an amount of Ampure XP Beads equal to 70% of the volume of the "**Pooled library**".
4. Vortex the "**Pooled library**" to mix well, spin down briefly and incubate at room temperature for 10 min.
5. Place the "**Pooled library**" on a magnetic stand for 2 min.
6. Remove the "**Pooled library**" and discard the supernatant.
7. With the "**Pooled library**" still on magnetic stand, add 200 uL of 70% ethanol.
8. Wait for beads to settle then remove the ethanol
9. Repeat Step 7-8 two more times for a total of 3 ethanol wash.
10. Evaporate the residual ethanol by allowing the beads to air dry for 10-15 min.
11. Elute the bead bound cDNA
    a. Add 20 uL of Nextera Resuspension Buffer to the dried beads.
    b. Remove the "**Pooled library**" from magnetic stand, vortex to mix well, spin down briefly and incubate at room temperature for 2 min.
    c. Place the "**Pooled library**" on magnetic stand for 2 min.
    d. Transfer the supernatant to a new tube labelled "**Sequence ready library**".
    e. Repeat Steps a. thru e. so that the final "**Sequence ready library**" yield is 40uL.
12. Pipette 14uL of water onto an empty well in the 384 well plate then add 1uL of the "**Sequence ready library**".
13. Check the profiles of "**Sequence ready library**" using the High Sensitivity LabChip (Fig 4).

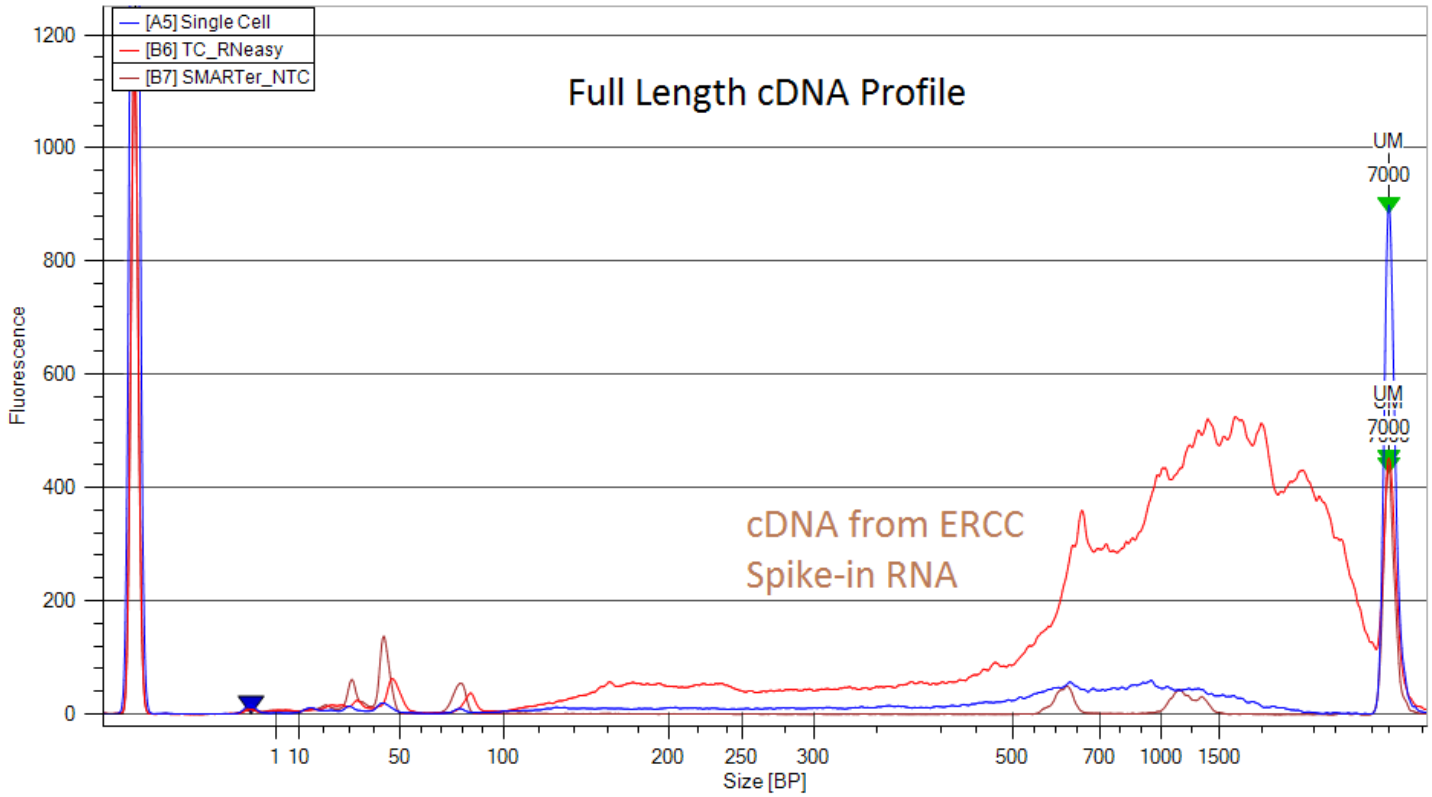


Fig 2. Full length cDNA profiles from a single cell (Blue), TC_RNeasy (Red) and SMARTER_NTC (Brown). Note the SMARTer_NTC profile shows 2 prominant peaks around 600bp and 1200bp. These peaks represent cDNA from the External RNA Controls Consortium (ERCC) Spike-in RNAs of which sequences are known. SMARTER_NTC is a positive control and tests for success of RT reaction and possible contaminating RNA.

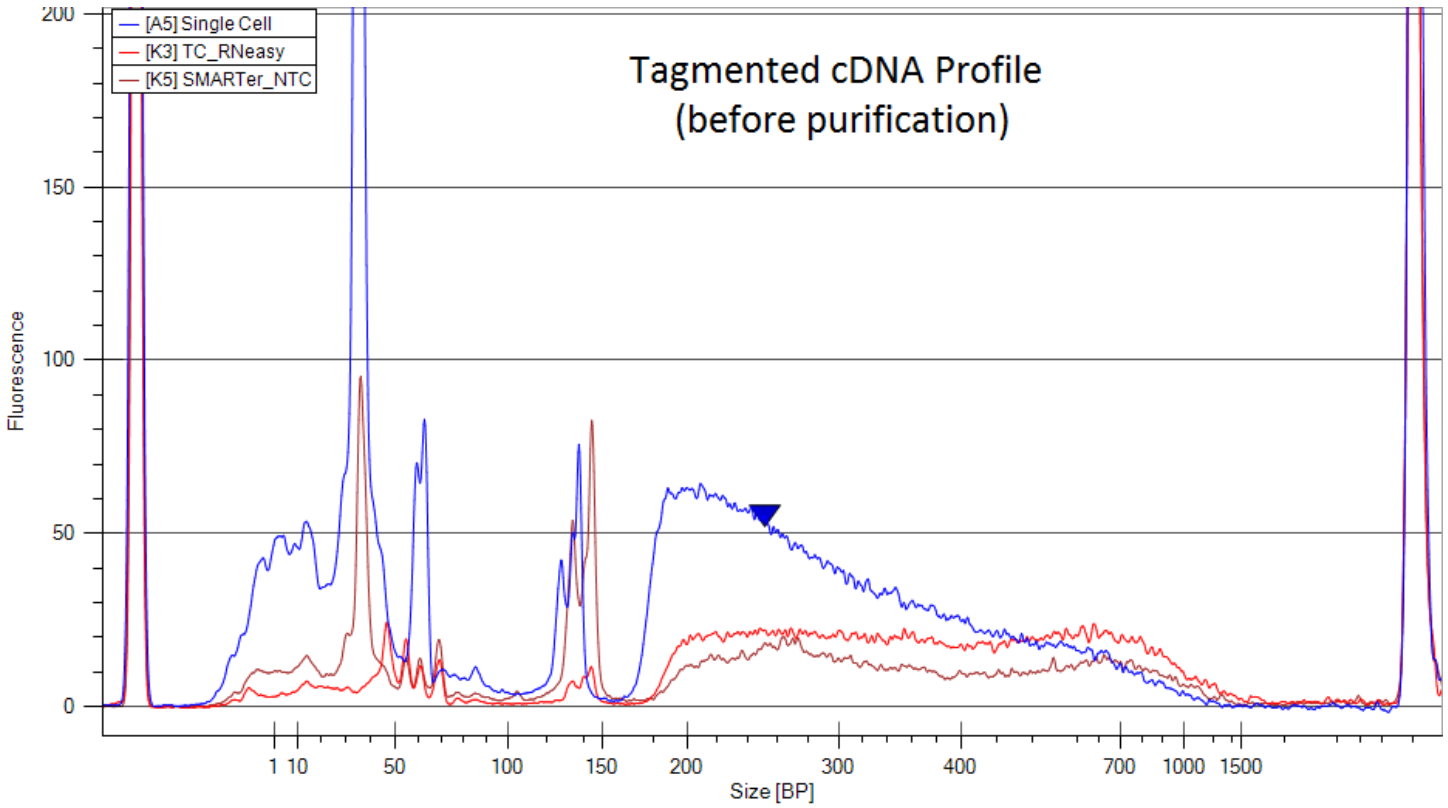


Fig 3. Tagmented cDNA profiles of samples from Fig 2.

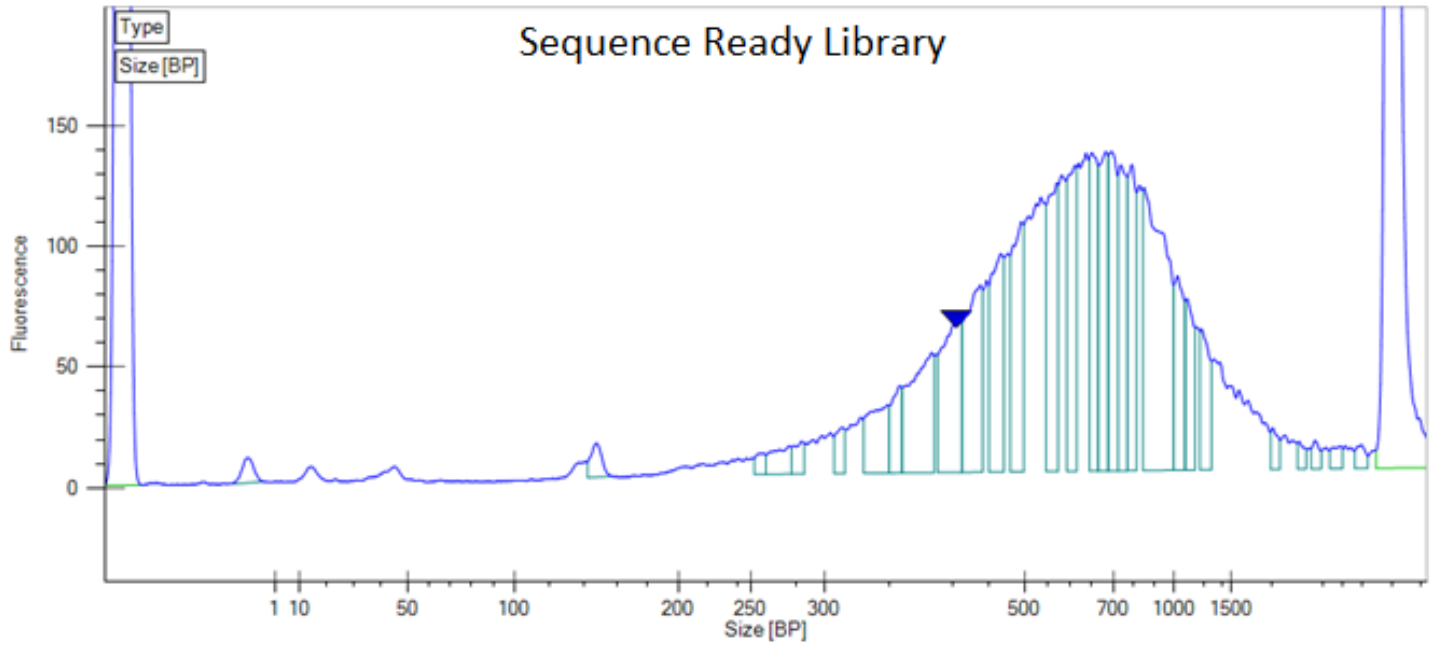


Fig 4. Profile of he final Sequence Ready Library.

9. Analysis of Protein Complexes using Affinity Purification and Mass Spectrometery (AP-MS)

Analysis of Protein Complexes using Affinity Purification and Mass Spectrometry (AP-MS)

Most cellular functions and biological processes require the coordinated action of a number of proteins that are usually assembled in multi-protein complexes or interacting in dynamic signalling pathways. The analysis of protein complexes and interactions is therefore of central importance in biological research.

Recent advances in mass spectrometry have turned it into the technology of choice for identifying and characterising the primary structure of proteins. Its advantages include sensitivity (being able to identify peptides present at femtomole levels), speed, compatibility with high-throughput strategies and easy automation. It also allows for the characterisation of post-translational modifications and can be adapted to perform quantitative analyses. Two techniques are used most commonly to generate molecular ions from peptides: matrix-assisted laser desorption/ionisation (MALDI) and electrospray ionisation (ESI). Based on these, a range of mass analysers and instruments, with different characteristics and performances, are available and suited to tackling the varied challenges of protein analysis (*Aebersold & Mann, 2003; Domon & Aebersold, 2006*).

The tandem affinity purification (TAP) method is a generic purification strategy that allows rapid and efficient purification of protein complexes (*Rigaut et al., 1999*). It requires fusion of a dual affinity tag, either N or C terminally, to the protein of interest. The original TAP tag consisted of two IgG binding domains of Protein A and a calmodulin binding peptide (CBP) separated by a TEV protease cleavage site. The use of a generic tag in principle allows for parallel sample preparation without the need to optimise the purification protocol for each protein complex, making it suitable for large-scale studies. We recently modified the original TAP tag by substituting the IgG binding domains for 3 copies of the FLAG peptide, and by introducing an extra TEV site (*Pardo et al., 2010*). We will use this new tag (FTAP) in the practical.

The TAP purification takes place in two steps, first through magnetic beads covalently linked to an anti-FLAG antibody, from which the bait (and associated proteins) are eluted with TEV protease, and then via the CBP on calmodulin-coated beads in the presence of calcium. This second step allows elution in physiological conditions with EGTA, making the protein complex also amenable to structural studies. Many other combinations of epitope tags have been used successfully. Although using a dual affinity tag, in the course we will perform single affinity purification. This is better suited to medium or high throughput studies, where tandem affinity purification would be impractical in terms of time, cost and starting material required.

Once the complex is isolated, a typical workflow involves partial separation of the interacting proteins by one-dimensional gel electrophoresis. Protein bands are excised, and proteins digested in-gel by sequence specific proteolysis (normally using trypsin). An alternative to the Gel-LC/MS approach is to directly digest the protein complex in solution. A third possibility is to forgo elution and perform the digestion of bound proteins on the beads. This is the approach we will follow in this course. It offers the advantage of bypassing lengthy gel running and post-processing steps.

The mixture of peptides is extracted, separated by HPLC and analysed by ESI-tandem mass spectrometry, which measures the mass of whole peptides and involves further isolating specific peptides and breaking them into smaller fragments that are measured to produce the tandem mass spectra. Detailed peptide structural features can be deduced from the analysis of the masses of the peptides and fragments ions subsequently generated during MS/MS. The tandem mass spectra are converted into peak lists, which are then matched to peptide sequences using a variety of algorithms (*see Marcotte, 2007* and *Nesvizhskii, 2010 in the folder for a quick intro to protein identification*).

The Proteomic Mass Spectrometry group at the Sanger Institute normally use Mascot (Matrix Science, UK) as the protein identification algorithm (Matrix Science Mascot Help at http://www.matrixscience.com/search_form_select.html is a good place to start learning about it). Mascot matches the experimental tandem mass spectra to theoretical spectra generated by *in silico* "digestion" of a protein database, and outputs a list of identified proteins together with the identifying peptides. A peptide ion score is assigned to each peptide match based on the probability that the observed match is a random event. The accuracy of peptide identifications can be assessed using a random database strategy. A random database is generated by shuffling the amino acids in each protein in the real database, and then used for the database searching. Mascot can output a false discovery rate (FDR), calculated based on the number of peptides identified in the random versus the real database, automatically in each search.

Confidence that a protein has been identified correctly generally comes from the confidence in correct peptide matches, represented by a low FDR and good peptide scores with top ranking matches, and from getting multiple matches to peptides from the same protein. In most published works only proteins identified by two or more confident peptides are reported.

Due to the extremely high sensitivity of mass spectrometry, it is important to perform a parallel analysis on a negative control sample. Even though bands may not be seen in the control purification, proteins are often identified from it! This control will help in discriminating true interactors from background binding.

USEFUL REFERENCES

General MS Proteomics

Aebersold & Mann, 2003. Mass spectrometry-based proteomics. Nature, 422: 198-207.

Domon & Aebersold, 2006. Mass spectrometry and protein analysis. Science 312: 212-217.

Marcotte, 2007. How do shotgun proteomics algorithms identify proteins? Nature Biotech 25: 755-757.

Walther & Mann, 2010. Mass spectrometry-based proteomics in cell biology. J Cell Biol 190: 491-500.

Quantitative mass spectrometry

Bantscheff et al., 2007. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 389: 1017-1031.

Wilm, 2009. Quantitative proteomics in biological research. Proteomics 9: 4590-4605.

Affinity purification-mass spectrometry

Rigaut et al., 1999. A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol. 17:1030-2.

Gingras et al., 2007. Analysis of protein complexes using mass spectrometry. Nature Reviews Mol Cell Biol 8: 645-654.

Pardo & Choudhary, 2012. Assignment of Protein Interactions from Affinity Purification/Mass
Spectrometry Data. J Proteome Res 11: 1462-74

Dunham et al., 2012. Affinity-purification coupled to mass spectrometry: basic principles and strategies. Proteomics 12:1576-90.

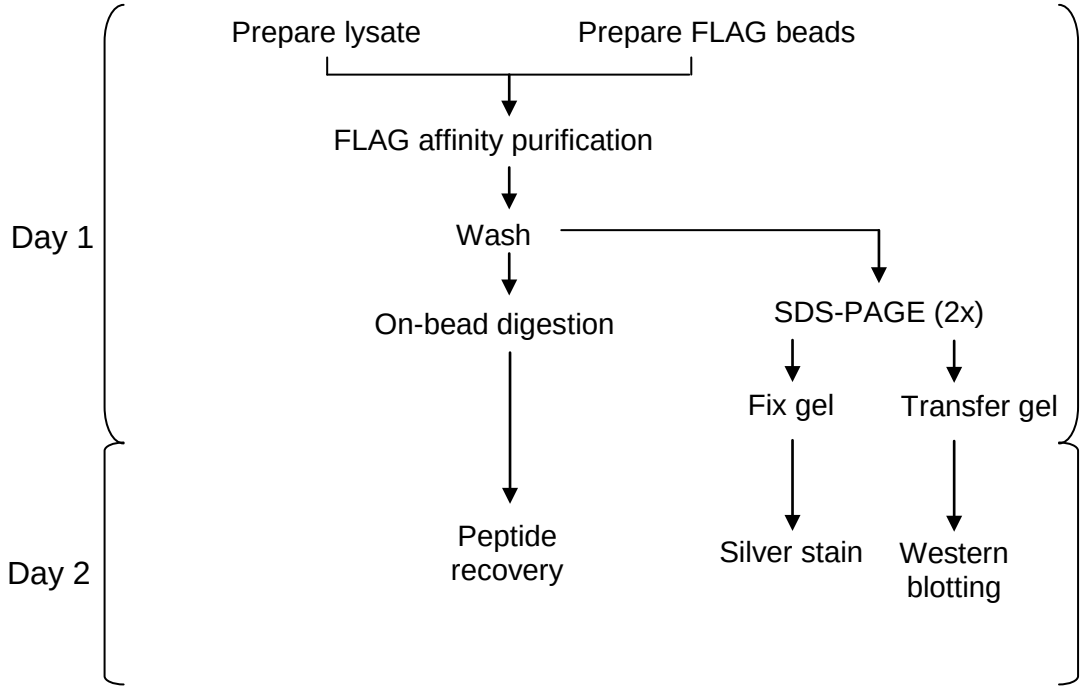
Gingras & Raught, 2012. Beyond hairballs: The use of quantitative mass spectrometry data to understand protein-protein interactions. FEBS Lett 586:2723-31.

The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. Nature Methods (2013) 10: 730–736.

Oeffinger, 2012. Two steps forward--one step back: advances in affinity purification mass spectrometry of macromolecular complexes. Proteomics 12:1591-608.

Proteomics Special Issue on Protein-protein interactions:
http://onlinelibrary.wiley.com/doi/10.1002/pmic.v12.10/issuetoc

Workflow


Prepare lysate
Prepare FLAG beads
FLAG affinity purification
Day 1
Wash
On-bead digestion
SDS-PAGE (2x)
Fix gel
Transfer gel
Peptide recovery
Silver stain
Western blotting
Day 2

Protocol 1: FLAG affinity purification from mouse embryonic stem cells

(*Adapted from Pardo et al., Cell Stem Cell, 2010, 6: 382-395)*

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical

Reagents

- Lysis buffer (50 mM Tris-HCl pH 8, 450 mM NaCl, 0.2% NP-40, 1 mM EDTA, 2 mM MgCl2, 1 mM DTT*, protease inhibitors* -Complete, EDTA-free, Roche)
- Dilution buffer (50 mM Tris-HCl pH 8, 0.05% NP-40, 1 mM EDTA, 2 mM MgCl2, 1 mM DTT*, protease inhibitors* -Complete, EDTA-free, Roche)
- Dynabeads Protein G (Life Technologies)
- Anti-FLAG antibody M2 (F1804, Sigma)
- IPP150 buffer (10 mM Tris-ClH pH 8, 150 mM NaCl, 0.1% NP-40)

* Protease inhibitors and DTT should be added immediately prior to use.

Procedure

All buffers should be ice-cold. Place ten 1.5-ml eppendorf tubes and a couple of 15 ml Falcon tubes in ice. Place homogeniser on ice. Tubes containing samples should always be kept in ice.

Preparation of anti-FLAG beads

1. Mix Protein G Dynabeads well and take 100 μl into eppendorf tube. Insert the tube in the magnet for 1 min. Remove liquid without touching the beads.
2. Re-suspend beads in 1 ml of PBS-0.01% Tween 20 by pipetting. Insert tube in magnet for 1 min, and remove liquid.
3. Re-suspend beads in 90 μl of PBS-0.01% Tween 20 and add 10 μl of anti-FLAG M2. Incubate in roller for 15 min at room temperature. Remove liquid by placing in magnet as before.
4. Re-suspend beads in 1 ml of PBS-0.1% Tween 20. Leave in ice until ready to start the affinity purification.

Preparation of lysate

5. To 10 ml of lysis buffer add 1 tablet of protease inhibitor cocktail (grinded to powder) and 10 ul of 1M DTT (final concentration 1 mM).
6. Incubate cell pellet briefly at 37°C until starting to thaw. Resuspend the cells thoroughly by gently tapping the tube. Leave in ice.

7. Add 5 ml of ice-cold lysis buffer to the cell suspension and mix by gently swirling the tube. Incubate in ice for 10 min.
8. Transfer cell suspension to cold homogeniser and lyse cells by grinding with the tight pestle, moving up and down at least 20 times (*keep homogeniser in ice whilst doing this and avoid making bubbles*). Do some more strokes if the lysate is still viscous.
9. Transfer lysate to cold eppendorf tubes. Pellet debris in a cooled centrifuge at 13000 rpm for 15 minutes.
10. Pool supernatants into a cold 15 ml Falcon tube (*Note: do not disturb the pellet. It is better to leave some supernatant behind, since the supernatant close to the pellet is a source of non-specific background contamination. Do not aspirate the viscous DNA "blob" if present*). Add twice the volume of dilution buffer to the lysate to bring the salt and NP40 concentrations down to 150 mM and 0.1% respectively. Collect a 30 μl aliquot of the diluted lysate and leave in ice.

Affinity purification set up

11. Place anti-FLAG beads tube in the magnet and remove the supernatant. Take 1 ml of lysate and use it to re-suspend the beads. Add this to the rest of the lysate. Add 1 ul/ml of benzonase.
12. Incubate lysate with beads in a rotating wheel at 4°C for 1 hour.
13. Place Falcon with lysate and beads in the magnet for 1 min. Collect a 30 μl aliquot of the supernatant. Discard the rest of the supernatant (*Note*: *bait and associated protein should now be stuck to the beads. In real experiments, store the supernatant until the purification has been checked, just in case something has gone wrong*).
14. Wash beads with 1 ml of ice-cold IPP150 buffer by pipetting up and down 4-5 times. Transfer the sample to a new cold 1.5-ml eppendorf tube.
15. Wash beads four more times with 1 ml of cold IPP150, try to keep the pipetting reproducible. Before you remove the supernatant of the last wash, collect 110ul aliquot of beads + buffer for analysis by PAGE/silver staining (Protocol 3) (*Note: this aliquot contains the undigested protein complex*).

Continue on to Protocol 2 (On-bead digestion) and Protocol 3 (PAGE).

Notes

- For soluble cytoplasmic proteins, reduce the amount of salt in the lysis buffer to 150 mM NaCl and detergent to 0.1%, and do not dilute the lysate.
- A protocol for TAP for membrane proteins can be found in Fernandez *et al*., Mol Sys Biol (2009) 5:269.

Protocol 2: On-bead digestion

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical

Reagents

- Trypsin sequencing grade (Promega): Stock solution 0.1 ug/ul in 0.5% formic acid
- 50 mM ammonium bicarbonate (prepared fresh; will last 2 days)
- IPP150 buffer no NP-40 (10 mM Tris-ClH pH 8, 150 mM NaCl)

Procedure

Day 1

1. Wash beads containing your protein complex three times with 1 ml of IPP150 (with no NP-40) as before.
2. Wash beads with 1 ml of 50 mM ammonium bicarbonate.
3. Resuspend beads in 100 ul of 50 mM ammonium bicarbonate. Add 2 ul of trypsin stock solution (0.1 ug/ul). Incubate with shaking at 37°C overnight.

Day 2

4. Pellet the beads with the magnet and collect the supernatant (*Note: This contains the peptides for MS*). Acidify the supernatant with 5 ul of 5% formic acid.
5. Add 100 ul of 1M AmBic to the beads. Resuspend by pipetting. Collect supernatant using the magnet as before and acidify with 5 ul of 5% formic acid.
6. Pool the supernatants collected in steps 7 and 8.
7. Dry the peptides in a SpeedVac at 45 °C (this will take approximately 1.5 hours).

Protocol 3: Denaturing Polyacrylamide gel electrophoresis (SDS-PAGE)

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical

Reagents

- 1x MOPS running buffer (made from 20x stock, Life Technologies)
- NuPAGE Novex Bis-Tris 4-12% gels, 1 mm (Life Technologies)
- 4x LDS sample loading buffer (Life Technologies)
- 1M DTT

Procedure

1. Prepare the IP sample: placing the tube with the aliquot of “pre-digest complex” in the magnet for 1 min and remove the supernatant. Place tube back in ice. Add ddH2O, sample buffer and DTT (as per table below - Beads).
2. Prepare input and supernatant samples:

Sample	**ddH2O**	**Sample**	**4x LDS sample buffer**	**1M DTT**
Input (Lysate)	10 ul	5 ul	5 ul	1 ul
Supernatant	10 ul	5 ul	5 ul	1 ul
Beads	15 ul	Beads	5 ul	1 ul

3. Vortex briefly. Incubate all samples at 70°C for 10 min. Leave at room temperature.
4. Wash two pre-cast gels with water, remove comb and white sticker. Flush the wells with 1x running buffer. Assemble the electrophoresis cell and fill the chambers with 1x running buffer (200 ml in the inner chamber, 600 ml in the outer chamber).
5. Briefly spin the samples at high speed.
6. Load samples in the two gels. (*Note: Place IP sample tube in magnet before loading*).

   Gel for silver staining: 5 ul of lysate and supernatant, 15 ul of IP, 3 ul of marker (S).

   Gel for Western Blotting: 15 ul of lysate and supernatant, 5 ul of IP, 3 ul of marker (WB).
7. Run the gels at 200V until the dye reaches the bottom (set the time of run to 50 min).
8. Proceed to silver staining and gel transfer.

Protocol 4: Protein silver staining

Hazards and Personal Protective Equipment

Lab coats and nitrile gloves must be worn at all times during the practical
Silver Quest Stainer –contains 30% silver nitrate – Very toxic/corrosive/very flammable
Silver Quest Sensitizer contains 30% NN Dimethylformamide and 15% Sodium Hydroxide – Carcinogen/Mutagen/Irritant - not to be handled by new or expectant mothers.
Silver Quest Developer Enhancer contains 30-60% formaldehyde and 10-30% methyl alcohol – Toxic/Corrosive/Teratogen/Irritant/Possible carcinogen - Irritant not to be handled by new or expectant mothers
Silver Quest Developer – contains 30% potassium Carbonate – Irritant
Ethanol – Flammable
Silver Quest Stopper – contains 10-30% Tris (hydroxymethyl) aminomethane and EDTA 10-30% - Irritant
Fixing solution – contains 40% ethanol/10% acetic acid – corrosive/Highly flammable
Acetic acid – corrosive – Use in fume hood

Reagents

- Gel fixing solution (40% ethanol, 10% acetic acid)
- 30% Ethanol
- SilverQuest silver staining kit (Invitrogen)

Procedure

All incubations should be performed on a rotary shaker gently rotating. Always use 100 ml of each solution per mini-gel. Please use the checklist at the end of protocol to keep track of the steps performed.

9. Remove the gel from the plastic cassette and place it in a clean staining tray containing 100 ml of MilliQ water. Rinse briefly and discard the water.
10. Incubate the gel in 100 ml of fixing solution with gentle rotation for 20 min.
11. Decant fixing solution. Incubate the gel in 100 ml of 30% ethanol for 10 min with shaking.
12. Decant liquid. Add 100 ml of sensitising solution (10 ml of Sensitiser + 30 ml of EtOH + 60 ml of ddH2O). Incubate gel in sensitising solution for 10 min.
13. Decant liquid. Incubate the gel in 100 ml of 30% ethanol for 10 min.
14. Decant liquid. Incubate the gel in 100 ml of ultrapure water for 10 min.
15. Decant liquid. Incubate the gel in 100 ml of staining solution (1 ml of Stainer + 99 ml of ddH2O) for 15 min.
16. Decant liquid. Wash the gel with 100 ml of ultrapure water for 20-60 seconds (equivalent to a brief rinse).

17. Incubate the gel in 100 ml of developing solution (10 ml of Developer + 90 ml of ddH2O + 1 drop of Developer enhancer) for 4-8 min until bands start to appear and the desired band intensity is reached.
18. Once the appropriate staining intensity is reached, immediately add 10 ml of Stopper directly to the gel still immersed in the developing solution. Incubate for 10 min.
19. Decant the solution and incubate the gel in 100 ml of ultrapure water for 10 min.
20. Scan the gel.

Checklist for silver staining

Step	Reagent	Protocol	Done
Fix	40% ethanol 10% acetic acid	20 min	
Wash	30% ethanol	10 min	
Sensitize	30% ethanol 10% Sensitizer	10 min	
First wash	30% ethanol	10 min	
Second wash	Water	10 min	
Stain	1% Stainer	15 min	
Wash	Water	20-60 sec	
Develop	10% Developer 1 drop Developer enhancer	4-8 min	
Stop	Stopper 10 ml	10 min	
Wash	Water	10 min	

Protocol 5: Western blot analysis

Hazards and Personal Protective Equipment

Methanol – toxic – Dispose of in dedicated container
ECL Prime – toxic, irritant
Lab coats and nitrile gloves must be worn at all times during the practical

Reagents

- 1x Transfer buffer (Made from 20x stock, Life Technologies; contains 10% methanol)
- Skimmed dried milk (Marvel)
- PBS-T (PBS containing 0.1% Tween-20)
- Anti-FLAG-HRP-linked antibody (Sigma)
- ECL Prime detection reagents (GE Healthcare)
- Nitrocellulose membranes and filter paper sandwich (BioRad)

Procedure

Day 1

1. Place the nitrocellulose membrane in 50 ml of transfer buffer.

2. Assemble a transfer sandwich as follows: filter paper / gel / membrane / filter paper. To do this open the plastic cassette and remove the wells by cutting with the knife. Wet a filter paper with 1x transfer buffer and place it on top of the gel. Invert the cassette to rest the paper side on your palm, and with the knife push the gel away from the plastic cassette by running the knife through the slot at the bottom of the cassette. Next place the paper/gel assembly on one of the plastic cassette sides. Cut the bottom thicker lip with the knife. Next place the nitrocellulose membrane on top of the gel (*roll a tube on the assembled sandwich to remove bubbles*). Place the second filter paper on top of the membrane and remove bubbles.

3. Wet three sponges with transfer buffer and place them on the anode electrode plate (the one that looks like a box). Place the transfer sandwich on top, as it was assembled. Place two more wet sponges on top. Place the cathode on top (*careful not to dislodge the sandwiches*), lift the whole assembly and press it together over a tray (*to collect excess transfer buffer that will drip*) and insert it into the tank. Hold it in place using the clamp. Fill the inner transfer chamber with 1x transfer buffer. Fill the outer chamber with water that will act as coolant.

4. Set the following transfer conditions (30 V, 200 mA, 20 W, 1 hour).

5. Add 30 ml of blocking solution (5% non-fat dried milk in PBS-T) to a 50 ml Falcon tube. Remove membrane from sandwich and cut upper left hand corner. Roll membrane on itself along the long axis with the proteins towards the inside, insert in the Falcon tube and unroll, making sure there are no big bubbles between membrane and tube. Incubate at 4°C overnight in a roller.

Day 2

6. Dilute primary antibody (anti-FLAG-HRP) 1:10,000 in 3 ml of blocking solution.

7. Discard blocking solution from the membrane and add the antibody solution. Incubate for 1 hour at room temperature in a roller.

8. Transfer the membrane to a plastic tray with 25 ml of PBS-T. Incubate with shaking for 10 min at room temperature. Discard the PBS-T and repeat the washing twice more. Leave the membrane in PBS-T until detection.

Detection

9. Mix 1 ml of ECL Prime reagent A with 1 ml of ECL Prime reagent B and pipette onto a plastic sheet. Drain excess PBS-T from membrane by holding membrane with forceps and touching the edge against a tissue, and place the membrane with the protein side down on the detection reagent. Incubate for 5 min.

10. Drain excess reagent from membrane and place side up between plastic sheets. Smooth out air bubbles and make sure the edges are dry.

11. Place blot in x-ray film cassette.

12. In the dark room, carefully place an autoradiography film on top of membrane and close the cassette. Expose blot for 1 min.

13. Remove film and insert in the developer. When the developed film comes out, assess results and repeat with different exposure time(s) if required.

10. Computational Analysis of Biological Pathways

Chapter 5

Computational Analysis of Biological Pathways

Anton J. Enright[1] & **Gary D. Bader**[2]

A major challenge in biology today is the integration of information from the various fields of molecular and cellular biology to accurately understand the workings of the cell. New fields of biology, such as functional genomics and proteomics, take advantage of the availability of completely sequenced genomes to enable large-scale mapping of cellular molecular networks. Large-scale cell mapping methods are being developed faster than ever before and are creating a tidal wave of new information about how the various parts of the cell fit together. Bioinformatics will play a vital role in overcoming this data integration challenge, as databases, visualization software, and analysis software must be built to enable data assimilation, as well as make the results accessible and useful for answering biological questions.

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom
[2]Terrence Donnelly Centre for Cellular and Biomolecular Research. University of Toronto

5.1 Introduction

A major challenge for biologists today is to gain understanding of the workings of the cell by integrating available information from the various fields of molecular and cellular biology into an accurate cellular model that can be used to generate hypotheses for testing. New subfields of biology, such as functional genomics and proteomics, take advantage of the availability of completely sequenced genomes to enable large-scale mapping of cellular molecular networks through new experimental methods. These methods are being developed faster than ever before and are creating a tidal wave of new information about cellular processes (Bader, Heilbut et al. 2003). Bioinformatics will play a vital role in overcoming this data integration and modeling challenge, as databases, visualization software, and analysis software must be built to enable data assimilation and make the results accessible and useful for answering biological questions. As experimental methods become more sensitive, more of the cell is being mapped. Decades ago, many metabolic pathways in organisms from bacteria to mammals were mapped because early biochemical methods could detect their abundant enzyme and small molecule components. These days, experimental methods, such as protein identification by mass spectrometry, are sensitive enough to detect molecules at only a few copies per cell. This allows biologists to further map cellular metabolism and importantly to see many regulatory, or cell signaling, networks for the first time. Cell signaling network mapping is particularly interesting because of its role in diseases with a high prevalence in the population, such as cancer, diabetes and Alzheimers disease. Molecular interaction networks generated by some of these experimental cell mapping methods provide a convenient and practical scaffold for integrating other types of data, since they represent the intricate connections that are a hallmark of metabolic processes. A molecules interacting and reacting partners define its function in a biological system. Thus, gene expression data derived from a cDNA microarray, when integrated with a molecular interaction network, place the data into a coherent functional context. The opportunity for gaining further understanding of biological systems from large amounts of new and existing data through intelligent data integration has served to drive biomolecular interaction and pathway analysis in bioinformatics.

Interestingly, the availability of new types of biological data has enabled new computational methods in pathway informatics which, in turn, enables new discoveries at the bench. Computational pathway and network analysis should be considered early on in the conception of cell mapping experiments to quicken the discovery cycle. For example, the computational reconstruction of metabolic pathways from a recently sequenced bacteria allowed investigators to discover that the organism could not synthesize specific amino acids and required an environmental source to live. This allowed the bacteria, which causes a gastrointestinal disease, to be cultured in a lab and studied, something that was previously not possible (Renesto, Crapoulet et al. 2003). The pathway informatics field is relatively new and is changing rapidly, in part due to the large amounts of cell map data now available. Initial work in this area involved the computational representation of metabolic pathways; more recently, it has focused on designing cellular signaling pathway databases. The available databases and tools in this area are evolving quickly, much like some other quickly developing areas of bioinformatics. For instance, only recently have the main protein-protein interaction database efforts developed common data exchange formats that make the data more accessible; pathway data exchange formats are only starting this process. Because of this rapid pace of change, this chapter covers a number of existing tools, but also focuses on fundamental theory that should be applicable to many new databases and tools as they become available. Specifically, a number of useful pathway and network resources are covered, focusing on freely available and/or very accessible tools. A selection of such tools are briefly described in Table 5.1. Pathway and molecular interaction databases and emerging common data exchange formats that multiple databases and tools are starting to support are covered first. Algorithms for predicting protein-protein interactions and metabolic pathway reconstruction are then described, along with a guide to various Web resources that implement these algorithms in a user-friendly manner. Next, a description of pathway and network visualization tools and important underlying concepts and algorithms for these tools are described. Finally, a special focus section is included, which examines some of the emerging analyses that are possible when integrating gene expression data with pathway and network information. This is meant to illustrate the interesting biological questions that can be answered and new hypotheses that can be generated by integrating existing data in the network context. Not all resources can be covered in depth, so the largest databases and most commonly used data analysis tools are featured here; pointers to online descriptions and lists of other pathway and network related databases and software are provided at the end of the chapter.

Before delving into databases, it is interesting to know where the data comes from. What would an

ideal biological experiment be able to tell us? The answer is no less than everything: what molecules are in the cell at what time and at what place, how many molecules are there, what molecules they interact with, and the specifics of their interaction dynamics. Ideally, one would want this information not only over the course of the cell cycle, but also in all important environmental conditions and under all known disease states. In relation to this huge amount of information, current experimental methods, while incredibly useful and growing orders of magnitude better every decade, only scratch the surface. A wide range of biochemical, molecular biological and genetic experiments have been invented to help elucidate cellular systems and determine which cellular parts are involved and how they fit together. Enzymatic reactions have been studied for centuries, initially examining processes such as fermentation. The basic principle of experimentally mapping metabolism, composed mainly of protein enzymes, is to identify an enzymatic process (e.g., the conversion of glucose to ethanol in yeast, progressively purifying cellular extracts to find the enzymes involved). Validation involves taking the purified enzyme to see if it can convert the given substrate to a product. Thus, the process requires protein separation and purification technology, as well as molecule identification methods to identify the enzyme, any cofactors, substrates, and products involved in the reaction. Major advances in this area have been made using various forms of chromatography, gel-based separation techniques, nuclear magnetic resonance (NMR), and mass spectrometry. Chromatography and gel separation work on the basic principle that a molecular mixture can be decomposed based on component physiochemical properties, such as size or charge. NMR and mass spectrometry can be used to directly identify small molecules and proteins based on atomic distance measurements and mass, respectively. Enzymologists further characterize the reaction rates of enzymes (kinetics) and the detailed enzymatic mechanism involved in catalysis (Voet and Voet 2004). Signaling pathways, which involve many more direct protein-protein relationships, such as phosphorylation of one protein by another protein (kinase), can be mapped using protein-protein interaction detection methods. Many other types of molecules are involved in signaling, so protein interactions only tell part of the story. Many techniques for determining protein-protein interactions have been developed over the past few decades. One popular class of experiments depends on co-purification, using the methods described above. It is reasoned that proteins that strongly interact will purify as a complex that can be further degraded using harsher purification conditions to finally separate and identify the complex components. Importantly, this means that the definition of a protein complex depends on the purification conditions used, which measure a continuum of protein-protein interaction strengths. An example of a modern biochemical co-purification uses affinity chromatography to purify a protein complex from a cellular extract then identifying the resulting complex components using mass spectrometry. Yeast two-hybrid methods are often used to determine if two proteins are interacting. An activation and DNA binding domain of a transcription factor are attached to each protein, respectively. If the two proteins of interest interact, the activation and DNA binding domains will also interact, forming a functional transcription factor that will express an engineered reporter gene. Presence of the reporter gene thus indicates binding. Another often-used method is molecular cross-linking, an experimental method where a linear molecule of defined length having two reactive ends is added to a mixture containing a potential complex, in order to cross-link proteins that are close together; the distance over which an interaction can be detected is determined by the length of the cross-linker being used. Subsequent purification and definition of the complex is easier, since the protein complex is covalently tied together instead of just being electrostatically bound. Many other protein-protein interaction determining experimental methods exist (Voet and Voet 2004), but almost all current experiments suffer from observer effect, whereby the conditions of the experiment disturb the natural biological process. Since each experiment has its own strengths and weaknesses in this regard, it is only after multiple types of experiments have been performed that one can really be sure of the result.

5.2 Pathway and molecular interaction databases

Given the breadth and depth of pathway and molecular interaction network information already available in the literature, as well as that being generated from large-scale experiments, it is no surprise that a number of databases have been built to try to represent and store this information. In fact, there are over 100 pathway and molecular interaction related database resources that are available via the Internet. These range widely in form and content and include full-featured pathway databases, ones that focus on protein-protein interactions, and organism- or disease-specific pathway databases. Some databases contain molecular interaction information, although they

Database	Scope	Data Model	Special Features	License
BIND	Many species	Binary molecular and genetic interactions, complexes and pathways	BLAST search, domain view, graphical view	Public domain
DIP	Many species	Binary protein-protein interactions	BLAST search, graphical view	Free for academics
GRID	Bidding yeast, fruit fly, worm	Binary protein-protein and genetic interactions	Osprey network analysis	Free web-based access
HPRD	Human	Binary protein-protein interactions	Curated protein records	Free for academics
IntAct	Many species	Sets of interacting proteins	Graphical view	Freely available
MINT	Many species	Sets of interacting proteins	Java viewer	Freely available
BioCyc	Many species	Metabolic pathway-based ontology	High quality pathways and full featured software	Free for academics
KEGG PATHWAY, LIGAND	Many species	Metabolic pathways	Pathway diagrams and compound database with 2D structure	Freely availably
Reactome	Many species	Biological pathways	Graphical view, Pathway diagrams, Cytoscape viewing	Freely available

Table 5.1: An overview of selected pathway databases

are not primarily focused on interactions. For instance, the DDBJ/EMBL/GenBank feature table, originally designed for annotating nucleic acid sequences , contains some information about binding sites of gene expression and translation control elements, such as ribosome binding sites (specified using the *RBS* key). This information is not widely used, however. This section will focus on the largest published and/or freely available database resources that are considered to be the most generally useful. It is likely that certain specialist pathway and molecular interaction databases will be better-suited to answering certain types of queries; URLs to these are provided at the end of this chapter. The cell is a large, complex and very dynamic connected network of molecules. Because of its complexity, biologists think of the cell as having substructures and subsystems, such as organelles, pathways, and complexes, so as to aid in understanding the overall picture. While organelles and complexes are structures that can often be seen under a microscope, pathways are (obviously) not so it is important to realize that pathways are human constructs and are just parts of a larger, fully-connected molecular interaction network. A working definition of a pathway is a series of molecular interactions and reactions, often forming a network. Sometimes human pathway organization is based on recognized biochemical or information processing phenomena. For instance, a series of metabolic reactions could start with the intake of a metabolite from the environment, which is converted quickly and irreversibly to something else, such as the glycolysis pathway breaking down glucose to generate energy (ATP). Also, signal propagation in a series of signaling pathway steps could be shown to follow a specific pathway, such as when a ligand binds to a cell surface receptor which results in signal propagation through a kinase cascade to the nucleus where transcription is activated. Often in these cases, the start and end points of a pathway is defined by observation of some readily detectable phenotype after stimulation or perturbation, such as observing gene expression after stimulating the cell with a peptide growth hormone. In other instances, pathway organization is based on the order that components of a pathway are discovered. In this case, it is possible for existing pathways to be merged if enough 'cross talk' between the existing pathways is found. Not only have biologists organized the cell into pathways and modules, but they have classified the pathways themselves into different types and each of the main types, generally has a different computational representation in the various existing pathway databases. The main types of biochemical/biophysical based pathway representations model metabolic, signal transduction (also called cell signaling), and gene regulation pathways. Metabolic pathways are generally defined by a series of chemical actions and the chemical results of those actions, for the purpose of changing one molecular species into another. Glycolysis is a typical example of a metabolic pathway that converts glucose to energy. Signal transduction pathways are usually defined by binding events, sometimes involving chemical actions (e.g., phosphorylation events), for the purpose of communicating information from one place in the cell to another. Often the binding events are protein-protein interactions. The MAP kinase cascade is a common example of a signal transduction pathway that conveys information from an externally activated cell surface receptor to the nucleus in order to effect change in gene expression in response to the external signal. Gene regulation networks involve transcription factors activating or repressing expression of genes, a simple case being activation of expression of a set of genes. If one or more of the set of genes being activated is itself a transcription factor that can act independently, then a more complex network results. Another type of pathway that is not generally biochemically/biophysically defined but is classically mapped is a genetic pathway, which is a series of genetic interactions. Genetic interactions are not physical binding events, but rather are defined by a change in phenotype caused by a change in genotype. A simple type of genetic interaction is called synthetic lethal, where two genes that are genetically altered (for instance knocked out) do not cause a phenotype change on their own, but when altered together, cause death of the organism. The implication of such a relationship, in a simple case, is that the genes are part of linear parallel pathways that both can carry out the same vital biological process where one pathway can compensate for the other if one is damaged. An example of this might be two parallel metabolic pathways that both build an essential amino acid from precursor molecules (Figure 5.1). Genetic interactions and pathways have always been important in biology, since most early knowledge of biological processes was defined genetically before the advent of molecular biology techniques starting in the 1950s. The completion of genomes of various model organisms has led to an increase in the use of genetics to examine certain biological processes and gene function more quickly or easily than some current biochemical techniques allow.

The distinction between these types of pathways here is related to human representation, so is purely virtual and does not relate to any intrinsic structure in the cell or organism. In fact, many pathways, or regions of the molecular interaction network can be clas-

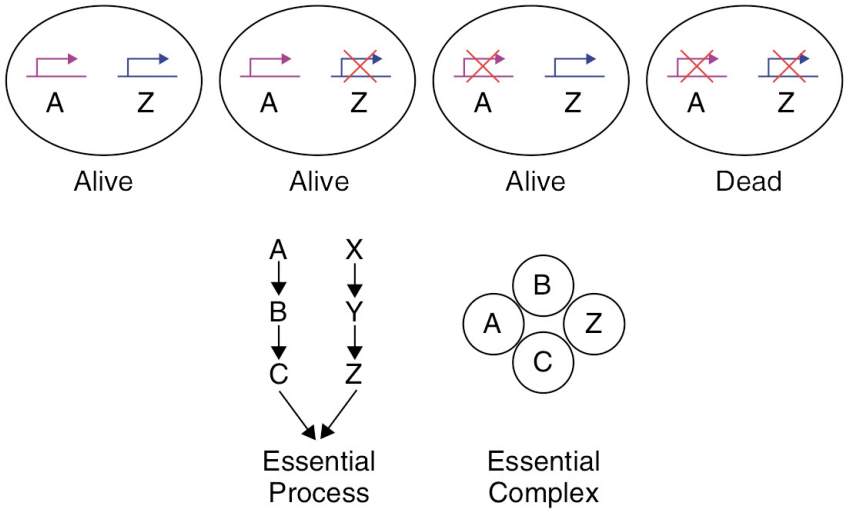


Figure 5.1: Biochemical meaning of a synthetic lethal genetic interaction. A genetic interaction occurs when two or more genes affect a phenotype when mutated together. In this case, a synthetic lethal interaction is shown. When genes A and/or Z are expressed, the cell is alive, but when both A and Z are not expressed (shown here) or mutated, the cell dies. In the simple biochemical case, this may be the consequence of genes A and Z being in parallel pathways that impinge on a common essential metabolite or process or perhaps genes A and Z are both part of a molecular complex that can not assemble without both gene products. For the parallel pathways shown here, synthetic lethal interactions between all combinations of genes (A,B,C) and genes (X,Y,Z) would be expected, not just the A-Z interaction shown here. Real synthetic lethal interactions may have more complex relationships to the underlying biochemical network and many other types of genetic interactions exist.

sified into multiple categories and represented in different ways, depending on the context. For instance, a metabolic pathway database might call a specific pathway metabolic and represent it one way and a signaling database might classify it as a signaling pathway and represent it in a different way. This can lead to difficulties when trying to integrate these pathways into the same system. Thus, it is important to understand how a database represents the information it stores to be able to query it and to understand its advantages and limitations. A representation system (also called a data model or abstraction) is an invention that can be used to describe and organize a set of observations. Often, many different representation schemes are possible for the same type of information, and two different people given the task to invent an abstraction independently can create different systems, especially for complex and partially undefined biological information. A single representation scheme must be agreed upon before it can be useful for data communication, although such a decision involves considering a number of tradeoffs. An ideal representation system compactly and efficiently describes exactly the information useful to the users of the system. Very compact and efficient representation often results when communicating amongst people with the same extensive common knowledge, such as scientists in a specific subfield who all understand the jargon and concepts of their field. Compactness can be achieved because common knowledge can be taken for granted, thus does not have to be explicitly represented each time information is communicated. This compactness can enormously reduce communication time and effort, making it very useful. Unfortunately, using a compact representation to communicate between people who do not share the same common knowledge does not work as well. These people will have trouble understanding each other unless common knowledge is explicitly represented. This frequently happens when people in different subfields in science communicate. Similarly, computer programs that are not programmed with extensive rules defining common knowledge can generally not properly *understand* very compact representation and require explicit coding of extra information and logic to perform actions such as querying the compact data. If enough information about data relationships is encoded, whether in software or in the data exchange format, a computer may be able to draw new conclusions using logical inference. This is a major basis for the field of artificial intelligence (AI). For example, if the computer program

knows that a kinase catalyzes the addition of a phosphate group, it can easily infer that protein X is a kinase if it adds a phosphate group to another protein in the molecular interaction network. Another related trade-off is between simplicity and complexity of representation. The advantage of having a simple model that captures the basic properties of the data is that it is easily created, understood and used, but it can not represent all detail that may be known about a system. The complex model may be able to represent everything that is known, but might be too unwieldy to be useful in some cases. Many aspects of biological systems that may be useful to represent can significantly add to the complexity of a representation scheme. Examples are level of detail, context and scope and what is sometimes called provenance in the database community. Each of these are dealt with individually below. Adding levels of detail in data modeling is useful for representing data at varying levels of knowledge or understanding, down to the limit of detail that is of interest. Certain types of biological information are understood in much more detail than others. When detail is known, a detailed data model should be able to represent it. With a model that includes multiple levels of detail, there is a choice between representing the same information at low, intermediate, or high detail. Depending on the use-case, more or less detail may be required. (Use-case, a technical term from the field of software engineering, describes how a user interacting with a software system from the point of view of different types of users). For instance, if the mechanism of a protein phosphorylation event by a tyrosine kinase is understood down to the movement of electrons in the chemical reaction, it might be useful to capture all of the known information for someone studying electron dynamics. Alternatively, someone studying the global properties of protein interaction networks might only be interested in the fact that one protein interacts with another and would find information on electron dynamics distracting. Adding to the complexity of biological knowledge representation, levels of detail in the cell map can be considered across large ranges (scales) of time and space where each level of the organizational hierarchy may require its own abstraction system. As an example across spatial scales, the molecular parts of the cell have widely established representation systems, such as the 20-letter amino acid code for protein sequence and the atoms, bonds and connectivity of atoms in a three-dimensional protein structure. Also, atomic bonds are measured in picometers (1×10^{-18} m). Neither of these abstractions work well in describing larger substructures of the cell. For instance, an organelle like the nucleus would simply be too difficult to examine if it was completely described by three-dimensional structures. An average cell is measured in micrometers (1×10^{-6} m), and neuron length in large organisms can extend well into the meter range. Similarly, across temporal scales, ultra-fast electron flow in a biochemical reaction, measured at attoseconds (1×10^{-18} s), can be described when it is known, but any useful abstraction to describe electron flow would not be useful for describing events on the time scales of the cell cycle, measured at minutes or hours. Context and scope of biological network information can also add complexity to a representation scheme because molecular interactions and reactions depend on the presence of the participating molecules at permissive conditions, such as being in the same place at the same time at a normal temperature. Thus, it may not suffice to record that a reaction occurs among participating molecules because that reaction may or may not occur with the same participants in different cells, developmental stages or in different organisms. Similarly, if the experimental methods used and observations that were made to discover pathway knowledge are of interest, it may be useful to model all different types of experiments that were performed as well as who did them and when. It may also be useful to track information was inferred from a similar pathway to the one of interest. This knowledge tracking information is sometimes referred to as provenance, which simply means proof of origin and authenticity, and can be used to track error propagation in a database. Provenance can also add significant complexity to any model. These fundamental trade-offs and issues in knowledge representation are important, since the multiple representation schemes present for pathways makes it difficult to integrate information from multiple different pathway sources into a single system. A universal language for pathways has not yet been developed. Such a language might not actually be efficient to use for day to day work because it would be too verbose, although it would be useful as a data model for a universal pathway repository and could be mapped to a user-preferred representation scheme for practical purposes. Development of a universal pathway language is difficult, especially when many aspects of biology are poorly understood. Paradigm shifts that significantly change our thinking about an aspect of biology could occur that call for fundamentally different ways of representing the data. In summary, the preferred representation of a pathway depends heavily on what the information will be used for (the use-case). Different representation schemes and the definition of different types of pathways have evolved in pathway informatics because different sets of common knowledge and different use-cases exist within different

communities (different subfields of biology). Apart from understanding the representation used by a database and why it was chosen, a few things should be kept in mind when using a pathway database. These include scope, data quality, freshness of data, data quantity, availability and technical architecture, each of which will be dealt with here. The scope of a database is important to know when searching for information. Pathway and interaction databases are springing up to collect information from the literature, but this is a difficult task because of the data complexity, thus databases often focus on a specific area. Knowledge of database scope can prevent wasted time searching for information of interest in the wrong database or otherwise misusing the database. For example, the GRID database contains information about protein-protein interactions and genetic interactions, two related data types with very different properties. It is possible for a user to search for protein interaction, only find genetic interactions and misinterpret them as protein interactions if they are not aware of the scope of GRID. The data quality of a database depends heavily on level of curation and validation and can be difficult to independently assess, unless one is already an expert. Pathway databases can range from those that store computed information with little or no validation to those resources with large teams of experienced biologists and complex validation systems. While expert curated databases are the gold standard, collections of lower quality information are still useful, but generally require that the user be an expert and have the time to sort through it. For instance, databases of protein-protein interactions created automatically by literature extraction techniques (text-mining) might only be 70% accurate, but might have some correct information that no other database contains. Importantly, many protein-protein interaction databases contain a large amount of information derived from relatively few large-scale or high-throughput experimental methods such as comprehensive yeast two-hybrid or large-scale biochemical purifications. It is known that these techniques can generate a large number of false positive interactions as well as miss correct information (false negatives); thus, in search results, it is important to determine which records of interest come from large-scale versus which ones are curated. Data freshness is also important, and databases that are well-maintained and store current information can indicate higher data quality. Some pathway and network databases, along with many other biological databases, are not being actively maintained, even though the Web site and database might still be available. Old data sets can be very useful, but it is important to be aware of the data set age. Users should make a point of looking for dates on the homepage of the database as well as in the records, or creation times of datasets available on FTP sites, if available, to find out how recent the data are. Another measure of the usefulness of a database is data quantity. Certain databases have been built and published without much data. Possible reasons for this include data model research or showcasing new technology. If a user is interested in the data contained in a database more than any other aspect, the statistics page, if available, should be examined before using the resource. Additionally, the user should never assume completeness of a data set, since most of the data in pathway and interaction databases is derived from the literature and it is difficult to guarantee that all known data about a subject is collected. Another issue to be aware of before using a database is availability. This comes in two forms: download restrictions and intellectual property (IP) restrictions. Obviously a database can not used if it can not be accessed. Download or electronic availability is simply whether or not the data can be accessed via some electronic medium (e.g., via the Web or by FTP). It is possible that a database only exists in paper form, in which case it is not very accessible (or useful). Both academic and corporate databases can be advertised on a Web site, but not actually be accessible either because they are sold or possibly under construction. IP restrictions are copyright statements and licensing terms that must be adhered to when using a given database. For pathway databases, this is generally freely available to all, free for academics only, or restricted use for all users. Also, some licensing terms of corporate databases reach through to cover inventions made using the data in the database (Greenbaum and Gerstein 2003). Finally, the technical architecture of a biological network database should be considered if the user is planning to work with the data in a more technical manner, such as writing a program to extract some information of interest. As with other biological databases, pathway databases come in a wide range of formats from simple text files to custom object-oriented databases. For technical work, the ease of access to the data from a program defines another type of data availability. Some databases simplify programmatic access by providing application programming interfaces (APIs) in multiple programming languages. Discovering the technical aspects of a pathway resource is generally more difficult than for simple user query access because fewer people are interested in the underlying software architecture than the data contained therein. Databases that are published often describe the architecture. If this is not the case, the database creators must be contacted

directly. In summary, to get the most value out of pathway and network databases, it is important to know the data, scope and representation scheme. The following section gives an overview of the most widely used interaction and pathway databases in alphabetical order within categories.

5.2.1 Primarily Molecular Interaction Databases

BIND

The Biomolecular Interaction Network Database (Bader, Betel et al. 2003) is currently, the largest collection of freely available information about pairwise molecular interactions and complexes. A small number of pathways are also available through BIND. At the time of this writing, BIND contains mainly yeast (*Saccharomyces cerevisiae*) protein-protein, protein-DNA, and genetic interactions, as well as protein complex data. Also contained are fruit fly (Drosophila melanogaster) and worm (*Caenorhabditis elegans*) protein-protein interactions from large-scale experiments, and a fair amount of curated information from these and over 800 other species as well. BIND also has the largest staff of curators of any public interaction database and is currently adding records to the database on a regular basis. A subproject of BIND is MMDBBIND (Salama, Donaldson et al. 2002), which stores all of the molecular interactions automatically extracted from PDB in BIND format. The biopolymer interaction subset of this was recently made available via the BIND Web interface. There are three main types of data objects in the BIND data model: interaction, molecular complex, and pathway. Interactions occur between two objects, which can be RNA, DNA, protein, small molecule, molecular complex, photon, and gene (for genetic interactions). Each object contains a description of its origin, whether organismal or chemical and references a more complete description of the object in a primary database for that object. For instance a protein might be further described in databases such as SwissProt or GenPept. Importantly, each interaction record is supported by at least one publication. Publications in BIND can describe the publication opinion, whether supporting or disputing, with respect to the information they are attached to. An interaction record may contain further description (all are optional):

1. A short description of the record
2. The cellular location of each object and of where the interaction takes place. Both a start and an end location can be stored.
3. The experimental conditions under which the interaction was observed to occur. This is only a description of the experiments that were done to show the interaction. Experimental data, such as gel images, are not stored in BIND.
4. A conserved sequence comment containing any potential conserved sequence that is known and that is functionally relevant (for instance, a conserved binding site). While potentially useful, this field is currently rarely used in BIND.
5. A list of binding sites on each object in the interaction and a list of pairings between these sites. For instance, an SH3 domain can be defined on one protein partner and be described as binding a proline rich region defined on the other protein partner.
6. A list of chemical actions and chemical states that describe any chemical reactions that occur while the objects are interacting. For instance, protein A may phosphorylate protein B to change the state of B from inactive to active.
7. An intramolecular interaction flag to store whether the interaction is occurring within a single molecule. For instance, calmodulin is found in both an extended calcium bound mode and a collapsed non-calcium bound mode where the N and C termini bind to each other.

BIND defines a molecular complex as two or more molecules that interact to form a stable complex and function as a unit. An example is the ribosome. A complex is defined as a collection of interactions that are already part of BIND along with some other information, such as assembly order, stoichiometry and complex topology. Similar to complexes, pathways in BIND are a collection of at least one interaction, but whose molecules form an ordered network and are generally free from each other. Examples of pathways are glycolysis and a MAP kinase cascade. BIND accepts user submitted records through either a Web-based form interface or through an import service, if the number of submitted interactions is large. The main searching interface is a full text search of all text fields in any BIND record, but there is also a BLAST vs. BIND tool that allows querying BIND for similar protein sequences to one of interest. The main BIND browse interface is a list of the records in the database, although there is also an interaction viewer Java applet that allows a user to traverse the molecular interaction network starting from a record of interest. A typical BIND record is shown in Figure 5.2. BIND is available for browsing and

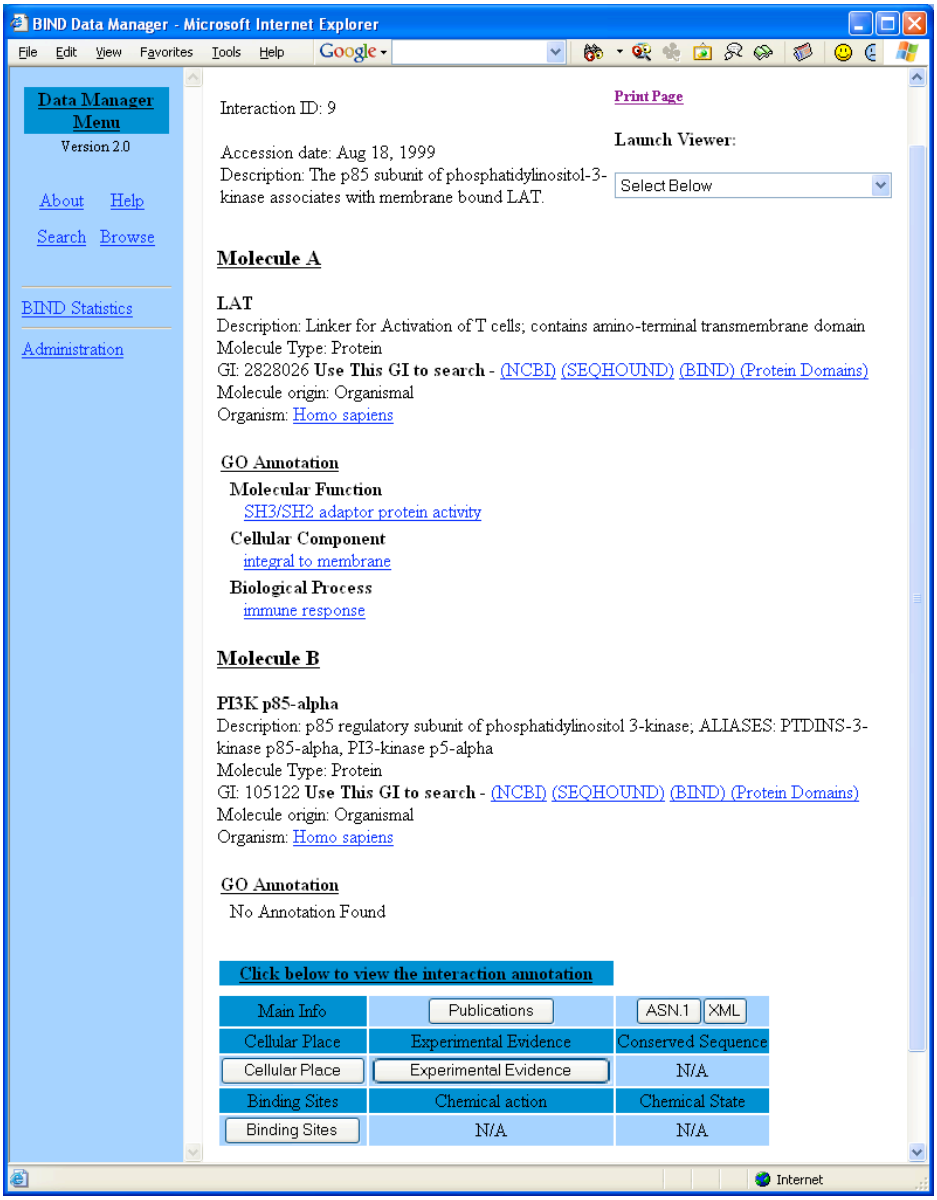


Figure 5.2: A typical BIND molecular interaction record is shown. This is a protein-protein interaction that contains extra annotation information about the cellular place, experimental evidence for and binding sites involved in the interaction. More information about this protein interaction can be found by clicking the various links. For instance, the *Launch Viewer* section at the top right of the figure can be used to launch a Java applet that allows visual navigation of the interaction network starting at the molecules in this record. The links within the molecule section (e.g., Molecule A) can be used to access the protein sequence (NCBI and SeqHound), what other BIND records contain this molecule (BIND), and what domains are on this protein (Protein Domains).

querying via the Web and can be freely downloaded without restrictions by FTP in the BIND format (Bader, Donaldson et al. 2001).

DIP

DIP, or the Database of Interacting Proteins(Xenarios, Salwinski et al. 2002), contains only experimentally derived protein-protein interaction data. It is one of the largest collections of publicly accessible protein-protein interaction information, containing data from over 100 organisms. The data is collected from large-scale protein interaction mapping experiments, as well as from expert curators. The DIP data model consists of binary protein interactions along with information on each protein, the method used to determine the interaction, and a set of publication references to support the record. DIP allows queries by proteins (called nodes), BLAST, protein sequence motifs, and by journal article. An analysis of the confidence level of interactions in the DIP database has been published (Deane, Salwinski et al. 2002), and two types of confidence of interaction scores are available:

- PVM, or Paralogous Verification Method, assigns a higher reliability score to an interaction whose paralogs are also seen to interact in DIP.
- EPR, or Expression Profile Reliability, index deems a set of protein-protein interactions more reliable if it has a similar expression profile as a high-quality subset of DIP.

These scores can be calculated only for budding yeast (*Saccharomyces cerevisiae*) protein interactions because they depend on the existence of a large amount of external information currently only available for budding yeast. A few satellite services are available as extensions to DIP. LiveDIP extends the DIP data model to describe protein-protein interaction processes using states and state transitions. Users can search for states or interactions as well as find paths through LiveDIP from one protein of interest to another. DLRP, or the Database of Ligand-Receptor Partners, contains a small set of ligand-receptor interactions for download, but does not provide search services. A BLAST query interface for DIP is also available. DIP is available via the Web for browsing, querying and downloading by academic users. Commercial users must acquire a license for use. A typical DIP record is shown in Figure 5.3.

GRID

GRID, or the General Repository for Interaction Datasets (Breitkreutz, Stark et al. 2003), contains protein-protein and genetic interactions currently for budding yeast (*Saccharomyces cerevisiae*) as YeastGRID, fruit fly (Drosophila melanogaster) as FlyGRID and *Caenorhabditis elegans* as WormGRID. GRID is actively being expanded and other species may be available in the near future. GRID provides a simple summary of each interaction along with gene names, Gene Ontology (GO) annotation and the experimental system that was used to determine the interaction. A network visualization tool called Osprey is also available to visualize, browse and analyze the interaction networks the various GRIDs.

HPRD

HPRD, or the Human Protein Reference Database(Peri, Navarro et al. 2003), is a recently released database of human proteins, but also contains a significant amount of protein-protein interactions. Information about the domain and region of interaction, if available, is present as well as the type of experiment done to detect the interaction. Expression, domain architecture and post-translational modifications are also curated for each protein. A number of curated pathways created from the interaction data are available as images. A typical HPRD record is shown in Figure 5.4. HPRD data can be browsed and searched by a number of common database fields as well as by BLAST over the Web. Data are freely available to academics in the PSI-MI format and commercial entities require a license.

IntAct

IntAct is a relatively new database of freely available protein interactions maintained by the European Bioinformatics Institute (EBI). An initial implementation available from mid-2002 focuses on protein-protein interactions collected from large-scale published studies and some literature curation and allows searching by protein name and browsing using a graphical network interface. One difference in the IntAct data model compared to many other protein-protein interaction databases is that interactions are not necessarily binary, but rather are sets. The advantage of using sets to store interactions is that they can represent certain types of protein complex data where information is not known about the exact physical interactions in the complex, but only that the set of proteins co-purifies. Representing information in this way has become more

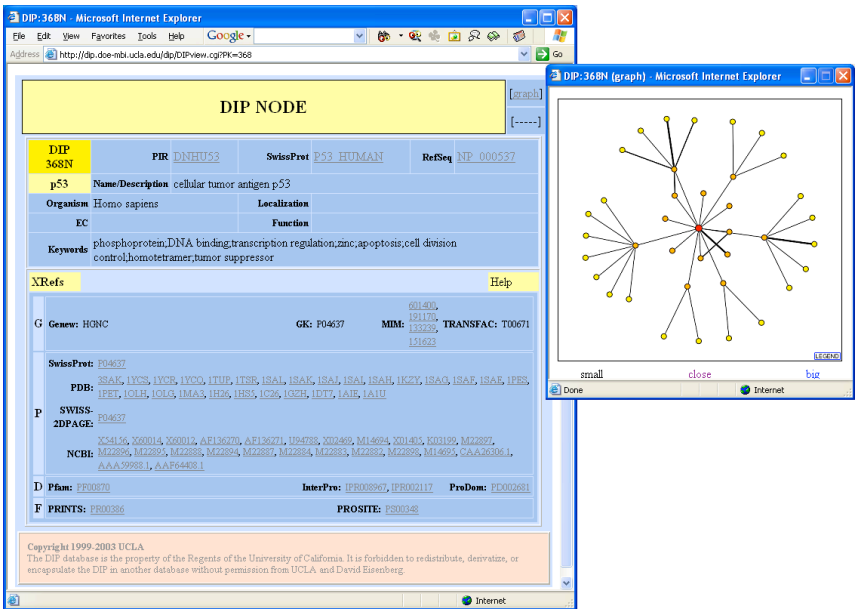


Figure 5.3: A DIP human protein-protein interaction is shown with associated protein interaction network static visualization diagram. The protein shown in p53, involved in a number of cancers. This is the protein centric viewing mode of DIP, which contains links to find more information about the protein in standard protein databases. The DIP network browser shows a static, clickable image of the neighborhood of the given protein. This view can be accessed using the *graph* link at the top right hand side of the view. An edge centric view is also available in DIP (not shown), which additionally shows the evidence for the interaction.

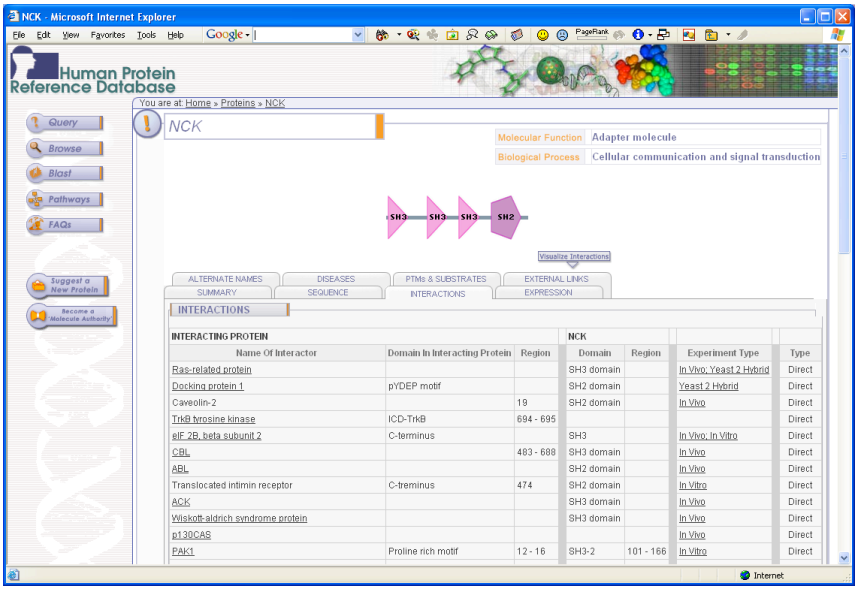


Figure 5.4: An HPRD protein information record showing a number of protein-protein interactions for NCK, the selected protein. NCK is involved in the regulation of actin cytoskeleton rearrangement. HPRD has many other curated fields and interactions are just a subset of the information provided. These other fields are accessed using the tabs shown under the domain view of the protein (graphic at the top of the record). The left part of the screen contains links to the main operations on the database. The query button leads to a form that allows searching by a number of database fields, such as protein name, post-translational modifications and diseases. Other buttons allow browsing, querying using BLAST and show pathway diagrams using HPRD information.

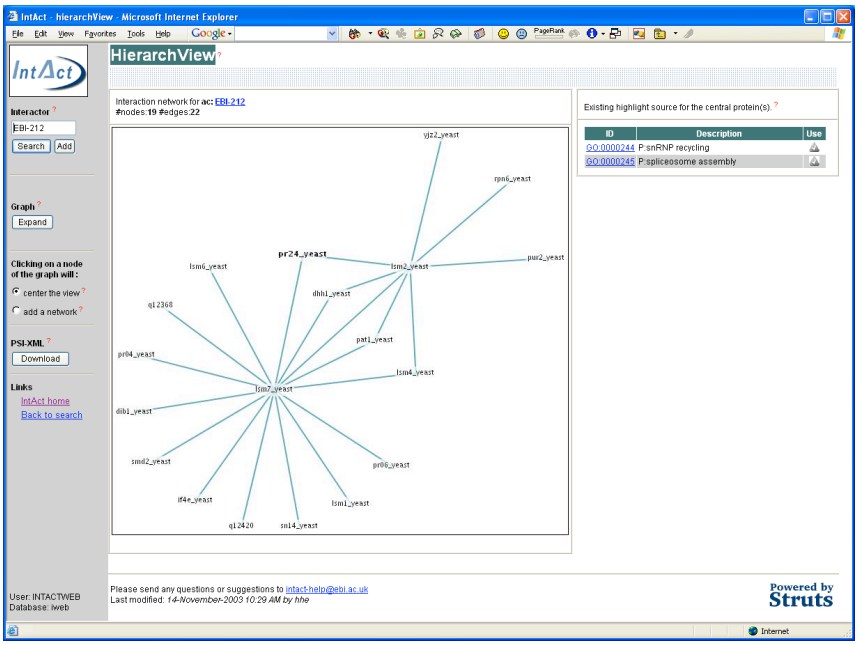


Figure 5.5: The graphical representation of a set of IntAct database yeast protein-protein interaction records. Lines represent protein-protein interactions. After IntAct can be queried by protein name or database identifier. The interactions that result from a query are displayed in a table and selected interactions from this table can be graphed using the *Graph* button at the bottom of the table to yield this network view. The user can click on the network to link to navigate to the corresponding IntAct record. A summary of network statistics is shown above the network.

important since the release of large-scale biochemical co-purification studies (Gavin, Bosche et al. 2002; Ho, Gruhler et al. 2002). Data is available in the PSI-MI XML format. A graphical visualization of a set of IntAct records is shown in Figure 5.5.

MINT

MINT, or the Molecular INTeraction database(Zanzoni, Montecchi-Palazzi et al. 2002), contains data on the functional interactions between proteins. This consists mainly of direct protein-protein interactions, but also includes indirect and genetic interactions for more than a dozen organisms. MINT can store enzymatic modifications and information on binding domains and kinetics. It has a similar data model to IntAct and is searchable by protein information, domain and PubMed identifier. Once viewing an interaction record, the user can launch the MintViewer Java application, which allows visual browsing of the database starting from a protein of interest. This viewer is unique in that it shows the sizes of proteins as well as if a protein can be phosphorylated and the number of interactions per edge. Users can submit their own data and the entire database is freely available for download in PSI-MI format. A typical MINT record is shown in 5.6.

5.2.2 Primarily Metabolic Pathway Databases

EcoCyc

EcoCyc (Karp, Riley et al. 2002) is a literature derived, curated encyclopedia of the *Escherichia coli* bacteria metabolism (strain K12). It contains among the highest quality annotation of metabolic pathways of any species. MetaCyc (Karp, Riley et al. 2002) is another literature-derived database that uses the same Pathway Tools software as EcoCyc, thus having the same user interface; it contains information about pathways in over 150 species, mainly microorganisms, with *E. coli* having the largest representation. BioCyc is a collection of pathway databases containing EcoCyc and MetaCyc and additionally contains metabolic pathway reconstructions for a number of organisms with sequenced genomes, including Human, using the Pathologic system. EcoCyc (or, more generally, the Pathway Tools software) uses a hierarchical class structure to define an ontology of metabolic pathways and their components. An ontology is a specification that de-

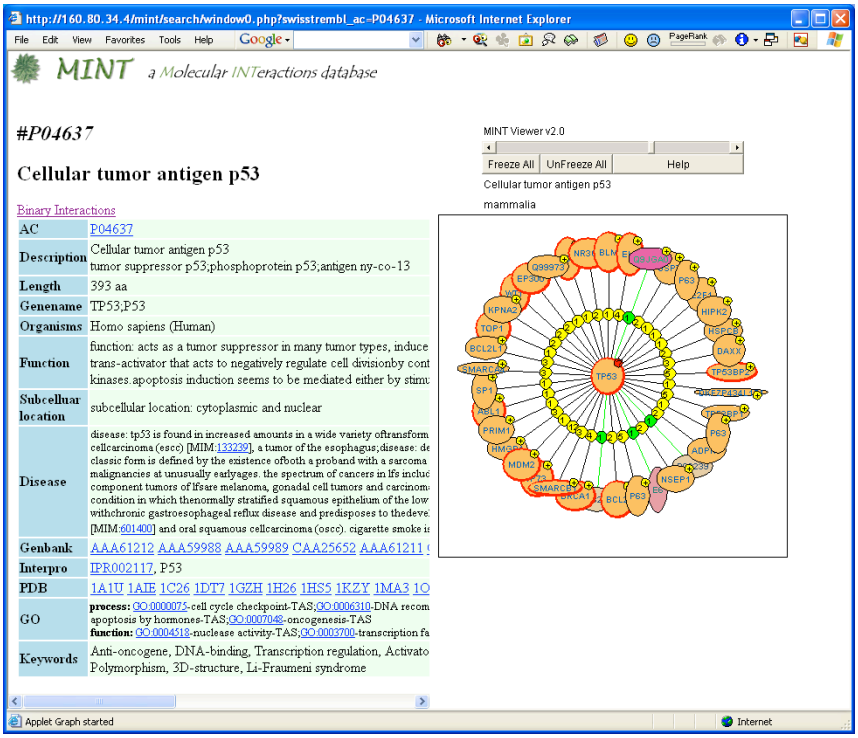


Figure 5.6: A MINT database protein record and associated graphical representation of the protein-protein interactions this protein is involved in. Larger circles in the graphical view indicate larger proteins while small positive symbols indicate more interactions can be viewed if that node is clicked on. The number of experiments in the database that describe an interaction is shown in the small circles in the middle of the lines that represent interactions.

scribes the concepts and their relationships from a specific knowledge domain. It can contain a list of terms and their definitions, a taxonomy of those terms, constraints, attributes and values. An example of a constraint is a value type constraint, where a field in a database record is limited to a specific data type, like a floating point number instead of being any data type. The presence of a class hierarchy allows a computer program to /emphunderstand more about the classification of data into a set of types. For instance, if a ribosomal RNA (rRNA) was classified as such, but the rRNA class is a subclass of RNA, which has children snRNA and tRNA, then a computer program would automatically know by very simple logical inference, in this case, that anything labeled as rRNA, tRNA or snRNA is also of type RNA. This allows certain complex queries to be more easily made, such as a search for all RNAs, even though nothing in the database is specifically annotated as RNA. Although this kind of search can be implemented on a case by case basis in software, the presence of a defined ontology for the data allows it to be done automatically. EcoCyc has classes for chemicals, anatomical structures, enzymatic reactions, and generalized reactions among other things. Chemicals and generalized reactions are the most complex class hierarchies. The top part of the Chemicals class hierarchy is shown in Figure 5.7.

EcoCyc models many aspects of metabolism. Instances of classes exist for all the reactions indexed by the Enzyme Commission and any others that have been added. Enzymes and the biochemical reactions they catalyze are represented as well as information on transcriptional regulation of gene expression and molecular transport processes and binding reactions. The network of *E. coli* metabolism is organized into Pathways and Super Pathways. The underlying Pathway Tools software is among the more feature rich tools for metabolic pathway informatics and allows querying and browsing in numerous ways. A user can query by any number and combination of fields and values using Web-based forms, choose from a list of pathways, browse the ontology, and choose specific classes to see instances of those classes and view a metabolic overview (Figure 5.8) that allows gene expression data to be viewed along with selected pathways. Users can also search for sequences in EcoCyc similar to one of interest using BLAST. Pathways, reactions, compounds and genes all have their own record views with links that allow easy traversal from one part of *E. coli* metabolism to another. Pathways, reactions, compounds, transcriptional units and gene regulation schematics, among other types of data are visualized as simple diagrams to aid understanding of these relationships. Users can also click on many parts of these images to get more information about what they clicked on. Practically, Pathway Tools databases can be downloaded for local installation and accessed using Lisp, Perl or Java. Lisp is the native language of Pathway Tools. Flat files can also be accessed for custom parsing.

Generally, all BioCyc databases that use the Pathway Tools software can be accessed in the same manner. EcoCyc, MetaCyc and some of the other BioCyc databases are only freely available to academics for research purposes, but some databases have used the Pathway Tools to curate their own organism-specific pathway databases, which are made available freely. An example is TAIR, the Arabidopsis Information Resource (Rhee, Beavis et al. 2003), which makes the AraCyc database available.

KEGG

The Kyoto Encyclopedia of Genes and Genomes (Kanehisa, Goto et al. 2002), pathway database contains curated metabolic and signaling pathways. Information on enzymatic reactions, enzymes, small molecules and genes is also available from KEGG. Pathways are available as searchable and clickable images called maps. Pathway maps can depict metabolism, regulatory pathways and large complexes, such as the ribosome. Each type of these maps has their own graphical representation style. Most metabolic pathway maps are reference maps, which depict generalized pathways. Generalized pathways are not species specific, thus might never be found in their entirety in a single species. The user can select to highlight the enzymes on the generalized map that are present in an organism of interest. Some pathway maps, such as a subset of the regulatory maps, are truly species specific. Pathway maps in KEGG link to the underlying LIGAND database comprising three main types of information, enzymes, reactions and compounds. Recently a glycan database was added to KEGG to store information about carbohydrates and their structures. Enzymes that are present in a pathway map are stored in the ENZYME table along with further information about them, which can link to other parts of KEGG. For instance, the reaction field in the enzyme record links to the REACTION table in KEGG, which stores information about each reaction. Individual compounds in the substrate and product fields link to the COMPOUND table. An enzyme record also links to all of the genes that encode that enzyme in the various species stored in the GENES table in KEGG. A typical enzyme record is shown in Figure 5.9. Pathways can be searched and browsed via the

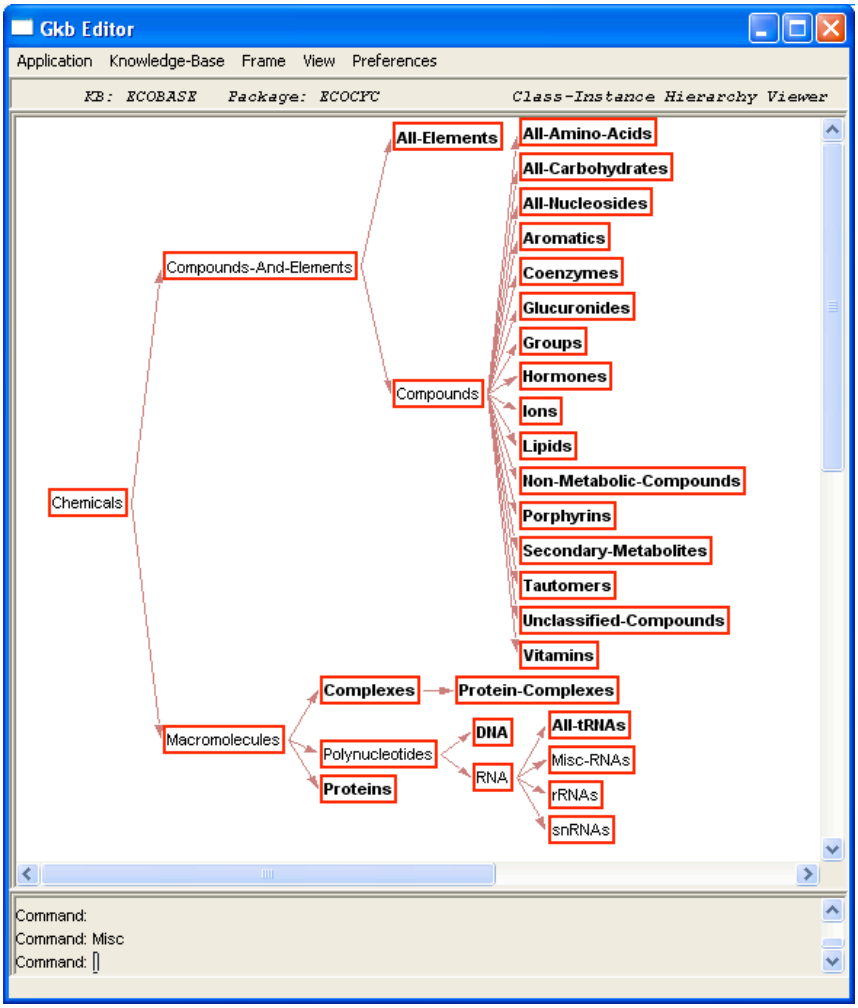


Figure 5.7: A portion of the Chemicals class hierarchy is shown for EcoCyc, which is similar across the BioCyc family of databases. Each red box represents a class of data. Classes are organized from most general at the top of the tree to most specific at the bottom of the tree. Bold font names for some classes indicate that subclasses exists which are not currently expanded. Each class contains a number of slots (also called fields) to hold data, which are not shown. Notice the specialization of data types from the left (root) to the right (leaves) of the class hierarchy.

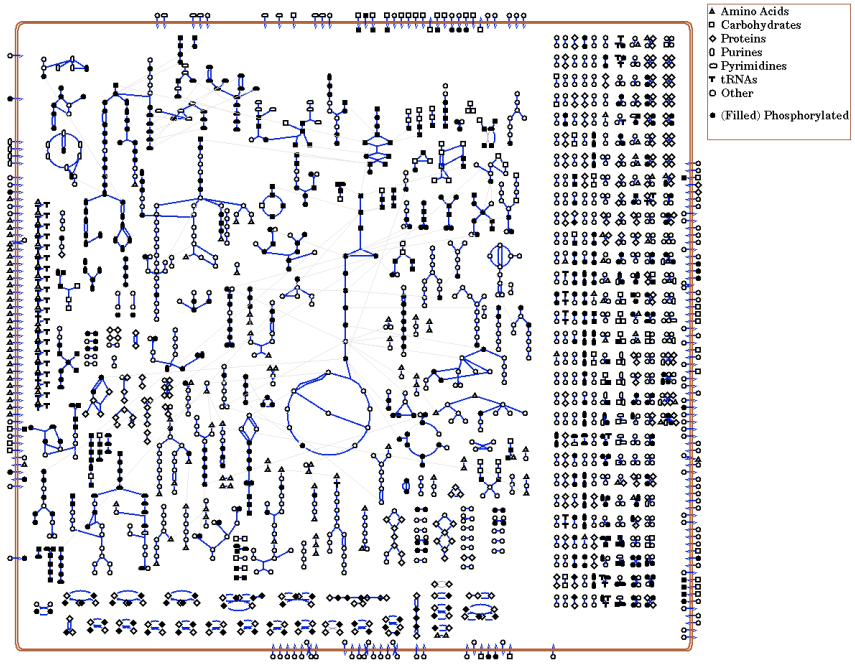


Figure 5.8: An automatically generated overview of all of the pathways and reactions in EcoCyc. The brown double line box around the figure represents the cell membrane of E.coli. The nodes of the network have different shapes depending on their molecular type, but generally represent metabolites, not enzymes. A legend of possible types is shown at the top right of the figure. The central network, which contains a cycle is the glycolysis pathway followed by the TCA cycle. Transport events across the membrane is represented by arrows.

KEGG web site. Enzymes, reactions and compounds can also be searched. The various interlinked KEGG databases are available via the Web and for download by FTP. The COMPOUND database maintains a list of curated chemical structures for most of the compounds available, although only two-dimensional structures are available on the FTP site.

5.2.3 Strategies for Navigating Interaction Databases

The number and various types of interaction databases can be bewildering. From a very general and practical perspective, users searching for the latest molecular interactions from large-scale studies and the literature should search BIND and DIP, since they are the largest resources of this type of information. If a protein name of interest is not found directly in either of these databases, use the BLAST search functionality with the protein sequence of interest. Users interested in human protein-protein interactions should search HPRD as well. Other interaction databases mentioned above should also be searched for completeness. Users interested in well-known metabolic pathways should try searching the BioCyc and KEGG databases. One useful tip is to browse the pathway hierarchy available in the BioCyc pathways to get a sense of the type of information available. Those curious about signal transduction pathways can get a lot of useful information from the BioCarta website even though the information is not in a computer readable form. These resources are likely the most widely used freely accessible interaction and pathway databases, although there are many specialized resources available.

5.2.4 Database Standards

Recently, the pathway informatics community has started to develop common data exchange formats. The Proteomics Standards Initiative (PSI) has developed the first version of an XML-based format for exchanging protein-protein interactions, called PSI-MI (PSI Molecular Interactions; (Hermjakob, Montecchi-Palazzi et al. 2004). The data model of the format is simple, containing an interaction record comprised of a set of proteins that interact, an experimental conditions controlled vocabulary, and information about publication references and protein features, such as binding sites and post-translational modification sites. Figure 5.10 shows the top level of a PSI-MI record. The

```
ENTRY       EC 5.4.2.2

NAME        phosphoglucomutase
            glucose phosphomutase
            phosphoglucose mutase

CLASS       Isomerases
            Intramolecular transferases (mutases)
            Phosphotransferases (phosphomutases)

SYSNAME     alpha-D-glucose 1,6-phosphomutase

REACTION    alpha-D-glucose 1-phosphate = alpha-D-glucose 6-phosphate

SUBSTRATE   alpha-D-glucose 1-phosphate

PRODUCT     alpha-D-glucose 6-phosphate

COMMENT     Maximum activity is only obtained in the presence of
            alpha-D-glucose 1,6-bisphosphate. This bisphosphate is an
            intermediate in the reaction, being formed by transfer of a
            phosphate residue from the enzyme to the substrate, but the
            dissociation of bisphosphate from the enzyme complex is much slower
            than the overall isomerization. The enzyme also catalyses (more
            slowly) the interconversion of 1-phosphate and 6-phosphate isomers
            of many other alpha-D-hexoses, and the interconversion of
            alpha-D-ribose 1-phosphate and 5-phosphate. Formerly EC 2.7.5.1.

REFERENCE   1. Joshi, J.G. and Handler, P. Phosphoglucomutase. I. Purification and
            properties of phosphoglucomutase from Escherichia coli. J. Biol.
            Chem. 239 (1964) 2741-2751.

            2. Najjar, V.A. Phosphoglucomutase, in Boyer, P.D., Lardy, H. and
            Myrback, K. (Eds.), The Enzymes, 2nd edn., vol. 6, Academic Press,
            New York, 1962, pp. 161-178.

            3. Ray, W.J. and Roscelli, G.A. A kinetic study of the
            phosphoglucomutase pathway. J. Biol. Chem. 239 (1964) 1228-1236.

            4. Ray, W.J., Jr. and Peck, E.J., Jr. Phosphomutases, in Boyer, P.D.
            (Ed.), The Enzymes, 3rd edn., vol. 6, Academic Press, New York ,
            1972, pp. 407-477.

            5. Sutherland, E.W., Cohn, M., Posternak, T. and Cori, C.F. The
            mechanism of the phosphoglucomutase reaction. J. Biol. Chem. 180
            (1949) 1285-1295.

PATHWAY     PATH: MAP00010  Glycolysis / Gluconeogenesis
            PATH: MAP00030  Pentose phosphate pathway
            PATH: MAP00052  Galactose metabolism
            PATH: MAP00500  Starch and sucrose metabolism
            PATH: MAP00521  Streptomycin biosynthesis
            PATH: MAP00522  Erythromycin biosynthesis
            PATH: MAP00530  Aminosugars metabolism

ORTHOLOG    KO: K01835  phosphoglucomutase

GENES       HSA: 5236(PGM1) RNO: 24645(Pgm1) DME: CG5165-PA(Pgm)
            CEL: R05F9.6 DRA: DR2071 TMA: TM0184 MMA: MM0301 MM1521

DISEASE     MIM: 171900  Phosphoglucomutase-1

MOTIF       PS: PS00710  [GSA]-[LIVMF]-x-[LIVM]-[ST]-[PGA]-S-H-[NIC]-P

STRUCTURES  PDB: 1JDY  1KFI  1KFQ  1LXT  1VKL  3PMG

DBLINKS     IUBMB Enzyme Nomenclature: 5.4.2.2
            ExPASy - ENZYME nomenclature database: 5.4.2.2
            WIT (What Is There) Metabolic Reconstruction: 5.4.2.2
            BRENDA, the Enzyme Database: 5.4.2.2
            CAS: 9001-81-4
///
```

Figure 5.9: A KEGG ENZYME database record is shown. A number of fields are present including the name, class, reactions, curated annotation, links to KEGG pathway maps and other external resources. Importantly, the enzyme record is for the general Enzyme Commission reaction EC 5.4.2.2, which is for phosphoglucomutase reactions, not any specific enzyme in an organism. The genes that encode this type of enzyme in a number of species is shown in the GENES field, some of which are removed here for brevity. The '///' string terminates the record to aid in parsing.

BIND, DIP, HPRD, MINT and IntAct databases make their data available for download as PSI-MI files. Also network visualization tools, such as Cytoscape, can read and write PSI-MI formatted XML.

Notably, two data exchange formats for exchanging mathematical pathway simulation models are available, Systems Biology Markup Language (SBML) (Hucka, Finney et al. 2003) and Cell Markup Language (CellML). An example of a mathematical pathway model is a system of ordinary differential equations (ODEs) that describe the rates of all of the reactions in a pathway. With the right parameters (for example, initial concentrations and kinetic constants) the computer can calculate the concentration of the various molecular species in a pathway over time. A number of simulation tools support these formats. SBML and CellML do not generally contain information relevant to databases, such as accession numbers for proteins and small molecules involved in the reactions. The scope of each data exchange format is shown in Figure 5.11.

5.3 Prediction Algorithms for pathways and interactions

A number of prediction algorithms for molecular interactions and pathways are available. This section will focus on the most generally accepted types of these algorithms, which attempt to predict protein-protein interactions and reconstruct metabolic pathways from genome sequences. A summary of the main methods is shown in Figure 5.12.

5.3.1 Methods to predict protein-protein interactions

The recently available large number of complete genome sequences has allowed researchers to find sequence based patterns that correlate with protein-protein interactions. Importantly, many of these protein interaction prediction methods predict functional relationships between proteins, which may or may not represent a direct protein-protein interaction. For instance, two proteins may be predicted to interact, but in reality they are only part of the same biological process, pathway or complex. Thus, these predictions have an associated confidence and must be validated experimentally. At the end of this section, a number of online resources that implement these methods and make available the predictions will be discussed.

Gene Neighborhood

Genome co-localization or gene neighborhood approaches were some of the first methods for predicting protein-protein interactions from the genomic context of genes. Such methods exploit the observation that genes whose protein products physically interact (or are functionally associated) are sometimes maintained in close physical chromosomal proximity to each other (Tamames, Casari et al. 1997; Dandekar, Snel et al. 1998; Overbeek, Fonstein et al. 1999). The most obvious case of this phenomenon are operons in bacteria and archaea, where genes whose protein products function in the same biological process are transcribed on the same polycistronic mRNA. Operons are rare in eukaryotic species (Zorio, Cheng et al. 1994; Blumenthal 1998); however, genes involved in the same biological process or pathway are frequently physically proximal on the chromosome (Dandekar, Snel et al. 1998). Thus, it is possible to predict functional or physical interaction between proteins encoded by genes that are repeatedly observed in close proximity (e.g., within 500 bp) across many genomes. This method has been successfully used to identify new members of metabolic pathways (Overbeek, Fonstein et al. 1999). This method is able to predict more functional associations with more complete genomes. In order to assess whether pairs of orthologous genes share a common gene neighborhood across multiple genomes one needs protein sequences and their genomic locations and an orthology mapping between proteins from the various genomes. Orthology mappings are generated by searching for pairs of close bi-directional best hits (PCBBHs). Bi-directional best hit is defined as the best BLAST hit for protein 1 in genome X is protein 1' in genome Y and the best BLAST hit for protein 2 in genome X is protein 2' in genome Y. Pairs of close bi-directional best hits extend this definition to those genes where proteins 1 and 2 are situated within 300 bp in genome X and the genes of proteins 1' and 2' are situated within 300 bp in genome Y. Genes that satisfy these criteria can be considered as having a conserved gene neighborhood across two genomes. When this procedure is repeated across multiple genomes, it becomes possible to identify genes which are statistically significantly co-localized across many genomes, and are hence likely to either physically interact or be functionally associated. These criteria are quite strict, and it is also possible to perform the procedure using pairs of close homologs (PCHs). Sets of PCBBHs or PCHs in multiple genomes are typically scored for significance based on the number and phylogenetic distribution of genomes in which they are co-localized. Phylogenetic distance can be estimated by examining a 16S rRNA

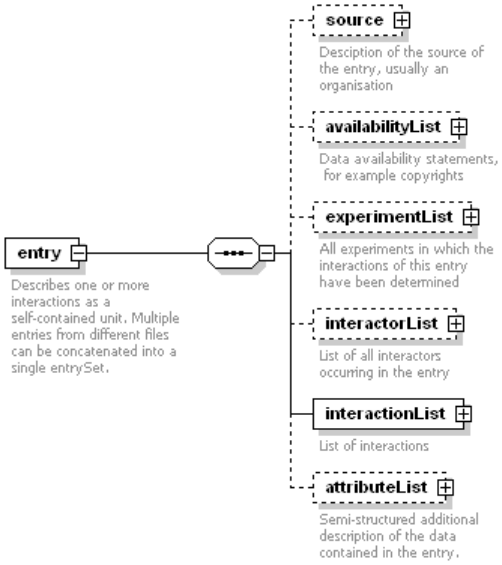


Figure 5.10: The main components of the PSI-MI data model for describing protein-protein interactions. Boxes represent defined XML data types. Dashed lines represent optional elements. The hexagonal box represents a collection of elements that are below it. Minus and plus symbols in small boxes represent expanded and collapsed views of each element, respectively. Collapsed boxes have more elements inside them that are not shown. The full schema for PSI-MI is on the PSI-MI Web site.

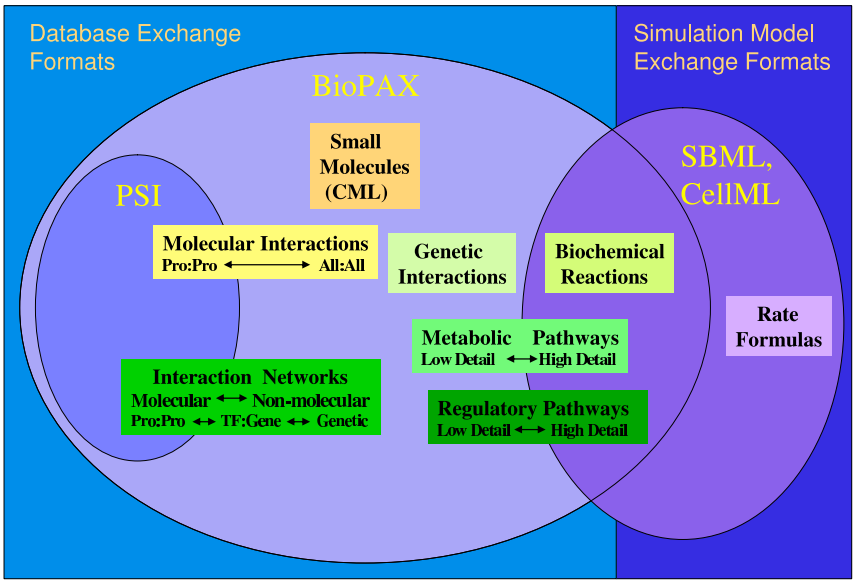


Figure 5.11: A diagram showing the scope of each of the data exchange formats discussed in this chapter. Pro:Pro indicates protein-protein interactions, All:All indicates interactions between many different types of molecules. TF stands for transcription factor. CML stands for Chemical Markup Language, which is an XML format for storing chemicals, generally small molecules. Database exchange formats are those that are primarily suited for data exchange and integration and include data elements like database identifiers. Simulation and model exchange formats are primarily suited for describing mathematical models of biological processes.

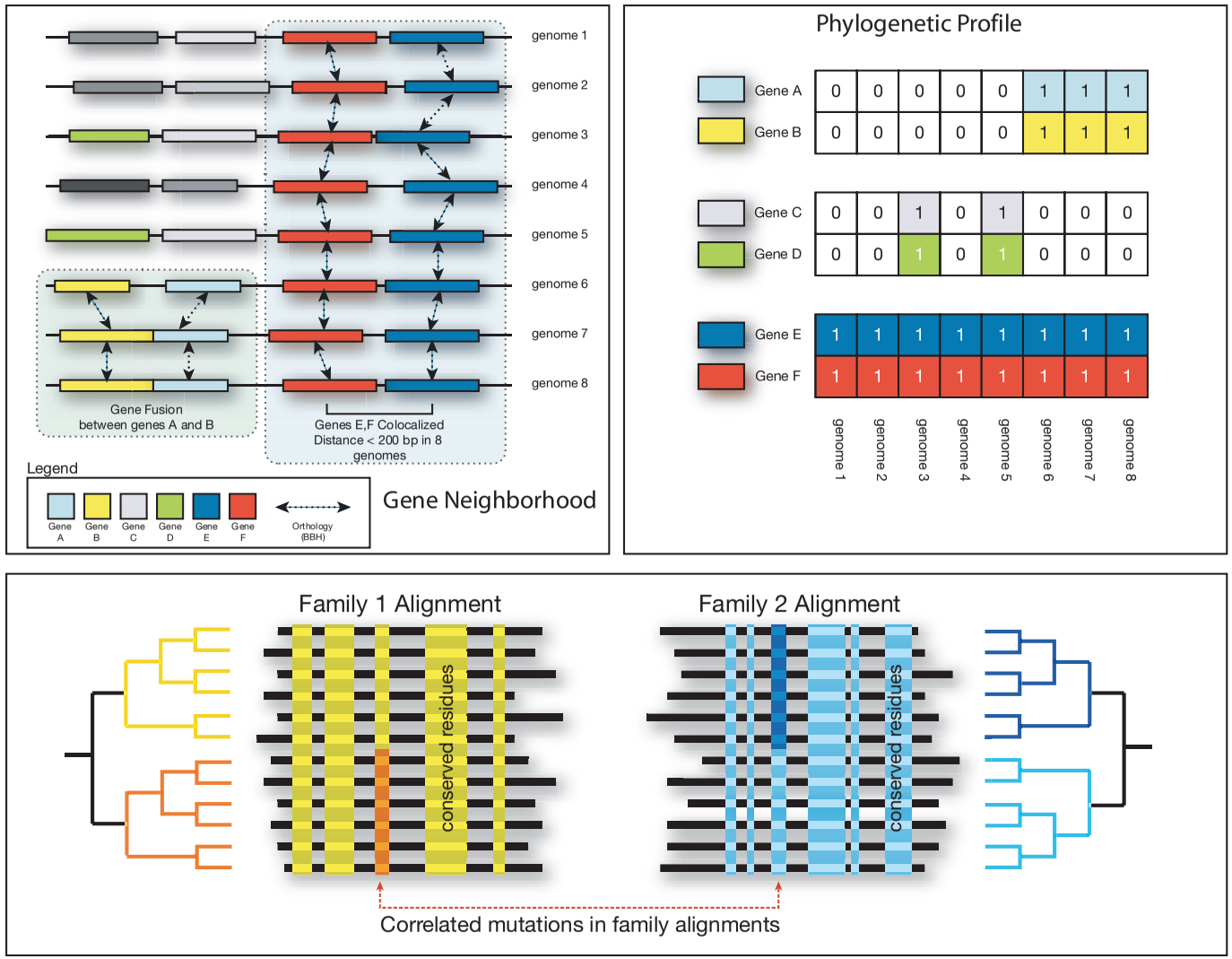


Figure 5.12: (Top left) Co-localization or gene neighborhood plots for eight complete genomes, showing a pair of genes (red and blue) which are in close physical proximity in all eight genomes. A gene fusion event between two genes (yellow and light blue) in two genomes is also shown. (Top right) Example phylogenetic profiles of selected genes from the previous panel. These three pairs of genes have the same patterns of co-occurrence in all eight genomes, and may physically interact based on this evidence. (Bottom) Two protein family alignments are shown with conserved regions highlighted (red and blue). Correlated mutations (green) are present in two identical sub-trees for each family, which indicates that these sites may have co-evolved and thus may be involved in mediating interactions between proteins from each family.

phylogenetic tree. A common score (coupling score) for the likelihood that two genes interact based on summing individual scores from multiple genomes is then calculated. Finally, candidate genes that have significant coupling scores are candidates for either physical interaction, or functional association (Figure 5.12, top left).

Phylogenetic Profiles

Two of the main driving forces in genome evolution are gene genesis and gene loss (Snel, Bork et al. 2002). The fact that a gene pair remains together across many different species often represents a concerted evolutionary effort to keep them together, as might be the case if they functioned in the same biological process. This criterion is less strict than gene co-localization, where gene pairs must not only be present, but also must be closely situated on the genome. Phylogenetic profiles show the presence or absence of genes across complete genomes from many species (Ouzounis and Kyrpides 1996; Rivera, Jain et al. 1998; Pellegrini, Marcotte et al. 1999). Pairs of genes that have very similar phylogenetic profiles are candidates for physical interaction or functional association. This method has been used to infer physical interaction (Pellegrini, Marcotte et al. 1999) and also to predict the cellular localization of gene products (Marcotte, Xenarios et al. 2000), since genes involved in the same biological processes are often co-localized. Disadvantages of this method include heavy dependence on the number and distribution of genomes used to build the profile. A pair of genes with similar profiles across many bacterial, archaeal and eukaryotic genomes is much more likely to interact than genes found to co-occur in a small number of closely related species. Another weakness is that evolutionary processes such as lineage-specific gene loss, horizontal gene transfer, non-orthologous gene displacement (Galperin and Koonin 2000) and the extensive expansion of many eukaryotic gene families can make orthology assignment across genomes very difficult. However, given the increasing number of completely sequenced genomes, the accuracy of these predictions is expected to improve over time. Phylogenetic profile based prediction of protein interactions has been shown to be an accurate and widely applicable method. One of the easiest ways to utilize this information for prediction of protein interaction is to use precomputed phylogenetic profiles for proteins of interest. The Clusters of Orthologous Groups (COGs) resource at the National Center for Biotechnology Information (NCBI) contains large numbers of profiles for various genomes (Tatusov, Fedorova et al. 2003). To construct a phylogenetic profile, an ortholog mapping must be made for all of the proteins in the genomes of interest. Orthologs can be defined using the bidirectional best hit (BBH) approach. A phylogenetic profile for a protein can then be constructed by representing the presence or absence of an ortholog for that protein across all genomes analyzed. This can be efficiently represented as a simple binary vector with '1' indicating presence and '0' representing absence of a gene in each genome. A score of the expectation of presence of a homolog in a genome can be used instead of a binary value. All profiles are compared to all other profiles using a clustering procedure. A distance measure (such as Pearson correlation coefficient or Euclidean distance) between each profile (vector) and all other profiles is used to group profiles according to how similar they are. This correlation calculation can easily be performed using the PEARSON function in Microsoft Excel. Protein profiles that are highly similar or identical to each other represent candidate proteins that physically or functionally interact (Figure 5.12, top right).

Gene Fusion

A gene fusion event represents the physical fusion of two separate parent genes into a single multi-functional gene. This is the ultimate form of gene co-localization: interacting genes are not just kept in close proximity on the genome, but are physically joined as a single entity. It has been suggested that the driving force behind these events is to lower the regulational load of multiple interacting gene products (Enright, Iliopoulos et al. 1999). Gene fusion events hence provide an elegant way to computationally detect functional and physical interactions between proteins (Enright, Iliopoulos et al. 1999; Marcotte, Pellegrini et al. 1999). This method is complementary to both co-localization of genes and phylogenetic profiles and uses both genome location and phylogenetic analysis to infer function or interaction. Gene fusion events are detected by cross-species sequence comparison. Fused (composite) proteins in a given reference genome are detected by searching for un-fused component protein sequences, that are homologous to the reference protein, but not to each other. These un-fused query sequences align to different regions of the reference protein, indicating that it is a composite protein resulting from a gene fusion event. A number of issues can complicate this analysis, the largest of which is the presence of promiscuous domains. These domains (such as helix-turn-helix and DnaJ) are highly abundant in eukaryotic organisms. The domain complexity of eukaryotic proteins coupled with the presence of promiscuous domains and large

degrees of paralogy can hamper the accurate detection of gene fusion events. Although the method is not generally applicable to all genes (i.e., it requires that an observable fusion event can be detected between gene pairs) it has been successfully applied to a large number of genomes, including eukaryotic genomes (Enright and Ouzounis 2001) (Figure 5.12, top left).

In-silico two-hybrid (i2h)

It has been shown that a mutation in the sequence of one protein in a pair of interacting proteins is frequently mirrored by a compensatory mutation in its interacting partner. The detection of such correlated mutations can be used to predict protein-protein interactions and also has the potential to identify specific residues involved at the interaction sites (Pazos and Valencia 2002). Previous analyses (Gobel, Sander et al. 1994) involved searching for correlation of residue mutations between sequences in the same protein family alignment (intra-family). The in-silico two-hybrid method extends this approach by searching for such mutations across different protein families (interfamily). Prediction of protein-protein interactions using this approach is achieved by taking pairs of protein family alignments and concatenating these alignments into a single cross-family alignment. A position-specific scoring matrix is then built from this alignment, and a correlation function is then applied to detect residues which are correlated both within and across families. Correlated sites that potentially indicate protein interaction are scored. Disadvantages of this method include the computational complexity of constructing the large numbers of multiple sequence alignments required. Poor quality alignments can dramatically increase noise in the procedure. One advantage is that a single accurate prediction of an interaction between two proteins can infer interaction between all members of both families used directly from the sequence alignment (Figure 5.12, bottom). A related method, called mirror-tree, predicts that two proteins functionally interact if the phylogenetic trees constructed from the multiple sequence alignments of each protein are similar (Pazos and Valencia 2001). Phylogenetic trees will be similar for proteins that co-evolve, thus this method is similar to phylogenetic profiles.

Other biological context approaches

A common inductive step in biology involves transferring the function from a known gene to an unknown gene that is similar, where similarity can be defined in many ways. This should be treated as a hypothesis that requires experimental validation/falsification. For example, gene functions are often transferred on the basis of sequence similarity. Since the molecular interactions involving a protein define its function, protein interactions may correlate among similar genes (again with similarity being defined in possibly many ways). Gene expression microarrays are often used to detect genes that have similar expression patterns, which therefore may have similar functions. It has been shown that many interacting proteins are co-expressed, based on microarray analyses (Ge, Liu et al. 2001; Grigoriev 2001; Jansen, Greenbaum et al. 2002). Although these methods cannot be used to determine whether or not two proteins directly interact, a number of computational approaches have been developed that use the correlation between correlated gene expression and interaction for protein-protein interaction or functional linkage prediction. This analysis becomes much more reliable with more expression data and genes that have high correlation across ten experiments are much more likely to be related functionally than genes correlating across two experiments. Another biological context based protein interaction prediction approach uses interologs, or orthologous interactions. If an interaction between two proteins is known in one organism, it may be possible to predict that their orthologs bind in another organism. This relationship has not been very well studied due to a lack of comprehensive interaction datasets across species, but it has been shown to be useful in some examples (Matthews, Vaglio et al. 2001; Tien, Lin et al. 2004).

5.3.2 Integrating Existing Datasets

A useful approach to protein interaction prediction would be to use the best predictions of each existing method and ignore the worst predictions. Without experimental validation, how is it possible to know which predictions are good? A recently described interaction prediction method combines information from multiple biological datasets that are known to be noisily correlated to protein-protein interaction to minimize the noise associated with each set in order to reliably predict protein co-complex interactions (both proteins are present in the same complex) in the budding yeast (Jansen, Yu et al. 2003). This method uses Bayesian networks to compare each source of interaction evidence against samples of known positive (proteins in the same complex) and negative (proteins in different cellular locations) interactions, allowing a statistical reliability score of interaction prediction to be calculated for each data source. A probability value for a protein interaction can be calculated given the set of noisy datasets it is present in and the confidence in each dataset. The

noisy data sources included large-scale protein interaction datasets, gene co-expression and similar biological functional annotation. Protein interactions predicted in this way have been shown to be as reliable as pure experimental techniques and cover a larger proportion of genes.

5.3.3 Summary

Each of the methods covered in this section has strengths and weaknesses. Gene neighborhood, phylogenetic, and in silico two hybrid profiles give better predictions the more completely sequences genomes they use. The gene fusion method predicts well, but is not general, since the actual number of detected fusion events is small in existing genomes. All of these methods are currently better suited for prokaryotic proteins, since because there are few completely sequenced eukaryotic genomes available. Co-expression analysis is limited by a low correlation to known protein interactions, although a lot of eukaryotic microarray data is available. With all of these complications to protein interaction prediction, where should one start? The next section addresses this question.

5.3.4 Resources for interaction prediction

A number of resources make available precomputed results using some of the methods described above for a number of genomes in a user friendly manner. These should be queried first when interested in predicted functional linkages among proteins. The STRING resource (von Mering, Huynen et al. 2003) makes available precomputed gene neighborhood, gene fusion, phylogenetic profile, co-expression, and co-mentioned in PubMed abstract based protein-protein functional associations combined with collected experimental and database evidence information for over 110 genomes in a very graphical and user friendly manner. Phylogenetic profiles are derived from the COG database and protein sequences are updated from Swiss-Prot. STRING allows searching by gene name, accession number and sequence of interest (as long as it is already present in the database). Results are graphically displayed and scored using a STRING specific scoring scheme that correlates with validated protein-protein functional associations. Predictions can be filtered by a user defined threshold score. Figure 5.13 shows a screenshot of STRING results. All STRING predictions can be downloaded for local use. Predictome (Mellor, Yanai et al. 2002) is a Web-based tool for visualizing predicted and experimentally determined interactions between proteins. Computational methods include gene fusion, gene neighborhood and phylogenetic profiles, while experimental methods are yeast two-hybrid, biochemical copurification and correlated gene expression. The precomputed analysis is available for over 40 genomes with some genomes, like budding yeast, containing more experimental information than others. Predictome can be searched using gene names or keywords and information is graphically visualized using a Java applet that allows browsing the network of stored interactions. All Predictome predictions can be downloaded for local use. The AllFuse database (Enright and Ouzounis 2001) provides a comprehensive set of inferred protein-protein interactions from gene fusions in 24 complete genomes (both prokaryotic and eukaryotic). This information can also be downloaded by FTP for local use. Finally, the GeneCensus site provides the results of the Bayesian network data integration protein interaction predictions in a searchable and graphical form for budding yeast. A user can search by yeast gene name and select a score cutoff to retrieve the predictions as a graphical network, centered around the gene of interest.

5.3.5 Metabolic Pathway reconstruction

Given a newly sequenced genome and a list of conserved metabolic pathways from a closely related species, it should be possible to predict metabolic pathways in the new genome. A few software systems attempt to do this metabolic pathway reconstruction from complete and almost complete genomes. Signaling pathways cannot currently be reconstructed in this manner because they seem to be much less conserved than metabolic pathways. Starting with a list of predicted ORFs, enzymatic function is assigned in an iterative manner from a list of known enzymes (Figure 5.14). Enzymatic functions are generally assigned by sequence similarity, but could also use any other technique for defining gene function until as many possible links to known pathways have been made. Confidence that a pathway is present in a given organism can be calculated from the number of enzymes that are unique to that pathway that are seen in the new genome. Enzymes that are part of multiple pathways cannot be considered to unambiguously indicate the presence of a pathway. Pathways are then validated by checking that they balance (that is, input compounds equal output compounds). If they don not balance because of missing enzymes, these enzymes can be more thoroughly searched in the genome being annotated. This process is termed hole-filling and can get quite complex. Results for the reconstruction will be

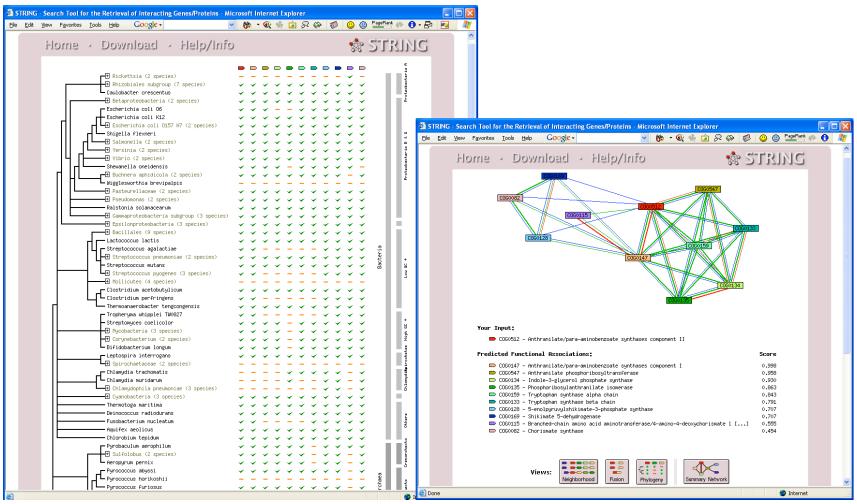


Figure 5.13: Two aspects of the STRING resource are shown. On the left, an overview of the phylogenetic profiling predictions results is shown. A phylogenetic tree is shown on the left side, with a general phylogenetic grouping shown on the right side of the columns. Each column represents a protein shown in the results (different colored pointed boxes as column headings). Green check boxes indicate the protein is present in the corresponding organism and red minus signs mean the protein is not present. Similar columns indicate functional interactions according to the prediction method (see text). The right screenshot shows the results summary view, with different colored lines indicating which method predicted each link. This view is useful to study directly after a query to get an idea of the types and strengths of the results. Red gene fusion, green gene neighborhood; blue - phylogenetic profiling.

better when the experimentally known pathways being used are from a species that is close to the one being annotated. Results will suffer when this is not the case. Additionally, biochemical activity may be observed and characterized in a reference pathway without identifying the enzyme involved. For these pathway steps, no reconstruction can take place. Interestingly, metabolic reconstruction can be performed on gapped, or unfinished genomes, as long as there is enough sequence to make functional enzyme assignments (Selkov, Overbeek et al. 2000). The final stages of metabolic pathway reconstruction can include manual curation of all functional assignments up to and including wet lab experiments to validate the results. Two tools that are available for metabolic pathway reconstruction are the Pathologic and WIT systems. Pathologic is a component of the Pathway Tools software of the BioCyc project, discussed above. Pathologic takes as input a sequenced genome in GenBank flatfile format as well as information on the ORFs and predicted EC numbers for each possible enzyme and uses the MetaCyc database to reconstruct probable metabolic pathways. A number of reconstructions for sequenced genomes are available on the BioCyc Web site.

WIT (for *What Is There?*) is another metabolic pathway reconstruction tool whose results are stored in the WIT database. WIT uses the EMP database of enzyme and metabolic pathways that is available on the same web site for pathway reconstruction given a genome of interest. Pathway reconstructions in WIT start from raw genomic sequence; functional annotation tools are directly used to infer enzymatic functions for predicted ORFs. Finally, each reconstruction is curated by an expert biologist. Currently, WIT contains pathway reconstruction information on 39 completely or partially sequenced genomes and can be searched, browsed via the Web. Reconstructed pathways can be visualized as static clickable images. WIT is scheduled to be replaced by an updated system called SEED in the near future.

5.4 Network and Pathway Visualization Tools

Biologists are very visually oriented people and often prefer studying diagrams over tables. For this reason, it is common to see biologists sketch networks of molecular interactions, although this activity is only practical with small networks. Network and pathway visualization tools are computer programs that automate this task and can draw a diagram of a network or pathway.

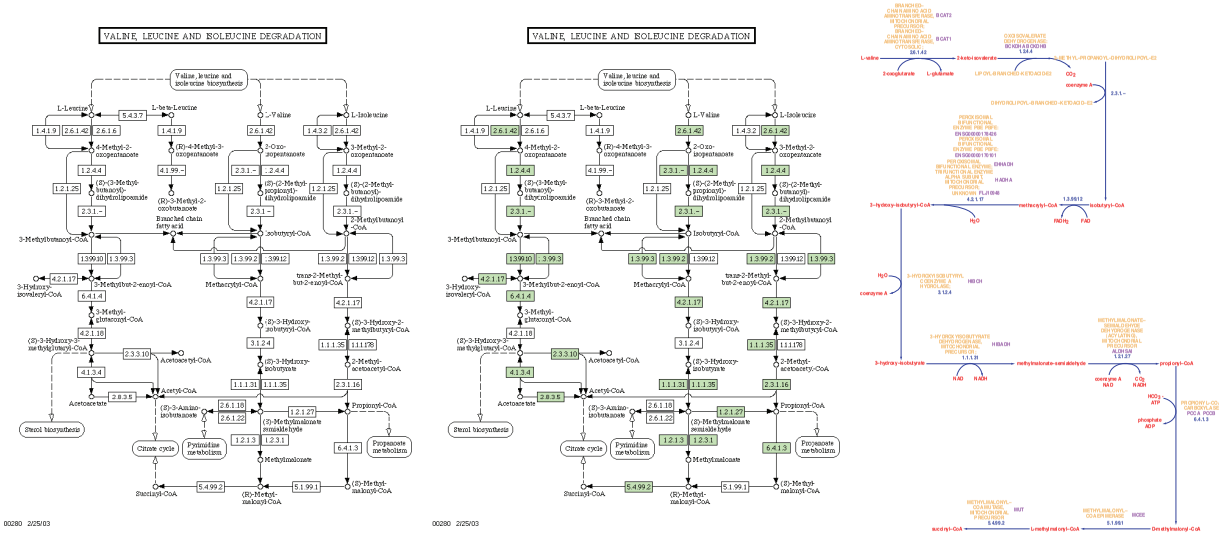


Figure 5.14: An example of metabolic pathway reconstruction from KEGG and BioCyc: The valine degradation pathway reconstructed in human. The left panel shows the reference valine degradation pathway in KEGG. The KEGG reference pathway is a superset of all known valine degradation pathway components from all organisms. The middle panel shows enzymes that KEGG has found to be present in the sequenced human genome highlighted in green. In KEGG, the enzymes are represented by their Enzyme Commission (EC) number (e.g., 2.6.1.42), which defines the enzyme function. The EC system is a hierarchy of enzyme functions similar to the Gene Ontology molecular function controlled vocabulary, but much older. Notice that not all enzymes from the reference pathway are highlighted in green. This is because KEGG was not able to find these enzymes in the human genome. A good example of this is the 3-hydroxyisobutyryl coenzyme A hydrolase (EC 3.1.2.4) that should exist in the human valine degradation pathway because there are no other enzymes from the reference pathway that can replace its function. Thus, this missing enzyme represents a /emphhole in the pathway. This does not mean that the enzyme doesn't exist in the human genome. It may not be easily recognized because of sequence divergence over evolution or because of inaccurate gene finding. The HumanCyc pathway reconstruction from the BioCyc family of databases is able to fill the hole (right panel). Notice that the EC 3.1.2.4 enzyme is present and linked to the HIBCH gene. Clicking on this gene in HumanCyc links to various sequence databases that contain this gene, as well as to publications that provide evidence that the HIBCH gene is an EC 3.1.2.4 enzyme. The extra computational and curatorial effort by HumanCyc is able to fill pathway holes.

Often these tools offer much more than just a pretty picture, and can incorporate data integration features and powerful data analysis modules. This section describes the basic types of tools available, gives some background on how they work and focuses on newly developed tools for network visualization and analysis.

Pathway visualization tools, especially for browsing metabolic pathways, have been around since soon after the first metabolic databases were built. For instance, a pathway drawing tool is present in the ACeDB database (Eeckman and Durbin 1995) and in EcoCyc (Karp, Riley et al. 2002). Many of these tools display static pictures that are clickable, so a user could use their mouse to click on a component of the pathway, such as an enzyme or small molecule, to get more information about it from a database. Examples of static clickable pathway images can be found on the KEGG (Kanehisa, Goto et al. 2002) (Figure 5.14) or BioCarta Web sites. More advanced tools are able to dynamically generate pathway diagrams from an underlying database that allow the user to change how the pathway is viewed. For instance, the EcoCyc database contains a pathway visualization tool that can display varying levels of detail about a pathway, from an overview to a detailed view showing all chemical structures of small molecules in the pathway (Figure 5.15). The WIT database (Overbeek, Larsen et al. 2000) also contains a dynamic pathway diagram generator. Pathway tools view data in manageable chunks, the pathways. While this is very useful for browsing and reference as a database curator generally predefines each pathway, it is not as amenable to analysis of large data sets. In order to map and study new pathways, biologists must possess tools that allow them to easily add their own novel data and expand beyond previously defined knowledge. Very recently, within the past one to two years, a number of biomolecular network visualization tools have been created and made available, fulfilling a need to analyze the large amount of molecular interaction data being generated by proteomics and other high-throughput or large-scale studies (see above). While new and often still under development, network visualization tools can be very useful for understanding relationships within large interconnected data sets. They also provide a practical framework for integrating other types of biological data, such as gene expression values (see below).

All network tools rely on concepts from the computer science field of graph theory, so a brief discussion of basic graph theory concepts is in order. Graph theory is based on the notion of a graph, a representation of connected data as a set of nodes (or vertices) and a set of connecting edges (Figure 5.16). Edges may be directed, in which case they may be called arcs. Nodes and edges may have associated weights or other data values. Different classes of graphs exist; for instance, a graph that does not contain any cycles is called acyclic (also called a tree). Tree graphs have a root node and leaf nodes, and a collection of trees is termed a forest. An example of a directed acyclic graph in bioinformatics is the Gene Ontology, where the most general annotation term is the root and the most specific terms are leaf nodes. The number of edges connected to a node is called the degree for an undirected graph. For a directed graph, in-degree and out-degree are the number of arcs input and output, respectively, from a node. Note that a graph is a completely abstract mathematical concept and can be mapped to any problem where a mapping can be imagined; thus, direction, weight, and connectivity do not have any biological (or other domain) specific meaning until a mapping is made. Intuitively, biomolecular interaction networks can be mapped to a graph, where biomolecules are represented as nodes and interaction information as edges. Other information could also be mapped; for instance, edge direction may be mapped from cell signaling and chemical action information and edge weight may be derived from reaction kinetics, publication opinion, experimental system type, or statistically derived confidence values for the data. Some types of biological interaction information can not be faithfully mapped to a graph, or there may be multiple ambiguous mappings or mappings that cause loss of information. For example, protein complexes larger than two molecules detected in a co-immunoprecipitation experiment cannot easily be described using the binary relationships in a graph; rather, they can only be accurately represented as a set, since the direct physical connections between the proteins in the complex are not known from the experiment. The reason graph theory is used to represent biological networks, is that it can be used to answer many interesting biological questions. For instance, if one wants to find out if one protein connects to another protein in a protein interaction network, an algorithm (called a breadth-first search) can be run that is guaranteed by a mathematical proof to find the shortest path between the two nodes, if it exists. Many other useful graph algorithms exist to manipulate, query, analyze, and visualize graphs. In a social network, where nodes represent people and edges their friendships, the shortest path between people on average is six, hence the famous saying that everyone on Earth is connected by six degrees of separation. While shortest-path is a relatively quick query, some graph algorithms are notoriously difficult to compute. For example, the traveling salesperson problem

Box 5.1:Advanced graph theory applications

Interestingly, any graph can be represented as an NxN matrix, called an adjacency matrix, where the rows and columns represent the nodes in a graph and a '1' is placed in the matrix at position (i, j) if node i connects with node j. If the edges in the graph are weighted, the weight can be recorded at position (i, j) instead of a '1'. Since many types of matrices in bioinformatics are NxN, or square, they can be represented as a graph and it is sometimes useful to make this conversion to visualize the matrix. One interesting example is a protein sequence similarity matrix, which records the sequence similarity (e.g., as calculated by BLAST), of a set of sequences in an all against all fashion. The rows and columns of a similarity matrix represent the set of things being compared; in this case, protein sequences and matrix position (i, j) records the similarity score of protein i compared with protein j. By visualizing this data as a network instead of a matrix, the connections between clusters of similar proteins are more visually apparent.Mathematicians may also convert a graph to the adjacency matrix to apply algebraic matrix operations to the matrix to solve specific graph problems. Sometimes the matrix operations are faster than the same operations performed directly on a graph. For instance, the entries (i, j) in the square of an adjacency matrix correspond to the number of paths of length two that exist in the graph between nodes i and j. This can be extended to higher powers of the adjacency matrix. Squaring the matrix quickly gives this answer if the matrix is sparse (filled with many zeros), but not as quickly if the graph is dense. Fortunately, many problems in biology translate to sparse graphs. One algorithm in bioinformatics that uses this mathematical problem solving tactic to cluster a similarity matrix is MCL (Enright, Van Dongen et al. 2002). Through a series of adjacency matrix multiplications of the similarity graph and other mathematical operations, clusters of similar proteins are detected. Proteins in a similarity cluster have more paths between them than to proteins in other clusters. The matrix squaring operations are involved in counting the number of paths from one protein to another.

aims to find the optimal path (e.g., the least expensive) for a salesperson to travel along in order to visit all cities and return to their starting point, given a number of cities and the cost between each pair. As more cities are considered, the number of possible cycles that visit all of them grows exponentially and finding the cycle of minimum cost using an exhaustive search quickly becomes unfeasible. Often, though, it is possible to approximate the optimal solution in a practical amount of time. More information about graph theory can be found in several well-written graph theory algorithm books (Bollobas 1998; Mehlhorn and Naher 1999; Cormen 2001). Network visualization tools additionally rely on algorithms from the computer science field of graph layout. Typically, graph layout algorithms try to make a graph look aesthetically pleasing; that is, they try to minimize the overlapping of nodes and edges so that as much of the graph as possible is clearly visible. Graph layout is practical and generally works well on small to medium sized graphs, such as those up to about 500 nodes for a typically sized viewing area, such as a computer monitor. Larger networks than this require larger than normal viewing areas, such as a multiple monitor desktop or large format printers. There are many types of graph layout algorithms, such as arranging nodes hierarchically, in a circular fashion or in less structured formats. Importantly, the type of graph layout algorithm that will work best depends heavily on the type of network that is input. For instance a highly connected network will not display well when laid out hierarchically; only a truly hierarchical graph, like a tree, will lay out well in this case. Thus, the most useful network visualization tools contain multiple layout methods that should all be tried to see which one generates the or most aesthetically pleasing layout. One of the most commonly used types of layout is called a spring-embedded algorithm, from the general class of force-directed layout algorithms, which contain many variations on a theme. In a typical case, the graph is modeled as a physical system where edges are springs and nodes are like-charged particles. The layout starts by placing all nodes randomly and then calculates the position of each node given that long edges are like stretched springs and will pull the connected nodes close together, while nodes will repel each other the closer they get. By iterating over time, the graph can stabilize on the final layout, which will have relatively short edges and relatively non-overlapping nodes. Think of this as taking a bunch of like-charged beads (nodes) connected by springs (edges), throwing them up in the air, and seeing what pattern they are in when they land.

5.4.1 BioLayout

BioLayout is a Java-based general network visualization tool with a custom layout algorithm that preferentially places functionally similar nodes together (Enright and Ouzounis 2001) so that biologically relevant functional clustering is more easily seen. For the BioLayout algorithm to function properly, a network has to be loaded with associated functional classes for the nodes. The BioLayout file format supports node classes and edge weights. BioLayout edges are directed by default, but this can be changed in the properties dialog box. A number of useful selection features are available to select nodes based on functional annotation and network topology properties, including a feature to select nodes by edge weight using a slider bar.

A new version of BioLayout has recently been released which improves speed and stability. This version uses the OpenGL toolkit for rapid display of very large graphs in a 3D enviroment (up to 10,000 nodes with 1,000,000 edges). This version includes the ability to automatically cluster graphs using the MCL algorithm and can automatically generate gene expression graphs from Affymetrix data. A screenshot of a large expression graph in 3D is shown in Figure 5.17.

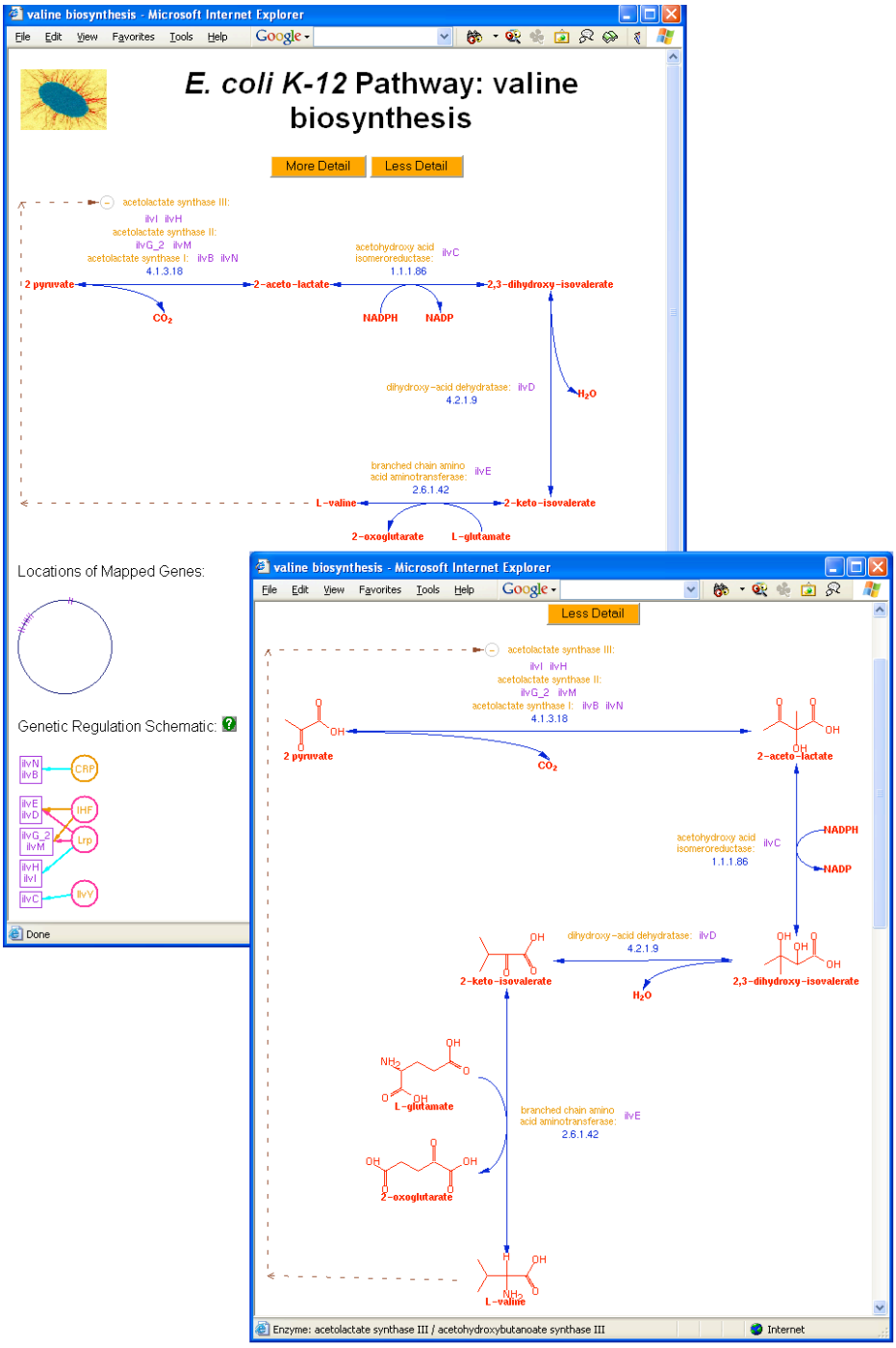


Figure 5.15: The valine biosynthesis pathway dynamically drawn by the Pathway Tools software that supports the BioCyc family of databases. The advantage of automatic pathway diagram layout is that the diagram can be drawn according to user preference. Here two views of the same pathway are shown, the bottom one in more detail than the top. Notice the presence of small molecule structures in the bottom view. The pathway diagram uses the same representation as KEGG, with nodes being metabolites and edges representing enzymes. The lower left side of the left window graphically shows the locations of the genes involved in the pathway on the *E. coli* chromosome. Notice that many of the genes for this pathway are located very close together on the chromosome. This type of information is used by the gene neighborhood protein interaction approach to predict protein functional relationships (see text). The lower diagram shows the genetic regulatory network for transcription factors (circles) that regulate the genes in this pathway (purple squares).

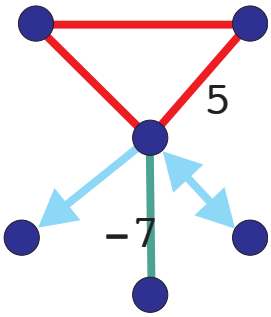


Figure 5.16: An introduction to terminology and visual notation in the computer science field of graph theory. Blue circles are nodes or vertices (singular is vertex), undirected lines (red, green) are called edges, directed lines are called arcs (cyan). Nodes or edges can have associated attributes, such as weights. Here two edge weights are shown, 5 and -7. A series of edges that form a closed loop is called a cycle (red lines). The colors are present in this figure solely to annotate the graph and are not part of normal visual notation. A graph is an abstract mathematical concept. Edge direction, weights and other attributes do not mean anything until mapped to a specific problem.

5.4.2 Cytoscape

Cytoscape is a freely available, open-source Java-based network visualization and analysis tool. Cytoscape's main strengths compared to other network visualization tools are its ability to analyze network data in the context of other types of data, a range of layout algorithms and the ability to add new features using plugins. For instance, gene expression values for specific genes can be mapped as node colors for a network. Cytoscape networks can be edited, nodes can be selected, dragged and rotated using the mouse. Also, complex node and edge selections can be made based on user defined combinations of loaded attributes and graph topology using *filter* functionality. Cytoscape uses the concepts of network attributes and visual attributes when integrating and visualizing information on the network. There are two types of network attributes: node and edge attributes. A node attribute is simply a data value that is loaded onto a node. If the node represents a protein, a node attribute could be the name of the protein, a term that describes the functional classification of that protein, perhaps from the Gene Ontology, or a protein abundance measurement. Similarly, an edge attribute is a data value that is loaded onto an edge. If the edge represents an interaction among two proteins, an edge attribute could be the strength of the interaction or the type of experimental method that was used to detect the interaction. Multiple types of node and edge attributes can be loaded simultaneously, as long as each type has a different name. Either attribute can be discrete or continuous. An example of a discrete attribute is a list of interaction detection experimental methods that could be edge attributes. An example of a continuous attribute is a set of gene expression values that ranges from 0.0 to 1.0. Cytoscape allows all numbers to be continuous and all numbers and strings to be discrete.

Visual attributes in Cytoscape are aspects of a network diagram that could be displayed in different ways (Figure 5.18). The seven types of node visual attributes in Cytoscape are currently node shape, size, label, font, color, border color and border type. The six types of edge visual attributes are currently edge label, font, color, line type (line, dashed line, etc.), target arrow, and source arrow. The last two types represent the arrow type at each side of an edge. Other types of visual attributes can be imagined, such as transparency and ones that display multidimensional data. Once a network is loaded into Cytoscape, any attributes that are loaded onto edges or nodes can be mapped to visual attributes using the Cytoscape visualization mapper. Multiple visual mappings, called visual styles by Cytoscape, can be created, and all are automatically saved in a preferences file whenever any changes are made. A specific example of a mapping for a protein interaction network would be to load up normalized gene expression values, which range from 0.0 to 1.0, with 1.0 being the highest gene expression values in the set, then creating a visual style that maps this to node color with 0.0 being green and 1.0 being red. Cytoscape will then color all nodes continuously according to the style and an expression value of 0.5 will be colored midway between green and red. Another strength of Cytoscape is the ability to add features by loading external plugins. Plugins can be written in Java and can add any type of feature, such as a new layout algorithm or analysis technique. The source code for

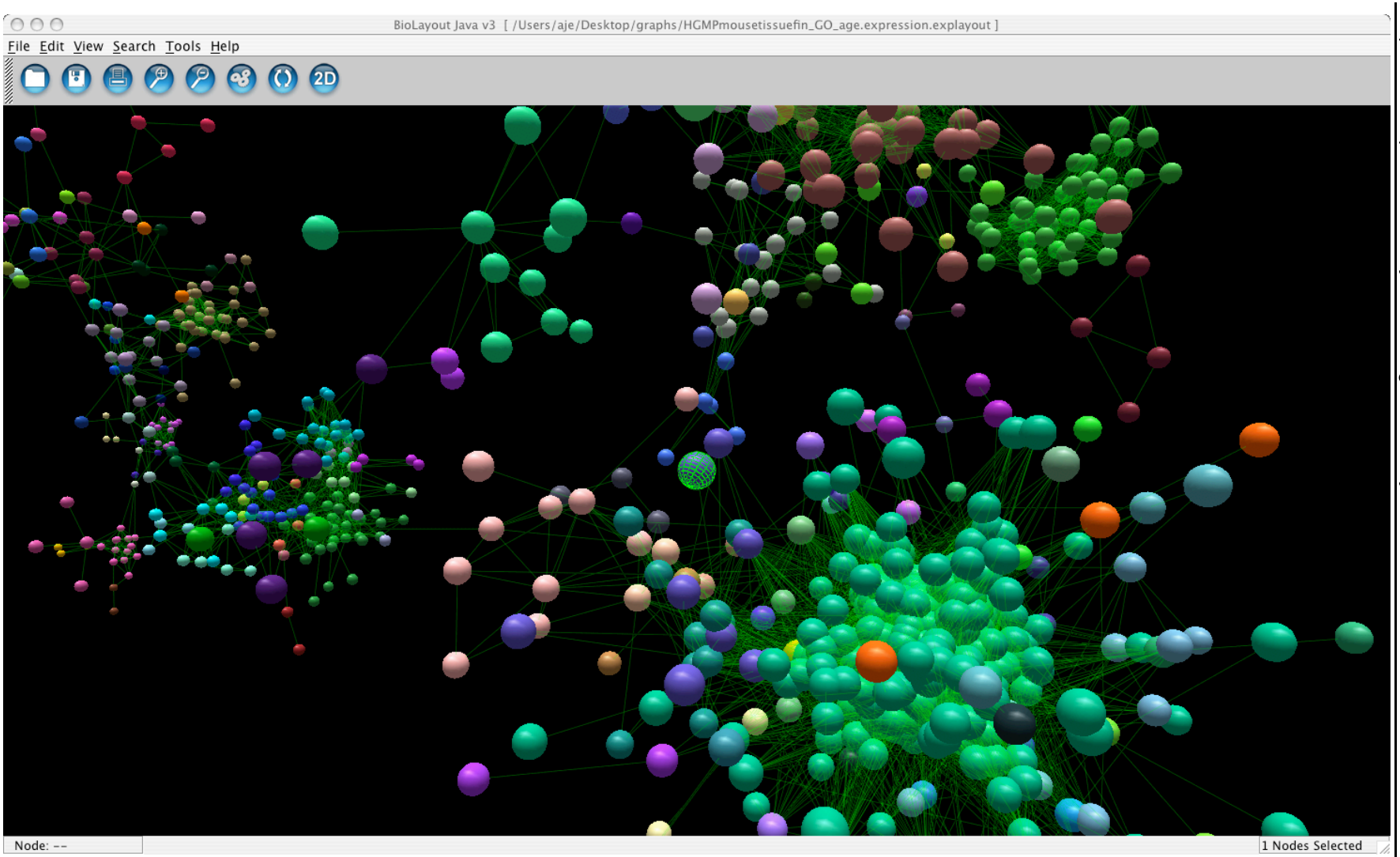


Figure 5.17: A BioLayout gene expression graph plotted in 3D. Nodes here represent probes which are connected to other probes due to their co-expression across a range of timepoints or tissues. Nodes are coloured according to their cluster as defined by using the MCL clustering algorithm.

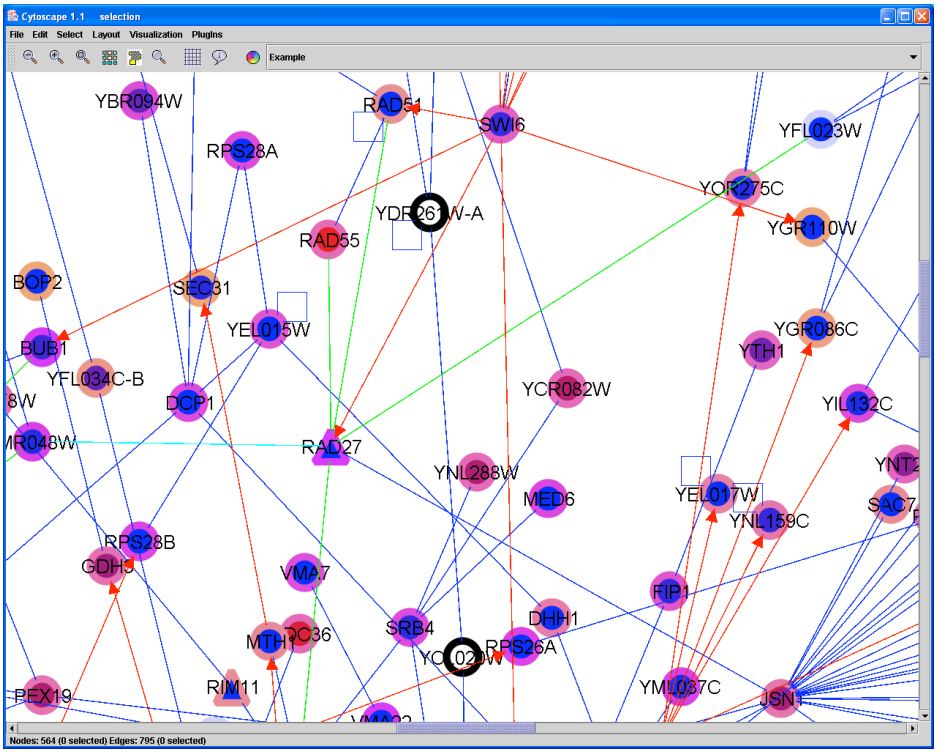


Figure 5.18: Zooming in on a network in Cytoscape shows part of a large connected network of protein and genetic interactions from budding yeast. This view is meant to emphasize the visual customization available in Cytoscape. Different color edges represent different types of interactions (blue is protein-protein, green = genetic synthetic lethal, cyan is genetic synthetic slow growth and red is protein-DNA). Arrows point from protein transcription factor to the genes they bind upstream of potentially regulating them. Nodes represent genes and are colored blue to red by most significant to least significant gene expression fold change from a gene knock-out condition (Gal4 KO) compared to normal control. These are statistically processed gene expression values indicating significance of fold change between two conditions. Node border color ranges from purple to yellow by least gene expression fold change to most gene expression fold change from a gene knock-out condition (Gal4 KO) compared to normal control. These are the raw gene expression values from a cDNA comparative hybridization chip. The shape of each node corresponds to functional annotation of interest with diamond as signal transduction, triangle as either meiosis, PolIII transcription, mating response or DNA repair. Circles are unassigned categories (other annotation). Importantly, this view in Cytoscape is not standard, but was derived by the user by selecting data attributes to map to visual attributes. For instance, the user could have chosen to represent protein-DNA interactions as yellow dashed-line arrows instead of red arrows simply by changing options in Cytoscape's visual mapper.

Cytoscape is freely available, since it is open-source; thus, any interested software developers can write new plugins to extend the functionality of Cytoscape for their own purposes. If these are published, it is hoped that they also would be made freely available to the community, although the availability of a plugin is decided by the plugin creator. Types of plugins that are currently available for Cytoscape include ActiveModules which finds regions of a molecular interaction network that are correlated across multiple gene expression experiments, PathBLAST, which find regions of two protein interaction networks where the protein sequences and connectivity is conserved (possibly with gaps), Biomodules which is a network clustering application for finding loose functional associations of connected molecules in a network. Plugins that add input functionality include a reader for PSI-MI and SBML files. The list of available plugins can be found on the Cytoscape Web page. Cytoscape also makes available a number of network layout algorithms. Useful layouts include circular, hierarchical, organic, embedded. Circular and hierarchical algorithms try to lay out the network as their names suggest. Organic and embedded are two versions of a force-directed layout algorithm.

5.4.3 Osprey

Osprey is a Java-based network visualization and analysis tool for protein-protein and genetic interaction networks(Breitkreutz, Stark et al. 2003). Traditionally, Osprey has been yeast focused, but since it connects to the GRID database, other species that GRID supports are available to it. Currently Osprey connects to yeast, worm, fly and human version of GRID. Osprey has a very user-friendly interface and a number of custom graph layouts that have been designed to better-represent complex biological networks. Osprey supports visualizing a network and assigning colors to the nodes based on function and edges by experimental interaction detection type. A small set of visual attributes can be custom defined for nodes and edges. These are label font and font type, node size, edge width, and edge arrow size. Networks can be edited to add new nodes and edges. When a new node is added, it requires a name, which is searched in the selected GRID database. Any database match for that protein name automatically imports the biological annotation for that protein. If no match is found, the node is added to the graph without annotation. The network analysis features of Osprey are filters for various types of information present as node and edge attributes (network filters) and topology properties (connection filters). For instance, using a network filter, the user can select to only show nodes that match specific GO process terms or edges that were determined using a given experimental system. Connection filters allow selected viewing of node by minimum degree and depth from a selected set of nodes of interest. Networks that are built in Osprey can be saved and loaded in the Osprey file format.

5.4.4 VisANT

VisANT (Hu, Mellor et al. 2004) is a Web-based Java applet for visualizing and analyzing biomolecular interaction data. Like the other tools discussed here, networks can be loaded and visualized with a number of layout schemes. VisANT is interesting in that it directly connects to the Predictome database, which makes it very easy to load and view wide range of existing interaction datasets, either by searching for a gene of interest or by experimental method. For instance, all of the known synthetic lethal interactions can be loaded directly using the Method table in the View menu. VisANT also provides many hyperlinks to web pages that have more information about a particular protein (through a context sensitive menu). Loading data into VisANT is species specific and over 40 species are supported, although most of them are prokaryotic. Interestingly, VisANT includes the ability to search the graph for feedback or feedforward loops and other cycles that may be involved in interesting regulatory processes in signaling pathways. The user can also answer the question *How are my selected proteins connected?* using the find shortest path feature. Even though VisANT runs through a Web browser, it allows the user to upload and save their own data as well as make specific datasets available on the server for others to share (although this requires the user to register with the system and login at the beginning of each session).

5.4.5 Summary

BioLayout useful for laying out large networks in a biologically meaningful way by keeping proteins with similar attributes together. Cytoscape excels at visualizing multiple data types and performing plugin analyses. Osprey is tailored for the biologist at the bench and is recommended for this user group. VisANT connects directly to Predictome, so has direct access to many predicted and experimental interactions, although VisANT currently only runs as a Java applet through a Web page.

5.5 Integrating gene expression data with pathway information

As already mentioned in this chapter, many types of biological data can be integrated with biological pathways or networks for the purpose of gaining biological context. Gene co-expression is correlated with protein interactions and pathways and the two types of information can be used together to help define gene function and further understand the dynamics of cellular pathways. This section will showcase some of the free tools that are available for analyzing gene expression in the context of networks and pathways. There are currently three main categories of tools available that integrate pathway and gene expression information:

- Tools that visualize expression on a pathway diagram
- Tools that perform over-representation analysis (ORA) using pathways
- Tools that co-cluster expression and pathway data

Many tools are available to visualize gene expression information on a pathway diagram. Both Pathway Tools and Pathway Processor (Grosu, Townsend et al. 2002) allow gene expression data to be visualized on predefined pathways from EcoCyc and KEGG, respectively. GenMAPP (Dahlquist, Salomonis et al. 2002) adds the ability to define pathways using basic drawing tools (Figure 5.19). Cytoscape allows visualization of gene expression data on any network 5.18. Generically, these tools must be able to load pathway information and gene expression data and match genes from one data set to the other. The general problem of automatically matching gene identifiers across datasets is an unsolved problem in bioinformatics, so these tools usually require the gene names to match. One tool available that tries to ease the problem of name or identifier matching for gene expression data sets is MatchMiner (Bussey, Kane et al. 2003), currently available for human genes. Once matched, gene expression data is mapped as colors on proteins in the pathway diagram, generally with deeper shades of red indicating overexpression and deeper shades of green indicating underexpression. GenMAPP only runs on Windows, but currently supports human, mouse, rat and budding yeast (new species are supported regularly), while pathway processor runs on any platform with a recent version of Java installed, but only supports budding yeast and *Bacillus subtilis*.

Over-representation analysis is a statistical analysis that determines if a list of items is significantly over-represented in a sample given the number and types of items that exist. For instance, if a sample has three blue items and three red items and there are known to be 500 blue items and 100 red items in the world, then red is over-represented and blue is under-represented in the sample, since it would be expected that there would be five blue items and one red item if items were randomly picked from the set of 600 blue and red items. This can be applied in biology if the samples are, for instance, sets of genes defined by gene expression clustering, by shared Gene Ontology annotation terms, or the set of genes in a pathway or region of a network. Typically, a hypergeometric distribution to model random sampling without replacement, with an optional multiple hypothesis correction (such as a Bonferroni correction), is used to calculate the probability that a set of genes are over-represented compared to chance (Robinson, Grigull et al. 2002). Sometimes, a Fisher's exact test is used to calculate a similar type of probability indicating if there are non-random associations between genes in the category of interest and those that are not in that category. The Fisher's exact test is technically better for small values, but both of these statistical methods can not perfectly deal with problems of errors and systematic bias in the sample (Zeeberg, Feng et al. 2003), so statistical values indicating over-representation should be interpreted carefully before basing further studies on the results. At least four tools are available that perform ORA on gene lists, which can be derived from gene expression datasets (for instance over or underexpressed genes) using a number of possible annotation categories. MAPPFinder (Doniger, Salomonis et al. 2003), from the GenMAPP project, tests whether a given GO term is significantly enriched for genes of interest from a MAPP file, using a hypergeometric distribution. Since MAPP files can represent any set of genes, including pathways and simple sets of genes with the same Gene Ontology annotations, MAPPFinder is quite general once a MAPP file is constructed. MAPPFinder is conveniently deals with gene expression datasets compatible with GenMAPP. GoMiner (Zeeberg, Feng et al. 2003) analyzes two sets of genes, one containing genes of interest (for instance overexpressed) and the other containing all known genes in a set (for instance on a microarray) and tests whether a GO annotation term is significantly enriched or depleted in interesting genes, using the Fisher's exact test method. Both MAPPFinder and GoMiner show their results in the context of the Gene Ontology, but MAPPFinder displays slightly more statistical information for each GO term and GoMiner con-

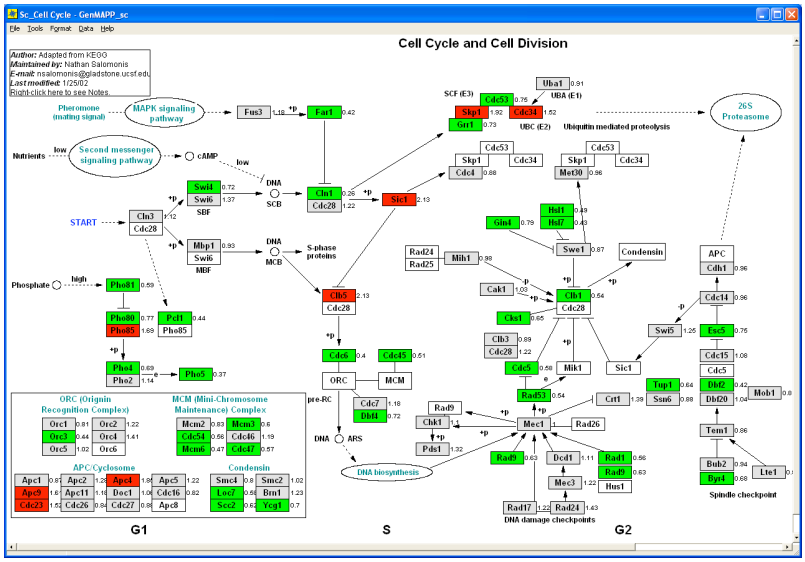


Figure 5.19: The GenMAPP program displaying a hand curated pathway of the yeast cell cycle overlaid with gene expression data from a single experiment. Proteins/genes are represented as boxes and relationships between them as lines. When creating a pathway diagram, the user can use a number of drawing tools to indicate more than just molecular interactions. For instance, large oval on the top left of this diagram represent signaling pathways and the oval on the top right represents the 26S proteasome, a large protein complex. Both of these are general concepts represented using the same shape and rely on previous biological knowledge to differentiate them. Up-regulated genes are colored red and down-regulated genes are colored green according to a gene expression experiment that was loaded in GenMAPP. Information about the maintainer of this pathway is shown on the upper left hand side of the figure.

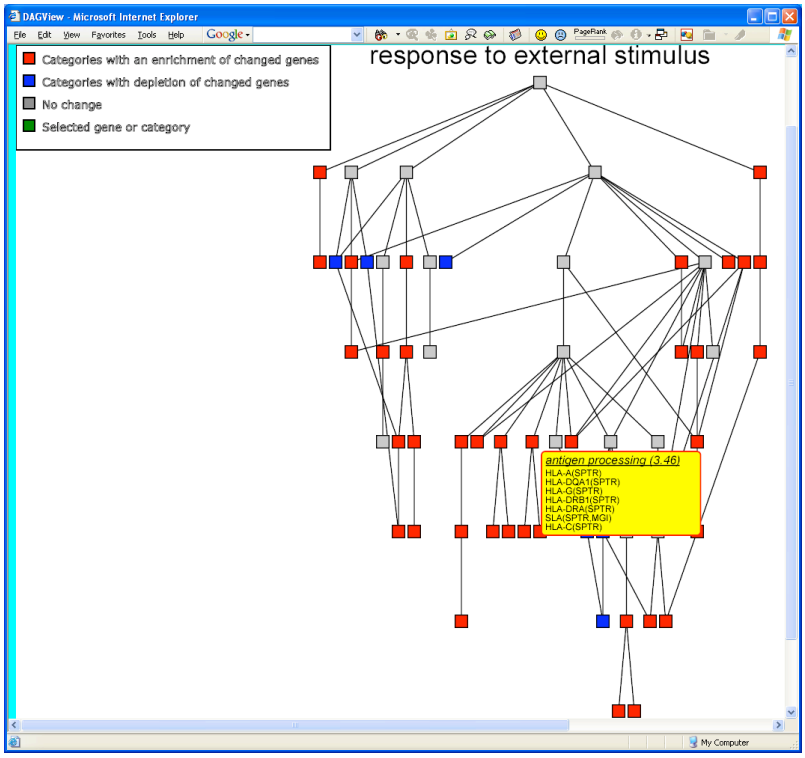


Figure 5.20: The Gene Ontology (GO) DAG summary view from GoMiner showing GO annotation terms under the general term *response to external stimulus*. GO terms are shown as boxes and lines between the boxes indicate parent-child relationships among terms. Enriched interesting terms (red) and depleted interesting terms (blue) are derived using over-representation analysis from the list of interesting genes that were input into GoMiner for this analysis. An example of an interesting gene list could be the set of genes that were found to be co-expressed in a microarray gene expression experiment. Running the mouse over specific categories activates display of a yellow box showing the GO term and the genes annotated with that GO term. Other views of the data in table form are available from the main GoMiner window.

tains a visualization of a large part of the GO network as a summary, with terms colored by significant enrichment or depletion of interesting genes (Figure 5.20). EASE (Hosack, Dennis et al. 2003) is a tool that performs ORA on a given list of genes using an easily customizable set of annotation categories. Predefined categories include GO, KEGG, PFAM, SMART and a number of others. EASE gene lists support numerous gene identifiers including Affymetrix IDs for easy analysis of gene expression results. Funspec (Robinson, Grigull et al. 2002) is another tool that performs ORA, but for any given set of budding yeast genes against a number of different types of annotation from GO, MIPS, SMART and PFAM domains and custom gene sets from large-scale interaction, localization and protein complex experimental mapping studies. Finally, a Cytoscape plugin called ActiveModules (Ideker, Ozier et al. 2002) is available to find regions of a given network that are co-regulated across multiple gene expression conditions. Co-regulated subgraphs are called *active modules* and are hypothesized to represent pathways or biological processes whose components are active at the same time. Gene expression data must be statistically processed to yield p-values of significance of fold change between two conditions in order to use ActiveModules, thus at least two gene expression conditions are required for Affymetrix style gene expression data, one condition of interest and one a control or at least one cDNA differential fluorescence hybridization chip condition, but ActiveModules was designed to analyze multiple sets of fold change p-values at once. It is interesting to note that while the tools mentioned here have mostly been designed for studying gene expression information, they can also be used to study other types of data, such as protein expression or other similar data that might be of interest to analyze. Also, gene expression is only mildly correlated with protein abundance, thus conclusions about the activity of biological processes from gene expression need to be further validated experimentally. Analysis that integrates transcriptional profiling with pathway information is currently being heavily researched. One example of this is regulatory network reconstruction, where gene expression data sets from multiple conditions as well as a number of known gene expression regulators, such as transcription factors and signal transduction proteins that activate or inactivate those transcription factors, are used to reconstruct portions of the gene regulatory network (Segal, Shapira et al. 2003). As this research progresses, more advanced tools for this analysis will undoubtedly be created.

5.6 Summary

Many other topics about protein-protein interactions and pathways exist beyond what has been covered in this chapter. A sample of these very interesting topics are mathematical pathway modeling (Bower and Bolouri 2001), molecular docking of proteins with proteins and proteins with small molecules (Ofran and Rost 2003), genetic interactions (Forsburg 2001), and molecular interaction network clustering (Bader and Hogue 2003).

5.6.1 Worked Example:

Given a protein sequence, its function and possible biomolecular interactions will be predicted using a number of contextual genomic approaches. This example, while not covering all available means, provides a reasonable overview of available methods.

This example will use the yeast tryptophan biosynthesis enzyme TrpCF (SWISSPROT ID: TRPC_ECOLI). This is a bi-functional enzyme with one domain consisting of an indole-3-glycerol phosphate synthase (EC 4.1.1.48) and the other an isomerase (EC 5.3.1.24). These two domains carry out two enzymatic steps of tryptophan biosynthesis and the substrate is passed between them. This protein is a typical example of gene fusion in metabolic enzymes.

Although the sequence, structure and function of the enzyme are well characterized, very little experimental information is available about which proteins this enzyme interacts with (both physical and function interactions). Going to the STRING Web site and typing the identifier of this gene (TRPC_ECOLI) into the identifier search field at the top page results in a summary of predicted interactions displayed from precomputed data.

The initial search returns predicted interaction data from a variety of sources and is shown in Figure 5.21.

Each row in Figure 5.21 indicates a potential interactor based on evidence from a number of sources. Evidence is scored by STRING from 0 to 1. In this case the threshold is set at 0.40, so only evidence scoring above this threshold is shown. For example, the TrpCF query protein is predicted to interact with TrpB by virtue of high-quality evidence from gene neighborhood, gene fusion and text-mining. Further medium quality evidence from gene concurrence (phylogenetic profile) adds to this, providing a total evidence score of 1.0. This evidence is obtained across multiple species using ortholog information.

Clicking on the evidence buttons displayed below the results provides details about where the evidence originated from each source. Clicking on the

		Neighborhood	Gene fusion	Cooccurence	Homology	Co-expression	Experiments	Databases	Text mining	Score
TRPB_ECOLI	Tryptophan synthase ? chain	O	O	O					O	1.00
TRPA_ECOLI	Tryptophan synthase ? chain	O		O		○	○	○	O	0.99
TRPE_ECOLI	Anthranilate synthase I	O		O					O	0.99
TRPG_ECOLI	Anthranilate synthase II	O	O	O					O	0.99
PABA_ECOLI	Para-aminobenzoate synthase II	O	O						O	0.90
LPW_ECOLI	Trp operon leader peptide	O							O	0.80
ARGB_ECOLI	Acetylglutamate kinase								O	0.70
PABB_ECOLI	Para-aminobenzoate synthaase I	O					○	○	O	0.60
MOAC_ECOLI	Molybdenum cofactor biosynthesis protein C	O							O	0.54
ILVC_ECOLI	Ketol-acid reductoisomerase			O						0.44

O, high-quality evidence; O, medium-quality evidence; ○, low quality evidence.

Figure 5.21: Initial Predicted Interactions from Worked Example Search:w!

fusion button, hence shows that predicted interactions between the query TrpCF (TRPC_ECOLI) and PabA (PABA_ECOLI) are due to these genes being fused in two yeast species. Similarly, the interaction predicted with TrpG, is due to these genes being fused in the organism Archaeoglobus fulgidus.

Clicking on the *Gene Neighborhood* button, shows that many genes in this pathway are kept together in close proximity on many genomes. The neighborhood plot generated shows that TrpCF is often located in close proximity to many other pathway members (e.g., TrpB and TrpA).

Finally, if one clicks on the *Summary Network* button, all of these predictions are overlaid in a graphical representation of this protein's predicted interactions with other proteins. Nodes in this network are connected according to evidence from an interaction prediction source. Evidence from multiple sources is shown as parallel edges connecting the same nodes and allows the highest quality interaction predictions to be easily located in the network.

Proteins involved in biological pathways or in the formation of protein complexes will often form *cliques* (highly connected subnetworks). The resulting graph for TrpCF is one such case. Changing the default *network depth* parameters to values greater than two, expands this network and allows interconnected pathways, and biological processes to be visualized. At a network depth of two, one should be able to see connections between these tryptophan biosynthesis genes other biosynthesis pathways including: aromatic amino acids, arginine, valine, isoleucine and histidine. An intriguing connection between TrpG (TRPG_ECOLI) and a group of highly-connected hypothetical proteins based on gene-concurrence (phylogenetic profile) should be visible in this network.

Using the STRING resource hence allows a gene to be placed in a biological context according to a large number of predicted functional associations and protein-protein interactions across many species. Given the large number of poorly characterized or unannotated (hypothetical) genes and proteins in complete genomes, this example shows how useful a resource like STRING is for biological discovery and the guiding of directed experiments.

5.6.2 Problem Set:

```
>Hypothetical protein
SFNTIIDWNSCTAVQQRQLLMRPAISASESITRTVNDILDNVKARGDEALREY
SAKFDKTTVTALKVSAEEIAAASERLSDELKQAMAVAVKNIETFHTAQKLPPV
DVETQPGVRCQQVTRPVASVGLYIPGGSAPLFSTVLMLATPARIAGCKKVVLC
SPPPIADEILYAAQLCGVQDVFNVGGAQAIAALAFGTESVPKVDKIFGPGNAF
VTEAKRQVSQRLDGAAIDMPAGPSEVLVIADSGATPDFVASDLLSQAEHGPDS
QVILLTPAADMARRVAEAVERQLAELPRAETARQALNASRLIVTKDSAQCVEI
SNQYGPEHLIIQTRNARELVDSITSAGSVFLGDWSPESAGDYASGTNHVLPTY
GYTATCSSLGLADFQKRMTVQELSKEGFSAVASTIETLAAAERLTAHKNAVTL
RVNALKEQA
```

A protein sequence that was derived as part of a bacterial genome sequencing project is shown. Using this sequence, perform the following steps:

- Using the STRING webserver, search using this sequence for genome context links and for the direct function of the protein.
- Build an initial summary network with default parameters to visualize functional links between this protein and other proteins.
- Go to the KEGG webserver and search this sequence against known metabolic pathways. Use BLASTP when searching the KEGG GENES database. Find the corresponding *E. coli* entry (identifier will start with 'eco') for this gene and its KEGG pathway. Remember that, frequently, the same gene may have different identifiers in different databases.
- Click on the KEGG pathway to draw the metabolic map of the pathway that this gene is involved in.

- Go back to the initial STRING network and try to map each member of the functional network to this metabolic pathway.

Based on this analysis, answer the following questions:

1. What is the function of this gene?
2. What is the core metabolic pathway that this gene is involved in?
3. How much of the core KEGG pathway is recovered by functional interactions?
4. What other KEGG pathways are linked to by functional interactions?
5. By varying the network depth and interactors shown fields of STRING, try to reconstruct a genomic network of pathways and complexes that link to this core pathway. Go to a maximum network depth of 5 and no more than 50 interactors shown.

Answers:

1. Histidinol dehydrogenase (HDH); HISX_ECOLI (hisD) EC:1.1.1.23.
2. Histidine metabolism (KEGG PATH:eco00340).
3. Most of the KEGG pathway should be recovered by the initial functional interactions: i.e. hisG, hisIE, hisA, hisH/hisF, hisC, hisD. Although hisB is not currently recoverable using genome context approaches.
4. You should find links from ARLY_ECOLI (argH) to Urea cycle and metabolism of amino groups [PATH:eco00220]; Alanine and aspartate metabolism [PATH:eco00252] and Arginine and proline metabolism [PATH:eco00330]. You will also find links to Valine, leucine and isoleucine biosynthesis [PATH:eco00290] through LEU3_ECOLI (leuB).
5. The network returned should look very dense and complicated. Obvious features in this genomic functional network are many amino-acid biosynthesis pathways including: Leucine, isoleucine and valine, tryptophan, arginine and methionine as well as purines. You should also find complexes involved in transcription and translation including RNA polymerase subunits (rpoA), elongation factors (EFG), helicases and ribosomal proteins. Iron (e.g. FecD) and amino-acid transporters (aroP) are also part of this complex network.

5.6.3 Web Resources

See Table 5.2 for a list of websites and resources described within this chapter.

5.6.4 Further reading

1. (Bader, Heilbut et al. 2003) Functional genomics and proteomics is producing enormous amounts of data quickly. This review discusses the various types of large-scale data available as of 2003, especially pathway related data, and the possibility of integrating it to better understand the workings of the cell. 2. (Uetz 2002) This book chapter discusses various ways of visualizing molecular interaction information. 3. (Karp 2001) This review discusses pathway databases and knowledge representation, focusing on descriptions of EcoCyc as an example. 4. (Phizicky and Fields 1995) This review is one of the best collections of descriptions of experimental methods to detect protein-protein interactions.

5.6.5 Acknowledgements

The authors would like to acknowledge Chris Sander for support during the writing of this chapter, Michael Cary for work on the Pathway Resource List and Debbie Marks for helpful editorial input.

Resource Name	URL
AllFuse	http://www.ebi.ac.uk/research/cgg/allfuse/
ayesian network data for yeast	http://bioinfo.mbb.yale.edu/genome/intint/
BIND	http://bind.ca
BioCarta	http://www.biocarta.com
BioCyc	http://biocyc.org/
BioLayout	http://biolayout.org/
BioPAX	http://www.biopax.org
CellML	http://www.cellml.org/
Cytoscape	http://www.cytoscape.org/
DDBJ/EMBL/GenBank Feature Table Definition	http://www.ncbi.nlm.nih.gov/projects/collab/FT/
DIP	http://dip.doe-mbi.ucla.edu/
Funspec	http://funspec.med.utoronto.ca/
GenMAPP and MAPPFinder	http://www.genmapp.org/
GoMiner	http://discover.nci.nih.gov/gominer/
GRID	http://biodata.mshri.on.ca/grid
HPRD	http://www.hprd.org/
IntAct	http://www.ebi.ac.uk/intact
KEGG	http://www.genome.ad.jp/kegg/kegg2.html
MINT	http://160.80.34.4/mint/
Osprey	http://biodata.mshri.on.ca/osprey
Pajek	http://vlado.fmf.uni-lj.si/pub/networks/pajek/
Pathway Resource List	http://www.cbio.mskcc.org/prl
Predictome	http://predictome.bu.edu
PSI-MI	http://psidev.sourceforge.net
SBML	http://sbml.org/
STKE	http://stke.sciencemag.org/
STRING	http://www.bork.embl-heidelberg.de/STRING/
WIT	http://wit.mcs.anl.gov/WIT2/

Table 5.2: Internet resources for analysis of pathways, interactions and function

5.7 References

Bader, G. D., D. Betel and C. W. Hogue (2003).."BIND: the Biomolecular Interaction Network Database." Nucleic Acids Res. 31(1): 248-250.
Bader, G. D., I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson and C. W. Hogue (2001).."BIND–The Biomolecular Interaction Network Database." Nucleic Acids Res. 29(1): 242-245.
Bader, G. D., A. Heilbut, B. Andrews, M. Tyers, T. Hughes and C. Boone (2003).."Functional genomics and proteomics: charting a multidimensional map of the yeast cell." Trends Cell Biol 13(7): 344-56.
Bader, G. D. and C. W. Hogue (2003).."An automated method for finding molecular complexes in large protein interaction networks." BMC.Bioinformatics. 4(1): 2.
Blumenthal, T. (1998).."Gene clusters and polycistronic transcription in eukaryotes." Bioessays 20(6): 480-7.
Bollobas, B. (1998). Modern graph theory. New York, Springer.
Bower, J. M. and H. Bolouri (2001). Computational modeling of genetic and biochemical networks. Cambridge, Mass., MIT Press.
Breitkreutz, B. J., C. Stark and M. Tyers (2003).."The GRID: the General Repository for Interaction Datasets." Genome Biol. 4(3): R23.
Breitkreutz, B. J., C. Stark and M. Tyers (2003).."Osprey: a network visualization system." Genome Biol. 4(3): R22.
Bussey, K. J., D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, W. Ajay and J. N. Weinstein (2003).."MatchMiner: a tool for batch navigation among gene and gene product identifiers." Genome Biol. 4(4): R27.
Cormen, T. H. (2001). Introduction to algorithms. Cambridge, Mass., MIT Press.
Dahlquist, K. D., N. Salomonis, K. Vranizan, S. C. Lawlor and B. R. Conklin (2002).."GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways." Nat Genet 31(1): 19-20.
Dandekar, T., B. Snel, M. Huynen and P. Bork (1998).."Conservation of gene order: a fingerprint of proteins that physically interact." Trends Biochem Sci 23(9): 324-8.
Deane, C. M., L. Salwinski, I. Xenarios and D. Eisenberg (2002).."Protein interactions: two methods for assessment of the reliability of high throughput observations." Mol.Cell Proteomics. 1(5): 349-356.
Doniger, S. W., N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor and B. R. Conklin (2003).."MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data." Genome Biol 4(1): R7.
Eeckman, F. H. and R. Durbin (1995).."ACeDB and macace." Methods Cell Biol. 48: 583-605.
Enright, A. J., I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis (1999).."Protein interaction maps for complete genomes based on gene fusion events." Nature 402(6757): 86-90.
Enright, A. J. and C. A. Ouzounis (2001).."BioLayout–an automatic graph layout algorithm for similarity visualization." Bioinformatics. 17(9): 853-854.
Enright, A. J. and C. A. Ouzounis (2001).."Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions." Genome Biol 2(9): RESEARCH0034.
Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002).."An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res. 30(7): 1575-1584.
Forsburg, S. L. (2001).."The art and design of genetic screens: yeast." Nat.Rev.Genet. 2(9): 659-668.
Galperin, M. Y. and E. V. Koonin (2000).."Who's your neighbor? New computational approaches for functional genomics." Nat Biotechnol 18(6): 609-13.
Gavin, A. C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, et al. (2002).."Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature 415(6868): 141-147.
Ge, H., Z. Liu, G. M. Church and M. Vidal (2001).."Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae." Nat.Genet. 29(4): 482-486.
Gobel, U., C. Sander, R. Schneider and A. Valencia (1994).."Correlated mutations and residue contacts in proteins." Proteins 18(4): 309-17.
Greenbaum, D. and M. Gerstein (2003).."A universal legal framework as a prerequisite for database interoperability." Nat Biotechnol 21(9): 979-82.
Grigoriev, A. (2001).."A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae." Nucleic Acids Res. 29(17): 3513-

3519.
Grosu, P., J. P. Townsend, D. L. Hartl and D. Cavalieri (2002).."Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks." Genome Res 12(7): 1121-6.
Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, et al. (2004).."The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data." Nat Biotechnol 22(2): 177-83.
Ho, Y., A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, et al. (2002).."Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." Nature 415(6868): 180-183.
Hosack, D. A., G. Dennis, Jr., B. T. Sherman, H. C. Lane and R. A. Lempicki (2003).."Identifying biological themes within lists of genes with EASE." Genome Biol 4(10): R70.
Hu, Z., J. Mellor, J. Wu and C. DeLisi (2004).."VisANT: an online visualization and analysis tool for biological interaction data." BMC BioinformaticsBMC Bioinformatics 5(1): 17.
Hucka, M., A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, et al. (2003).."The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models." Bioinformatics 19(4): 524-31.
Ideker, T., O. Ozier, B. Schwikowski and A. F. Siegel (2002).."Discovering regulatory and signalling circuits in molecular interaction networks." Bioinformatics. 18 Suppl 1: S233-S240.
Jansen, R., D. Greenbaum and M. Gerstein (2002).."Relating whole-genome expression data with protein-protein interactions." Genome Res. 12(1): 37-46.
Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt and M. Gerstein (2003).."A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science 302(5644): 449-53.
Kanehisa, M., S. Goto, S. Kawashima and A. Nakaya (2002).."The KEGG databases at GenomeNet." Nucleic Acids Res. 30(1): 42-46.
Karp, P. D. (2001).."Pathway databases: a case study in computational symbolic theories." Science 293(5537): 2040-2044.
Karp, P. D., M. Riley, S. M. Paley and A. Pellegrini-Toole (2002).."The MetaCyc Database." Nucleic Acids Res. 30(1): 59-61.
Karp, P. D., M. Riley, M. Saier, I. T. Paulsen, J. Collado-Vides, S. M. Paley, A. Pellegrini-Toole, C. Bonavides and S. Gama-Castro (2002).."The EcoCyc Database." Nucleic Acids Res. 30(1): 56-58.
Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates and D. Eisenberg (1999).."Detecting protein function and protein-protein interactions from genome sequences." Science 285(5428): 751-753.
Marcotte, E. M., I. Xenarios, A. M. van Der Bliek and D. Eisenberg (2000).."Localizing proteins in the cell from their phylogenetic profiles." Proc Natl Acad Sci U S A 97(22): 12115-20.
Matthews, L. R., P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent and M. Vidal (2001).."Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or."interolog."." Genome Res 11(12): 2120-6.
Mehlhorn, K. and S. Naher (1999). Leda : a platform for combinatorial and geometric computing. New York, Cambridge University Press.
Mellor, J. C., I. Yanai, K. H. Clodfelter, J. Mintseris and C. DeLisi (2002).."Predictome: a database of putative functional links between proteins." Nucleic Acids Res 30(1): 306-9.
Ofran, Y. and B. Rost (2003).."Analysing six types of protein-protein interfaces." J Mol Biol 325(2): 377-87.
Ouzounis, C. and N. Kyrpides (1996).."The emergence of major cellular processes in evolution." FEBS Lett 390(2): 119-23.
Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch and N. Maltsev (1999).."The use of gene clusters to infer functional coupling." Proc Natl Acad Sci U S A 96(6): 2896-901.
Overbeek, R., N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov, Jr, N. Kyrpides, M. Fonstein, N. Maltsev and E. Selkov (2000).."WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction." Nucleic Acids Res 28(1): 123-125.
Pazos, F. and A. Valencia (2001).."Similarity of phylogenetic trees as indicator of protein-protein interaction." Protein Eng 14(9): 609-14.
Pazos, F. and A. Valencia (2002).."In silico two-hybrid system for the selection of physically interacting protein

pairs." Proteins 47(2): 219-27.
Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates (1999).."Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc Natl Acad Sci U S A 96(8): 4285-8.
Peri, S., J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. K. Gandhi, M. Gronborg, et al. (2003).."Development of Human Protein Reference Database as an initial platform for approaching systems biology in humans." Genome Res 13(10): 2363-71.
Phizicky, E. M. and S. Fields (1995).."Protein-protein interactions: methods for detection and analysis." Microbiol Rev 59(1): 94-123.
Renesto, P., N. Crapoulet, H. Ogata, B. La Scola, G. Vestris, J. M. Claverie and D. Raoult (2003).."Genome-based design of a cell-free culture medium for Tropheryma whipplei." Lancet 362(9382): 447-9.
Rhee, S. Y., W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, et al. (2003).."The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community." Nucleic Acids Res 31(1): 224-8.
Rivera, M. C., R. Jain, J. E. Moore and J. A. Lake (1998).."Genomic evidence for two functionally distinct gene classes." Proc Natl Acad Sci U S A 95(11): 6239-44.
Robinson, M. D., J. Grigull, N. Mohammad and T. R. Hughes (2002).."FunSpec: a web-based cluster interpreter for yeast." BMC.Bioinformatics. 3(1): 35.
Salama, J. J., I. Donaldson and C. W. Hogue (2002).."Automatic annotation of BIND molecular interactions from three- dimensional structures." Biopolymers 61(2): 111-120.
Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman (2003).."Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet 34(2): 166-76.
Selkov, E., R. Overbeek, Y. Kogan, L. Chu, V. Vonstein, D. Holmes, S. Silver, R. Haselkorn and M. Fonstein (2000).."Functional analysis of gapped microbial genomes: amino acid metabolism of Thiobacillus ferrooxidans." Proc Natl Acad Sci U S A 97(7): 3509-14.
Snel, B., P. Bork and M. A. Huynen (2002).."Genomes in flux: the evolution of archaeal and proteobacterial gene content." Genome Res 12(1): 17-25.
Tamames, J., G. Casari, C. Ouzounis and A. Valencia (1997).."Conserved clusters of functionally related genes in two bacterial genomes." J Mol Evol 44(1): 66-73.
Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, et al. (2003).."The COG database: an updated version includes eukaryotes." BMC Bioinformatics 4(1): 41.
Tien, A. C., M. H. Lin, L. J. Su, Y. R. Hong, T. S. Cheng, Y. C. Lee, W. J. Lin, I. H. Still and C. Y. Huang (2004).."Identification of the substrates and interaction proteins of aurora kinases from a protein-protein interaction model." Mol Cell Proteomics 3(1): 93-104.
Uetz, P. S., B.;Ideker, T. (2002). Visualization and Integration of Protein-Protein Interactions. Protein-Protein Interactions: A Molecular Cloning Manual. E. Golemis. Cold Spring Harbor, NY, CSHL Press.
Voet, D. and J. G. Voet (2004). Biochemistry. New York, J. Wiley & Sons.
von Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel (2003).."STRING: a database of predicted functional associations between proteins." Nucleic Acids Res. 31(1): 258-261.
Xenarios, I., L. Salwinski, X. J. Duan, P. Higney, S. M. Kim and D. Eisenberg (2002).."DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." Nucleic Acids Res. 30(1): 303-305.
Zanzoni, A., L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich and G. Cesareni (2002).."MINT: a Molecular INTeraction database." FEBS Lett. 513(1): 135-140.
Zeeberg, B. R., W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, et al. (2003).."GoMiner: a resource for biological interpretation of genomic and proteomic data." Genome Biol 4(4): R28.
Zorio, D. A., N. N. Cheng, T. Blumenthal and J. Spieth (1994).."Operons as a common form of chromosomal organization in C. elegans." Nature 372(6503): 270-2.