Module 7

Variation, Function and Disease

Bert Overduin, Ph.D.

Edinburgh Genomics
The University of Edinburgh
Edinburgh EH9 3FL
Scotland, United Kingdom

Overview

- Tools to predict variant effects
- Gene expression databases
- “Disease” / phenotype databases
- Ontologies

Variant Effects

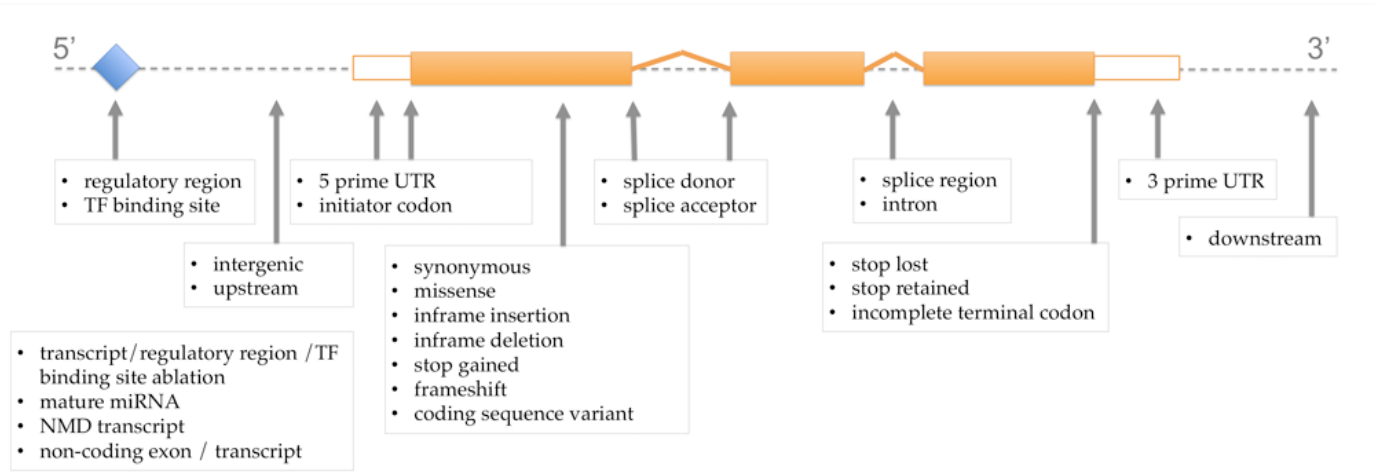
Tools to categorise and prioritise newly discovered variants:

- Variant Effect Predictor (Ensembl)
- Variation Annotation Integrator (UCSC)
- PolyPhen-2
- SIFT

The Open Door Workshop

Variant Effect Predictor (VEP)

- Predicts functional consequences of known and unknown variants
- For substitutions, insertions, deletions and structural variants



VEP Output

- Affected genes / transcripts / regulatory features / motifs
- Gene symbols
- IDs from Ensembl, CCDS, UniProt, HGVS
- Consequence (e.g. missense, stop gained, stop lost)
- Location of variant
- Co-located known variant(s)
- Minor allele frequencies from the 1000 Genomes Project
- PolyPhen and SIFT prediction and score

etc. etc.

Polyphen-2 and SIFT

- Predict the effect of missense variants

PolyPhen-2 (Polymorphism Phenotyping)

- Uses physical and comparative considerations
- Scale from 0 (benign), via possibly damaging to 1 (probably damaging)

SIFT (Sorting Intolerant From Tolerant)

- Uses the degree of conservation of amino acid residues
- Scale from 0 (deleterious) to 1 (tolerated)

Warning

- All these tools do is make predictions!
- Findings should always be confirmed by doing experiments!

Gene Expression Databases

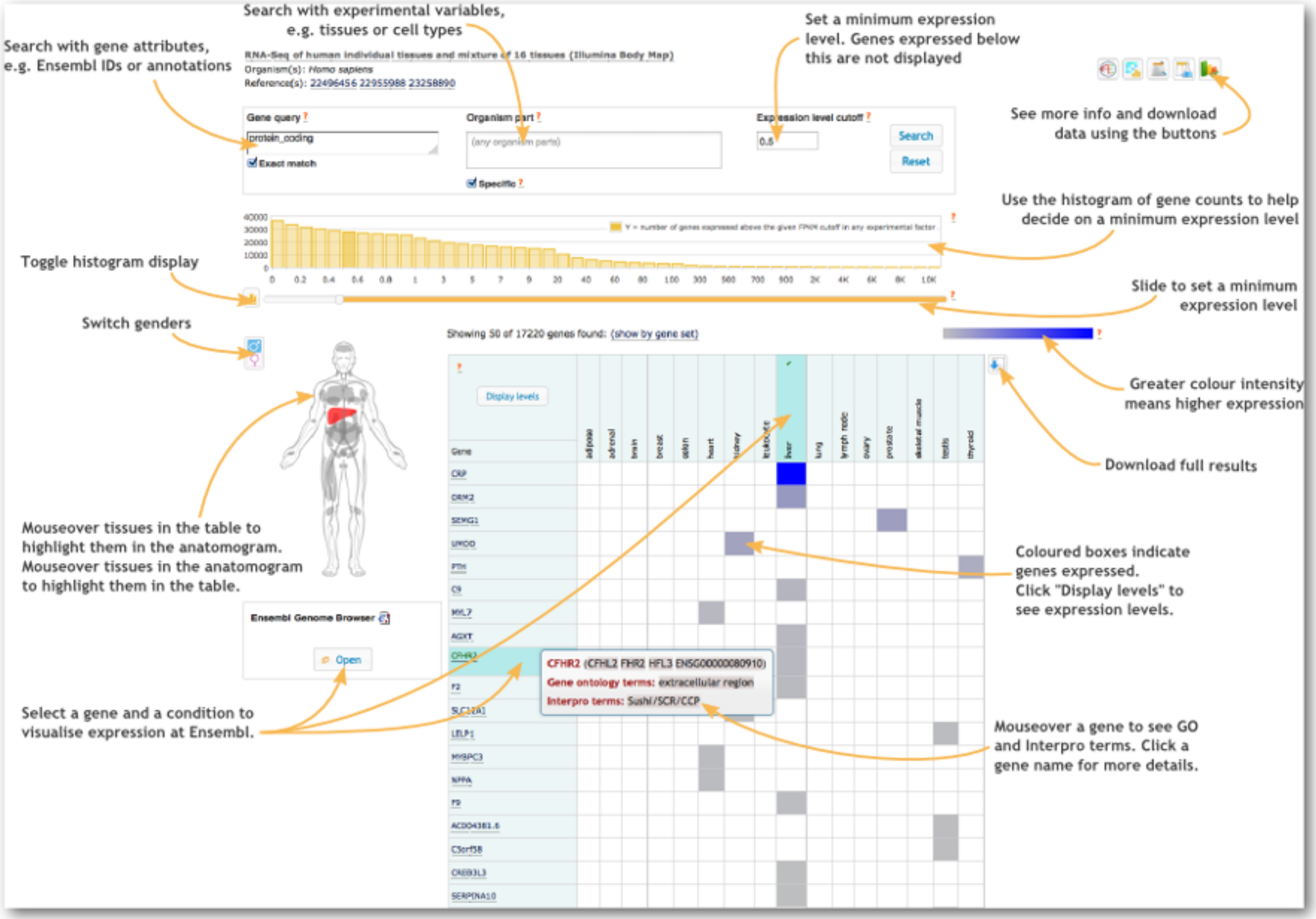
GEO Profiles (NCBI)

- Gene expression profiles
- Derived from GEO (Gene Expression Omnibus)

Expression Atlas (EBI)

- Baseline Atlas: shows which gene products are present (and at what abundance) in "normal" conditions
- Differential Atlas: shows genes that are up- or down-regulated in a wide variety of different experimental conditions
- Derived from ArrayExpress

Baseline Atlas



Search with gene attributes, e.g. Ensembl IDs or annotations

Search with experimental variables, e.g. tissues or cell types

Set a minimum expression level. Genes expressed below this are not displayed

See more info and download data using the buttons

RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)
Organism(s): *Homo sapiens*
Reference(s): 22496456 22955988 23258890

Gene query: protein_coding
☑ Exact match

Organism part: (any organism parts)
☑ Specific

Expression level cutoff: 0.5

Use the histogram of gene counts to help decide on a minimum expression level

Toggle histogram display

Slide to set a minimum expression level

Switch genders

Showing 50 of 17220 genes found: (show by gene set)

Gene	adipose	adrenal	brain	breast	colon	heart	kidney	leukocyte	liver	lung	lymph node	ovary	prostate	skeletal muscle	testis	thyroid
CRP									■							
ORM2									■							
SEMG1													■			
UMOD							■									
PTH																■
C9									■							
MYL7						■										
AGXT									■							
CFHR2									■							
F2									■							
SLC12A1							■									
LELP1															■	
MYBPC3						■										
NPPA						■										
F9									■							
AC004381.6															■	
C3orf58															■	
CREB3L3									■							
SERPINA10									■							

CFHR2 (CFHL2 FHR2 HFL3 ENSG00000080910)
Gene ontology terms: extracellular region
Interpro terms: Sushi/SCR/CCP

Greater colour intensity means higher expression

Download full results

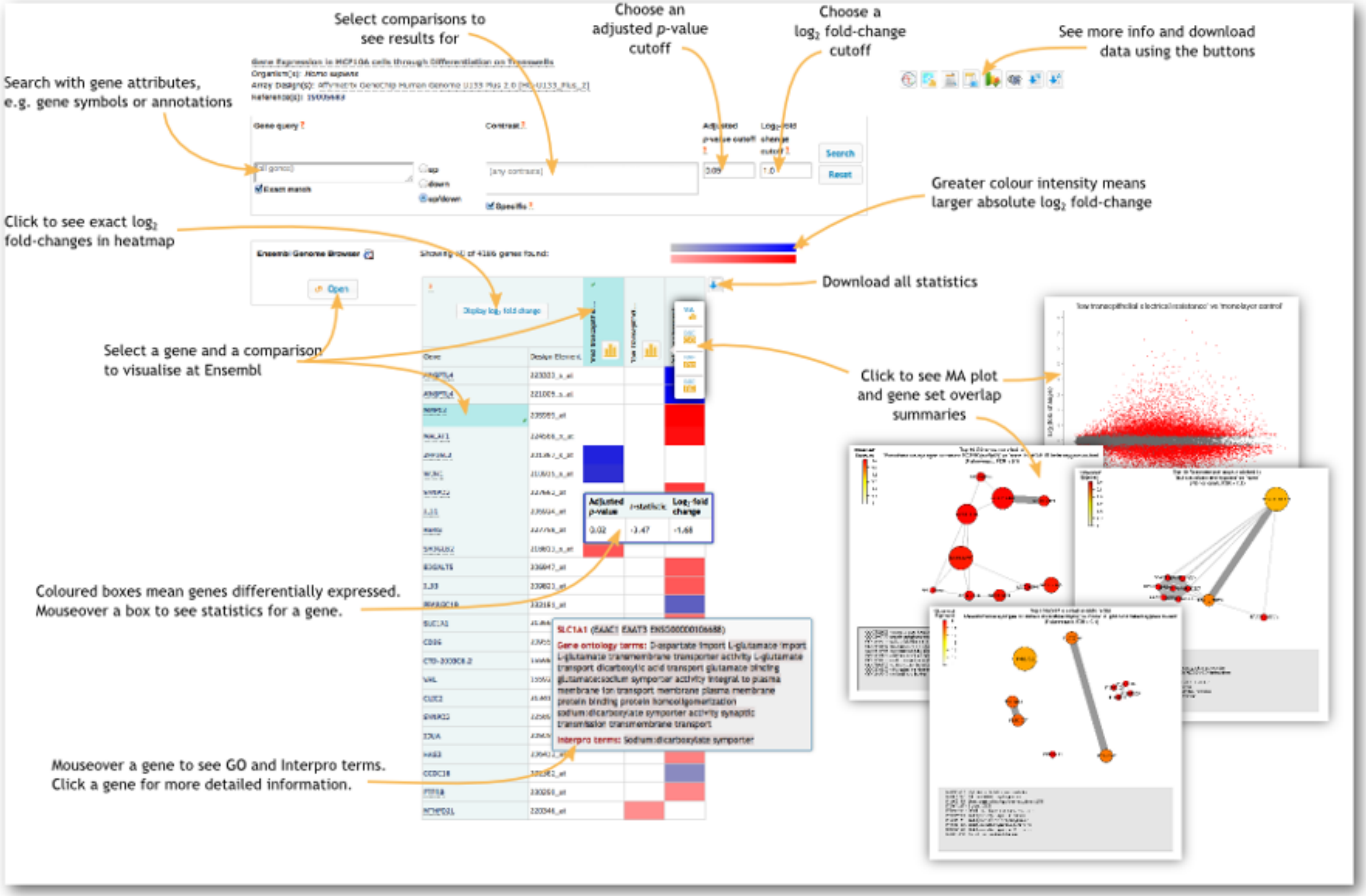
Mouseover tissues in the table to highlight them in the anatomogram. Mouseover tissues in the anatomogram to highlight them in the table.

Coloured boxes indicate genes expressed. Click "Display levels" to see expression levels.

Select a gene and a condition to visualise expression at Ensembl.

Mouseover a gene to see GO and Interpro terms. Click a gene name for more details.

Differential Atlas



Select comparisons to see results for

Choose an adjusted *p*-value cutoff

Choose a $\log_2$ fold-change cutoff

See more info and download data using the buttons

Search with gene attributes, e.g. gene symbols or annotations

Greater colour intensity means larger absolute $\log_2$ fold-change

Click to see exact $\log_2$ fold-changes in heatmap

Download all statistics

Select a gene and a comparison to visualise at Ensembl

Click to see MA plot and gene set overlap summaries

Coloured boxes mean genes differentially expressed. Mouseover a box to see statistics for a gene.

Mouseover a gene to see GO and Interpro terms. Click a gene for more detailed information.

Gene Expression in MCF10A cells through Differentiation on Transwells
Organism(s): Homo sapiens
Array Design(s): Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133_Plus_2]
Reference(s): 15000683

Gene query: [all genes]
☑ Exact match
○ up ○ down ● up/down

Contrast: [any contrast]
☑ Specific

Adjusted *p*-value cutoff: 0.05
Log_2-fold change cutoff: 1.0

Search
Reset

Ensembl Genome Browser
Open

Showing 40 of 4186 genes found:

Display $\log_2$ fold change

Adjusted *p*-value	*t*-statistic	Log_2-fold change
0.02	-3.47	-1.68

SLC1A1 (EAAC1 EAAT3 ENSG00000106688)

Gene ontology terms: D-aspartate import L-glutamate import L-glutamate transmembrane transporter activity L-glutamate transport dicarboxylic acid transport glutamate binding glutamate:sodium symporter activity integral to plasma membrane ion transport membrane plasma membrane protein binding protein homooligomerization sodium:dicarboxylate symporter activity synaptic transmission transmembrane transport

Interpro terms: Sodium:dicarboxylate symporter

Gene
Design Element
'low transepithelial electrical resistance' vs 'monolayer control'

Disease / Phenotype Databases

OMIM

- Online Mendelian Inheritance in Man
- Catalog of human genes and genetic disorders

COSMIC

- Catalog Of Somatic Mutations In Cancer
- Database of somatically acquired mutations found in human cancer

DECIPHER

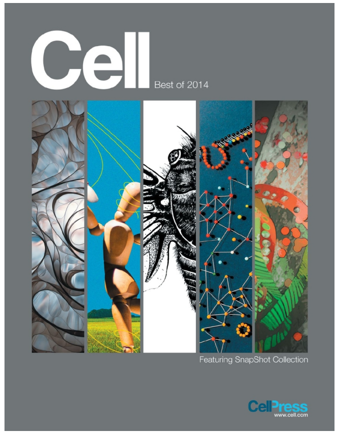
- DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources)
- Database of genomic variation data from analysis of patient DNA

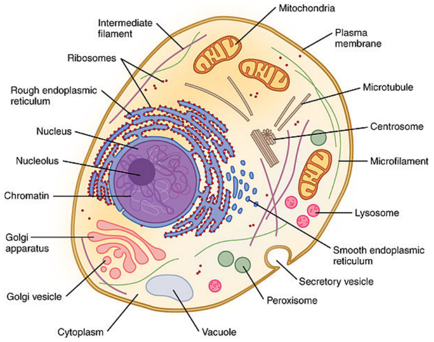
Ontologies

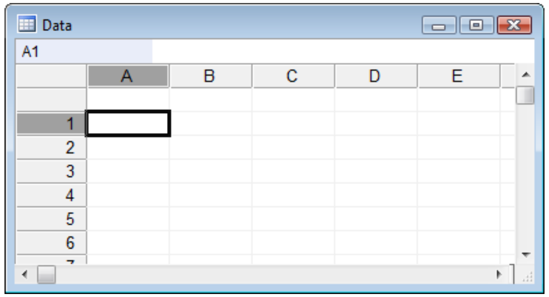
What do you think about when you hear the word “cell”?

Ontologies

What do you think about when you hear the word “cell”?







Ontologies

Gene Ontology (GO)

- Describes gene products in terms of their:
  - Associated biological processes (what?)
  - Cellular components (where?)
  - Molecular functions (how?)

Sequence Ontology (SO)

- Describes features and attributes of biological sequence

Worked examples

Worked Examples URLs

- Ensembl Variant Effect Predictor: http://www.ensembl.org/info/docs/tools/vep/index.html
- UCSC Variation Annotation Integrator: http://genome.ucsc.edu/cgi-bin/hgVai
- PolyPhen-2: http://genetics.bwh.harvard.edu/pph2
- SIFT: http://sift.jcvi.org
- OMIM: http://omim.org
- GEO Profiles: http://www.ncbi.nlm.nih.gov/geoprofiles
- Expression Atlas: http://www.ebi.ac.uk/gxa/home
- COSMIC: http://cancer.sanger.ac.uk/cosmic/
- DECIPHER: https://decipher.sanger.ac.uk/
- Gene Ontology: http://geneontology.org/
- Sequence Ontology: http://www.sequenceontology.org/