Name: _____

Genome Informatics

Wellcome Genome Campus Conference Centre, Hinxton, Cambridge, UK
17 – 20 September 2018

Scientific Programme Committee:

Alicia Oshlack
Murdoch Children's Research Institute, Australia

Aaron Quinlan
University of Utah, USA

Melissa Wilson Sayres
Arizona State University, USA

Tweet about it: #GI2018

@ACSCevents /ACSCevents /c/WellcomeGenomeCampusCoursesandConferences

Wellcome Genome Campus Scientific Conferences Team:

Rebecca Twells
Head of Advanced Courses and Scientific Conferences

Treasa Creavin
Scientific Programme Manager

Nicole Schatlowski
Scientific Programme Officer

Jemma Beard
Conference & Events Organiser

Lucy Criddle
Conference & Events Organiser

Beccy Jones
Conference & Events Assistant

Laura Hubbard
Conference & Events Manager

Sarah Offord
Conference & Events Office Administrator

Zoey Willard
Conference & Events Organiser

Dear colleague,

I would like to offer you a warm welcome to the Wellcome Genome Campus Advanced Courses and Scientific Conferences: Genome Informatics. I hope you will find the talks interesting and stimulating, and find opportunities for networking throughout the schedule.

The Wellcome Genome Campus Advanced Courses and Scientific Conferences programme is run on a not-for-profit basis, heavily subsidised by the Wellcome Trust.

We organise around 50 events a year on the latest biomedical science for research, diagnostics and therapeutic applications for human and animal health, with world-renowned scientists and clinicians involved as scientific programme committees, speakers and instructors.

We offer a range of conferences and laboratory-, IT- and discussion-based courses, which enable the dissemination of knowledge and discussion in an intimate setting. We also organise invitation-only retreats for high-level discussion on emerging science, technologies and strategic direction for select groups and policy makers. If you have any suggestions for events, please contact me at the email address below.

The Wellcome Genome Campus Scientific Conferences team are here to help this meeting run smoothly, and at least one member will be at the registration desk between sessions, so please do come and ask us if you have any queries. We also appreciate your feedback and look forward to your comments to continually improve the programme.

Best wishes,

Dr Rebecca Twells
Head of Advanced Courses and Scientific Conferences
rebecca.twells@wellcomegenomecampus.org

General Information

Conference Badges
Please wear your name badge at all times to promote networking and to assist staff in identifying you.

Scientific Session Protocol
Photography, audio or video recording of the scientific sessions, including poster session is not permitted.

Social Media Policy
To encourage the open communication of science, we would like to support the use of social media at this year's conference. Please use the conference hashtag **#GI2018**. You will be notified at the start of a talk if a speaker does not wish their talk to be open. For posters, please check with the presenter to obtain permission.

Internet Access
Wifi access instructions:

- Join the 'ConferenceGuest' network
- Enter your name and email address to register
- Click 'continue' to send an email to the registered email address
- Open the registration email, follow the link 'click here' and confirm the address is valid
- Enjoy seven days' free internet access!
- Repeat these steps on up to 5 devices to link them to your registered email address

Presentations
Please provide an electronic copy of your talk to a member of the AV team who will be based in the meeting room.

Poster Sessions
Posters will be displayed throughout the conference. Please display your poster in the Conference Centre on arrival. There will be two poster sessions during the conference.

Odd number poster assignments will be presenting in poster session 1, which takes place on Tuesday, 18 September, at 18:15 – 19:30.

Even number poster assignments will be presenting in poster session 2, which takes place on Wednesday 19 September, at 18:15 – 19:30.

The abstract page number indicates your assigned poster board number. An index of poster numbers appears in the back of this book.

Conference Meals & Social Events
Lunch and dinner will be served in the Hall, apart from on Monday 17 September when it will be served in the Conference Centre. Please refer to the conference programme in this book as times will vary based on the daily scientific presentations.

All conference meals and social events are for registered delegates.Please inform the conference organiser if you are unable to attend the conference dinner.

The Hall Bar (cash bar) will be open from 19:00 – 23:00 each day.

Please note there are no lunch or dinner facilities available outside of the conference times.

Dietary Requirements
If you have advised us of any dietary requirements, you will find a coloured dot on your badge. Please make yourself known to the catering team and they will assist you with your meal request.

If you have a gluten allergy, we are unable to guarantee the non-presence of gluten in dishes even if they are not used as a direct ingredient. This is due to gluten ingredients being used in the kitchen.

For Wellcome Genome Campus Conference Centre Guests
Check in
If you are staying on site at the Wellcome Genome Campus Conference Centre, you may check into your room from 14:00. The Conference Centre reception is open 24 hours.

Breakfast
Your breakfast will be served in the Hall restaurant from 07:30 – 09:00

Telephone
If you are staying on-site and would like to use the telephone in your room, you will need to contact the Reception desk (Ext. 5000) to have your phone line activated - they will require your credit card number and expiry date to do so.

Departures
You must vacate your room by 10:00 on the day of your departure. Please ask at reception for assistance with luggage storage in the Conference Centre.

For Holiday Inn Express & Red Lion, Whittlesford Bridge Hotel Guests
Check in
If you are staying on site at the Holiday Inn Express you may check into your room from 14:00. Hotel staff are on hand 24 hours a day.

Breakfast and Dining
Your breakfast will be served in the hotel, Great Room from 06:30 – 10:00.

The hotel also offers a relaxed licensed bar and lounge area.

Telephone and Internet
A telephone and free wireless internet access is available in your room, wireless is complimentary.

Departures
You must vacate your room by 12:00 on the day of your departure. A luggage store is available at the Conference Centre.

Wellcome Genome Campus Scientific Conferences guests receive a 15% discount on food at the Red Lion, Whittlesford Bridge Hotel.

Transfers
If you are staying off site, a complimentary shuttle service has been organised with Richmond's Coaches. The shuttle service is as follows:

Monday, 17 September
Holiday Inn Express – Wellcome Genome Campus 13:30
Wellcome Genome Campus – Holiday Inn Express 21:00

Tuesday, 18 September

Holiday Inn Express – Wellcome Genome Campus	08:30
Wellcome Genome Campus – Holiday Inn Express	21:00

Wednesday, 19 September

Holiday Inn Express – Wellcome Genome Campus	08:30
Wellcome Genome Campus – Holiday Inn Express	21:30

Thursday, 20 September

Holiday Inn Express – Wellcome Genome Campus	08:30

Taxis

Please find a list of local taxi numbers on our website. The conference centre reception will also be happy to book a taxi on your behalf.

Return Ground Transport

Complimentary return transport has been arranged for 13:30 on Thursday 20 September to Cambridge station and city centre (Downing Street), and Stansted and Heathrow airports.

A sign-up sheet will be available at the conference registration desk from 16:20 on Monday, 17 September. Places are limited so you are advised to book early.

Please allow a 30 minute journey time to both Cambridge and Stansted Airport, and two hours to Heathrow.

Messages and Miscellaneous

Lockers are located outside the conference centre toilets and are free of charge.

All messages will be posted on the registration desk in the Conference Centre.

A number of toiletry and stationery items are available for purchase at the conference centre reception. Cards for our self-service laundry are also available.

Certificate of Attendance

A certificate of attendance can be provided. Please request one from the conference organiser based at the registration desk.

Contact numbers

Wellcome Genome Campus Conference Centre – 01223 495000 (or Ext. 5000)
Wellcome Genome Campus Conference Organiser (Laura) – 07733 338878

If you have any queries or comments, please do not hesitate to contact a member of staff who will be pleased to help you.

Conference Summary

Monday, 17 September

13:00 – 14:50	Registration with lunch
14:50 – 15:00	Welcome and Introductions
15:00 – 16:20	Session 1: Data Curation, Integration, and Visualization
16:20 – 17:00	Afternoon Tea
17:00 – 18:20	Session 1 continued
18:20 – 19:00	Drinks reception
19:00	Dinner

Tuesday, 18 September

09:00 – 09:55	Keynote Lecture *by Katie Pollard*
09:55 – 10:00	Comfort break
10:00 – 11:20	Session 2: Personal and Medical Genomics
11:20 - 11:50	Morning Coffee
11:50 – 13:10	Session 2 continued
13:10 – 14:30	Lunch
14:30 – 15:50	Session 3: Comparative, Evolutionary, Metagenomics
15:50 – 16:30	Afternoon Tea
16:30 – 17:50	Session 3 continued
17:50 – 18:15	Lightning talks
18:15– 19:30	Drinks reception and Poster session II (Odd numbers)
19:30	Dinner

Wednesday, 19 September

09:00 – 09:55	Keynote Lecture *by Rafael Irizarry*
09:55 – 10:00	Comfort break
10:00 – 11:20	Session 4: Transcriptomics, Alternative Splicing and Gene Predictions
11:20 - 11:50	Morning Coffee
11:50 – 13:10	Session 4 continued
13:10 – 14:30	Lunch
14:30 – 15:50	Session 5: Epigenetics and non-coding genome
15:50 – 16:30	Afternoon Tea
16:30 – 17:50	Session 5 continued
17:50 – 18:15	Lightning talks
18:15– 19:30	Drinks reception and Poster session II (Even numbers)
19:30	Conference Dinner, Silver Service

Thursday, 20 September

09:00 – 10:20	Session 6: Variant Discovery and Genome Assembly
10:20 - 11:00	Morning Coffee
11:00 – 12:20	Session 6 continued
12:20 – 12:25	Closing remarks
12:20 – 13:30	Lunch
13:30	Close of conference, coaches depart to Cambridge, Stansted and Heathrow airports

Conference Sponsors

We would like to acknowledge the generous support from the following organisations:

Genome Informatics

Wellcome Genome Camps Conference Centre,
Hinxton, Cambridge

17 – 20 September 2018

Lectures to be held in the Francis Crick Auditorium
Lunch and dinner to be held in the Hall Restaurant
Poster sessions to be held in the Conference Centre

Spoken presentations - If you are an invited speaker, or your abstract has been selected for a spoken presentation, please give an electronic version of your talk to the AV technician.

Poster presentations – If your abstract has been selected for a poster, please display this in the Conference Centre on arrival.

Conference programme

Monday, 17 September

13:00 – 14:50 **Registration with lunch**

14:50 – 15:00 **Welcome and Introductions**
Aaron Quinlan, University of Utah USA

15:00 – 16:20 **Session 1: Data Curation, Integration, and Visualization**
Session chairs: Casey Green & Sarah Teichmann

15:00 Immunogenomics one cell at a time
Sarah Teichmann
Wellcome Sanger Institute, UK

15:20 Flexible and interactive visualization of GFA sequence graphs
Giorgio Gonnella
University of Hamburg, Germany

15:40 Using clustering trees to visualise single-cell RNA-sequencing data
Luke Zappia
Murdoch Children's Research Institute, Australia

16:00 Expression Atlas: exploring gene expression at tissue and single cell level across species and biological conditions
Laura Huerta
EMBL-EBI, UK

16:20 – 17:00 **Afternoon Tea**

17:00 – 18:20 **Session 1 continued**

17:00 Interpreting high-throughput genomic analyses with the het.io knowledgebase
Casey Green
University of Pennsylvania, USA

17:20 Pathogens in Ensembl: Enabling the march against biotic threat
Nishadi De Silva
EMBL-EBI, UK

17:40 Butler enables rapid analysis of thousands of human genomes on the cloud.
Sergei Yakneen
EMBL, Germany

18:00 InterMine: widening integrative data analysis
Rachel Lyne
InterMine, UK

18:20 – 19:00 **Drinks reception**

19:00 **Dinner**

Tuesday, 18 September

09:00 – 09:55 **Keynote Lecture**
Session chair: Melissa Wilson Sayers, Arizona State University, USA

A population genetic view of human chromatin organization
Katie Pollard
Gladstone Institute of Data Science & Biotechnology / UCSF, USA

09:55 – 10:00 Comfort break

10:00 – 11:20 **Session 2: Personal and Medical Genomics**
Session chairs: Sri Kosuri & Kaitlin Samocha

10:00 The Impact of Rare Genetic Variation on Pre-mRNA Splicing
Sri Kosuri
UCLA, USA

10:20 Integration of whole genome, whole exome, and transcriptome sequencing pipelines for comprehensive genomic profiling of 55 pediatric cancer subjects
Patrick Brennan
Nationwide Children's Hospital, USA

10:40 Non-driver somatic alterations confer good prognosis in lung cancer patients
Dennis Wang
NIHR Sheffield BRC, UK

11:00 Resolving tumor heterogeneity at single cell resolution
Shamoni Maheshwari
10x Genomics, USA

11:20 - 11:50 **Morning Coffee**

11:50 – 13:10 **Session 2 continued**

11:50 Evaluating the role of rare variation in children with developmental disorders
Kaitlin Samocha
Wellcome Sanger Institute, UK

12.10 Inference of mutational status, loss of heterozygosity, and clonality in tumor-only data
Hossein Khiabanian
Rutgers University, USA

12.30 Modelling double strand break susceptibility to interrogate structural variation in cancer
Tracy Ballinger
IGMM, UK

12.50 Retrieving Charras genomic tracts from the Uruguayan admixed population
Lucia Spangenberg
Institut Pasteur de Montevideo

13:10 – 14:30 **Lunch**

14:30 – 15:50 **Session 3: Comparative, Evolutionary, Metagenomics**
Session chairs: Mario Caccamo & Ellen Leffler

14:30 Paired sequencing of host and parasite genomes in severe malaria cases
Ellen Leffler
University of Oxford, UK

14:50 PPanGGOLiN: Depicting microbial diversity via a Partitioned Pangenome Graph
Guillaume Gautreau
Genoscope, France

15:10 The role of structural variants in the adaptive radiation of African Cichlids
Luca Penso Dolfin
Earlham Institute, UK

15:30 Coevolution of chromosome changes and gene regulation in ruminants
Marta Farre Belmonte
Royal Veterinary College

15:50 – 16:30 **Afternoon Tea**

16:30 – 17:50 **Session 3 continued**

16:30 Understanding the genetic components controlling apomixis
Mario Caccamo
NIAB, UK

16:50 Querying colored and compacted de Bruijn graphs of thousands of related genomes
Nina Luhmann
University of Warwick

17:10 Genome mining for metabolic gene clusters in yeast
Christopher Pyatt
NCYC - Quadram Institute, UK

17:30 Comparative analysis of hundreds of vertebrate genomes in Ensembl
Carla Cummins
EMBL-EBI, UK

17:50 – 18:15 **Lightning talks**

18:15– 19:30 **Drinks reception and Poster session I** (Odd numbers)

19:30 **Dinner**

Wednesday, 19 September

09:00 – 09:55 **Keynote Lecture**
Session chair: Alicia Oshlack

Understanding variability and systematic bias in highthroughput data
Rafael Irizarry
Dana-Farber Cancer Institute, USA

09:55 – 10:00 Comfort break

10:00 – 11:20 **Session 4: Transcriptomics, Alternative Splicing and Gene Predictions**
Session chairs: Barbara Englehardt & Mark Robinson

10:00 On the analysis of long-read sequencing data for gene expression
Mark Robinson
University of Zurich, Switzerland

10:20 Single-cell isoform RNA sequencing (ScISOr-Seq) across thousands of cells reveals isoforms of cerebellar cell types.
Hagen Tilgner
Weill Cornell Medicine, USA

10:40 Bootstrapping Biology: Quick and easy de novo genome assembly to enable single cell gene expression analysis
Nikka Keivanfar
10x Genomics, USA

11:00 Discrete and continuous differential expression analysis for single-cell RNA-seq data
Koen Van den Berge
Ghent University, Belgium

11:20 - 11:50 **Morning Coffee**

11:50 – 13:10 **Session 4 continued**

11:50 A generative model for single-cell RNA-sequencing
Barbara Englehardt
Princeton University, USA

12.10 TALC: Transcriptome-aware Long Read Correction
Lucile Broseus
CNRS, France

12.30 Constraint for mRNA structure in human synonymous mutations
Jeff Gaither
Nationwide Children's Hospital, USA

12.50 Differential isoform usage in Parkinson's disease
Fiona Dick
University of Bergen, Norway

13:10 – 14:30 **Lunch**

14:30 – 15:50 **Session 5: Epigenetics and non-coding genome**
Session chairs: Jordana Bell & Alexander Suh

14:30 Interpreting variation in the human methylome
Jordana Bell
King's College London, UK

14:50 Delineation and annotation of the human regulatory landscape across 400+ cell types and states
Wouter Meuleman
Altius Institute for Biomedical Sciences, USA

15:10 Tissue-specific enhancer and promoter evolution in mammals
Maša Roller
EMBL-EBI, UK

15:30 Exploratory analysis of retrotransposon activity in the octopus brain
Massimiliano Volpe
Stazione Zoologica Anton Dohrn, Italy

15:50 – 16:30 **Afternoon Tea**

16:30 – 17:50 **Session 5 continued**

16:30 Mind the gap - interrogating the non-coding genome with single-molecule technologies
Alexander Suh
Uppsala University, Sweden

16:50 Identification of genes escaping X-inactivation and the variability of escape across cells, tissues and twin pairs
Antonino Zito
King's College London, UK

17:10 DNA methylation changes as a marker of senescing leaves in Arabidopsis thaliana.
Minerva Susana Trejo Arellano
Swedish University of Agricultural Sciences, Sweden

17:30 Recent evolution of the epigenetic regulatory landscape in human and other primates
Raquel Garcia Perez
Institute of Evolutionary Biology, Spain

17:50 – 18:15 **Lightning talks**

18:15– 19:30 **Drinks reception and Poster session II** (Even numbers)

19:30 **Conference Dinner, Silver Service**

Thursday, 20 September

09:00 – 10:20 **Session 6: Variant Discovery and Genome Assembly**
Session chairs: Jeff Kidd & Melissa Wilson Sayers

09:00 Sex differences in reference genome affect variant calling and differential expression
Melissa Wilson Sayers
Arizona State University, USA

09:20 Encoding yeast genomic diversity using variation graphs
Prithika Sritharan
Quadram Institute Bioscience, UK

09:40 Genome analysis in a polymorphic moss with large, ancient sex chromosomes
Sarah Carey
University of Florida, USA

10.00 ScaffHiC Genome Scaffolding by Modelling Distributions of Hi-C Paired-end Reads
Zemin Ning
Wellcome Sanger Institute, UK

10:20 - 11:00 **Morning Coffee**

11:00 – 12:20 **Session 6 continued**

11:00 De novo assembly and analysis of a canine genome
Jeff Kidd
University of Michigan, USA

11:20 Pandora variation inference for pangenomes from Nanopore or Illumina data
Rachel Colquhoun
University of Oxford, UK

11:40 Direct measurement of spontaneous structural variation through whole-genome sequencing of three generation human pedigrees
Jonathan Belyeu
University of Utah, USA

12:00 VarTrix is an open-source software tool for assigning variants to individual cells
Ian Fiddes
10x Genomics, USA

12:20 – 12:25 **Closing remarks by the Scientific Programme Committee**

12:20 – 13:30 **Lunch**

13:30 **Close of conference, coaches depart to Cambridge, Stansted and Heathrow airports**

These abstracts should not be cited in bibliographies. Materials contained herein should be treated as personal communication and should be cited as such only with consent of the author.

Spoken Presentations

Immunogenomics one cell at a time

Sarah Teichmann

Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK

Dissecting immune responses by next generation sequencing methods at single cell resolution provides novel insights into cell states, including antigen receptor diversity, cell-cell interactions in tissues, and how cellular phenotypes change in the context of adaptation from lymphoid to non-lymphoid tissues.

Here I will describe our recently developed computational methods for single cell trajectory and bifurcation inference of transcriptomes and antigen receptor sequences. I will also present applications of these methods to reconstruct tissue adaptation trajectories of regulatory T cells in mouse and man, and application to analysis of immune cells within barrier tissues such as lung and decidua.

Notes

Flexible and interactive visualization of GFA sequence graphs

Giorgio Gonnella, Niklas Niehus, Stefan Kurtz

Universität Hamburg, MIN-Fakultät, ZBH - Center for Bioinformatics, Bundesstraße 43, 20146 Hamburg, Germany

GFA (https://github.com/GFA-spec) is an emerging format for representing sequence graphs, including assembly de Bruijn and string graphs, variant graphs and gene splicing graphs. A growing list of software tools supporting the format is available at https://github.com/GFA-spec/GFA-spec#implementations and includes sequence assemblers, read mappers, sequence variant analysis tools, and scripting language libraries, such as GfaPy (Gonnella and Kurtz, 2017, Bioinformatics, 33(19):3094-3095) and RGFA (Gonnella and Kurtz, 2016, PeerJ, 4:e2681). Bandage (Wick et al., 2015, Bioinformatics, 31(20), 3350-3352), an interactive visualization tool for assembly graphs, supports, among other formats, GFA1. However, the latest version of the format specification (GFA2; https://github.com/GFA-spec/GFA-spec/blob/master/GFA2.md) introduced new powerful features, such as the support of generalized local alignments (i.e. not necessarily end-to-end), representation of alignments of reads to contigs, gaps, and subgraphs including any specified subset of the graph. These features allow extending the prospective use case of the format to mapping and assembly of long reads (e.g. PacBio and Nanopore), scaffolding graphs, representation of variant graphs and, through customization, possibly other yet unforeseen applications. Here, we present GfaViz, a tool for the interactive visualization of GFA sequence graphs. The tool was implemented in C++ based on the Qt framework (https://www.qt.io) and the OGDF library (http://www.ogdf.net). It is, to our knowledge, the first visualization tool supporting the GFA2 format, including all new features of the revised format. It allows the user to select among different force-based layout algorithms. The visual representation of the graph can be fully customized (e.g. colors, proportions, font of single elements or the whole graph) and exported to vector and raster image formats. The layout and customizations are saved in the GFA file itself as application-specific metainformation. Thus no external configuration files are required.

Notes

Using clustering trees to visualise single-cell RNA-sequencing data

Luke Zappia 1,2, Alicia Oshlack 1,2

1 Bioinformatics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia 2 School of Biosciences, Faculty of Science, The University of Melbourne, Melbourne, Victoria, Australia

Single-cell RNA-sequencing is commonly used to interrogate complex tissues in order to identify and compare the cell types present. This type of experiment is particularly prevalent in the developmental setting. A key step in this approach is assigning cells to different clusters that are assumed to be distinct cell types. Although this can be done by comparison with reference datasets, cells are more routinely grouped using unsupervised clustering and we have catalogued more than 60 currently available scRNA-seq clustering methods. Most clustering methods have parameters which affect the number of clusters produced, either by specifying an exact number, a parameter which controls the clustering resolution or indirectly through other parameters. The clustering resolution that is chosen can have a profound effect on further analysis but it is unclear how to make this choice. Existing clustering metrics often score only single clusters or resolutions, or require datasets to be perturbed and clustered multiple times which can be infeasible for large datasets.

Here we present clustering trees, a visualisation that shows the relationship between clusters as the clustering resolution increases. In a clustering tree each cluster is represented as a graph node with edges representing the overlap in samples between clusters at different resolutions. These trees can highlight instability that may indicate over clustering and help choose which resolution to use, particularly when combined with existing domain knowledge such as the expression of marker genes. More generally, clustering trees are a compact, information-dense visualisation that can serve as an alternative to plotting cells in reduced dimensions such as t-SNE. Importantly clustering trees display information across resolutions, in contrast to more common visualisation which only show results of a single clustering. Here we explain how clustering trees are produced using the clustree R package (http://cran.r-project.org/package=clustree) and illustrate how they can be used with an example of scRNA-seq data from kidney organoids.

Notes

Expression Atlas: exploring gene expression at tissue and single cell level across species and biological conditions

Laura Huerta, Nuno A. Fonseca, Anja Fullgrabe, Nancy George, Haider Iqbal, Monica Jianu, Jonathan Manning, Suhaib Mohammed, Pablo Moreno, Alfonso Munoz-Pomer, Irene Papatheodorou

EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

Expression Atlas (www.ebi.ac.uk/gxa) is an open science resource at EMBL-EBI that selects, curates, re-analyses and displays gene expression data in a baseline context, e.g. to find genes expressed in human brain development, and in a differential context, e.g. to find up-regulated genes in Alzheimer's disease. Experiments from ArrayExpress, GEO and SRA/ENA/DDBJ are selected for curation and analysis. Data curation involves enriching sample annotation with additional metadata, annotating metadata with Experimental Factor Ontology (EFO) terms and deciding comparisons for differential expression analysis based on associated publications and correspondence with the original researchers. Data analysis is performed using open source tools for microarray data and our standardized pipeline iRAP (github.com/nunofonseca/irap) for RNA-seq data.

Currently, we provide gene expression analysis results for more than 3300 experiments across 50 different species. Expression Atlas can be searched by gene, gene set and biological condition queries. The use of EFO annotations allows efficient search via ontology-driven query expansion and facilitates data integration across multiple experiments. We offer downstream analysis and visualization such as gene co-expression, biological variation among replicates, transcript quantification, visualization of gene expression in Ensembl genome browser and enrichment of Gene Ontology terms and Reactome pathways.

Single Cell Expression Atlas (www.ebi.ac.uk/gxa/sc) is a new component of Expression Atlas that provides information on where and under what conditions different genes are expressed at single cell level. To achieve this, we curate single cell datasets following the single cell metadata standards that we have developed, annotate metadata with EFO terms and systematically re-process each dataset using a new version of iRAP for single cell RNA-seq data. The first release of Single Cell Expression Atlas displays analysis results for more than 48,000 cells from four different species (human, mouse, zebrafish and fruit fly) and enables visualisation of clusters of cells and their annotations. Finally, Single Cell Expression Atlas supports searches for both gene and marker gene expression within and across 24 single cell datasets covering more than 200 different cell types.

Notes

Interpreting high-throughput genomic analyses with the het.io knowledgebase

Casey Greene

University of Pennsylvania, USA

Many different types of analyses lead to a list of biomedical entities and associated scores. For example, differential expression, factorization, and other methods for analysis of genome-wide datasets produce a score for each gene. We describe methods that map these findings into other concepts (pathway, disease, etc) using the het.io knowledgebase. These methods can be used to rapidly assess the biological plausibility of findings.

Notes

Pathogens in Ensembl: Enabling the march against biotic threat

Nishadi De Silva, Helder Pedro, Manuel Carbajo, Paul Kersey, Andy Yates

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

The Ensembl portals for bacteria (over 44,000 genomes), fungi (over 800 genomes) and protists (over 180 genomes) integrate genomic, transcriptomic, variation and comparative data sets; disseminated through a common suite of open-source tools, programmatic interfaces and training programmes. These portals allow for exploring microbial pathogenesis by (a) determining orthologous relationships between well-studied pathogens to those under-studied and under-funded, with the potential to establish new drug targets, (b) generating alignments between closely-related pathogenic and non-pathogenic species to help visualize unique sequences in the pathogen that point to forces that govern its evolution, (c) calculating potential effects of variants acquired by different pathogen populations that impact virulence, resistance and fitness.

We present two further aspects of the Ensembl microbial suite that are relevant to the study of pathogens. Firstly, microbial genes in Ensembl are routinely annotated with experimentally verified functions within their host(s), during an infection, through collaborations with pathogen-host interactions resources such as PHI-base. Genes with key behaviour can be searched for, visualised on the genome and projected to orthologues. This increased understanding of the multitude of ways that a pathogen can exploit or disrupt host cell function is crucial in the design of drugs and fungicides, and elucidating resistance mechanisms of the host.

Secondly, we have facilitated community curation projects for chosen pathogens. It was clear that some species had discrepant gene sets circulating among different research communities, and gaps in the annotation. Through the use of Apollo, an online gene editing tool, we have empowered researchers spread across multiple countries to collaborate and redefine the de facto gene set for their species of interest within a few months. We offered training, infrastructure deployment, quality checking and dissemination of the new annotations through Ensembl. We anticipate that the cumulative effect of improved gene sets overlaid with rich annotations and the unifying of communities in this way will accelerate the discovery of novel therapeutics, and plant and animal breeding strategies.

Notes

Butler enables rapid analysis of thousands of human genomes on the cloud.

Sergei Yakneen, Sebastian M. Waszak, Michael Gertz, Jan O.Korbel

European Molecular Biology Laboratory, Institute of Computer Science, Heidelberg University

Genomics researchers increasingly turn to cloud computing as a means of accomplishing large-scale analyses efficiently and cost-effectively. Successful operation in the cloud requires careful instrumentation and management to avoid common pitfalls, such as resource bottlenecks and low utilization that can both drive up costs and extend the timeline of a scientific project.

The Butler framework for large-scale scientific workflow management in the cloud has been developed to meet these challenges. The cornerstones of Butler design are: ability to support multiple clouds, declarative infrastructure configuration management, scalable, fault-tolerant operation, comprehensive resource monitoring, and automated error detection and recovery. Butler relies on industry-strength open-source components in order to deliver a framework that is robust and scalable to thousands of compute cores and millions of workflow executions.

Butler has been used within the context of several international large-scale genomics projects, including the International Cancer Genomics Consortium's (ICGC) Pan Cancer Analysis of Whole Genomes Project (PCAWG) and the European Open Science Cloud Pilot projects to commandeer multiple cloud computing environments with up to 3,200 CPU cores and 12 TB of RAM, and carried out over 2.5 million compute jobs. Successful analyses pursued with the aid of Butler include the discovery and joint genotyping of germline SNPs, Indels, and structural variants across the PCAWG cohort. Butler's error detection and self-healing capabilities ensured that these analyses were carried out with minimal human intervention.

Butler has been proven to add significant value to large-scale cancer genomics analyses in the cloud, delivering results within a matter of weeks that have previously taken months under similar circumstances. The flexible design of this framework allows easy adoption within other fields of Life Sciences and ensures that it will scale together with the demand for scientific analysis in the cloud for years to come.

Notes

InterMine: widening integrative data analysis

Rachel Lyne1,2, Daniela Butano1,2, Justin Clark-Casey1,2, Sergio Contrino1,2, Julie Sullivan1,2, Yo Yehudi1,2 and Gos Micklem1,2

1Cambridge Systems Biology Centre, University of Cambridge, Cambridge, United Kingdom. 2Department of Genetics, University of Cambridge, Cambridge, United Kingdom.

InterMine is an open source data warehouse built specifically for the integration and analysis of complex biological data. Integrative analysis is a powerful approach that allows many sources of evidence to be analysed together. Complementary data from model organisms is often useful as part of this process.

There is a broad selection of InterMine databases worldwide, covering many organisms, including HumanMine, PhytoMine (over 87 plant genomes), the Legume federation InterMines (Chickpea, Soy, Legume, Peanut, Bean), MedicMine (Medicago), ThaleMine (Arabidopsis), budding yeast, rat, zebrafish, mouse, fly, nematode, flatworm, xenopus and also drug targets.

The InterMine framework includes a user-friendly web interface as well as a powerful web service API, with multiple language bindings including Python and R. An advanced query builder supplements keyword search, and results can be interactively explored and refined. The interface is designed to allow flexible and iterative querying in which items collected as results in one step are used as input in the next. A set of graphical analysis tools provide a rich environment for data exploration including statistical enrichment of sets, and visualisations, such as expression graphs and interaction networks. Apps for both iOS and Android allow InterMine access "on the go".

Recent work includes a major rewrite of the user interface to exploit the latest technologies, providing an improved user experience and enabling better integration with third party tools.

InterMine is now starting a new project in collaboration with the Sansone Group in Oxford to bring InterMine functionality more readily to bench scientists, empowering them to conform to the FAIR data principles. New features will include tools to help transform research data files into cloud-hosted InterMine databases, enhancing ISAtools for metadata capture, persistent stable URIs, RDF export, together with tools to support data deposition, dissemination and publication.

Notes

A population genetic view of human chromatin organization

Katherine S. Pollard

Gladstone Institute of Data Science & Biotechnology, 1650 Owens Street, San Francisco, CA, USA

Genetic variation affects cellular and organismal biology through modifying genes and regulatory elements. But the degree to which it influences or is influenced by chromatin structure is less well understood. I will discuss several projects that bring together large-scale human genome sequencing, chromatin capture (Hi-C), and functional genomics (ChIP-seq) data to explore the interplay of genome evolution and chromatin organization. We show that topological domain (TAD) boundaries are under strong negative selection across primates and in healthy people–but not in patients with autism, developmental delay, and cancer–suggesting a broad role for “enhancer hijacking” in human disease. Secondly, we explore concordance between genetic architecture (linkage disequilibrium (LD)) and chromatin organization in 22 cell types and find no correlation between LD maps and chromatin interaction maps at any distance scale above the 5 kilobase resolution of the Hi-C data. Finally, we demonstrate that many transcription factors bind genome sites containing specific local DNA structures that do not match out current concept of sequence motifs, which opens the door to an expanded view of how mutations can alter regulatory elements. Together these studies illustrate how biophysics and genetics can be brought together through computational modeling to shed new light on genome function.

Notes

The Impact of Rare Genetic Variation on Pre-mRNA Splicing

Sri Kosuri

Mutations that cause an exon to be skipped can have severe consequences on gene function and cause disease. Here we explore the extent to which human genetic variation results in loss of exon recognition by developing a Multiplexed Functional Assay of Splicing using Sort-seq (MFASS). We assayed variants in the Exome Aggregation Consortium (ExAC) within or adjacent to thousands of human exons. We found that an unexpectedly large fraction of variants lead to almost complete loss of exon recognition. We also found that most of these exon-disrupting variants are located outside of canonical splice sites, are distributed evenly across distinct exonic and intronic regions, and are difficult to predict a priori.

Notes

Integration of whole genome, whole exome, and transcriptome sequencing pipelines for comprehensive genomic profiling of 55 pediatric cancer subjects

Patrick Brennan, Ben Kelly1, Greg Wheeler1, James Fitch1, Kyle Voytovich1, Anna Spencer1, Elizabeth Varga1, Kristen Leraas1, Tara Lichtenberg1, Vincent Magrini1,3, Dan Koboldt1,3, Julie Gastier-Foster1,2,3, Richard Wilson1,3, Elaine Mardis1,3, Catherine Cottrell1,2,3, Peter White1,3

1 The Institute for Genomic Medicine, Nationwide Children's Hospital
2 Department of Pathology, Nationwide Children's Hospital
3 Department of Pediatrics, The Ohio State University

Pediatric Cancers are the leading cause of death by disease past infancy for children in the United States. We performed extensive genomic characterization of 55 pediatric cancer patients, requiring us to develop novel computational approaches to integrate and interpret these data. Medically meaningful research results were clinically verified and returned to the treating clinician to inform diagnosis, prognosis, and potential efficacy of targeted therapeutics.

Each subject underwent comprehensive genomic profiling of both tumor and matched normal samples, using a combination of WES, WGS and RNA-Seq. High depth WES identified germline and somatic variants, and copy number variation (CNV), while WGS was utilized to additionally identify structural variants (SV), non-coding variants, and CNV. RNA-Seq data were used for variant calling, expression profiling, and fusion detection. Our fusion detection pipeline uses an ensemble of seven fusion detection tools and ranks fusions based on the consensus. These fusion calls are fed into a database to track frequencies across all subjects and cancer types.

During this study, we utilized a WES capture reagent (IDT xGen Lockdown) spiked with additional CNV specific probes distributed across the genome along with enhancement in clinically relevant cancer loci. To integrate our transcriptome and WES pipelines, the RNA-Seq variants were merged with our somatic results to confirm variant expression. A curated list of cancer genes was integrated with our CNV pipeline to visualize cancer genes that are in regions of amplification or loss.

At present, of the 55 pediatric cancer patients, five patients with pilocytic astrocytomas have been identified with a BRAF fusion; all of which have been clinically confirmed. Fusions with diagnostic, prognostic or therapeutic utility have been found in ~33% of patients. Germline variants causing cancer predisposition have been found in ~22% of patients, and approximately half of our patients have harbored somatic SNV of prognostic or therapeutic relevance. Overall our comprehensive genomic profiling approach yielded clinically relevant results for ~75% of our patients, impacting patient diagnosis, prognosis and eligibility for targeted therapeutics and clinical trials.

Notes

Non-driver somatic alterations confer good prognosis in lung cancer patients

Dennis Wang 1, Nhu An Pham 2, Timothy Freeman 1, Frances A. Shepherd 2, Ming-Sound Tsao 2

1 NIHR Sheffield Biomedical Research Centre, University of Sheffield, Sheffield, UK; 2 Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada

Genomic profiling of patient tumors has linked somatic driver mutations to survival outcomes of non-small cell lung cancer (NSCLC) patients, especially for those receiving targeted therapy. However, it remains unclear whether non-cancer associated mutations have any utility as prognostic markers. From NSCLC xenograft genomes, we used penalised regression to identify 865 genes for which high burden of somatic copy number alterations and point mutations are associated with longer disease-free survival (HR=0.153, P=1.48x10-4) in patients. Only 5% of somatic alterations in these genes have been previously implicated in cancer, and by integrating with gene expression, we were able to validate their prognostic value in three independent patient datasets. Patients with high alteration burden could be further stratified based on the presence of immunogenic mutations, revealing another subgroup of patients with even better prognosis (85% with >5 years survival). We show that the presence of immunogenic mutations in these passenger mutations were more prognostic and associated with CD8 expression than using total mutation burden. 95% of these 865 genes lack documented activity relevant to cancer, but are in pathways regulating cell proliferation, motility and immune response were implicated. Our results demonstrate that non-driver somatic alterations may influence the outcome of cancer patients by increasing beneficial immune response and inhibiting processes associated to tumorigenesis.

Notes

Resolving tumor heterogeneity at single cell resolution

Shamoni Maheshwari, Kamila Belhocine, Rajiv Bharadwaj, Claudia Catalanotti, Ian Fiddes, Lance Hepler, Vijay Kumar, Andrew Price, Joe Shuga, Deanna M Church and Sarah Garcia.

10x Genomics, Pleasanton, California, USA

"Genome sequencing" of tumor samples is a complex task because it cannot reasonably be assumed that all cells within the tumor share the same underlying genome sequence. Somatic variation, drift, and selection rapidly create heterogeneity within the tumor. This intra-tumor genetic heterogeneity is a key driver of tumor evolution and presents a formidable challenge to cancer therapeutics. Single cell DNA (scDNA) sequencing is emerging as a powerful tool to quantify genetic heterogeneity and reconstruct the evolutionary history of cancer. However, currently available workflows for single cell genomics suffer from low throughput, require cumbersome equipment, and remain only accessible to a few select expert laboratories.

We have developed a microfluidic, partitioning-based solution that allows sensitive copy number profiling of thousands of single cell genomes concomitantly. Accompanying this platform, we have also developed a scalable bioinformatics pipeline for genome alignment, normalization and integer-scaled copy-number detection. We present an evaluation of key performance metrics such as coverage, uniformity, and reproducibility on a diploid human cell line.

Furthermore, we demonstrate the power of this technology by generating sparse WGS of 10,000+ nuclei from five sections of a frozen breast tumor. Histopathology analyses reported 75% tumor purity. However, when we used scDNA to quantify the mixture of tumor-normal cells, we observe a gradient of tumor purity with diploid cells ranging from 91 to 18% of the total biopsy. We characterized the CNV mutational spectra of the non-diploid cancer cells, identifying 122 polymorphic CNV events that ranged in size from 3-148 Mbp and used these events to cluster cancer cells into 7 genetically distinct subclonal populations. Finally, we find that the tumor mass is composed of spatially separated expansions of divergent sub-clones, evidence that supports a branched polyclonal model of tumor evolution. In conclusion, we show that our novel droplet-based system enables cancer genomics at single cell resolution and provides insight into the evolutionary history of tumor progression.

Notes

Evaluating the role of rare variation in children with developmental disorders

Kaitlin E. Samocha, Joanna Kaplanis, Giuseppe Gallone, Matthew E. Hurles on behalf of the Deciphering Developmental Disorders study

Wellcome Sanger Institute

Over recent years, many novel developmental disorders have been discovered, driven primarily by large-scale sequencing projects of affected children and their parents. In the Deciphering Developmental Disorders (DDD) study, we exome sequenced over 13,500 children and their parents, when available, from across the UK and have determined a genetic diagnosis in ~35% of these children, with the vast majority of diagnoses being explained by a de novo variant.

In order identify more developmental disorder associated genes, we have taken a multi-pronged approach. Firstly, we evaluated de novo variation in the ~10,000 parent-child trios in the DDD study. We implemented a novel simulation-based method to evaluate significant excesses of de novo variants in individual genes, which provides increased power to detect associated genes. Using this novel method, which scores all classes of variants (e.g. protein-truncating, missense) on a unified severity scale, we identified dozens of novel developmental disorder genes.

Our second approach to understand the potential genetic contributions in the undiagnosed cases in the DDD study has been to evaluate the burden of inherited variation. We jointly called the exome data from DDD with ancestry-matched controls from the INTERVAL and UK10K cohorts (total n = 37,898). We find that the DDD cases have significantly more rare, inherited protein-truncating variants than controls, specifically in genes previously associated with developmental disorders ($p < 10\text{-}5$) and constrained genes (those intolerant of protein-truncating variation; $p < 10\text{-}9$). However, we find no evidence of an increase in biparental inheritance. Our current analyses point to a complex role for such rare, inherited variation, which may be working via oligogenic mechanisms.

Notes

Inference of mutational status, loss of heterozygosity, and clonality in tumor-only data

Hossein Khiabanian, Gregory Riedlinger, Mohammad Hadigol, Kim Hirshfield, Lorna Rodriguez, Shridar Ganesan

Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ, USA

Recent advances in clinical sequencing technologies have resulted in unprecedented access to the genomes of individual tumors. High-depth, hybrid-capture-based assays aim to identify somatic mutations in cancer cells for accurate diagnosis and treatment. However, most clinical-grade implementations lack patient-matched germline DNA and additional analysis is needed to infer variants' mutational status. For instance, when a potentially pathogenic mutation in a tumor suppressor gene is detected, it is imperative to determine whether it is germline or somatic and resolve any evidence for the loss of the wild-type copy, which may be pertinent for treatment efficacy. Genomic heterogeneity in both tumor and non-tumor cell populations also confounds distinguishing sub-clonal tumor alterations from those possibly originating from the non-tumor component in the tumor microenvironment. Therefore, as the use of tumor-only sequencing assays increases, systematic interpretation of clinical-grade data is necessary to accurately describe the genomics of a single tumor. Here, we present LOHGIC (LOH-Germline Inference Calculator), developed on a model-selection scheme using Akaike Information Criterion weighting. LOHGIC finds the most consistent model for mutations' germline-versus-somatic status, and infers loss of heterozygosity (LOH), mutated allele's copy-number, and cancer cell fraction, incorporating inherent biases in clinical sequencing and sample purity estimation. We used LOHGIC to assess BRCA1/2 alterations in 1,636 solid tumors and evaluated its results with genomic testing data, demonstrating 93% accuracy, 100% precision, and 96% recall. This analysis highlighted a differential tumor spectrum associated with BRCA1/2 mutations under LOH: germline BRCA1 mutations were exclusively found in women and were breast cancer or of mullerian origin, whereas germline BRCA2 mutations led to a wider spectrum of cancers in both sexes. Moreover, inferring clonality raised the hypothesis that certain sub-clonal mutations did not reflect alterations arising in cancer cells, but might be present in hematopoietic cells infiltrating the tumor microenvironment. Indeed, sequencing peripheral blood and macrodissected lymphocytes showed that 79% of sub-clonal TET2 and DNMT3A mutations found in the original specimens were not tumor mutations, but instead were detected due to the presence of clonal hematopoiesis of indeterminate potential (CHIP). Analyzing 113,079 solid tumors from 21 cancer types further indicated that many sub-clonal alterations in CHIP-associated genes detected in solid tumor specimens possibly arose from their presence in admixed hematopoietic elements. Our results demonstrate that inference of mutational signatures and dissection of heterogeneity can generate diagnostic hypotheses that when clinically tested, lead to improved prognostication, ensuring that treatment strategies are correctly focused on tumor alterations.

Notes

Modelling double strand break susceptibility to interrogate structural variation in cancer

Tracy J Ballinger, Britta Bouwman, Nicola Crosetto, Colin A. Semple

MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Science for Life Laboratory, Karolinska Intitutet

Structural variants (SVs) are known to play important roles in a variety of cancers, but their origins and functional consequences are still poorly understood. Many SVs are thought to emerge via errors in the repair processes following DNA double strand breaks (DSBs) and previous studies have experimentally measured DSB frequencies across the genome in cell lines. Using these data we derive the first quantitative genome-wide models of DSB susceptibility, based upon underlying chromatin and sequence features. These models are accurate and provide novel insights into the mutational mechanisms generating DSBs. Models trained in one cell type can be successfully applied to others, but a substantial proportion of DSBs appear to reflect cell type specific processes. Using model predictions as a proxy for susceptibility to DSBs in tumours, many SV enriched regions appear to be poorly explained by selectively neutral mutational bias alone. A substantial number of these regions show unexpectedly high SV breakpoint frequencies given their predicted susceptibility to mutation, and are therefore credible targets of positive selection in tumours. These putatively positively selected SV hotspots are enriched for genes previously shown to be oncogenic. In contrast, several hundred regions across the genome show unexpectedly low levels of SVs, given their relatively high susceptibility to mutation. These novel 'coldspot' regions appear to be subject to purifying selection in tumours and are enriched for active promoters and enhancers. We conclude that models of DSB susceptibility offer a rigorous approach to the inference of SVs putatively subject to selection in tumours.

Notes

Retrieving “Charrúas” genomic tracts from the Uruguayan admixed population

Lucia Spangenberg, Maria Ines Fariello, Monica Sans, Hugo Naya

Institut Pasteur de Montevideo, Facultad de Ingeniería (UDELAR), Facultad de Humanidades (UDELAR), Montevideo, Uruguay

During almost two centuries, the native population of Uruguay was not mentioned, except in few historical revisions. Those revisions referred to the two main Indian groups, as the macro-ethnic Charrúa entity (including Guenoas and other minor groups), and the Guaraníes (with amazonic affinities). In 1831 in the massacre of "Salsipuedes" the military forces of the government of the country killed most of the natives, surviving only a small fraction that were able to escape. The implications were huge ethically, socially and also genomically. Regarding the social aspect, native contribution is overwhelmingly underestimated: in the population census of 1852 natives were not even mentioned (categories were only "Whites", "Mulattoes", "Blacks", "foreigners"); On the census of 2011 only 4.9% Uruguayans admit to have at least one native ancestor even when recent studies on mitochondrial genomes revealed that native contribution might be as high as 31-37%, and when inferred to the nuclear genome it could be 10 to 14% (Hidalgo et al 2005; Sans 2009; Bonilla et al. 2015). Regarding the genomic consequences, Uruguayan native genomes are almost lost. Ancient DNA studies could shed some light in this aspect, but until today only one complete ancient American genome was sequenced (Anzick-1, Rasmussen et al 2014). In Uruguay, only few prehistoric mitochondrial genomes have been sequenced (Sans, 2015). Then, the possibility of identifying native segments of DNA from present population could be a useful proxy to reconstruct native genomes and, when possible, to identify different native ethnic origins. In the present work we retrieve for the first time, from whole genome sequencing of 10 individuals with known (self-declared) Uruguayan native ancestry, native genomic DNA fragments to get some insight into the long lost "Charrúa" genome. Genome wide ancestry proportions, differences in self-declared and genomic estimation, homogeneity of the native Uruguayan population and similarity to other native tribes in the region were determined. A "proto-Charrúa" genome was constructed in order to improve comparisons to other populations and to determine frequent and variable haplotypes.

Notes

Paired sequencing of host and parasite genomes in severe malaria cases

Ellen Leffler, Gavin Band, Jim Stalker, Thuy Nguyen, Kirk Rockett, and Dominic Kwiatkowski

Genome-wide association studies in humans have uncovered a handful of loci robustly associated with severe P. falciparum malaria, but have so far not considered the relevance of parasite genetic variation. Several stages of infection involve molecular interactions between parasite and host proteins, raising the question of whether specific genetic combinations might influence the course of disease. To begin to look for such effects, we generated high-coverage DNA sequencing data (Illumina HiSeq X Ten) for 1095 blood samples from highly parasitaemised children with severe symptoms of malaria from The Gambia and Kenya, and aligned it to a joint human-P. falciparum reference. To access parasite genomes from 3423 additional malaria cases with lower parasitaemia, we used selective whole genome amplification with primers enriched for binding sites in the P.falciparum genome relative to the human genome. Both approaches yielded substantial coverage of the parasite genome (mean 197x and 143x coverage, respectively) and we used GATK HaplotypeCaller to obtain variant calls. Consistent with previous reports, we found little evidence for population structure within parasite populations, and also observed no overall correlation between host and parasite genomes. Finally, we report an initial scan for interactions between parasite genotypes at merozoite genes and host genotypes at invasion receptor genes.

Notes

PPanGGOLiN: Depicting microbial diversity via a Partitioned Pangenome Graph

Guillaume Gautreau1, Christophe Ambroise2, Catherine Matias3, Amandine Perrin 4, Valentin Sabatet1, Rémi Planel1, Marie Touchon4, Claudine Médigue1, Eduardo Rocha4, Stéphane Cruveiller1 and David Vallenet1

1 : Laboratoire d'Analyse Bioinformatique en Génomique et Métabolisme (LABGeM) - CEA, Genoscope : DRF/IBFB/Gen, CNRS : UMR8030, Université d'Evry - Université Paris-Saclay
2 rue Gaston Crémieux - France
2 : Laboratoire de Mathématiques et Modélisation d'Evry
CNRS : UMR8071, Université d'Evry-Val d'Essonne, Institut national de la recherche agronomique (INRA)
3 : Laboratoire de Probabilités et Modèles Aléatoires (LPMA) - Site web
Université Pierre et Marie Curie (UPMC) - Paris VI, CNRS : UMR7599, Université Paris VII - Paris Diderot
4 : Génomique évolutive des Microbes - Site web
Institut Pasteur Paris, Centre National de la Recherche Scientifique : UMR3525 - Département Génomes et Génétique - 25-28 rue du docteur Roux, F-75724 Paris Cedex 15 - France

By collecting and comparing all the genomic sequences of a species, pangenomics studies focus on overall genomic content to understand genome evolution both in terms of core and accessory parts. The core genome is defined as the set of genes shared by all the organisms of a taxonomic unit (generally a species). Accessory part is crucial to understand the adaptive potential of bacteria and contains genomic regions that are exchanged between strains by horizontal gene transfers. A consensus representation of multiple genomes would provide a better analytical framework than using individual reference genomes. Here, we introduce this concept, giving it a formal representation using a graph model built up from genes clustered into families.
Pangenomes are generally stored in a binary matrix denoting the presence or absence of each gene family across organisms. However, this structure does not handle the genomic organization of gene families in each organism. In our approach, we propose a graph model where nodes represent families and edges chromosomal neighborhood information. Indeed, it is known that core gene families share conserved organizations along genomes.
Based on this data structure, our method classifies gene families through an Expectation-Maximization algorithm based on Bernoulli mixture model and in order to take into account the genomic context of gene families, we smooth the classification with neighborhood information using Markov random field model. This approach splits pangenomes into three partitions: (1) persistent genome, equivalent to a relaxed core genome (genes conserved in all but a few genomes); (2) shell genome, genes having intermediate frequencies corresponding to moderately conserved genes potentially associated with environmental adaptation capabilities; (3) cloud genome, genes found at very low frequency. Finally, the partitions are then overlaid on the neighborhood graph in order to obtain what we called the Partitioned Pangenome Graph. Thanks to this graphical structure and the associated statistical model, the pangenome is resilient to randomly distributed errors (e.g. an assembly gap in one genome can be offset by information from other genomes, thus maintaining the link in the graph).
Pangenomics is a relevant paradigm for very large scale comparative genomics. Using this approach the overall microbial genomes of refseq have been used to build Partitioned Pangenome Graphs. Thereby we establish the sets of persistent and variable paths specific to each species. The method (https://github.com/ggautreau/PPanGGOLiN) is integrated into the MicroScope platform to detect the regions of genomic plasticity.

Notes

The role of structural variants in the adaptive radiation of African Cichlids

Luca Penso-Dolfin, Wilfried Haerty, Federica Di Palma

Earlham Institute

African Lakes Cichlids are one of the most impressive example of adaptive radiation. Independently in Lake Victoria, Tanganyika and Malawi several hundreds of species arose within the last 10 million to 100,000 years. Whereas most analyses in cichlids focused on nucleotide substitutions across species to investigate the genetic bases of this explosive radiation, to date, no study has investigated the contribution of structural variants (SVs) to speciation events (through a reduction of gene flow) and adaptation to different ecological niches. Here, we explore the repertoires and evolutionary potential of different SV classes (deletion, duplication, inversion, insertions and translocations) in five cichlid species (Astatotilapia burtoni, Metriaclima zebra, Neolamprologus brichardi, Pundamilia nyererei and Oreochromis niloticus). We investigated the patterns of gain/loss evolution across the phylogeny for each SV type enabling the identification of both lineage specific events and a set of conserved SVs, common to all four species in the radiation. Both deletion and inversion events show a significant overlap with SINE elements, while inversions additionally show a limited, but significant association with DNA transposons. Genes lying inside inverted regions are enriched for "behavior" (GO:007610), "retina development in camera-type eye" (GO:0060041) and "embryonic skeletal system development" (GO:0048706). Moreover, we find that duplicated genes show enrichment for "antigen processing and presentation" (GO:0019882) and other immune related categories. These results point at a possible role of genome restructuring in the evolution of adaptive traits. Altogether, we provide the first, comprehensive overview of rearrangement evolution in East African Cichlids, and some initial insights into their possible contribution to adaptation.

Notes

Coevolution of chromosome changes and gene regulation in ruminants

Marta Farré, Jaebum Kim, Anastasia A. Proskurjakova, Yang Zhang, Anastasia I. Kulemzina, Qiye Li, Yang Zhou, Yingqi Xiong, Jennifer L. Johnson, Polina Perelman, Warren E. Johnson, Wes Warren, Anna V. Kukekova, Guojie Zhang, Stephen J. O'Brien, Oliver A. Ryder, Alexander S. Graphodatsky, Jian Ma, Harris A. Lewin, Denis M. Larkin

1Royal Veterinary College, London, UK. 2Konkuk University, Seoul, Korea. 3Institute of Molecular and Cellular Biology, Novosibirsk, Russia. 4Novosibirsk State University, Novosibirsk, Russia. 5Carnegie Mellon University, Pittsburgh, PA, USA. 6BGI-Shenzhen, Shenzhen, China. 7University of Illinois at Urbana-Champaign, Urbana, IL, USA. 8Dobzhansky Center for Genome Bioinformatics, St. Petersburg, Russia. 9Smithsonian Conservation Biology Institute, Front Royal, VA, USA. 10McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. 11Kunming Institute of Zoology, Kunming, China. 12Centre for Social Evolution, University of Copenhagen, Copenhagen, Denmark. 13Institute for Conservation Research, San Diego Zoo, CA, USA. 14University of California, Davis, CA, USA.

The role of chromosome rearrangements in driving evolution has been a long-standing question of evolutionary biology. Here we focused on ruminants as a model to assess how these rearrangements have contributed to modifications of gene regulation in evolution. Using reconstructed ancestral karyotypes of Cetartiodactyls, Ruminants, Pecorans, and Bovids, we traced patterns of gross chromosome changes. The lineage leading to the ruminant ancestor after the split from other cetartiodactyls, was characterized by mostly intrachromosomal changes while the lineage leading to the pecoran ancestor (including all livestock ruminants) included multiple interchromosomal changes. We then determined that the functional enhancers in the ruminant evolutionary breakpoint regions are highly enriched for DNA sequences under selective constraint acting on lineage-specific transposable elements and a set of 25 specific transcription factor motifs associated with recently active transposable elements. Using liver gene expression data from five species (cattle, pig, cat, human, and mouse) we found that genes near ruminant breakpoint regions are characterized by more-divergent expression profiles among species, particularly in cattle, which is consistent with the phylogenetic origin of these breakpoint regions. This divergence was significantly greater in genes with enhancers that had at least one of the 25 specific transcription factor motifs in their regulatory domains and located near bovidae-to-cattle lineage breakpoint regions. Therefore, by combining ancestral karyotype reconstructions with analysis of enhancer and gene expression evolution, we show that lineage-specific regulatory elements colocalized with gross chromosome rearrangements provided valuable functional modifications that helped promote ruminant evolution.

Notes

Understanding the genetic components controlling apomixis

Mario Caccamo, NIAB, Cambridge

The intrinsic complexity of plant genomes due to the high content of repetitive elements, its polyploid nature, the high heterozygosity level and its large size are the biggest challenges for sequence assembly and annotation. In the past a key limiting factor to capture all of these features in a genome assembly project was the short length of sequencing reads. The rencent availability of long-range sequencing and mapping technologies provide sufficient information to get high quality contiguous reference sequences for complex genomes. We have used a strategy that combines long/medium read length sequence with chromosome conformation capture to get a *de novo* high-quality genome assembly of the forage grass *Eragrostis curvula*. The *E. curvula* complex has a basic chromosome number of X = 10 and includes cytotypes with different ploidy levels (from 2X to 8X) that may undergo sexual reproduction and facultative or obligate apomixis (asexual reproduction). The availability of this genome has already allowed us to identify candidate regions hypothesised to harbour the apomixis control region/s and will help to establish evolutionary relationships with other members of the family Poaceae, unravelling the taxonomy of the *Eragrostis curvula* complex. This genomic tool will also help us to understand the molecular pathways of important traits to improve forage quality.

Notes

Querying colored and compacted de Bruijn graphs of thousands of related genomes

Nina Luhmann[1], Guillaume Holley[2], Páll Melsted[2], Mark Achtman[1]

[1]Warwick Medical School, Warwick University, Coventry, UK; [2] University of Iceland, Reykjavík, Iceland

Efficient algorithms and data structures are essential to scale genomic analyses to the amount and variety of sequencing data available today. This is especially true now that large databases as EnteroBase[1], containing thousands of genomes from bacterial pathogens, have become publicly available and can be used for extensive comparative analyses.
De Bruijn graphs are typically used for genome assembly, but they are also potentially ideal for other interesting problems such as variant detection in large datasets. In a colored de Bruijn graph, input genomes can be reconstructed from the graph through additional coloring of its nodes. Bifrost[2], recently developed by Holley *et al.*, efficiently builds compacted coloured de Bruijn graphs directly from genome assemblies or sequencing reads, and thereby allows to construct such a graph for 120.000 *Salmonella enteric*a assemblies in ~3.5 days using 106GB of main memory.
We developed an efficient BLAST-like search of a Bifrost graph by querying k-mers from input sequences, similar to the approach of querying large datasets with k-mers in BIGSI by Bradley *et al*.[3] The node coloring allows to estimate an alignment score of these k-mer hits for all genomes in the graph, and we are evaluating whether this measure indicates the presence or absence of a query.
We applied this search to thousands of bacterial pathogen genomes available in EnteroBase. Searching a coloured de Bruijn graph of such large collections of closely related genomes allows within seconds to investigate features inherent in these genomes, e.g. the distribution of pathogenicity islands over different subspecies or within single genomes. Such an efficient search will also be useful when comparing new samples against a large number of related reference genomes, e.g. for aligning ancient DNA sequences against multiple instead of a single related genome. We will also discuss the natural limitations of such k-mer based methods imposed by the sequence diversity in the data in the context of the presented applications of k-mer based graph search.

[1]http://enterobase.warwick.ac.uk
[2]https://github.com/pmelsted/Bifrost
[3] Bradley, Phelim, et al. "Real-time search of all bacterial and viral genomic data." *bioRxiv* (2017): 234955.

Notes

Genome mining for metabolic gene clusters in yeast

Christopher Pyatt, Steve James1, Adam Elliston2, Jo Dicks1, Ian Roberts1

1National Collection of Yeast Cultures, Quadram Institute, Norwich Research Park, Norwich, NR4 7UA, UK, 2The Biorefinery Centre, Quadram Institute, Norwich Research Park, Norwich, NR4 7UA, UK

Biosynthetic gene clusters are co-located and co-expressed groups of non-homologous genes that co-ordinately encode the biosynthetic pathway of a specialised metabolite. They are a curious quirk of genome organisation, particularly with respect to their apparent resistance to being split up by recombination and their assembly through the duplication and relocation of primary metabolism genes. They produce a wealth of secondary metabolites in a wide variety of organisms, from bacteria to plants and fungi. These secondary metabolites are useful in a range of applications, from biofuels and industrial processes to food additives and medical treatments. Yeasts are fungal organisms used extensively in the industrial-scale production of secondary metabolites (e.g. biosurfactants, flavour compounds). Several have recently been shown to possess metabolic gene clusters, of a number of types, but yeasts on the whole have not yet been seriously examined in terms of gene cluster content. As such, they are prime candidates for gene cluster mining, both to further exploit the metabolic potential of these organisms (particularly with regards to silent, or facultatively expressed, pathways that are missed in metabolic screens) and to gain a greater understanding of the evolutionary processes leading to this unusual genome organisation phenomenon.

The genomes of 792 strains from the UK National Collection of Yeast Cultures (NCYC; http://www.ncyc.co.uk) were searched for industrially-relevant gene clusters using both established bioinformatics tools (e.g. BLAST, HMMER, antiSMASH) and ad hoc methods (e.g. spatial clustering of enzyme-coding genes), enabling pros and cons of the two approaches to be inferred. Two gene clusters producing glycolipid biosurfactants were investigated in detail in a group of closely related basidiomycetous yeasts. Contrasting patterns of evolution were discovered between the two gene clusters, particularly with regards to individual gene homology, overall cluster structure and prevalence within the genomes of close relatives.

This ongoing study is highlighting the biosynthetic diversity of a previously unexplored yeast collection. It is providing new insights into the processes by which these industrially useful genetic elements, metabolic gene clusters, are formed, organised and maintained within the genome, despite conflicting evolutionary forces.

Notes

Comparative analysis of hundreds of vertebrate genomes in Ensembl

Carla Cummins, Mateus Patricio, Wasiu Akanni, Matthieu Muffato, Paul Flicek

European Molecular Biology Laboratory,European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

A major goal of the Ensembl project is to accommodate the introduction of hundreds of new genomes with each new update, in order to create the most effective resources for using genomic data in biology. Due to its quadratic nature, comparative analysis of many genomes simultaneously can be a serious bottleneck in many workflows; Ensembl being no exception. Therefore, as data introduction accelerates, it becomes essential to integrate highly scalable methods into our comparative protocols. Here, we describe a multifaceted approach.
Firstly, complexity reduction measures have been put in place. Ensembl's species sampling approach is clade-based, which lends itself well to projection of annotation from a reference species to members of the clade. Traditional methods for clustering of genes into homologous groups are compute-intensive. By performing a first-pass clustering of projected genes (i.e. genes annotated via projection are added to their source gene's cluster by default), we can reduce the number of sequences that need to pass through the comprehensive clustering step resulting in lower overall compute requirements. For example, in Ensembl release 94, planned for September 2018, this method will be applied to 18 mouse genomes, 19 primates, 3 mammals and 45 fish, yielding an expected 25% decrease of the search space.
Secondly, our methods for multiple-sequence alignment, gene tree and homology inferences have been overhauled to include the use of profile Hidden Markov Models (HMMs). HMMs provide a fast and sensitive means of classifying sequences and can largely replace alternative sequence searching algorithms. For gene families, we use the TreeFam library of HMMs, which shifts the algorithmic complexity of classifying the genes into families from quadratic (all-vs-all BLAST comparisons) to linear. For whole-genome alignments, we are benchmarking nhmmer as a possible replacement for exonerate to rapidly align the millions of sequences required to build an Enredo synteny graph.
These changes, along with others, represent a move towards more scalable pipelines, purpose-built for growing, multi-genome datasets.

Notes

Understanding variability and systematic bias in highthroughput data.

Rafael Irizarry

In this talk I will demonstrate the presence of bias, systematic error and unwanted variability in next generation sequencing. I will show the substantial effects these have on downstream results and how they can lead to misleading biological conclusions. I will do this using data from the public repositories as well as our own. We will then describe some preliminary solutions to these problems.

Notes

On the analysis of long-read sequencing data for gene expression

Prof. Dr. Mark Robinson

University of Zurich, Winterthurstrasse 190, IMLS, Zurich, Switzerland

Sequencing technologies are continuing to evolve rapidly and the so-called third generation sequencing platforms, such as Pacific Biosciences (PacBio) or Oxford Nanopore (ONT) are now primed for sequencing of full-length cDNA or RNA on a transcriptome-wide scale. In a first project, we sequence samples from the same biological system using several different ONT protocols as well as a traditional (Illumina) short-read protocol. We investigated different approaches to quantifying transcript abundance from the long-read data and evaluate the correlation with abundances derived from the short-read data. Despite some coverage biases that may affect quantification (short transcripts covered better by individual reads than longer transcripts), the correlation of abundance across biological replicates was high. As expected, the correlation between abundances from long-read and short-read technologies was higher at the gene level than at the individual transcript level. However, the much lower depth from ONT led to low power in differential expression analyses. In a second project, PacBio was used for targeted measurement of the mixture of isoforms for a single gene. Here, we compared methods to cluster reads into distinct isoforms, in the context of much higher (than Illumina) error rates.

Notes

Single-cell isoform RNA sequencing (ScISOr-Seq) across thousands of cells reveals isoforms of cerebellar cell types.

Hagen Tilgner, Ishaan Gupta1,*, Paul G Collier1,*, Bettina Haase2, Ahmed Mahfouz1,3,4, Anoushka Joglekar1, Taylor Floyd1, Frank Koopmans5, Ben Barres6,&, August B Smit5, Steven Sloan6, Wenjie Luo1, Olivier Fedrigo2, M Elizabeth Ross1

1 Weill Cornell Medicine; 2 The Rockefeller Unicersity; 3 Leiden Computational Biology Center; 4 Delft bioinformatics Lab; 5 Amsterdam Neuroscience, VU University; 6 Stanford University

Abstract
Full-length isoform sequencing has advanced our knowledge of isoform biology. However, apart from applying full-length isoform sequencing to very few single cells, isoform sequencing has been limited to bulk tissue, cell lines, or sorted cells. Single splicing events have been described for <=200 single cells with great statistical success, but these methods do not describe full-length mRNAs. Single cell short-read 3' sequencing has allowed identification of many cell sub-types, but full-length isoforms for these cell types have not been profiled. Using our new method of single-cell-isoform-RNA-sequencing (ScISOr-Seq) we determine isoform-expression in thousands of individual cells from a heterogeneous bulk tissue (cerebellum), without specific antibody-fluorescence activated cell sorting. We elucidate isoform usage in high-level cell types such as neurons, astrocytes and microglia and finer sub-types, such as Purkinje cells and Granule cells, including the combination patterns of distant splice sites, which for individual molecules requires long reads. We produce an enhanced genome annotation revealing cell-type specific expression of known and 16,872 novel (with respect to mouse Gencode version 10) isoforms (see isoformatlas.com).

Notes

Bootstrapping Biology: Quick and easy de novo genome assembly to enable single cell gene expression analysis

Nikka Keivanfar, Ian Fiddes1, Jamie Schwendinger-Schreck1, Stephen Williams1, Stephane Boutet1, Donald Miller2, Doug Antczak2, Deanna M. Church1

1: 10x Genomics, Inc. Pleasanton, CA 94566
2: Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY USA

The availability of a high-quality draft assembly is a critical component for addressing biological questions about an organism. Until recently, development of such an assembly has been costly and time intensive. We recently described a protocol for simple genome assembly from a single library that utilizes cost-effective short read sequencers called Supernova (Weisenfeld et al., 2017). Here we demonstrate a rapid workflow wherein Supernova assemblies enable discovering new biology by supporting single cell gene expression (scRNA-Seq) analysis.

Using peripheral blood from a male Equus asinus (donkey), we first generate a phased, diploid assembly using Chromium Linked-Reads and the Supernova assembler. Our assembly has superior scaffold N50 [Supernova: 49.6 Mb; reference: 3.8 Mb] and contig N50 [Supernova: 494 Kb; reference: 66.7 Kb] compared to the E. asinus reference (GenBank accession GCF_001305755.1). Next, we use Cactus, a multi-genome alignment tool, followed by the Comparative Annotation Toolkit (CAT), to generate an annotated draft reference. CAT leverages the comprehensive annotation of horse to generate high-quality annotation on the E. asinus genome, keeping track of orthology relationships. Overall, our annotated assembly has extremely high number (>91% of horse genes in Ensembl) of orthologous vertebrate genes.

We next performed scRNA-Seq on peripheral blood lymphocytes from the same donkey, resulting in analysis of over 7,000 cells, which are clustered based on gene expression profiles. Further, CAT-based annotation enables annotation of clusters representing all expected major cell types, including subsets present at less than 1%.

We also demonstrated this ability of creating a reference for subsequent gene expression analysis using multiple human assemblies and a single lot of peripheral blood mononuclear cells. We compared the performance of these assemblies against the human reference assembly GRCh38 and show that all expected major subpopulations are identified.

We demonstrate an approach that allows for the rapid generation of genomic resources (assembly and annotation) that can support molecular studies. We demonstrate this using scRNA-Seq, and anticipate that this approach could support other types of analysis.

Notes

Discrete and continuous differential expression analysis for single-cell RNA-seq data

Koen Van den Berge1,2, Kelly Street3, Nathan Grinsztajn4, Sandrine Dudoit3,5, Lieven Clement1,2

1Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium; 2Bioinformatics Institute Ghent, Ghent University; 3Division of Epidemiology and Biostatistics, School of Public Health, University of California, Berkeley, USA; 4École Polytechnique, Paris, France; 5Department of Statistics, University of California, Berkeley, USA

Transcriptomics has become a standard tool in modern biology for unraveling the molecular basis of biological processes and diseases. Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of gene expression by characterizing its heterogeneity at the single-cell level within and across tissues and biological conditions. The characterization of particular cell types or states can be achieved through the discovery of marker genes by performing differential expression (DE) analysis between discrete groups of cells (inferred or known a priori).

However, DE analysis in scRNA-seq is non-trivial, in particular, due to zero inflation. Specifically, there are two types of zeros: biological zeros, when a gene is simply not expressed in the cell, and technical zeros (dropouts), when a gene is expressed in the cell but not detected. This phenomenon seems to be particularly problematic for full-length sequencing protocols (e.g SMART-Seq2), and precludes data analysis with bulk RNA-seq tools. Moreover, scRNA-seq data are often used to fit continuous trajectories that model developmental processes from progenitor cells to differentiated cells. These can be single trajectories, e.g., the development of stem cells to one cell type, or branching trajectories, where a progenitor cell population gives rise to multiple distinct cell types. Hence, a flexible modeling framework is required that can accommodate excess zeros.

We first show that bulk RNA-seq DE tools can be leveraged to scRNA-seq, improving upon dedicated state-of-the-art scRNA-seq methodology. We adopt the ZINB-WaVE method (Risso et al., 2018) to fit zero-inflated negative binomial (ZINB) models for every gene, allowing us to downweight excess zeros in subsequent analyses using the posterior probability that a count belongs to the NB count component. We show that these weights restore the performance of bulk RNA-seq tools in the presence of excess zeros and unlocks them for scRNA-seq applications.
Next, we extend our framework towards inference within and between trajectories. By estimating smooth functions of gene expression along developmental pseudotime, we are able to discover genes that are DE between branching trajectories, resulting in more informed results as compared to discrete DE.

Notes

A generative model for single-cell RNA-sequencing

Barbara Englehardt

Single cell RNA-sequencing data poses unique challenges to the genomics community in terms of normalization and RNA quantification, batch correction, visualization, and clustering. We address some of these challenges with a Gaussian process latent variable model (GPLVM), which is a nonlinear manifold that can be used to describe the generative process of each cell in an RNA-sequencing data set. We adapt this GPLVM by adding some non-canonical kernels to capture non-local and non-smooth structure in the data, and we use a heavy tailed t-distribution for the residuals to avoid filtering out cells that may have outlying read counts. We show on a number of data sets how this generative model can be used for normalization, batch correction, filtering, pseudotime inference, and visualization, and compare these results to other models for this purpose that have been discussed in the literature. We conclude with open problems and challenges in this space.

Notes

TALC: Transcriptome-aware Long Read Correction

Lucile Broseus[1,2], William Ritchie[1]

[1]Institut de Génétique Humaine, CNRS, Montpellier, France; [2]Université de Montpellier, France

Third generation sequencing technologies such as Oxford Nanopore and PacBio IsoSeq provide full-length RNA Long Reads, enabling the identification of complex transcript forms expressed in a sample. Yet, the sequences they generate display a high error rate (up to 15%), which can lead to many ambiguous alignments and impair downstream analyses such as transcriptome annotation and precise quantitation. In addition, quantification of transcript isoforms is difficult to assess with third generation sequencing technologies because they generally produce less reads. Hybrid correction solves the long-read accuracy problem by using a matched dataset of RNA-seq short reads (e.g. Illumina) as a reference to correct Long Reads, essentially applying a scheme of tuned alignments.

Currently, most correction methods are dedicated to DNA-sequencing data and thereby rely on computational approximations that do not hold in the the study of RNA isoform abundance. This results, for instance, in poorer performances on highly expressed genes or close to splice junctions. In addition, we observed the introduction of structural errors (e.g. insertion or deletion of exons or introns) which can prejudice the correct discovery of new transcripts. Based on these considerations, we have developed and implemented a transcription aware algorithm that takes into account RNA-seq data specificities to improve Long Read sequence quality and thus enhance the recovery of the true expressed forms. Our method is based on coloured De-Bruijn graphs that include splice junction information and harmonious walks through this graph that are consistent with coverage information contained in RNA-seq short reads. This increases the accuracy and representation of isoform calls in the sample.

Notes

Constraint for mRNA structure in human synonymous mutations

Jeff Gaither, James Li, David Gordon, Grant Lammi, Ben Kelly, Peter White

The Institute for Genomic Medicine, Nationwide Children's Hospital (all) and Department of Pediatrics, The Ohio State University (last supporting author).

We perform a comprehensive study of how synonymous single nucleotide variants (SNVs) may be pathogenic to humans by deforming mRNA structure. We used the Vienna software package to compute mRNA-structural metrics for 16,000,000 synonymous SNPs in 18,000 canonical human transcripts, and then correlated these metrics to population frequencies obtained from the gnomAD database.

We observed that mRNA structure explains some of the variance in gnomAD population frequencies. This is true even after all other natural variables (such as reference and alternate alleles, GC content and tRNA availability) have been taken into account. A striking example of this is given by CpG dinucleotides near the beginning of a transcript - SNVs of this type have a 50% smaller chance of appearing in the population when they are correlated to structure. SNVs for which structure has predictive power also tend to exhibit higher values of conservation and predictive pathogenicity metrics. We also show that constraint from mRNA structural deformation tends to go in the thermodynamically expected direction, with naturally stabilizing mutations A/U -> C/G tending to be constrained in the direction of excessive stability. Notably, we observe constraint for structure in most reference/alternate allele contexts. We can also view the effect of mRNA structure more directly by viewing principal components which are built almost entirely of structure, but which correlate significantly with population frequencies.

Computationally, our procedure rests on the construction of two separate models, one which combines our Vienna metrics with a set of potentially confounding parameters like GC content and sequence context, and another which includes only the confounding variables. In both cases we first tie all our variables together using a multiple factor analysis, which is a generalization of PCA which can be simultaneously applied to categorical and numeric data. Then we correlate each set of orthogonal components to allele frequencies obtained from gnomAD using a general logistic model. A SNP for which the models disagree is judged to be under selection for structure. Our results suggest that synonymous SNPs which cause significant mRNA disruptions are in some cases selected against wholly on the basis of their structural implications.

Notes

Differential isoform usage in Parkinson's disease

Fiona Dick, Gonzalo S. Nido, Charalampos Tzoulis

Department of Clinical Medicine, University of Bergen, Norway

Parkinson's disease (PD) is a major cause of death and disability with a worldwide socioeconomic impact. In spite of more than 200 years of research, the etiology and underlying molecular pathogenesis of PD remain unknown. While transcriptomic studies have helped identify and elucidate key molecular processes in PD, very few employ whole RNA-sequencing and all of these are based on poly-A capture of the mRNA. This approach dramatically increases the 3'- coverage bias in post-mortem tissues, limiting the accuracy of differential isofom usage estimation.
To overcome these limitations, we studied differential transcript usage (DTU) in the prefrontal cortex of individuals with PD (n = 27) and neurologically healthy controls (n = 22), using whole RNA-sequencing after ribosomal RNA depletion. RNA was extracted from fresh-frozen post-mortem brain tissue using standard methods, underwent active ribosomal RNA depletion and was sequenced on an Illumina HiSeq 4000 (stranded 125PE). For comparison, we re-analyzed an RNA-Seq dataset from a published study of the same brain region that employed poly-A selection.
For the DTU analyses, we implemented three pipelines based on alignment-free quantification using state-of-the-art recommendations. (1) DRIMSeq for model fitting + stageR for a two-stage testing; (2) DRIMSeq for filtering + DEXSeq for DTU analysis (instead of differential exon usage analysis); and (3) IsoformSwitchAnalyzeR as standalone, including its functional consequence prediction utility. Transcript abundances were estimated using Salmon. Additionally, we used DEXseq to estimate differential exon usage (with an alignment-based quantification using Hisat2 + HTSeq).
Our preliminary results suggest that several DTU events occur in the brain of individuals with PD. This is currently ongoing work and it remains to be determined whether these findings can be replicated across aforementioned workflows.

Notes

Interpreting variation in the human methylome

Jordana Bell

King's College London, UK

Recent large-scale epigenetic studies of the human methylome have identified multiple drivers of DNA methylation variability, and characterized methylation dynamics during ageing and in age-related disease. An overview of the extent methylome variation that is estimated to be under genetic control will be presented, based on recent large-scale efforts from human cohorts, followed by twin-based results identifying an enrichment of local genetic impacts on DNA methylation at enhancers and insulators profiled on the Infinium MethylationEPIC BeadChip. Further results will be presented along a discussion of DNA methylation dynamics in the context of human ageing and age-related disease across human cohorts.

Notes

Delineation and annotation of the human regulatory landscape across 400+ cell types and states

Wouter Meuleman, Alexander Muratov, Eric Rynes, John Stamatoyannopoulos

Altius Institute for Biomedical Sciences

The human genome encodes vast numbers of non-coding elements whose combined actuation patterns reflect regulatory processes across cellular states and conditions. Despite large-scale technology development for interrogating non-coding parts of the genome, pragmatic annotated high-resolution maps of regulatory regions and their inter-cell type dynamics have been lacking.

To address this issue, we applied a joint experimental and computational approach, integrating 733 deeply sequenced DNase I hypersensitivity assays spanning more than 400 distinct human cell types and states. These data enable a systematic and principled approach to studying regulatory architecture and dynamics on a global scale. We define a common coordinate system for regulatory DNA marked by DNase I hypersensitive sites, encompassing over 3 million elements defined and annotated with unprecedented resolution and detail.

Through systematic analysis of the dynamics of these regulatory regions across cell types and states, we derive a collection of Regulatory Components, providing a novel multi-component annotation of the human regulome. Using admixtures of multiple components, we show that it is possible to decompose biological features of cell and tissue samples and define the extent to which individual regulatory elements contribute to broader cellular regulatory programs.

These previously unappreciated features allow us to characterize the functional properties of genes and pathways. For instance, based solely on their regulatory landscape, we readily identify genes coding for lineage specifying factors. Moreover, we associate specific regulatory structures with distinct binding site motifs, as well as with gene expression patterns across cell types. Moreover, our Regulatory Components provide a fundamentally new framework for understanding how disease-associated variation maps to genome function, not otherwise appreciated.

Taken together, through integrative analysis across hundreds of cell types and states, we provide a novel multi-component annotation of the human regulatory landscape. Our Regulatory Components are predictive for functional and regulatory characteristics of genes, pathways and genetic variants. As such, they open up new horizons on the architecture of human genome regulation and function.

Notes

Tissue-specific enhancer and promoter evolution in mammals

Maša Roller1, Ericca Stamper1,2, Louise Harewood2, Diego Villar2, Aisling Redmond2, Duncan T. Odom2,3, Paul Flicek1,3

1European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
2University of Cambridge, Cancer Research UK Cambridge Institute, Robinson Way, Cambridge, CB2 0RE, UK
3Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Enhancers and promoters are important cis-regulatory elements that establish tissue-specific gene expression. They can be experimentally determined through the presence of specific histone modifications. Regulatory elements were traditionally also categorised as promoters if proximal to a TSS and capable of initiating transcription, and as enhancers if they can increase the utilisation of promoters. Recent work has shown that mammalian enhancers and promoters share many similar features and that some promoters can have enhancer activity. To understand the tissue specific components of regulatory evolution across the mammalian lineage, we have experimentally determined extensive regulatory profiles of four tissues in ten species of mammals. Specifically, we mapped in vivo occupancy of key histone modifications histone 3 lysine 27 acetylation (H3K27ac), histone 3 lysine 4 trimethylation (H3K4me3) and histone 3 lysine 4 monomethylation (H3K4me1) as a means to annotate active promoters and enhancers. We have also sequenced matched RNA-seq in all species and tissues to correlate regulatory information with gene expression.
The regulatory profiles of somatic tissues were markedly different than those of the testes, and this trend was consistent across all species studied. In all species, enhancers were mostly tissue-specific in their activity while promoters were tissue-shared. Using this data, we confirmed previous work that enhancers evolve more rapidly than enhancers. However, incorporating information on their activity across tissues revealed that both enhancers and promoters with tissue-specific activity evolve far more rapidly than those active across multiple tissues. We leveraged this dataset to also asses the prevalence of regulatory elements that act as promoters in one tissues and enhancers in another. Such dynamic promoter-enhancer elements were very rare in all species. Within a species they resembled promoters, but between species they evolved at rates that are more similar to enhancers. These results suggest that promoters and enhancers are rarely interchangeable, but when they are these dynamic regions show mixed characteristic of both regulatory elements. This work provides important insights into the evolution of tissue-specific regulation in mammals.

Notes

Exploratory analysis of retrotransposon activity in the octopus brain

Massimiliano Volpe (1), Giuseppe Petrosino (2), Giovanna Ponte (1), Oleg Simakov (3), Stefano Gustincich (4,5), Graziano Fiorito (1), Remo Sanges (1,5)

(1) Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Napoli, Italy; (2) Bioinformatics Core Facility, Institute of Molecular Biology gGmbH (IMB), Mainz, Germany; (3) Department of Molecular Evolution and Development, University of Vienna, Vienna, Austria; (4) Department of Neuroscience and Brain Technologies, Italian Institute of Technologies (IIT), Genova, Italy; (5) Area of Neuroscience, Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy.

Long interspersed nuclear elements (LINEs) are a class of autonomous retrotransposons. They can alter the genome by inducing mutations but they also play an active role in the evolution by providing genomic sequences that might be exapted by the host in order to evolve genomic novelties. LINEs are active in the human brain and were proposed to cause genomic mosaicism at the basis of the diversification of neuronal functions. LINEs are present in many copies in their hosts, for example they account for almost 20% of the human genome. Recently, LINEs expansion has been observed in the Octopus bimaculoides genome. Octopuses possess a complex nervous system making them the most intelligent invertebrates. We assembled a full-length potentially active LINE element in the Octopus vulgaris neural transcriptome suggesting that LINEs could be potentially active in the octopus brain. To search for somatic retrotransposition events we developed a pipeline able to identify full-length LINEs in the Octopus bimaculoides' genome and transcriptome and identify marks of retrotransposon activity such as age distribution and 3-prime enriched expansion. We have identified two putative LINEs showing a full-length ORF codifying for the expected domains, endonuclease and reverse transcriptase. One of them, belonging to the same group of the O. vulgaris brain expressed element, displays an age distribution consistent with recent activity and the 3-prime enriched expansion. To identify non-reference somatic insertions in the O. bimaculoides, we have performed an insertion-detection analysis on whole genome sequencing data from two different tissues sampled from the same individual: gonads and optic lobe. According to our results, the potentially active element shows a significantly higher number of non-reference integration sites in the optic lobe with respect to gonads therefore supporting the existence of somatic retrotransposition in the octopus brain.

Notes

Mind the gap – interrogating the non-coding genome with single-molecule technologies

Alexander Suh, Uppsala University

Non-coding DNA makes up a significant fraction of any eukaryotic genome. The “non-coding genome” comprises many different types of sequence categories ranging from introns, regulatory elements, telomeric tandem repeats, centromeric tandem repeats, to transposable element insertions. Previously, many of these regions have been notoriously difficult to sequence, assemble, and annotate either due to their repetitiveness or their base composition. Here I review how recent advances in single-molecule technologies such as long-read and linked-read sequencing, optical mapping, and chromosome conformation capture (Hi-C) are helping to overcome these issues and are starting to fill important knowledge gaps in genome research. I further present a multiplatform road map for how to minimize assembly gaps in *de-novo* assemblies while taking into account various assembly problems that may be caused by specific types of repetitive elements.
Identification of genes escaping X-inactivation and the variability of escape across cells, tissues and twin pairs

Notes

Identification of genes escaping X-inactivation and the variability of escape across cells, tissues and twin pairs

Antonino Zito (1), Julia El-Sayed Moustafa (1), Timothy J. Vyse (2), Kerrin S. Small (1).

(1) Department of Twin Research & Genetic Epidemiology, King's College London, SE1 7EH, UK (2) Department of Medical & Molecular Genetics, King's College London, SE1 9RT, UK

The X-linked dosage imbalance between the sexes is compensated by inactivation of one X chromosome during early female embryogenesis. The Lnc-RNA XIST is uniquely expressed from the inactive X (Xi) and drives the X-chromosome inactivation (XCI) process. XCI is incomplete; 15-20% of genes escape from XCI and maintain bi-allelic expression. Escape from XCI has high biomedical relevance and may underlie sexual dimorphisms in autoimmunity and cancer incidence, and clinical phenotypes in poly-X karyotypes. To date, the extent to which genetic and environmental factors influence escape from XCI, and the variability of escape dosages across individuals and cell types are not well characterized. To address these, we designed a metric (EscScore) that quantifies escape from XCI from paired RNA-seq and WGS data. Our score combines a gene's allele specific expression with the XCI-skew of the sample; values range from 0 to 1, with 0 = no escape (no expression from Xi) and 1 = complete escape (equal expression from the Xa and Xi). We applied our score to skewed samples (XCI-skew > 80%) from the TwinsUK cohort (age 38-85, median age = 60), including LCLs (N = 145), whole blood (N = 20), fat (N = 45), skin (N = 54) and purified monocytes, B, CD4+, CD8+ and NK cells from a highly-skewed monozygotic twin pair. Using previously published XCI status for X-linked genes, we find that EscScore distinguishes between genes designated as silenced (mean EscScore = 0.2, IQR = 0.24), variably escape (mean EscScore = 0.28, IQR = 0.46) and fully escape (mean EscScore = 0.57, IQR = 0.41). We identify 51 genes with novel tissue-specific escape patterns, 27 of which were previously classified as silenced and 24 classified as escapees but with no previous evidence of tissue-specificity. Within the purified immune cells, we identify 25 new escapees, of which 14 show cell-type specific escape patterns. We observe a subset of genes exhibiting high degree of variability of escape across individuals. We are investigating concordance of escape within monozygotic and dizygotic twin pairs and correlation with concurrently measured biomedical phenotypes. In this study, we comprehensively quantify escape in multiple immune cell types and tissues, and identify new putative escapees. We conclude that escape from XCI is a complex and widespread event, responsible for the modulation of X-linked functional allelic dosages in a cell-type and tissue restricted manner.

Notes

DNA methylation changes as a marker of senescing leaves in Arabidopsis thaliana.

Minerva S. Trejo-Arellano1, Saher Mehdi2, Jennifer de Jonge1 and Lars Hennig1

1 Swedish University of Agricultural Sciences, Department of Plant Biology and Forest Genetics and Linnean Center for Plant Biology, PO-Box 7080, SE-75007 Uppsala, Sweden.
2 Lab of Experimental Cardiology, Department of Cardiovascular Sciences, KU Leuven, Belgium

During normal development, changes in the epigenome are programmed to surpass developmental barriers. For instance, after DNA methylation patterns have been stablished during embryogenesis, they are dynamically redefined and maintained throughout development. However, little is known about the configuration of the methylome during plant senescence. With the aim of characterising the DNA methylation landscape in senescent Arabidopsis, bisulphite sequencing was performed on young and senescent leaves. Mapping of the methylation values across genes and TEs did not point to global changes in any methylation context. A more refined analysis showed that the majority of DMRs in the CG context were hypermethylated, while DMRs in CHG and CHH context lost methylation and mapped to TEs, being LTR-gypsy family the most affected one. Interestingly, most of the DMRs overlapping genes were hypomethylated. To address whether this correlation would induce expression changes, RNA sequencing was performed. Less than 10% of genes showed a significant change of expression in senescent conditions. However, none of them carried DMRs in the gene body, but rather CHH DMRs in the promoter. Among those, the GO class of H3K9 methylation was significantly enriched, suggesting the involvement of the RdDM pathway upon senescence induced stress.

Notes

Recent evolution of the epigenetic regulatory landscape in human and other primates

Raquel Garcia-Perez1,2, Gloria Mas-Martin2,3, Martin Kuhlwilm1,2, Meritxell Riera1,2, Antoine Blancher4, Marc Marti-Renom2,3,5, Luciano Di Croce2,3,5, Jose Luis Gómez-Skarmeta6, Tomas Marques-Bonet1,2,5, David Juan1,2

1Institut de Biologia Evolutiva, CSIC-UPF, Barcelona, Spain
2Universitat Pompeu Fabra UPF, Barcelona, Spain,
3Centro Nacional de Análisis Genómico - Centro de Regulación Genómica , CNAG-CRG, Barcelona, Spain
4Laboratoire d'Immunogenetique moleculaire , Faculte de Medecine Purpan, Universite Toulouse 3, Toulouse, France
5Catalan Institution of Research and Advanced Studies, ICREA, Barcelona, Spain
6Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Sevilla, Spain

Evolutionary biologists have long sought to discern the molecular basis of phenotypic variation. Changes in gene regulation are thought to play a major role in evolution and speciation, particularly in primates. Over the last decade, the field has experienced a major shift towards inter-species comparative epigenomics in search of a conceptual step-forward in our understanding of evolution. However, the lack of coherent multi-omic datasets has hindered the integrative study of the interplay between epigenomic and genomic evolution in different species. Our study aims to characterize the evolutionary dynamics of regulatory elements in the primate lineage. To that end, we have comprehensively profiled lymphoblastoid cell lines (LCLs) from human, chimpanzee, gorilla, orangutan and macaque, and we have created a suitable resource to investigate regulatory differences in human evolution. On the one hand, our results indicate that genes that are differentially expressed between species are enriched in those with non-conserved gene regulatory architectures, and also point towards a key role of intronic enhancers. On the other hand, the study of novel regulatory elements reveals that lineage specific regulatory elements are enriched in different sets of TFBS, which suggests that different TF are involved in the emergence of regulatory elements at different evolutionary stages. Finally, we are investigating the functional consequences of non-coding lineage specific mutations within these TFBS.

Notes

Sex Differences in Reference Genome Affect Variant Calling and Differential Expression

Melissa Wilson Sayres

Arizona State University, PO Box 874501, Tempe, AZ, USA

The human X and Y chromosomes evolved from a pair of homologous autosomes, but today have vastly different gene content and structure. Curiously, despite tremendous sex-bias in human disease, the sex chromosomes are rarely included in genome-wide analyses of human health and disease. One of the reasons for this exclusion is that the X and Y chromosomes don't follow autosomal patterns of inheritance. However, even when they are included, technical biases resulting from aligning all sequences to a single sex-averaged reference genome can result in erroneous mapping to X and Y. I will present results that failing to account for the ancestral sequence similarity between the human X and Y can affect variant calling and inference of gene expression.

Notes

Encoding yeast genomic diversity using variation graphs

Prithika Sritharan[1], Katharina T. Huber[2], Ian N. Roberts[1], Jo Dicks[1]

[1]National Collection of Yeast Cultures, Quadram Institute, Norwich, Research Park, Norwich, NR4 7UA, UK [2]School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK

Linear reference genomes have been used traditionally to guide the assembly of sequence reads, in particular prior to variant detection. However, the exclusion of common variants from the reference sequence poses fundamental limitations when studying entire species and therefore may be inadequate for organisms such as yeast which display a high level of sequence diversity. Alternative scaffolds have been introduced to supplement the reference genome yet the plethora of currently available mapping softwares are unable to handle alternate locus sequences which will be treated as paralogous duplications within the genome. Variation graphs provide an enriched reference structure in which the genomes of many individuals within a species population can be incorporated as variants forming embedded paths within a bi-directed sequence graph. The use of variation graphs has been shown to mitigate reference allele bias and improve both the accuracy and precision of read mapping, thereby increasing the detection of true, de novo variants.

The National Collection of Yeast Cultures (NCYC; http://www.ncyc.co.uk) contains approximately 4,000 diverse strains from over 530 species. A recent project has led to the whole genome sequencing of ~1,000 NCYC strains, with the largest species group attributed to Saccharomyces cerevisiae. Here, we describe an evaluation of variation graphs as yeast reference structures. In particular, we use Illumina sequence read sets for both NCYC and third-party S. cerevisiae strains to quantify read mapping and variant calling, comparing the use of linear genomes, single-strain variation graphs and multi-strain variation graphs (i.e. pan genomes) as reference structures. In all experiments conducted, we show that multi-strain variation graphs improve both the quantity of sequence read mapping to the reference structure and the quality of the alignment itself. These findings support the future use of variation graphs as reference structures for yeast genomes.

Notes

Genome analysis in a polymorphic moss with large, ancient sex chromosomes

Sarah B. Carey1, Adam C. Payton1, Tikahari Khanal1, Joan Glenny-Pescov1, Kerrie Barry2, Jane Grimwood2,3, Jerry Jenkins3, Jeremy Schmutz2,3, and Stuart F. McDaniel1

1University of Florida, Gainesville, FL, USA
2Department of Energy Joint Genome Institute, Walnut Creek, CA, USA
3HudsonAlpha Institute of Biotechnology, Huntsville, AL, USA

Sex chromosomes have evolved several times across the tree of life. The bryophytes, whose ancestor is thought to be dioecious (i.e. separate sexes), provide novel systems for understanding the evolution of ancient sex chromosomes. Recent data have shown bryophyte sex chromosomes have evolved over the last several hundred million years, meaning they have highly-variable levels of divergence between male and female individuals. Sex chromosomes are also riddled with tandem repeats and transposable elements, making them a challenge to assemble in a genome. The moss Ceratodon purpureus has sex chromosomes that constitute ~100 megabases (MB) of the 360 MB genome and is highly polymorphic (in single nucleotide polymorphisms, copy number variants, and structural variants) making assembly and analyses of its genome complex. Using a combination of Illumina and PacBio data we have assembled genomes of a male (R40) and a female (GG1) isolate of C. purpureus into 731 and 637 contigs, respectively. Using these data, we have analyzed patterns of structural variation between R40 and GG1 on autosomes and sex chromosomes. We also generated RNAseq data of 8 male/female sibling pairs sampled from across a latitudinal distribution of the species (Fairbanks, Alaska to Otavalo, Ecuador). We used these data to examine patterns of nucleotide diversity between the autosomes and sex chromosomes and differential expression between the sexes at two stages of moss development. Collectively, our results 1) show PacBio reads are able to assemble long contigs of the C. purpureus autosomes and sex chromosomes, although not to chromosome scale, 2) show structural variation can be assessed in recently captured portions of the sex chromosomes but ancient regions do not align well, 3) illustrate the importance of SNP-correcting reference genomes when comparing nucleotide diversity and gene expression among distinct isolates, and 4) and highlight the complexity of sex chromosome evolution in species with haploid sex chromosomes.

Notes

ScaffHiC – Genome Scaffolding by Modelling Distributions of Hi-C Paired-end Reads

Zemin Ning1, Shane McCarthy1,2, William Chow1, Jonathan Wood1, James Torrance1, Kerstin Howe1 and Richard Durbin1,2

1Wellcome Sanger Institute, Hinxton, Cambridge, UK
2Department of Genetics, University of Cambridge, Cambridge, UK

Chromosome-scale scaffolding is the ultimate task in creating de novo genome assemblies. With the ability to probe the three-dimensional architecture of whole genomes, Hi-C data derived from high-throughput sequencing offers exciting prospects in genome scaffolding. Due to a certain degree of noise in the Hi-C data, it is not always the contig pair with the largest number of mapped paired-end reads that should be joined. Various studies suggest that contig join probability is related to contig length, the density of mapped reads and most importantly, the distributions of mapped ends over the entire contigs.

We present a new algorithm based on mathematically modelling the distributions of mapped Hi-C reads over the assembled contigs from long read platform data, such as PacBio or Oxford Nanopore. We define an
index - contig distance index (CDI) to quantify the likelihood for each pair of contigs with significant mapped ends. A partner matrix is constructed to store the CDI indexes and is used for a layout of scaffold structure after some filtering.

We report scaffolding results with samples from the Genome10K Vertebrate Genomes Project (VGP) including human, fish and birds. To explore scaffolding differences between HiC and 10X Chromium datasets, we compare ScaffHiC to Scaff10X, a scaffolding tool we developed for 10X Genomics Chromium data.

Notes

De novo assembly and analysis of a canine genome

Dr Jeffrey Kidd

University of Michigan, 1241 Catherine Street, Ann Arbor, MI, USA

Dogs are an emerging model system for the study of genome evolution and human disease. Ongoing studies using short-read re-sequencing data often rely on an incomplete and fragmented reference assembly. The latest version of the dog genome reference, CanFam3.1, contains 19,553 gaps on primary chromosomes as well as 3,000 unplaced contigs with a combined length of 83Mbp. To complement this assembly, we recently completed a de novo assembly of a Great Dane dog based on 50x PacBio long-read sequence data. Assembly resulted in a contig N50 of 4.4 Mbp with the longest contig of 28 Mbp. The assembly results demonstrate a vast improvement with 20-fold increase in continuity and a drastically reduced number of gaps and unplaced contigs. We closed most of the gaps in the CanFam3.1 assembly, finding that the missing sequence is enriched for high GC content that is poorly represented in both Sanger and Ilumina data. We additionally identified 2,489 non-gap insertions of novel, non-repeat sequence missing from the CanFam3.1 assembly. Gene annotation identified 24,891 gene models, including 97 which are completely missing in the current assembly and 667 which have partial hits.

Notes

Pandora – variation inference for pangenomes from Nanopore or Illumina data

Rachel Colquhoun1, Michael Hall2, Derrick Crook3, Zamin Iqbal2

1University of Oxford, 2EMBL-EBI, 3NDM University of Oxford

Bacterial genetic variation originates through multiple mechanisms, including mutations during replication, movement of mobile elements, and various forms of recombination. As a result, genomes can be highly divergent with only a small fraction of genes shared by all; the union of all observed genes is the pan-genome. In this context, the ability to accurately detect genetic variation throughout the pan-genome and compare many genomes remains a difficult problem.

We present a novel reference graph structure, designed to allow approximation of a sequenced genome as a recombinant of genomes in the reference panel. We use this pan-genome reference graph to allow genotyping and discovery of SNPs and larger variants across the pan-genome from both long and short read data. We construct a pan-genome reference graph for E. coli from 23,000 genes and 15,000 intergenic sequence clusters and demonstrate high quality variant calls and sequence inference using nanopore or illumina sequence data from an E. coli outbreak. We demonstrate that we are able to achieve 99.996%/99.97% precision with 95.2%/88.3% recall for SNP genotyping from just 30X coverage of illumina/nanopore data. We show that we provide a systematic framework for analysing diverse sets of samples where a single reference would be inappropriate.

Notes

Direct measurement of spontaneous structural variation through whole-genome sequencing of three generation human pedigrees

Jonathan R. Belyeu (1,2), Ryan M. Layer (1,2), Julie Feusier (1,2), Lynn Jorde (1,2), Aaron R. Quinlan (1,2,3)

1. Department of Human Genetics, University of Utah. Salt Lake City, UT
2. USTAR Center for Genetic Discovery, University of Utah. Salt Lake City, UT
3. Department of Biomedical Informatics, University of Utah. Salt Lake City, UT

Measurement of de novo structural variation (SV) rates in human genomes is difficult owing to a lack of genome sequences from nuclear families, the complexity of structural variant calling, and the inherently lower mutation rate as compared to single-nucleotide mutation. Callsets are also dominated by false positives that are difficult to remove without employing aggressive filtering techniques that have a side effect of removing real variants. These complications have led to uncertainty in the spontaneous SV rate, with estimates of the number of events per live birth, ranging from ~1/6 to ~1/12.
We present our efforts to measure spontaneous SVs from 33 three-generation CEPH pedigrees with ~8 children per F2 generation (603 individuals in total). Multi-generation pedigrees provide an ideal experimental design for validating putative spontaneous SVs, as true mutations in the F1 generation can be distinguished from spurious predictions based upon transmission to at least one of the many children in the F2 generation. We curated the resulting set of candidate SVs using our tool SV-plaudit to visualize the evidence supporting the variant calls in each generation. This allowed us to remove false positives and finalize a set of accurate SV calls. We identified candidate SVs in the F2 generation, and again manually curated variant calls. With the F2 SVs, we will report a revised estimate of the rate of de novo SV mutation, as well as the prevalence of SVs arising from germline mosaicism. Finally, we used results from both sets of curated SV calls to create reproducible filters that identify true de novo SVs in other family-based WGS datasets. We are applying these filters to >3000 family "quartets" from the Simons Foundation Autism Research Initiative (SFARI) to explore the spontaneous SV rate in a larger cohort, assess how the rate changes with parental age, and compare the rate in affected vs unaffected siblings.

Notes

VarTrix is an open-source software tool for assigning variants to individual cells

Ian Fiddes, Stephen R. Williams1, Kamila Belhocine1, Christopher Miller3,4, Katie Sullivan-Bibee1, Robert Fulton2,4, Allegra Petti3,4, Timothy Ley2,3,4, Deanna M. Church1, Patrick Marks1

110x Genomics, Inc., Pleasanton, CA, USA; 2Department of Genetics, Washington University School of Medicine, St Louis, MO, USA; 3Department of Medicine, Division of Oncology, Washington University School of Medicine, St Louis, MO, USA; 4McDonnell Genome Institute,
Washington University School of Medicine, St Louis, MO, USA.

The proliferation of single cell genomics platforms in recent years has allowed for new insights in tissue heterogeneity and biology. Single cell RNA sequencing (scRNA-seq) has become a common experimental method used in molecular, cellular, and developmental biology as well as immunological and cancer studies. Currently, most scRNA-seq analysis focuses on expression estimates. However, scRNA-seq libraries also provide the ability to both call and detect known variants on expressed sequence. Here we present a method and open source tool called VarTrix, implemented in Rust, that takes as input a set of variants called from either scRNA-seq data or from matched WGS data and assigns them to single cells. Both single nucleotide and small indel variants are capable of being evaluated through a process of re-aligning each read to the possible haplotypes using a local Smith-Waterman alignment. This allows for the construction of a variant-cell matrix that is analogous to the gene-cell matrix used in scRNA-seq experiments, as well as associating individual cells with expression-based clustering. To evaluate this technique, we performed single cell gene expression (GEX) analysis in five Acute Myeloid Leukemia (AML) samples and assigned individual cells to previously identified somatic mutations, associating these with the expansion of subclonal populations. In addition, we used a standard cancer enrichment panel (IDT) to enrich a single cell whole genome (scDNA-seq) sequencing library constructed from a mix of cancer and normal cells. Copy number analysis of the unenriched data identified 335 cancer and 508 normal cells. Enrichment saturated the libraries, with an average depth of 317x over the enrichment loci. After segmenting the cells based on copy number, we identified 408 (86%) of SNVs in the cancer cell line and 180 (85%) in the normal cell line. Taken together, VarTrix provides a powerful new tool to interrogate single cell RNA and DNA datasets and evaluate heterogeneity in tumors.

Notes

Poster Presentations

Identification of Variation in Murine Tumour Developmental Pathways

Craig Anderson[1], The Liver Cancer Evolution Consortium, Martin Taylor[1]

[1] MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, United Kingdom

Despite well-regimented exposures and the strong sequence-specific biases associated with mutational accumulation, there remains opportunity for variation in tumourigenesis among carcinogen-induced mouse models of liver cancer. For example, mutagenesis in cells at different stages of the cell cycle and spatial dynamism of mutation and repair processes will introduce variation that can lead to alternative pathways for tumour development. Differentiation between developmental pathways can expose alternative factors that drive cancer progression and are potentially evident before the onset of hepatocellular carcinoma (HCC). These inferences could be highly relevant to prognoses, provided that there is little impact of evolutionarily-derived variation upon oncogenic risk.
We therefore sought to identify developmental variation of diethylnitrosamine-induced dysplastic nodules among five murine models, using whole genome sequencing. Furthermore, we aimed to infer the role of specific cellular processes among tumour developmental subgroups, using tumour-matched transcriptome sequencing.
We've leveraged a novel technique for determining the frequency of haplotypes over short genomic scales, enabling classification of hundreds of pre-cancerous nodules by clonal composition. Haplotypes further enabled dissection of variant allele frequency distributions for defining a scale that resolves the varying contribution of subclonal variants. Tumours adhering to particular developmental pathways reflected significant differences in gross mutational burden, as well as in the frequency of variation among putative oncogenic drivers. Transcriptomic variation between our tumour developmental pathway classifications highlight the role of alternative processes implicit in tumourigenesis, including genes previously implicated in HCC, as well as those associated with transcriptional and epigenetic regulation.

Transposable elements expression in the C. elegans early embryo

Federico Ansaloni (1), Elia Di Schiavi (2), Stefano Gustincich (1,3), Remo Sanges (1,4)

(1) Area of Neuroscience, SISSA, Trieste, Italy.
(2) Institute of Biosciences and BioResources, IBBR, Naples, Italy.
(3) Department of Neuroscience and Brain Technologies, IIT, Genova, Italy.
(4) Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Napoli, Italy.

Transposable Elements (TE) are repetitive elements that spread among the genomes through a cut-and-paste (DNA transposons) or a copy-and-paste mechanism (retrotransposons - LINE, SINE and LTR). TEs (mostly LTRs) are needed by mammalian embryos for processes like pluripotency maintenance, embryo viability and innate immune system priming. Moreover expression and activity of LINE L1 in human and mouse neurons determine brain somatic mosaicism which has been correlated with the evolution of cognitive capabilities but also with neurodegenerative disorders. Caenorhabditis elegans is a worm used as model organism for its simplicity: the adult is composed by about 1000 somatic cells of which 302 are neurons. About 15% of its genome derives by TEs and the Tc/Mar family (DNA TEs) is the most active, while retrotransposition was never observed under laboratory conditions. Taking advantage of a C. elegans public single cell RNA sequencing dataset we analyzed "if", "when" and "where" TEs are expressed during its development. To analyze TE expression we developed a bioinformatics pipeline able to quantify reads mapping specifically on TEs avoiding taking into account TE fragments embedded in other transcripts. Our results suggest high and stage-specific expression of TEs belonging to the LTR and SINE classes. LTRs are the TEs showing the highest expression levels in the C. elegans embryo. Expression of LTR elements CER1 and LTRCER1 is high in the first stages and decreases in the 16-cells stage when the number of pluripotent cells drops down and cells start to differentiate suggesting LTR elements involvement in the maintenance of pluripotency, as described in mouse embryonic cells. LTRs expression in the first embryonic stages was also suggested to prime the innate antiviral response. SINEs are non-autonomous elements that usually take advantage of the machinery encoded by LINEs for their retrotransposition. CELE45 is the only SINE element expressed in the C. elegans embryo and it is highly expressed in the 16-cells stage AB cells that give rise mostly to neurons. This may suggest a role played by SINEs in the development of the CNS and is in line with studies reporting high expression of retrotransposons in the brains of mammals and other model organisms.

Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: diffTF

Christian Arnold1, Ivan Berest1, Armando Reyes-Palomares1, Giovanni Palla1, Kasper Dindler Rasmussen2,3, Kristian Helin2,3 & Judith B. Zaugg

Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg
2 Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen
3 Novo Nordisk Foundation Center for Stem Cell Biology, Copenhagen

Transcription factor (TF) activity constitutes an important readout of cellular signalling pathways and thus for assessing regulatory differences across conditions. However, current technologies lack the ability to simultaneously assess activity changes for multiple TFs and in particular to determine whether a specific TF acts as repressor or activator. To this end, we introduce a widely applicable genome-wide method diffTF to assess differential TF binding activity and classifying TFs as activator or repressor by integrating any type of genome-wide chromatin with RNA-Seq data and in-silico predicted TF binding sites (available at https://git.embl.de/grp-zaugg/diffTF). We apply diffTF to a large ATAC-Seq dataset of mutated and unmutated chronic lymphocytic leukemia and identify dozens of TFs that are differentially active. Around 40% of them have a previously described association with CLL while ~60% constitute potentially novel TFs driving the different CLL subtypes. Finally, we validated the method experimentally using the well studied system of hematopoietic differentiation in mouse.

pyranges: efficient comparisons of genomic intervals in Python

Endre Bakken Stovner1,2, Pål Sætrom1,2,3

1Department of Computer Science, 2K. G. Jebsen Center for Genetic Epidemiology, Department of Public Health, 3 Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

Comparing sets of intervals is a fundamental task in genomics, and a few basic operations allow for answering complex questions such as finding overlaps between regions with different chromatin modifications and the genes that are closest to such overlapping regions.

Current tools that allow operations on genomic intervals include command line applications, like bedtools and bedops, and application programming interfaces (APIs), such as GenomicRanges available in R. There is also an API wrapper for bedtools called PyBedtools which allows easy use from Python, but relies on hard drive reads and writes for each operation. In R, the data structure GenomicRanges allows for efficiently representing and operating on genomic intervals, thereby allowing library authors to immediately begin solving their problem of interest. Indeed, the foundational GenomicRanges library is a cornerstone of genomics packages in the R BioConductor project.

Python is the fourth largest programming language in the world, and it is widely used in bioinformatics, yet Python lacks a GenomicRanges implementation. The PyRanges library remedies this. PyRanges is a high-performance datastructure for representing and manipulating genomic intervals and their associated data in Python. PyRanges is as least as fast and memory-efficient as its R counterpart for most operations. As PyBedtools uses a read and write step for each operation, direct comparisons are harder, but PyRanges is usually as fast for a single operation and much faster for series of operations as PyRanges only needs to read the data once. In summary, PyRanges allows efficient analyses of genomic data, is directly compatible with Python's wealth of data science libraries, and is therefore well suited for genomic analyses.

pyranges is available at https://github.com/endrebak/pyranges

Network and Pathway Analysis of Toxicogenomics Data

Gal Barel, Ralf Herwig

Max-Planck-Institute for Molecular Genetics, Dep. Computational Molecular Biology, Ihnestr. 73, 14195 Berlin

Toxicogenomics studies measure changes in the biological system due to a perturbation that is induced by a chemical molecule that is considered to be toxic for the system. The molecular effects are then analyzed in order to elucidate the mechanisms that are causing the toxicological response. Such toxic responses are often an undesired result of a drug treatment, and therefore toxicogenomics studies can help in better predicting and preventing them. Most of the toxicogenomics studies measure the transcriptomics levels upon drug introduction, at different dosages and time points. These are usually collected using microarrays and RNA-seq, and large collections of such data have been made publicly available in many databases and other resources. Several bioinformatics approaches allow for the analysis of these data, starting from the identification of differentially expressed genes between cases and controls, and expanding into molecular networks and pathways analysis, which allow for a broader understanding of the underlying mechanisms that lead to toxicity. In this work we make use of public toxicogenomics data and describe different tools and approaches for analyzing it. We demonstrate how gene expression values can be combined with biological pathways and networks by applying methods such as over-representation analysis and network propagation. We highlight how to extract subnetworks that represent functional modules that are involved in causing the toxic effects. We exemplify the results on four different drugs of the anthracyclines class: doxorubicin, epirubicin, idarubicin and daunorubicin. These drugs are commonly used as chemotherapy agents and have been shown to be highly efficient in cancer patients, however in many cases they also lead to severe cardiotoxicity. We summarize the results using the different approaches and compare the information provided by them. With this information, further studies could be designed such that ultimately the prediction and prevention of anthracycline induced cardiotoxicity will be improved.

What Does It Take to Sequence 100,000 Genomes?

Ewa Bergmann, Fred Farrell, Liam Paul, Illumina Laboratory Services Team

Illumina Inc.

Illumina Laboratory Services (ILS) is the accredited (ISO 15189:2012) sequencing provider to Genomics England for the 100,000 Genomes Project. Since 2014 ILS has built, operated and improved the necessary infrastructure to deliver 100,000 genomes. The sequencing service covers all steps from receiving extracted DNA to variant calling and annotation. As part of the project, ILS has already sequenced and delivered over 70,000 whole genomes of cancer and rare genetic disease patients and their families, and is aiming to deliver the remaining genomes by the end of the year.

Here, we present some of the logistical and technical challenges involved in such an unprecedentedly large-scale sequencing effort. The informatics service is built on Illumina's BaseSpace® Sequence Hub architecture. We describe the resources in terms of sequencing instruments, computing power and storage required, and our use and development of bespoke software to perform analysis, sample tracking, quality control, scheduling and delivery. We also share our plans to further increase throughput and decrease turnaround times, and highlight what this change means for an accredited Medical Laboratory.

Integrating Transcription Factor binding and gene expression to deconvolute cancer regulomes

Dóra Bihary, Shamith Samarajiwa

MRC Cancer Unit, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK.

The ability to integrate, model and data-mine complex high-throughput cancer related data sets across different biological scales, in both time and space is a major challenge to fully understand aetiology progression and dynamics of carcinogenesis. Integrating different types of functionally related regulatory data sets enables a better understanding of how these distinct regulatory layers interact and contribute to different carcinogenic phenotypes. We hope to gain a better understanding of carcinogenesis by exploring how direct targets of cancer related transcription factors (TFs) modulate the cancer hallmarks and how the epigenome impacts on this transcriptional regulation.

Computationally integrating TF binding (ChIP-seq) and gene expression (RNA-seq or microarray) data enables the identification of TF direct target genes. TP53 is one of the most well studied and the most frequently mutated genes in cancer. TP53 response (just like any other TF) depends on the cellular context and is achieved by regulating the expression of a specific subset of its targets genes. We integrate more than 50 TP53 ChIP-seq data sets with gene expression data to identify the universe of TP53 direct targets.

A new statistical framework for association testing based on aggregation of rare variants

Simon Boutry, Raphaël Helaers and Miikka Vikkula

Human Molecular Genetics, de Duve Institute, Université catholique de Louvain, Belgium
WELBIO (Walloon Excellence in Lifesciences and Biotechnology), de Duve Institute, Université catholique de Louvain, Brussels, Belgium

There are currently >7000 rare diseases, for over half of which the underlying genetic model is still unknown. Moreover, several common diseases previously thought to be uniform, complex genetic entities, are now considered to be genetically heterogeneous collections of rare, monogenic disorders for which, again, the causes are largely unknown. One major impediment to disease gene-discovery, even in this age of high-throughput, genome-wide sequencing, is the limited statistical power of most studies on rare diseases: e.g., small sample-sizes, or dilution of effect-sizes when the genetic cause is distributed over several changes, each of which does not account for a substantial number of patients.

The objective of this work is to build a powerful, flexible statistical software for association-testing of aggregated rare variants (obtained using annotation and filtering tools built into the Highlander software). We already compared statistical tests and packages, focusing on those that test for association of rare variants after aggregation into genetic regions (genes, group of genes, ...). Using our NGS data, we performed an extensive battery of analyses on these tests, classifying them into well-defined categories.

We found that the ability to rank regions correctly is completely dependent on the choice of test and its parametrization (e.g. pre-filtering, case-control design, type of region, genetic model). This requires considerable prior knowledge, in practice often unavailable. To overcome this, we built a statistical framework based on a representative subset of the most optimal association tests (least computationally expensive, and representative of all categories). We propose that, when in doubt, investigators use a cumulative, global ranking from these tests to prioritize candidate regions. Because this framework is directly connected to the Highlander database, investigators simply use the Highlander GUI to define cases versus controls, and select pre-filters (e.g, minor allele frequency, restricted gene-list, etc). The framework then automatically runs optimal tests for all parameter values. The results, classified per test, are presented in excel tables and graphs showing p-values. Genetic regions (e.g. genes) are ranked by p-value within each test, and an overall-ranking is also provided. We showed that this framework improves results when compared to a "one shot" use of any single test, and greatly facilitates interpretation. By virtue of its integration with Highlander (offering all the power of variant-filtering), we have rendered the best of the aggregation-based association tests accessible to biologists who might otherwise be reluctant to engage with data processing and programming (e.g. R, python, etc).

Standardizing gene names in key vertebrate species

Bryony Braschi, Paul Denny, Kristian Gray, Tamsin Jones, Ruth Seal, Susan Tweedie, Bethan Yates, Elspeth Bruford

HGNC, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom

Standardized gene nomenclature provides an essential resource for all researchers. The Vertebrate Gene Nomenclature Committee (VGNC), operating in parallel with the HUGO Gene Nomenclature Committee (HGNC), was established in 2016 to approve consistent gene names and symbols across vertebrate species that lack their own nomenclature group.

Our naming strategy for each selected vertebrate species starts by identifying a high confidence set of genes with consistently predicted 1:1 human orthologs identified using a subset of data from our HCOP (HGNC Comparison of Orthology Predictions) tool. This tool combines orthology assertions made by fourteen resources into a single tool (https://www.genenames.org/cgi-bin/hcop). These orthologs are assigned the human gene nomenclature using an automated pipeline. This strategy has resulted in >12K genes being named in each of chimpanzee, cow, dog and horse.

We are now focusing on naming the non-consensus orthologs, many of which belong to complex gene families and require careful manual review across multiple species. For these cases we have introduced a manual curation step to assess synteny, which then enables us to confirm 1:1 orthology in many cases. Phylogenetic analysis helps us when naming more complex gene families across vertebrates, including those with 1:many, many:1 and many:many homologous relationships.

We will present a phylogeny of the alpha-2-macroglobulins and closely related homologs across selected vertebrate species where phylogeny has proved informative and has helped us to assign appropriate nomenclature. While this phylogeny was manually constructed, we plan to develop a semi-automated pipeline for this step and this will be discussed.

Some complex gene families will require a higher level of manual curation and input from specialist advisors. We will continue to add new species to VGNC based on the quality of genome assembly and annotations, perceived importance as a model for humans and demand from the research community. Please email us if you have expertise in a particular species or gene family you could help us to name: vgnc@genenames.org

Using SMRT technology to map the genome-wide distribution of Base J

Benedikt G. Brink[1,2], Robert P. Sebra[3,4], Nicolai T. Siegel[1,2]

[1]Department of Veterinary Sciences, Experimental Parasitology, Ludwig-Maximilians-Universität, Munich, Germany; [2]Biomedical Center Munich, Physiological Chemistry, Ludwig-Maximilians-Universität, Munich, Germany; [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA; [4]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

The extracellularly parasite Trypanosoma brucei is responsible for animal trypanosomiasis, also called nagana in cattle, and African trypanosomiasis, or sleeping sickness, in humans. The parasite is injected into the bloodstream of mammals by the tsetse fly, where it is constantly exposed to the host's immune response. Thus, trypanosomes have evolved a coat of Variant Surface Glycoproteins (VSG) that shields the cell surface and invariant surface proteins from antibody attacks. Only one VSG is expressed at any given time, a tightly controlled mutually exclusive gene expression. Once recognized by the host's antibodies, the parasites are able to switch to a different VSG in order escape the immune response. How a switch in VSG expression is triggered and how mutually exclusive VSG expression is ensured remains unknown. Intriguingly, in the genome transcriptionally repressed regions are marked by a unique DNA modification of thymidine called Base J. Base J, also referred to as a 5th base, is enriched across transcriptionally repressed genes but absent from the gene coding for the active VSG, implying a role in the mutually exclusive expression of VSG genes.

During single molecule real time (SMRT) sequencing, DNA polymerases catalyse the incorporation of fluorescently labelled nucleotides into complementary nucleic acid strands. The arrival times and durations of the resulting fluorescence pulses yield information about polymerase kinetics and allow direct detection of modified nucleotides in the DNA template. The kinetics of Base J have been explored in Leishmania, a different Kinetoplastid, but not yet in T. brucei. We obtained SMRT sequencing data from T. brucei with 200x coverage and are aiming to pinpoint the exact locations of Base J in the genome, elucidating the role it might play in antigenic variation and the mechanisms behind the modification itself.

Regulation of the splicing of ultraconserved poison exons in SR splicing factors during mammalian neurogenesis

Carlos F Buen Abad Najar 1, Monica E Graham 2, Emily Powers 2, Nir Yosef 1,3, Liana F Lareau 2

1 Center for Computational Biology,
2 California Institute for Quantitative Biosciences,
3 Department of Electrical Engineering and Computer Science, University of California, Berkeley

Ultraconserved 'poison' exons regulate expression of RNA-binding proteins (RBP) including the important SR splicing factors. Inclusion of these alternative exons into mRNAs results in mRNA degradation. Misregulation of the expression of some SR genes causes severe disruptions in brain development, and our observations show that the alternative splicing of their poison exons varies across the differentiation of stem cells into neurons. We hypothesize that the ultraconserved sequences function as splicing regulatory elements that mediate an interconnected network controlling expression of many components of the splicing machinery.

To test this hypothesis, we measured changes in the inclusion of poison exons, measured as percent spliced in (ψ), in single cell and bulk RNA-seq data from human and mouse stem cells induced to neurogenesis. To identify trans-acting regulators of ψ, we used publicly available CLIP-seq data from 157 RBPs to train SVM classifiers and predict binding in and near the poison exons that show ψ changes. Enrichment analysis of RBP binding in and near exons that show similar temporal behaviour allowed us reconstruct a splicing regulatory network that controls poison exon splicing in neurogenesis.

In single cell data, changes in poison exon ψ often appear as an increase in variation in the ψ of individual cells. This points to an underlying heterogeneity in splicing regulation within cell subpopulations. Strikingly, we observed that groups of splicing factors are correlated to this heterogeneity, suggesting that the analysis of splicing at the single cell level can reveal splicing regulatory factors responsible for cryptic variation within subpopulations.

Our results indicate that the alternative splicing of SR poison exons changes in neurogenesis, and that this change is likely regulated by splicing factors and mediated by the ultraconserved sequences of these exons.

RNA-seq transcriptome profiles of whole blood associated to IGF-1 in cattle: primiparous vs multiparous

Laura Buggiotti1, Zhangrui Cheng1, Mazdak Salavati2, Haruko Takeda3, Claire Wathes1 and Genotype plus Environment Consortium4.

1Royal Veterinary College, London, United Kingdom; 2The Roslin Institute, Edinburgh, United Kingdom; 3University of Liège, Liège, Belgium; 4Ghent University, Brussels, Belgium.

Insulin-like growth factor 1 (IGF-1) is an important growth factor and plays an important physiological role in growth-promoting activities. Circulating concentrations of IGF-1 has been suggested as a potential physiological marker for improved efficiency and profitability in cattle production. IGF-1 is indeed a heritable trait in cattle and it has been associated with feed efficiency, growth performance, age at first calving, conception rate and other traits of importance in breeding schemes.
A total of 207 cows (41 primiparous -PP, and 166 multiparous -MP) were recruited from experimental farms and blood samples were drawn from the tail vein. Total RNA was isolated using the Tempus Spin RNA isolation Kit following the manufacturer's instructions. Library preparation was conducted by using 0.75 ug of total RNA with the Illumina TruSeq Stranded Total RNA Ribo-Zero Gold Sample Preparation kit and sequenced on the Illumina NextSeq 500 sequencer, producing on average 30 million single end reads of 75 nucleotides length per sample. Reads were trimmed according to base quality and Bos taurus assembly (UMD3.1.1) and its corresponding gene set was used as reference to map reads by the splice aware aligner HISAT2. Next, SAM files were converted to BAM files and coordinate sorted with SAMtools. BAM files were further processed with Picard Tools in order to remove PCR duplicates, add read group information, sort by chromosome and create indexes. Reads per gene were counted with cufflinks and normalized FPKM were then used in a correlation analysis with IGF-1, PP and MP were analyzed separately throughout. Correlation analysis highlighted 10279 and 9394 genes to be significantly correlated with IGF-1 in MP and PP, respectively. Moreover, high concentration of IGF-1 is correlated with lower gene expression which represented 93% and 87% in MP and PP of the total genes significantly correlated, respectively. Interestingly, in both, PP and MP, pathway analysis show an enrichment on the metabolic and cellular processes. GO enrichment analysis of the same data highlighted few GO terms PP or MP specific and further gene investigation is needed to better understand the complexity of IGF-1 actions in primiparous and multiparous cows.

Development of 'Targeted' Genotyping-by-Sequencing in Atlantic salmon (Salmo salar)

Alex Caulton1,2, John C McEwan2, Andrew S. Hess2, Rayna M Anderson2, Tracey C van Stijn2 ,Theódór Kristjánsson3, Eduardo Rodriguez3, Rudiger Brauning2 Neil Gemmell1 and Shannon M Clarke2

1University of Otago, Dunedin, New Zealand; 2AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand; 3Stofnfiskur, Staðarberg 2-4, Hafnarfjörður IS-221, Iceland

Genotyping-by-sequencing (GBS) is a reduced representation sequencing technology that employs restriction enzymes to sample a small proportion of the genome for variant discovery and genotyping. To minimise costs, low-depth (2-4x) sequencing is often used, however this generates a level of uncertainty in the SNP genotype scores obtained. This is particularly evident when distinguishing between homozygous and heterozygous genotypes at loci where only one allele is observed for an individual. Therefore, methods for genome-wide association studies (GWAS) that account for uncertainty in GBS genotype calls are desirable when searching for loci associated with a particular trait.

Here we demonstrate the utility of low depth GBS for GWAS and QTL mapping in a farmed population of Atlantic salmon. The trait under analysis; resistance to infectious pancreatic necrosis (IPN), has previously been mapped to a putative causal gene, epithelial cadherin (cdh1-1). A total of 1,488 Atlantic salmon, selected for either susceptibility or resistance to IPN were genotyped at 27,251 polymorphic sites across the genome using GBS, with an average read depth of 4.15 reads at each locus. To maximise data utility, SNPs with low coverage were retained for analysis and genotype probabilities were calculated to account for uncertainty. Using this approach we successfully identified the QTL for IPN resistance.

For traits that are controlled by a locus of large effect, it may be of interest to specifically genotype the casual variant(s) en masse in a large number of samples, while still capturing variants in the rest of the genome. Our GWAS identified a peak near the previously identified cdh1-1 gene. To specifically target the causative mutations in this gene we developed a high-throughput, targeted genotyping assay for Atlantic salmon to test the effects of these mutations on resistance to IPN in our population. The targeted assay also included variants associated with other key production-relevant traits in Atlantic salmon including size and age-at-maturity and sex determination. The targeted approach can be incorporated into our GBS workflow to realise the benefits of both specific and genome-wide genotyping or run as a standalone test for relatively small panels of variants of interest. We successfully obtained genotypes at the cdh1-1 locus using this approach, which subsequently allowed us to confirm association of this causative mutation with resistance to IPN in the farmed Atlantic salmon population under study.

Stage-wise testing to boost the power of differential analysis at the transcript-level and in single cell applications

Lieven Clement 1,2, Koen Van den Berge 1,2

1 Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium, 2 Bioinformatics Institute Ghent, Ghent University, Belgium

Characterizing gene expression through sequencing (RNA-seq) has provided a unique opportunity to unravel the molecular processes in biological tissues. Revolutions in sequencing technology and fast accurate transcript quantification disrupted the RNA-seq field by enabling researchers to interrogate the transcriptome at the single cell level and at the transcript level, respectively. Both settings imply researchers to assess multiple hypotheses for every gene, i.e. by assessing differential expression between multiple cell types (scRNA-seq) or transcripts (transcript-level analysis), leading to challenging multiple testing problems. Conventional approaches control the false discovery rate (FDR) on the individual hypothesis level and fail to establish proper gene-level error control, which compromises downstream validation experiments.

We introduce stageR (https://github.com/statOmics/stageR), a two-stage procedure that leverages the increased power of aggregated hypothesis tests while maintaining high biological resolution by post-hoc analysis of genes passing the screening hypothesis. The two-stage procedure is a general paradigm that can be adopted whenever individual hypotheses can be aggregated and achieves an optimal middle ground between biological resolution and statistical power. In this contribution, we show that a) it provides gene-level FDR control in scRNA-seq studies while boosting power for interaction effects without compromising the discovery of main effects and that b) stage-wise testing gains power and provides better FDR control in transcript level analysis workflows by aggregating hypotheses at the gene level, while providing transcript-level assessment of genes passing the screening stage. We will illustrate the flexibility of the approach by comparing different strategies for p-value aggregation in the screening stage as well as different packages and workflows for transcript level inference, and by discussing interesting single cell RNA-sequencing applications.

Go Get Data: making data great again

Michael Cormier1, Brent Pedersen1, Jonathan Belyeu1, Johannes Köster2, Aaron Quinlan1,3,4

1Department of Human Genetics, University of Utah, Salt Lake City, UT, USA
2Algorithms for reproducible bioinformatics, Institute of Human Genetics, University of Duisburg-Essen, Essen, NRW, Germany
3USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA
4Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

A frustrating, yet common challenge in genomics is identifying, collecting, and standardizing the annotations and datasets germane to one's experiment. There are many causes for this complexity, including multiple, disparate data and annotation repositories (e.g., GTEx, ENCODE, 1000 Genomes), genome build differences, chromosome labeling inconsistencies, and coordinate system differences across standard format such as SAM/BAM, VCF, GTF, FASTQ, and BED. This complex and difficult curation process hinders integrative analysis. Researchers consequently waste valuable time identifying where and how to collect and process data, thereby inhibiting reproducibility and constraining research creativity through inertia.

Inspired by modern software package managers, we have developed Go Get Data (GGD) to overcome these challenges. GGD leverages the Conda package management system and the Bioconda infrastructure to minimize the complexity and hassle of obtaining and processing data, providing a standardized data management system and "cookbooks" of data packages with automatic data curation. Data packages contain a set of instructions used to obtain, transform, and standardize -omics data, providing the user easy access to processed data for their analyses. GGD packages are organized by species, genome build, and data provider, thereby providing a simple ontology for package storage, with quick and easy package searching using GGD's search tool. GGD provides a stable source of dataset and annotation reproducibility through Conda's naming, version tracking, and dependency handling structure. This structure allows for researchers to use existing GGD data packages and add new packages as needed, with each package containing necessary versioning, extraction, and processing information. Packages can be cached for rapid download and can be easily added to a cookbook by using GGD's continuous testing and integration system. GGD is structured for -omics data management, allowing for extensive use with multiple data structures and types. We will present the existing GGD framework and describe plans for expanding functionality and the collection of recipes. As data cookbooks expand over the coming months, we anticipate that GGD will become a standard, community-drive ecosystem for quick, easy, and reproducible access to -omics data.

Assembly and phase of extreme long haploid-like subtelomeres in the parasite Trypanosoma brucei combining SMRT sequencing and Hi-C data

Raul O Cosentino (1), Laura SM Müller (1), Konrad U Förstner (2), T Nicolai Siegel (1)

(1) Experimental Parasitology, Department of Veterinary Sciences, Ludwig-Maximilians-Universität München, Munich, Germany; (2) ZB MED - Information Centre for Life Sciences, 50931 Cologne, Germany

Many pathogens evade the host immune response by periodically altering their surface proteins, a strategy known as antigenic variation. Often, its underlying mechanism relies on recombination between different genomic loci, a process affected by 3D genome architecture and local DNA accessibility. However, the factors affecting both genome architecture and antigenic variation have not been identified in any organism. One of the major obstacles in studying the role of genome architecture in antigenic variation has been the highly repetitive nature and heterozygosity of antigen arrays, which has precluded complete genome assembly in many pathogens.
To study the mechanisms underlying antigenic variation, we choose Trypanosoma brucei, a diploid, unicellular parasite causing sleeping sickness in human that contains a vast repertoire of more than 2000 genes coding for antigens.
To obtain a de novo phased T.brucei genome assembly, we combined SMRT sequencing and genome-wide chromosome conformation capture (Hi-C) data. The SMRT reads were assembled into contigs and the Hi-C data was used for scaffolding.
The resulting scaffold of the eleven chromosome genome has a size of ~42 Mb, 30% larger than the available reference genome. While on one hand, as expected, the central 'core' is highly similar and syntenic to the reference genome, validating our assembly approach; on the other hand, the subtelomeres are highly divergent and represent most of the extension observed in our genome assembly. We found that the antigen genes are arranged in extremely long haploid-like subtelomeric arrays, in some cases longer than the homozygous central 'core' region which harbours the housekeeping genes. Surprisingly, we found that for some chromosomes for which the location of the centromeres was elusive, the centromeres were located in the subtelomeric region, in the middle of antigen arrays. Finally, the addition of Illumina gDNA-seq data to correct errors significantly increased the number and length of annotated protein coding genes, pointing out that the error-rate, especially by the introduction of small INDELs, is still a problem in SMRT sequencing based assemblies, but it can be reduced using a hybrid approach.
In this study we demonstrate how the combination of SMRT sequencing with evolutionarily conserved features of the genome architecture can be exploited for the de novo assembly and phasing of complex diploid genomes.

Data mining of unmapped Illumina reads in BAM files

Anthony J. Cox, Julian Gehring, Thomas Krannich

Illumina Cambridge Ltd. (Cox, Gehring), Berlin Institute of Health (Krannich)

Most human WGS datasets on public archives comprise BAM format files (Li et al., 2009) that encode the results of a mapper such as bwa-mem (Li, 2013) applied to a set of Illumina reads. Typically, a small percentage of reads fail to map and so do not contribute to variant calls made from the reference alignments. Nevertheless, tools such as Expansion Hunter (Dolzhenko et al., 2017) demonstrate that valuable information about human genetic variation is contained within such reads.

While unmapped reads are often present in BAM files, they are included in an unordered way that makes further analysis impossible without extracting the unmapped read sets in their entirety, which is onerous at scale or if the data is hosted remotely. By contrast, BAM's sorting and indexing capabilities allow reads that map to a given genomic range to be efficiently extracted in a targeted fashion, even over HTTP or S3.

Here we contribute to the active topic of using approximate data structures such as Sequence Bloom Trees (Solomon and Kingsford et al., 2016) and minhash sketches (Jain et al., 2017) to facilitate search within large sequence datasets. We take a deliberately simple approach that avoids holding a large index in RAM and, like BAM's reference-based indexing, needs only an HTTP server supporting Range requests to handle remote queries.

We have developed an open-source package URCHIN that annotates each unmapped read in a BAM file with a short hash code, such that identical codes are given to reads that are identical, nearly identical or reverse complementary to one another. Sorting by these codes groups similar reads into clusters and, since the output is compatible with BAM, finding and extracting the reads associated with a code can be handled entirely by samtools and htslib.

Moreover, the hash codes themselves can serve as markers to track the presence or absence of particular non-reference sequences across samples. We demonstrate this by typing insertions from the PopIns (Kehr et al., 2017) and Sniffles (Sedlazeck et al., 2017) studies across a cohort of 150 WGS samples, and we also discuss applications to cancer sequencing.

ChroKit: a web-based computational framework for interactive NGS data mining

Ottavio Croci, Stefano Campaner

Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia, Milan, Italy

The advent of next generation sequencing has greatly improved our capability to dissect biological mechanisms at genome-wide level at progressively decreasing costs. First-level analyses of NGS data usually involve two steps: first, raw reads are mapped to a reference genome to obtain alignment files (in BAM or WIG format); second, dedicated statistical methods are applied on the alignment files to detect genomic regions of interest (in BED or GTF format). Once these analyses are completed, researchers need to carry out higher-level analyses which entail the integration of several NGS experiments/datasets in order to identify relationships between epigenetic marks, transcription factor binding and gene regulation. While there is a growing need for the development of new methods to analyse these kinds of data, existing tools suffer from several pitfalls, including the difficulty of use, the limitation to specific platforms or the lack of functionalities, thus raising the need for the implementation of new programs.
In response to these needs, we developed ChroKit (the Chromatin toolKit), a Shiny-based computational framework that allows a comprehensive processing of NGS data in an easy and interactive way. As input, ChroKit takes genomic regions and alignment files from first-level NGS analyses; the program is then able to perform several operations on regions of interest, including making overlaps with other genomic regions and subset them based on user-defined criteria. The strength of this application is the possibility to produce a wide variety of plots in a matter of few seconds and interact with them with an intuitive graphical interface to carry out deeper analyses. Working sessions can be exported as R-files, enabling easy sharing of analyses and precise annotation for full reproducibility of the workflows. This program can be deployed on a server to provide computational speed and multiple-user access from different devices, such as smartphones, tablets or PCs.
In summary, the ChroKit web application represents an effective solution to speed-up and facilitate integrated analyses of NGS data.

Identification of SNVs in the Moroccan population by whole-genome sequencing

Lucy Crooks, Paul Health2, Ahmed Bouhouche3, Elmostafa El Fahime3, Azedine Ibrahimi3, Mimoun Azzouz2, Youssef Bakri3, Mohammed Adnaoui3, Saïd Amzazi3, Rachid Tazi-Ahnini2,3

1Sheffield Hallam University, Sheffield, South Yorkshire, UK; 2University of Sheffield, South Yorkshire, UK; 3Mohammed-V University, Rabat, Morocco

Large-scale human sequencing projects have described around a hundred-million single nucleotide variants (SNVs). However, they have predominately focused on individuals with European ancestry. Homo sapiens evolved in Africa about 200,000 years ago. They are thought to have migrated from northeastern Africa around 60,000 years ago and then dispersed across southern Asia and into Europe. In line with this, genetic diversity is highest in African populations. 86% of SNVs from the 1000 Genomes Project (1000GP) were seen in only one geographical region and the highest proportion was in African populations. The 1000GP examined five resident African populations from West to East across the centre of the continent. Hence, to comprehensively define human genetic variation, further African populations from a broader geographic distribution should be sequenced. Identification of genetic differences between populations could indicate local adaptation or disparities in variants contributing to disease. It could increase our understanding of human evolutionary history and population movements within and adjoining Africa.
We present the results from a pilot study for Moroccan whole-genome sequencing. Morocco is on the northwest coast of Africa, most of the country is north of the Sahara desert. The 1000GP and the African Genome Variation Project concentrated on African populations south of the Sahara. Three individuals were sequenced to depths of 16-30X. We detected 5.9 million SNVs. These will be compared to SNVs from 1000GP and other large-scale projects. To explore differences between Moroccan and other African populations, a PCA was performed on 1000GP data and the Moroccan samples were projected onto the same principle components. The Moroccans were placed in the middle of the cline from European to African populations.

A Pan-Cancer Transcriptome Analysis Reveals Pervasive Regulation through Tumor-Associated Alternative Promoters

Deniz Demircioglu, Martin Kindermans, Tannistha Nandi, Engin Cukuroglu, Claudia Calabrese, Nuno Fonseca, Andre Kahles, Kjong Lehmann, Oliver Stegle, Alvis Brazma, Angela Brooks, Gunnar Rätsch, Patrick Tan, Jonathan Göke

Genome Institute of Singapore, Singapore; Duke-NUS Graduate Medical School, Singapore; EMBL-EBI, Hinxton, UK, ETH Zürich, Computer Science Dept, Switzerland; Memorial Sloan Kettering Cancer Center, New York, USA; University of California, Santa Cruz, USA

Cancer is a disease of the genome where alterations in the DNA lead to uncontrollable cell proliferation and division. These modifications in a cell's behavior are reflected at the transcriptome level. Transcriptional regulation, whose central element is the promoter, is responsible for controlling these changes in the expression. International consortia efforts such as TCGA and the ICGC produced vast amounts of publicly available RNA-Seq data to map changes in the cancer transcriptome. However due to lack of ChIP-Seq and CAGE data the role of the promoters in controlling transcriptional changes in cancer is still mostly unexplored.
Here, we developed a framework for estimating promoter activity from RNA-Seq data and used this framework to study the transcriptional regulatory changes that are associated with cancer. We have analyzed 1359 samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) and 1831 samples from the GTEx projects encompassing 27 cancer types. We demonstrated that our approach accurately identifies active promoters by comparing our promoter activity estimations with H3K4me3 data from the ENCODE project. We found hundreds of tissue specific alternative promoters that are not observable at the gene expression level. Furthermore, we identified promoters with significant activity changes in cancer compared to normal samples for individual cancer types. We examined the associations between noncoding promoter mutations and promoter activity levels, and mutational heterogeneity per cancer type and pan-cancer.
In summary, we showed that promoter activity can be estimated using RNA-Seq data and used this approach to identify cancer associated alternative promoters for 27 cancer types. We anticipate that the promoter activity estimation using RNA-seq data will broaden our understanding of the promoters' role in cancer by enabling the use of widely available RNA-Seq data. Furthermore, the catalogue of cancer associated promoters identified here will be a useful resource to uncover the regulatory and transcriptional changes in cancer.

Calling Copy Number Variants from Genotyping Arrays

Joe Dennis, Douglas Easton

Department of Public Health and Primary Care, University of Cambridge

More than 200,000 breast cancer cases and controls have been genotyped using Illumina genome-wide arrays (iCOGS and Oncorray). These experiments have been successful in identifying common SNPs and Indels associated with the risk of breast cancer. It is also possible to assess the contribution of common copy number variants (CNVs) by imputation. To assess the contribution of rare CNVs (frequency <1%) we developed methods to detect CNVs from the raw intensity data and B allele frequency (BAF). Intensity data from genotyping arrays contains a high level of noise. We applied a principal component adjustment to the intensities that reduced the noise level and enabled more accurate calling of CNVs using PennCNV. We also applied strict sample and locus quality control, excluding probes and samples with excessive variance. As a separate approach we calculated z-scores for each sample at each probe, adjusting for the distribution of intensities for each study and genotyping batch. CNVs were then called for each sample where z-scores indicated a clear shift in intensity from the sample average across each chromosome and BAF scores showed loss of heterozygosity. To assess the sensitivity and specificity of the calling we compared the methods on a large subset of samples that have clinically validated deletions or duplications in the BRCA genes.

Visual analytics of genome coordinates data

Diana Domanska, Aman Kumar and Geir Kjetil Sandve

Department of Informatics, university of Oslo, Oslo, Norway

While current genome browsers are very powerful for the detailed study of genomic features in specific regions, their capabilities for showing information at the genome-wide scale is mostly limited to showing broad variation in frequency along the genome. Genome-scale analysis is thus mostly dominated by hypothesis testing or the computation of genome-wide descriptive statistics.
Visual analytics provides an interesting alternative approach. In contrast to existing visualization approaches that aim to plot the genomic data of interest per set, the aim of visual analytics is to provide visual presentations tailored to specific analytical questions. In addition to frequency variation, one could for instance be interested in variation in length of feature occurrences, tendencies of relative positioning of occurrences, or in the overlap and relative positioning of occurrences for different features.
We will present an interactive visual analytics tool that allows zooming across all scales, showing several properties of feature occurrences at any selected scale. We demonstrate the approach on a variety of data sets, including mutation and histone modification data.

Transcriptomic analysis of the blood immune response to rVSV-ZEBOV vaccination

Alessia Donato1, Alice Gerlini2, Simone Lucchesi1, Sara Sorgi1, Donata Medaglini1, Gianni Pozzi1, Francesco Santoro1

1Lab. di Microbiologia Molecolare e Biotecnologia (LAMMB), Dipartimento di Biotecnologie Mediche, Università di Siena
2Microbiotec srl, Siena

Background
Ebola is a hemorrhagic fever caused by Ebolavirus (EBOV), belonging to the Filoviridae family. VSV-ZEBOV, a live-attenuated recombinant vesicular stomatitis virus vaccine which expresses the EBOV glycoprotein G, is the only vaccine with demonstrated clinical efficacy. Here we studied the transcriptomic response after injection of a single dose of this vaccine.
Materials and Methods
Blood from 64 healthy volunteers, 51 vaccinated either with 107 or 5x107 PFUs of rVSV-ZEBOV and 13 placebo, was collected at different time points after vaccination. RNA extracted from whole blood was sequenced on an Ion Proton instrument, performing a targeted sequencing on 20,812 genes. Differentially expressed genes (DEGs) between two time points were identified using the R package edgeR, the day of the vaccination was used as baseline and compared against the other time points. Functional analysis was conducted on the DEGs using tmod package which assesses the activation of 346 blood transcription modules of co-expressed genes, in order to find an immunological signature.
Results
Differential analysis between baseline and Day1 after vaccination showed 5,469 DEGs with a fold change greater than 1.2 and lower than 0.83. The number of DEGs decreased over time from vaccination: at Day35 only 10 genes were differentially expressed. Functional analysis revealed 135 different modules affected in response to vaccination. Pathways related to interferon and to viral receptors were markedly activated from Day 1 to Day 14. At days 2 and 3, neutrophils were inhibited and complement was activated, while from Day 28 no modules were activated. Correlation analyses of gene expression with anti-glycoprotein antibody titers identified 15 strongly correlated genes at day 14 after vaccination (absolute Spearman's Rho > 0.5, $p < 0.001$).
Conclusion
Vaccination with rVSV-ZEBOV produced a significant modulation of gene expression over time. This live viral vector induced a strong and durable modulation of genes associated with innate response, with downregulation of neutrophil-associated genes and upregulation of complement-related genes in blood at days 2 and 3, followed by upregulation of T cell- and cell-cycle-associated genes at days 7 and 14. An algorithm strongly correlating with antibody titers one year after vaccination was developed based on the expression levels of 15 genes.

Assessing Mutational Load at Regulatory Regions in Cancer

Kevin Donnelly, Lana Talmane, Martin Taylor, and Colin Semple

Institute of Genetics and Molecular Medicine, MRC Human Genetics Unit, University of Edinburgh, UK

Within cancer research, there now exists a large body of work concerned with somatic variation in coding regions of the genome devised to identify the driver mutations responsible for initiation of carcinogenesis and enhanced tumour growth. Regulatory regions remain comparatively understudied, however, this is changing with the increasing availability of whole genome sequencing (WGS) data for cancer genomes. In this ongoing study, we employ a kmer-based approach to ascertain which transcription factor binding sites are subject to increased mutational burden in differing cancer and tissue types. Using DNase hypersensitivity data derived from 125 ENCODE cell types, we define constitutively open and closed regions of the genome expected to be respectively enriched or depleted for factor binding sites. Paired tumour-normal WGS data obtained from ICGC is used to determine mutation rates for all possible kmers of a given length; by contrasting these rates across open and closed regions of the genome, we can identify kmers with a higher than expected burden in regulatory regions. The co-occurrence of these kmers with putative binding sites is determined using JASPAR motifs and the covariance of mutation rate of the kmer and the predicted binding strength of the motif is then used to establish whether it is association with the factor binding site that drives the increased mutation rate observed. In applying this methodology across different cancers, we can examine the similarity and disparity of mutational burden at different factor binding sites, and develop insight to the nature and extent of regulatory site mutation that may be tolerated in cancer.

Structural variants in Graphtyper: population-scale genotyping using pangenome graphs

Hannes P. Eggertsson[1,2], Snædís Kristmundsdóttir[1,3], Doruk Beyter[1], Helga Ingimundardóttir[1], Hákon Jónsson[1], Daníel F. Guðbjartsson[1,2], Kári Stefánsson[1,4], Páll Melsted[1,2], Bjarni V. Halldórsson[1,3]

[1]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland; [2]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland; [3]School of Science and Engineering, Reykjavik University, Reykjavik, Iceland; [4]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland.

Analysis of sequence diversity in the human genome is important for genetic studies. While structural variations may account for more base pair changes in the human genome than other types of sequence variations, they remain poorly understood and are often omitted in large-scale sequence analysis. Thus, there is a need for efficient and accurate methods to genotype structural variations. Here we present a structural variation extension to our previously published software Graphtyper, a publicly available tool for genotyping. Graphtyper realigns short-read sequence data to a pangenome, a variation-aware graph data structure that reflects sequence variations within a population by representing possible haplotypes as graph paths. Based on the realignments, variations within the pangenome is genotyped. We show that Graphtyper genotypes accurately across the sequence variation spectrum and can simultaneously genotype tens of thousands of whole-genomes. We believe that Graphtyper is a valuable tool for characterizing variations in genomic sequences and can assist in understanding how structural variations impact disease and other phenotypes.
Graphtyper is available at https://github.com/DecodeGenetics/graphtyper.

Kmer Based ReferenceFree Detection of FamilyPrivate Variants Reveal the Genetic Complexity of HHT

Andrew Farrell, W. WooderchakDonahue2,3, M. Velinder1, A. Ward1, P. Johnson3, J. McDonald2,4, P. BayrakToydemir1,2, G. Marth1

1) Department of Human Genetics, USTAR Center for Genetic Discovery, University of Utah; 2) Department of Pathology, University of Utah; 3) ARUP Institute for Clinical and Experimental Pathology; 4) Department of Radiology, Hereditary Hemorrhagic Telangiectasia Center, University of Utah

We have previously shown that our reference free, kmer based, variant detection method RUFUS has extremely high specificity and sensitivity for de novo variations of all types including SNPs, INDELs, and structural variations. Here we present a substantial extension of this method to identify low populationfrequency, familial inherited variations, which allows us to accurately track diseasecausing mutations through pedigrees, and pinpoint family-private diseasecausing variants that segregate with affected/unaffected status. We applied this novel method for analyzing patients with hereditary hemorrhagic telangiectasia (HHT), an inherited disease known to be caused primarily by mutations in the genes ENG, ACVRL1, and SMAD4 (in addition to BMP9, which is associated with a phenotype similar to HHT). However, the genetic cause of the disease remains unexplained in approximately 15% of individuals identified as having HHT, despite extensive efforts to identify the causative variants with stateoftheart existing tools. Here we present the results of our analysis of the 60X coverage Illumina whole genome sequencing data collected for 35 individuals from 13 distinct families, where previous causative variant identification methods have failed. To date, RUFUS was able to identify clear causative mutations in 7 of the 13 families: three families had a causative noncoding variant in the ENG or ACVRL1 genes that was missed by previous analyses. Two families had a deleterious variant in ACVRL1 intron 9 that ultimately disrupted splicing (confirmed by RNA sequencing), including one family with an ACVRL1 intron 9:chromosome 3 translocation (confirmed by PCR). Further confirmations are currently underway to identify additional HHT causative genes and genetic modifiers in the remaining 6 families. This means that our method was able to "solve" over half of the nondiagnostic cases, with several additional, promising hits being currently pursued, including novel mobile element insertions and small INDELs, missed by other methods, that may be disrupting splicing and gene regulation. Our methodological advances also reveal that noncoding variation plays a larger role in HHT than previously appreciated, and this is the first report to show the role of chromosomal translocation as a mechanism for the development of HHT.

Comparison of B-cell receptor sequences from multiple individuals

Anna Fowler, Gerton Lunter

Department of Biostatistics, University of Liverpool
Wellcome Trust Centre for Human Genetics, University of Oxford

B cell receptors (BCRs) are a component of the adaptive immune system that recognise and bind antigens. In order to have a healthy immune system, a diverse set of BCRs, capable of recognising many different antigens, is required. This BCR diversity is generated through a complex process of somatic recombination and hyper-mutation, thought to be capable of generating over 1013 unique BCRs. This is far greater than the 106 unique B cells estimated to be present in a single individual, resulting in very little overlap between any two BCR repertoires. NGS has allowed us to capture these somatic differences at the resolution of individual B cells, through targeted mRNA sequencing of the variable region of the BCRs.

The 3rd Complementarity Determining Region (CDR3) is the most variable region of BCRs and thought to be the most important region in determining antigen binding. There is no reference sequence, or set of reference sequences, for this region and it may differ in length between cells, as well as in composition. Comparison of this region is required for removing read errors, identifying clones, and determining antigen binding. Read errors are likely to be differences of single nucleotides; removing these will improve accuracy in analysis. Large clones are typically generated as part of an effective immune response and consist of cells that differ through hyper-mutation; their identification is important in understanding the immune system and its response to a stimulus such as a vaccine. Finally, it is possible that sequences may be substantially different in their nucleotide sequence but still bind the same antigen, this might be due to shared motifs, synonymous mutations, or amino acids with shared properties. Identifying these sequences is an important step in understanding the variability in BCR repertoires between individuals and linking receptor sequence to function.

We present a fast method for clustering together highly similar sequences, which aims to correct for read error and to partially identify clones. This algorithm is applied to a large data set consisting of 5 individuals, sampled at 5 different time points. We then explore computationally intensive distance metrics capable of capturing complex relationships and apply these to a subset of the data. This shows that sequences thought to be responding to the same stimulus have a greater degree of similarity than other sequences in the repertoire.

Improving genome annotation with long read transcriptomics increases diagnostic utility

Adam Frankish1, Charles A. Steward2, Jolien Roovers3, Marie-Marthe Suner1, Jose M. Gonzalez1, Peter De Jonghe3, Paul Flicek1

1European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK; 2Congenica Ltd, Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK; 3Neurogenetics Group, Center for Molecular Neurology, VIB, Antwerp, Belgium

The accurate identification and description of all human genes is essential for analysis of data informing both genome biology and clinical genomics. Even well-characterized protein-coding genes have potentially significant unannotated features, often because they show tissue- or developmental stage-specific expression. This may have consequences in the annotation of variants identified in clinical sequencing and go some way explaining why a causative variant is identified in only ~40% of patients with a genetic disorder. While the wealth of RNA-seq data that has been generated can help to identify many such features, transcripts reconstructed from short-read data are less reliable than those derived from longer reads, reducing their utility in functional annotation. More recently long read technologies such as SLRseq, PacBio and Oxford Nanopore Technologies have emerged with the promise of delivering high throughput single molecule sequencing ideal for supporting the annotation of full-length transcript models.

We produce the Ensembl/GENCODE reference gene annotation for human and mouse as part of the Ensembl project and aim to identify and classify all gene features with high accuracy based on biological evidence. We will describe the incorporation of long transcriptomic data into the Ensembl/GENCODE gene annotation using both expert manual gene annotation and a custom computational pipeline that maintains the stringency of a manual approach. To investigate the utility of long transcriptomic data we have updated the Ensembl/GENCODE gene annotation for 191 genes associated with early infantile epileptic encephalopathies (EIEE) a group of rare neurodevelopmental disorders. Using predominantly SLRseq and PacBio data we added 3550 novel alternatively spliced transcripts containing 1586 completely novel exons and 795 novel splice sites in existing exons, significantly extending the genomic coverage of exonic sequence. Furthermore, we then screened a cohort of 122 patients with the EIEE Dravet syndrome for the novel regions of the nine genes known to harbour causative variants and identified two de novo variants in SCN1A in two patients demonstrating the value of our effort to improve diagnostic yield.

Identifying genomic regions susceptible to systematic sequencing error to improve variant detection.

Timothy M. Freeman (1), Dennis Wang (1), Jason Harris (2)

1) NIHR Sheffield Biomedical Research Centre, University of Sheffield, UK;
2) Personalis Inc., Menlo Park, CA, USA

The ability to accurately call the correct nucleotide at each genomic locus is key to both assembling reference genomes and to identifying variants that occur between individuals (e.g. SNPs) and within individuals (e.g. somatic mutations). One of the major challenges to highly accurate variant calling is that certain regions of the genome are likely to have higher rates of systematic sequencing or alignment errors that may falsely appear to suggest the presence of low-level somatic variants. One simple way to prevent this is to set a threshold on the number of reads that show evidence for the alternative allele, but a high genome-wide threshold may erode sensitivity at quiescent loci which are relatively free of systematic errors. This reduction in sensitivity is a particular problem for sequencing samples where low-level somatic variants are expected, such as tumour samples.

A locus-specific coverage threshold based upon knowledge of the prevalence of systematic errors at different genomic loci would improve variant calling accuracy by minimising both types of errors. We have systematically catalogued all genomic loci which persistently present a minor allele at a low allele fraction, in 150 non-cancerous, normal human samples, as well as surveying other sequenced populations from the 100K Genomes Project. Our genome-wide statistical analyses revealed loci with mean allelic frequency and standard deviation which sharply deviate from the expected Hardy Weinberg allelic frequency distributions assuming no sequencing errors (suggestive of loci prone to systematic sequencing or alignment errors), and identified regions in which these loci are significantly enriched, such as GC-rich regions. We have demonstrated that these loci are widespread (covering 6.93-18.29% of autosomal chromosomes) and consistently show low-level variant allelic coverage in individual patients.

By identifying regions with high systematic errors in this way, we have revealed clues to understanding the underlying biochemical and computational causes of the limitations of sequencing approaches, and how these may be reduced in silico to improve variant detection, especially in samples where low-level coverage of variant alleles is expected. The results of this work form a resource that can provide significant utility to a wide range of scientists who rely on accurate variant calling in their research.

High resolution genetic mapping of causal regulatory interactions in the human genome

Daniel Gaffney, Natsuhiko Kumasaka, Andrew Knights,

Wellcome Sanger Institute

Physical interaction of distal regulatory elements in three-dimensional space poses a challenge for studies of common disease because noncoding risk variants may be substantial distances from the genes they regulate. Experimental methods to capture these interactions, such as chromosome conformation capture (CCC), usually cannot assign causal direction of effect between regulatory elements, an important component of disease fine-mapping. Here, we developed a Bayesian hierarchical approach that models causal interactions between regions of open chromatin using two-stage least squares. We applied our model to a novel ATAC-seq data from 100 individuals mapping over 15,000 high confidence causal interactions. Strikingly, the majority (>60%) of interactions we detected were over distances of <20Kb, a range where CCC-based methods perform poorly. For a fraction of loci, we identify a single variant that alters accessibility across multiple peaks, and experimentally validate one, the BLK/FAM167A locus associated with risk for multiple autoimmune disorders, using CRISPR engineering. Our study highlights how association genetics of chromatin state provides a powerful complement to CCC-methods to map regulatory variants to genes.

Integrating coding exon realignment into Ensembl clade-based vertebrate annotation

Carlos García Girón, Konstantinos Billis, Thibaut Hourlier, Leanne Haggerty, Osagie Izuogu, Denye Ogeh, Bronwen Aken, Fergal J. Martin, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom

Efforts to sequence and assemble all vertebrate species are already underway. Timely and consistent annotation of these assemblies is essential in order to create robust genomic resources that can be a platform for genomics-based discoveries in vertebrate evolution, ecology and biology.

At Ensembl we are addressing this need by continually improving our annotation methods. For vertebrates, we have developed and implemented a clade-based strategy that simultaneously produces a consistent annotation for multiple genomes of closely related species.

The transfer of genes, via a pairwise whole genome alignment mapping, from a well-annotated reference to a target species forms part of Ensembl's annotation strategy. This transfer is particularly useful in cases where the target species lacks species-specific RNA-seq data. The method, like any alignment-based approach, decreases in accuracy with increasing evolutionary distance. To help mitigate this, we generally choose an appropriate reference for each clade, but there are many vertebrate species without a closely related well-annotated genome.

To help map annotation across longer evolutionary distances we have integrated CESAR 2.0 (developed by the Hiller group, PMID: 28961744) into our annotation pipeline. CESAR 2.0 is a fast method of mapping exons and genes that can deal with splice sites that have shifted significantly. We tested mapping annotation from the human Ensembl/GENCODE gene set to a variety of different species at different evolutionary distances. We found that CESAR 2.0 performs significantly better than our existing methods for mapping annotation at longer distances.

As an example of this approach, for Ensembl release 94, we have used CESAR 2.0 to map reference annotation from the zebrafish GRCz11 assembly to 41 other fish genomes. These mapped annotations complement the annotation resulting from alignment of RNA-seq, cDNA and protein sequence. All of these annotations are integrated together to create the final Ensembl gene set for each genome.

Linking chromatin accessibility and gene expression in a cohesin-mutant leukaemia context

Greg Gimenez, Jisha Antony1,2, Julia A. Horsfield

1 Department of Pathology, Dunedin School of Medicine, University of Otago, New Zealand
2 Maurice Wilkins Centre for Molecular Biodiscovery, New Zealand

Within the nuclei of cells, DNA is modified and organised on many different levels, influencing accessibility of the genome to transcriptional machinery. Genome organisation includes DNA methylation, the wrapping of DNA into nucleosomes, and higher order chromatin structure. The three dimensional (3D) chromatin structure of DNA is organized by the cohesin complex. This complex is composed notably of STAG1, STAG2, RAD21 and SMC proteins. A better understanding of chromatin structure will lead to a better understanding of the regulation of gene expression.
The ATAC-Seq technique provides information on which parts of the genome are in open and accessible chromatin, and can therefore provide a measure of how the genome changes when chromatin organisation is disrupted. However, a major challenge is then to link chromatin accessibility with the transcriptome, the RNA output from the genome.
In this project we compared changes in chromatin accessibility (as measured by ATAC-Seq) with changes in transcriptome (measured using RNA-Seq) in a leukaemia cell line, K562, mutant for the STAG2 subunit of cohesin. Analysis of the biological pathways enriched when linking the chromatin accessibility and the gene expression profiles in a cohesin mutant context will be discussed. The results provide insight into how chromatin organization affects transcription in leukaemia cells.

Development of a low-frequency variant calling tool for the analysis of tumor samples processed with ultra-deep next-generation sequencing

Mădălina Giurgiu[1], Florentine Scharf[1], Anke Arnold[1], Andreas Laner[1], Julia Romic-Pickl[1], Anna Benet-Pagès[1], Elke Holinski-Feder[1]

[1]Medical Genetics Center, Munich, Bavaria, Germany

The recently defined Big-Bang Model of tumor development explains the intra-tumor heterogeneity by the formation of subclones during growth; tumor samples will always represent a mixture of these. Such mechanisms need to be taken into consideration for the detection of somatic mutations in tumor samples due to their presence in a small fraction of the cell population. Thus, variants of interest might occur at very low frequencies. Ultra-deep next-generation sequencing (NGS) is the most promising technology for *de novo* mutation detection, thanks to the huge amount of reads that sequencers can generate. Theoretically, all mutations regardless of the variant allele frequency (VAF) or genomic region can be *observed* given enough read depth. However, *calling* them with confidence is not trivial due to noise in the reads. We developed a low-frequency variant calling tool, LowFreq, which was integrated in our germline NGS analysis pipeline, allowing the detection of both high- and low-frequency variants. Higher sensitivity on low-frequency variant calling was achieved by integrating additional information describing the read mapping (e.g. uniformity and correlation scores) and quality parameters. Two reference DNAs (NA12560, NA12561) were mixed at seven different dilutions and sequenced on an Illumina NextSeq System (>1000x mean coverage over 1000 protein-coding genes). The simulated tumor data was used as training data for the method development. The pipeline performance was evaluated using in silico simulated tumor data with a mean coverage of >500x, covering 93 cancer-related protein-coding genes (RM8398 and NA12889 mixture). We show that we could reliably detect variants at a frequency down to 3%. One remaining central issue is the distinction of true low-frequency variants from artifacts, which are known to often be platform-specific sequencing errors. The aim is to increase the specificity of the low-frequency variant calling by a better filtering of artifacts. In a diagnostic application setting, 21 real tumor samples were sequenced on an Illumina MiSeq System with >2000X coverage for the target region of 6 cancer-related protein-coding genes. These data were thoroughly evaluated as a proof of concept to assess the mutation detection rate and to improve the artifact filtering based on real platform-specific data.]

Epigenetic dimorphism by feto-placental sex in human and its implications for the risk of adverse pregnancy outcome

Sungsam Gong, Ulla Sovio, Irving LMH Aye, Francesca Gaccioli, Justyna Dopierala, Michelle D Johnson, Emma Cook, Miguel Constância, D Stephen Charnock-Jones, Gordon CS Smith

Department of Obstetrics & Gynaecology, University of Cambridge

Complications of pregnancies, e.g preeclampsia and fetal growth restriction (FGR), are major causes of 5-10% of the global burden of disease and are associated with placental dysfunction. Epidemiological studies have demonstrated that fetal sex is associated with different patterns of placental pathology, the risk of perinatal death and the risk of preeclampsia, but the mechanisms are unclear. Here we argue that sex-related differences in placental function are associated with the risk of placentally-related complications of human pregnancy. Firstly we investigated how the placental methylation patterns are different by fetal sex based on whole genome oxidative bisulfite sequencing data (n=4). We found most highly ranked differentially methylated regions (DMRs) were located on the X chromosome, which we discuss further later, but there was a 225Kbp sex-specific DMR in the CSMD1 (CUB and Sushi Multiple Domains 1) gene on the chromosome 8. This sex-specific DMR was validated in additional placenta samples (n=8). The RNA-seq data set (64 female and 67 male placenta tissues) showed that CSMD1 mRNA was 1.8 fold higher in male placentas (P=8.5x10-7). We demonstrate a likely placenta-specific CSMD1 transcript variant not detected in the 21 somatic tissues analysed. Secondly we further analysed methylome and transcriptome data sets and found that placenta has twice as many as female-biased X chromosome genes (n=47) than other 19 human tissues we compared. This suggests more genes escape from X chromosome inactivation (XCI) in the placenta. We found 22 placenta-specific escapees, i.e. not female-biased in other tissues, and their promoter methylation patterns are different (P=0.001) from genes subject to XCI. We further illustrate that spermine synthase (SMS), one of the placenta-specific escapees, and its metabolite confer a fetal sex-dependent sensitivity to polyamine depletion and argue that the polyamine metabolism differs by fetal sex and it is related with placentally-derived complications of human pregnancy.

The European Variation Archive: a database of all types of genetic variation data from all species

Cristina Yenyxe Gonzalez, Jose Miguel Mut, Pablo Arce, Sundararaman Venkataraman, Andres Silva, Hannah McLaren, Thomas Keane

EMBL-European Bioinformatics Institute, Cambridge, United Kingdom

The European Variation Archive (EVA, https://www.ebi.ac.uk/eva) is a primary open repository for archiving, accessioning, and distributing genome variation including single nucleotide variants, short insertion and deletions (indels), and larger structural variants (SVs) in any species. Since launching in 2014, the EVA and sister project DGVa have archived approximately 700 million unique variants across 351 studies and 46 species.
A key function of the EVA as a long term data archive is to provide standard format, stable identifiers so that studies and discovered variants can be referenced in publications, cross-linked between databases and integrated with successive reference genome builds.
The EVA currently peers with the NCBI-based dbSNP and dbVar databases to form a worldwide network for exchanging and brokering. From 2017, issuing and maintaining locus identifiers is divided by taxonomy: the EVA is responsible for non-human species and dbSNP for human. Since then, the EVA has imported approximately 230 million identifiers from dbSNP and issued 200 million new ones.
Other services to researchers include: standard variant annotation, calculation of population statistics, and an intuitive browser to query and view variants from studies or across an entire species. The EVA currently offers a comprehensive REST API to query and export data that supports the htsget streaming protocol defined by the Global Alliance for Genomics and Health (GA4GH). The API is species agnostic and is already extensively used by translational and species-specific resources including Ensembl, Ensembl Genomes, Open Targets, WheatIS and the 1000 Sheep Genomes Project.
The EVA also contributes to maintaining the Variant Call Format (VCF) specification and has implemented a validation suite to ensure correctness of all the submissions made to the archive. This suite, as well as the rest of our software, is freely available on GitHub (https://github.com/ebivariation).

Measurement uncertainty in clinical whole genome sequencing tests

Mar Gonzalez-Porta, Ben Moore, Peter Krusche

Illumina Laboratory Services, Wellcome Trust Genome Campus, Hinxton CB10 1DR, UK
Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, CB10 1XL, UK

Accurate benchmarking of pipeline changes is key to the successful delivery of whole genome sequencing results within clinical laboratories. One of the most widely used strategies for benchmarking variant calls is to evaluate calling performance against a set of well characterised samples or truthset, typically reported as precision/recall estimates across the entire genome. To specifically evaluate the clinical impact of changes, it becomes relevant to further stratify performance metrics into subsets matching genomic regions of interest (e.g. protein coding exons or disease genes). In addition, repeat per-sample measurements are also needed to characterise variation in laboratory processes (precision studies). Altogether, the range of subset sizes combined with replicate variability pose a challenge for the interpretation of results, as well as for the reporting of measurement uncertainty, a common requirement for clinical tests under ISO standards such as ISO15189:2012.

Here we present our R package happyCompare [1], which comprises a set of statistical methods to quantify uncertainty in stratified performance metrics from hap.py outputs [2,3]. By modelling variant counts using a Bayesian Beta-Binomial model, we are able to compute highest-posterior density intervals that simultaneoulsy take into account subset size and replicate variability. We show how we have validated our method using simulations, and illustrate its use by comparing stratified recall of small variants in PCR-Free vs. Nano libraries for NA12878, an internationally recognised reference sample [4,5].

References

[1] happyCompare: A reporting toolbox for hap.py outputs - https://github.com/Illumina/happyCompare
[2] Haplotype VCF comparison tools - https://github.com/Illumina/hap.py
[3] Krusche, P et al., 2018. Best practices for benchmarking germline small variant calls in human genomes. bioRxiv doi: 10.1101/270157
[4] Eberle, M.A. et al., 2017. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Research, 27(1), pp.157-164.
[5] Zook, J.M. et al., 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology, 32(3), pp.246-251.

Splice-QTLs in the context of predisposition to colorectal cancer

Toby Gurran, Victoria Svinti MacLeod, Maria Timofeeva, Malcolm Dunlop, Li Yin Ooi, Peter Vaughan-Shaw, Alison Meynert, Susan Farrington, Colin Semple

MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Crewe Road, Edinburgh, EH4 2XU, United Kingdom

The heritability of colorectal cancer (CRC) has been estimated between 7.4% and 26% from a range of analyses based on family lineages and genetic similarity. Certain rare, high penetrance variants are well characterized; though these are estimated to account for only ~5% of all CRC cases. The majority of GWAS-identified risk SNPs for CRC fall within non-coding regions, and the mechanisms by which the majority of these variants contribute to disease predisposition are yet to be elucidated. However, recent data has suggested that variants altering splicing patterns may play roles in both cancer predisposition and progression.

This study has analysed RNA-seq from 221 samples of colonic mucosa (the precise tissue of origin of CRC) from a cohort of Scottish CRC patients and controls. All individuals were genotyped from blood samples via SNP-chips and imputation increased the number of testable variants to 14 million. The sQTLseekeR package was used to identify splice-QTLs based on alignment-independent transcript quantification via Salmon. Over 2,300 SNPs falling either within gene bodies or 5Kbp from a gene's 5' and 3' ends were identified as splice-QTLs associated with a change in the ratio of expression of transcripts from a gene.

Genes previously implicated in CRC predisposition from the NHGRI-EBI GWAS Catalog were found to have associated splice-QTLs, and pathway analysis revealed an enrichment of sQTL events in immune-related genes. The splice-QTLs were found to be significantly enriched within SNPs quantified by a large GWAS meta-analysis for CRC predisposition comprising 20,000 cases and 37,000 controls. These findings suggest that alternative splicing contributes to the functional mechanisms underlying the influence of non-coding SNPs in predisposition to CRC.

Trans-NanoSim: Characterizing and simulating Oxford Nanopore cDNA/dRNA reads using statistical models

Saber HafezQorani [1], Chen Yang [1], René Warren [1], Inanç Birol [1]

[1] BC Cancer Agency Genome Sciences Centre, Vancouver, BC, Canada

Recently, long read technologies are proving useful in interrogating transcriptomes for their gene expression and the transcript isoform signatures not accessible through short reads. Two Oxford Nanopore Technology (ONT) RNA sequencing (RNA-seq) protocols, using complementary DNA (cDNA) or direct RNA (dRNA) libraries, have been demonstrated to generate valuable data for studying complex mammalian transcriptomes. To better realize the potential of these data types, they need to be coupled with bioinformatics tools that are tuned to their platform-specific characteristics. Development of these tools would highly benefit from datasets with known ground-truth. Generating simulated data would be a cost-effective means to accomplish this, and is a widely used strategy in genomics research. Currently, there is an unmet need in the field for a tool that simulates RNA-seq experiments using ONT instruments. In this study, we are responding to this need.

We present Trans-NanoSim, a two-stage pipeline that (1) captures the technology-specific features of ONT transcriptome reads, and (2) simulates reads with similar characteristics. In the modelling stage, it utilizes state-of-the-art tools to align reads to a reference transcriptome, and generates statistical models that describe the read profiles, such as their error modes and length distributions. It also models features of the library preparation protocols used, including intron retention (IR) events in cDNA and dRNA reads. Further, it optionally profiles transcript expression patterns. Next, these models are used to produce in silico reads for a given reference transcriptome.

Although there are several read simulators specifically created for the ONT sequencing technology, none of them specifically address the complexity of ONT RNA-seq reads. We demonstrate the performance of our tool using publicly available experimental Oxford Nanopore cDNA and dRNA reads on H.Sapiens and M.Musculus transcriptomes. As a fast and scalable simulator that captures the features of an input transcriptome, we think that Trans-NanoSim will be of wide interest for the genomics community. As part of our software package, we also make available statistical models of cDNA and dRNA protocols using the latest chemistry.

Fish clade annotation in Ensembl

Leanne Haggerty, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, Osagie Izuogu, Denye Ogeh, Fergal J. Martin, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, United Kingdom

The actinopterygians, or ray-finned fish, account for nearly half of all extant vertebrates and exhibit a high level of phenotypic diversity. Previous and continuing studies have revealed divergent features of actinopterygian genomes. Specific phenotypes represented in such a diverse group can aid in extricating the set of common genes from those genes that evolved in specific lineages in response to environmental pressures.

In order to fuel research into both the evolution of fish and their genome biology we have recently annotated 41 fish species. The Ensembl gene annotation is based on a combination of RNA-seq data (where available), annotation mapping from zebrafish and protein-to-genome alignments using selected UniProt proteins. By generating the annotations in parallel we produce gene sets in a consistent and efficient manner.

For species with RNA-seq data we have generated sample-specific RNA-seq gene tracks in addition to generating the main gene set. Depending on the species, the samples may be different tissues, development stages and/or different environmental conditions. These additional gene tracks provide a unique window into the transcriptional profile of the fish, giving indications of what genes are expressing in each sample and also the dominant transcript structure present at each locus.

The collection of new fish annotations and other fish genome resources such as updated comparative genomics and variation data will be released in Ensembl version 94 (expected September 2018). Ensembl data is available in a variety of ways including our genome browser at http://www.ensembl.org, Perl and REST APIs, FTP site, publicly accessible MySQL databases and BioMart.

Cell type specific enhancer-promoter network inference based on single omic data

Tom Aharon Hait1,2, David Amar3, Ron Shamir1 and Ran Elkon2,4

1The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. 2Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel. 3Stanford Center for Inherited Cardiovascular Disease, Stanford University, Stanford, CA 94305, USA. 4Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 69978, Israel.

Recent sequencing technologies enable joint quantification of promoter and enhancer activities, allowing inference of enhancer-promoter (E-P) links. We show that current E-P inference methods produce a high rate of false-positive links. We introduced FOCS[1], a new inference method, and by benchmarking against ChIA-PET, HiChIP and eQTL data show that it obtains lower false-discovery rate and at the same time higher inference power. By applying FOCS to 2,630 samples taken from ENCODE, Roadmap Epigenomics, FANTOM5, and a new compendium of GRO-seq samples, we provide extensive E-P maps (http://acgt.cs.tau.ac.il/focs).

The next pressing challenge is to identify which of those E-P links is cell type specific and functional. To expand the FOCS methodology for this challenge we used linear mixed effects models, which account for samples' cell type labels. This framework infers cell type-specific active/repressed E-P links. Unlike extant methods, which use a large number of epigenetic and transcriptomic profiles, ours uses only one type of omic data (DNase hypersensitive sites profiles). Motif finding analysis applied to cell-type specific enhancers from active E-P links that were inferred by our method detected significant enrichments of TFs that are known master regulators of the respective cell types, including enriched motifs of GATA factors in enhancers active in acute myeloid leukemia (K562) cell line, hepatocyte nuclear factor 4 (HNF4) in liver hepatocellular carcinoma (HepG2) cell line, and ZEB1/CTCF factors in epithelial-like cell lines (e.g., MCF-7, T-47D, HCT116) that are known to be involved in the epithelial-mesenchymal transition process leading to metastasis.

1. Hait TA, Amar D, Shamir R, Elkon R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol 2018, 19:56.

Improving automated gene annotation for plant genomes.

John P. Hamilton, C. Robin Buell

Department of Plant Biology, Michigan State University, East Lansing, MI, USA

A core characteristic of plant genome architecture is duplication and is observed as whole genome duplication (polyploidy, segmental duplication) and extensive tandem arrays of duplicated genes. Commonly, gene families of intense research interest such as secondary metabolic pathways producing compounds with medicinal properties or disease resistance genes are present in tandem arrays. Genome assemblies from the short read era generally have these regions mis-assembled or collapsed, preventing the correct annotation of gene models in these regions. With the arrival of plant genomes assembled with long reads, it is possible to accurately resolve haplotypes of homologous/homoeologous chromosomes and the correct assembly of tandemly duplicated gene arrays. However, popular gene annotation pipelines often perform poorly on these improved plant genome assemblies with rampant merging of gene models in tandem gene arrays and incorrect prediction of internal gene structure for homoeologous genes. To correct the annotation of the gene families, multiple rounds of manual curation and reviews of the annotation by gene family and biochemical pathway experts are needed.

We will describe the modifications to our annotation pipeline that improves the automated structural annotation of complex plant genomes and minimizes or eliminates the manual curation needed. The results of the annotation of a tetraploid mint species, several medicinal plant species, and the annotation of the genes of the diterpene metabolic pathway in the teak tree with the improved annotation pipeline will be presented.

ELIXIR: Providing a coordinated European Infrastructure for Genomic Data and Services

Jennifer Harrow, on behalf of ELIXIR

ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

ELIXIR unites Europe's leading life science organisations in managing and safeguarding the increasing volume of data being generated by publicly funded research. It coordinates, integrates and sustains bioinformatics resources across its member states and enables users in academia and industry to access services that are vital for their research. There are currently 22 countries involved in ELIXIR, bringing together more than 200 institutes and 600 scientists.

ELIXIR's activities are coordinated across five areas called 'Platforms', which have made significant progress over the past few years. For instance, the Data Platform has developed a process to identify data resources that are of fundamental importance to research and committed to long term preservation of data, known as core data resources. The Tools Platform has services to help search appropriate software tools, workflows, benchmarking as well as a Biocontainer's registry, to enable software to be run on any operating system. The Compute Platform has services to store, share and analyse large data sets and has developed the Authorization and Authentication Infrastructure (AAI) single-sign on service across ELIXIR. The Interoperability Platform develops and encourages adoption of standards such as FAIRsharing, and the Training Platform helps scientists and developers find the training they need via the Training e-Support System (TeSS).

ELIXIR has also established a number of 'Communities', based around a specific research area, major technology, or specialist user interest group. The Communities drive and support the development and integration of strategically important areas with the Platforms and this covers research fields such as Human Data, Metagenomics, Plant Sciences and Proteomics. They do this bringing together domain experts from within ELIXIR and external collaborators. ELIXIR also has the capability to fund short technical pilot studies, called Implementation Studies, with the aim to drive service development and drive standard adoption. Successful Implementations Studies such as Beacons and AAI then lead to adoption and further collaborations with the wider research communities, for example with the Global Alliance for Genomics and Health.

Pathoscore: a tool for unbiased evaluation of variant pathogenicity metrics

James M Havrilla[1], Brent S Pedersen[12], Ryan M Layer[12], Edgar J Hernandez[12], Mark Yandell[123], Aaron R Quinlan[123]

1 Department of Human Genetics, University of Utah, Salt Lake City, UT. 2 USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT. 3 Department of Biomedical Informatics, University of Utah, Salt Lake City, UT.

Deciding upon the best metric to evaluate the pathogenicity of new genetic variation observed in studies of rare human disease can be a complex, and often biased task. Many variant pathogenicity predictors exist, yet determining which is most appropriate to use for a particular phenotype and its corresponding set of genes is difficult. One major complication in fairly evaluating different predictive metrics lies with ClinVar. ClinVar has numerous limitations, particularly that the criteria for establishing pathogenicity is inconsistent and benign variants are primarily common variants from population-scale datasets. Furthermore, ClinVar is highly curated but despite this bias, nearly all variant pathogenicity predictors evaluate and compare their metrics to others with ClinVar. However, such analyses often employ different versions or subsets of ClinVar variants, which complicates metric performance comparisons.

To address these difficulties, and to aid in the evaluation of new and existing methods, we have created pathoscore, which acts both as an evaluation tool and a collection of curated truth sets and metrics. The truth sets are filtered to exclude variants reported to be pathogenic, yet are common in the gnomAD population database. Pathoscore creates an interactive HTML page for evaluating the performance of variant pathogenicity predictors on different truth sets. The website contains interactive ROC curves, metric distributions for benign and pathogenic variants, Youden's J statistic curves, precision-recall curves, the number of variants each metric was able to score, and clinical utility scores. Our goal is not only to provide bioinformaticians with a fair and consistent way to evaluate their new tool against existing approaches, but also to provide clinicians and diagnosticians with an easy, comprehensive approach to identifying the best metric for their patients' datasets.

We demonstrate that most metrics perform best on autosomal dominant variation, while few perform well on compound heterozygous or autosomal recessive variants. We also illustrate that some metrics are best suited to scoring variants in specific disease contexts (e.g. cancer genes). While many metrics' scores are correlated and could therefore substitute for one another when evaluating most variant sets, pathoscore reveals the few metrics that complement one another. In principle, such complementary metrics can be used together to improve variant interpretation, especially when prioritizing variation among different heterogeneous sets of disease genes. Pathoscore illustrates the need to evaluate variant prediction metrics in multiple contexts, and facilitates the critical evaluation of such metrics in both clinical and research settings.

Highlander: variant filtering made easy

Raphael Helaers, Miikka Vikkula

Human Molecular Genetics, de Duve Institute, Université catholique de Louvain, Belgium

A substantial amount of data is being produced by NGS at ever-increasing rates. The technology generates considerable numbers of false positives, and their differentiation from true mutations is difficult. Moreover, the identification of changes-of-interest among thousands of variants requires annotation from various sources and advanced filtering capabilities. We developed Highlander, a Java software coupled to a local database, in order to centralize all variant data and annotations, and to provide powerful filtering tools that are easily accessible to the biologist. Data can be generated by any NGS machine and most variant callers. Variant calls are annotated using DBNSFP (providing predictions from 6 different programs, splicing predictions, prioritization scores from CADD and VEST, MAF from 1000G and ESP), ExAC, GoNL and SnpEff, subsequently imported into the database. The database is used to compute global statistics, allowing for the discrimination of variants based on their representation in the database. The GUI easily allows for complex queries to this database, using shortcuts for criteria such as "sample-specific variants", "variants common to specific samples" or "combined-heterozygous genes". Users can browse through query results using sorting, masking and highlighting of information. Highlander also gives access to useful additional tools, including visualization of the alignment, an algorithm that checks all available alignments for allele-calls at specific positions, and a module to explore the 'variant burden' gene by gene. Highlander is Open-Source (available at http://sites.uclouvain.be/highlander/) and also used by genetic centers of two university hospitals in Brussels, as well as in the BridgeIris project (http://bridgeiris.ibsquare.be/).

Bifrost - Highly parallel construction and indexing of colored and compacted de Bruijn graphs

Guillaume Holley, Páll Melsted, Trausti Sæmundsson

University of Iceland

De Bruijn graphs are the core data structure for a wide range of assemblers and genome analysis software processing High Throughput Sequencing datasets. For population genomic analysis, the colored de Bruijn graph is often used in order to take advantage of the massive sets of sequenced genomes available for each species.

However, memory consumption of tools based on the de Bruijn graph is often prohibitive, due to the high number of vertices, edges or colors in the graph. In order to process large and complex genomes, most short-read assemblers based on the de Bruijn graph paradigm reduce the assembly complexity and memory usage by compacting first all maximal non-branching paths of the graph into single vertices. Yet, de Bruijn graph compaction is challenging as it requires the uncompacted de Bruijn graph to be available in memory.

We present a new parallel and memory efficient algorithm enabling the direct construction of the compacted de Bruijn graph without producing the intermediate uncompacted de Bruijn graph. Our method relies on a space and time efficient data structure, the Blocked Bloom filter, enhanced with minimizer hashing and false positive balancing in order to increase performance. Despite making extensive use of a probabilistic data structure, our algorithm is deterministic and guarantees that the produced compacted de Bruijn graph is exact. Furthermore, Bifrost maintains in memory a fully incremental index of the compacted de Bruijn graph based on minimizer indexing. The index features a broad range of functions such as sequence querying, storage of user data alongside vertices and graph editing that automatically preserve the compaction property. Bifrost makes full use of the index efficiency and proposes a graph coloring method mapping efficiently each k-mer of the graph to the set of genomes in which it occurs.

Experimental results show that our algorithm is competitive with state-of-the-art de Bruijn graph compaction and coloring tools. Bifrost is able to build the colored and compacted de Bruijn graph of about 118,000 Salmonella genomes on a mid-class server in 3.5 days using 103 GB of main memory.

Availability of Bifrost and its C++ API: https://github.com/pmelsted/bifrost

JBrowse Connect: Launch Analysis Tasks from the Browser

Rob Buels[1], Eric Yao[1], Colin Diesh[1], Scott Cain[2], Lincoln Stein[2], Ian Holmes[1]

[1] University of California, Berkley, USA
[2] Ontario Institute for Cancer Research, Toronto, Canada

JBrowse is a widely used HTML5 genome browser with a thriving ecosystem of third-party contributed plugins. As one of the first JavaScript client genome browsers with a web API that can work directly from a static set of files, JBrowse can be run using any web server, including very lightweight web servers. The communication of information is all one-way; there is no concept of a session or persistent connection, and no way for clients to push messages back to server, or to each other. This minimalism can be advantageous (for example, it is secure by design) but for some purposes it is limiting.

We have developed JBrowse Connect, an optional server extension to JBrowse that runs simultaneously as a JBrowse plugin (on the client) and as an extensible back-end (on the server). JBrowse Connect is designed to serve client requests for management and analysis of genomic data; to broker messaging between client and server (for example, to notify logged-in users of newly-added annotation tracks); and to allow straightforward application-specific extensions (“hooks”). It leverages Sails, a JavaScript model-view-controller framework with object-relational business logic and websocket integration.

Like the JBrowse client, JBrowse Connect offers a plugin system for extensibility. As an illustration of this plugin system, we further present JBlast, a JBrowse Connect hook that bridges a JBrowse client and a running Galaxy instance. BLAST is implemented as an example analysis task, demonstrating the API for developing a JBrowse plugin that triggers, tracks, and then collects the results of an analysis workflow in Galaxy (or another workflow manager).

We further describing ongoing developments in the main JBrowse code base including support for CRAM format, CSI index files, text search, iframe-free embedding, and JBrowse Desktop.

Investigating splicing factor function using unannotated splice junction reads

Jack Humphrey, Kitty Lo, Pietro Fratta & Vincent Plagnol

UCL Genetics Institute, University College London, London, UK; University of Sydney, Sydney, Australia; Sobell Department of Motor Neuroscience and Movement Disorders, University College London, London, UK

RNA splicing requires a delicate interplay between cis-acting RNA sequence features and trans-acting splicing factors to distinguish exons from introns. Whether a potential exon is included or not depends on the balance of splicing factors binding to either enhancing or silencing features. The fidelity of the splicing reaction is constantly challenged by mutations, retrotransposed elements and unstable microsatellites, which can introduce cryptic or decoy sequence features. Splicing factors must therefore play a dual role: 1) maintaining canonical splicing, and 2) surveilling the transcriptome and repressing cryptic splicing. A failure on either front creates splicing noise, which can disrupt gene expression by introducing premature stop codons and frameshifts to target transcripts for degradation.

A number of RNA-binding proteins (HNRNPC, MATR3, TDP-43, PTBP1) have been shown to bind to and repress specific sequence features, as the depletion of these proteins leads to the inclusion of cryptic exons. Conversely, we identified a phenomenon in a TDP-43 gain-of-function mutant mouse where certain constitutively included exons were skipped in a process we name skiptic splicing (Fratta et al, 2018). We consider cryptic exon inclusion and constitutive exon skipping a spectrum of splicing noise dependent on the levels and preferences of different splicing factors. It is not known how many other splicing factors act in a similar fashion and whether splicing noise levels change in aging and disease.

RNA sequencing and modern split-read alignment software can detect novel splicing junctions that are not found in annotation databases. These junctions are likely the result of noisy splicing. We present Fidelio, an R package for annotating and quantifying splice junctions. By only using spliced RNA-seq reads it is possible to rapidly analyse very large datasets and we present the results of three such projects.

Using a large dataset of RNA-binding protein knockdowns, we use novel splice junctions to cluster proteins, identifying new functions in splicing. We demonstrate using human tissue data that splicing noise is highly variable between tissues and appears to increase with age. Finally, we use large neurological disease cohorts to evaluate sources of splicing noise in the brain.

Measuring Genome-Wide Changes in RNA Stability upon Nonsense Mediated Decay Abrogation

Kathryn Jackson-Jones, Robert Young, Dasa Longman, Javier Caceres, Martin Taylor

MRC Human Genetics Unit, MRC IGMM, University of Edinburgh, Edinburgh, UK

The nonsense mediated decay (NMD) pathway is a translation-dependent quality control pathway responsible for identifying and degrading mRNAs that would lead to truncated peptides. However, NMD has also increasingly been shown to be a regulator of post-transcriptional gene expression by degrading ~10% of normal transcripts and in this way fine-tune many physiological processes including development, cell cycle, cell stress and tumorigenesis. The RNA helicase and ATPase UPF1 is the key NMD factor in human cells whilst the protein NBAS has recently been shown to be essential for NMD in Caenorhabditis elegans, zebrafish and humans cell lines.
Traditionally NMD target mRNAs have been identified using differential expression analysis, however, NMD decreases the stability of transcripts rather than the expression directly. Therefore, to identify true NMD targets, we have measured change in stability, in conjunction with expression of RNA in a genome-wide manner upon depletion of both UPF1 or NBAS. We labelled HeLa cells with the uridine analogue 4sU and after extracting the RNA fractionated total RNA and nascent RNA containing 4sU.
We carried out standard differential expression analysis on total RNA using the kallisto-sleuth pipeline and found 457 genes significantly upregulated by 2-fold when UPF1 was depleted. Using the ratio of counts for each transcript in nascent compared to total RNA we calculated the half-life for each transcript in control and depleted cells. We found 112 genes whose stability was significantly increased by 2-fold when UPF1 was depleted. Of these 27 genes were found to have increased expression and increased stability resulting in three distinct classes of RNAs; those whose expression increases but stability is unchanged, those whose stability is increased but expression is unchanged and bona fide NMD targets whose stability and expression are both increased.
Interestingly, we found a large amount of intronic reads in the nascent RNA suggesting splicing is not simultaneous with transcription as has been suggested. We will next modify the kallisto-sleuth pipeline to further investigate this transcript processing as well as investigating characteristics of genes in each of the three groups.

Coloc-stats: a unified web interface to perform colocalization analysis of genomic features

Chakravarthi Kanduri (1,2,*), Boris Simovski (1,*), Sveinung Gundersen (1,3,*), Dmytro Titov (1,3), Diana Domanska (1), Christoph Bock (4-6), Maria Chikina (7), Alexander Favorov (8,9), Ryan Layer (10), Aaron Quinlan (10), Nathan Sheffield (11), Gosia Trynka (12), Geir Kjetil Sandve (1,2)

(1) Department of Informatics, University of Oslo, Oslo, Norway
(2) K. G. Jebsen Coeliac Disease Research Centre
(3) Elixir Norway, Oslo node
(4) CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
(5) Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria
(6) Max Planck Institute for Informatics, Saarbrücken, Germany
(7) University of Pittsburgh School of Medicine, Pittsburgh, PA, USA
(8) John Hopkins University School of Medicine, Baltimore, Maryland, USA
(9) Vavilov Institute of General Genetics, Moscow, RF
(10) Department of Human Genetics, University of Utah, Salt Lake City, UT, USA
(11) Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA
(12) Wellcome Trust Sanger Institute, Hinxton, UK
(*) Equal contribution

Functional genomics assays produce sets of genomic regions as one of their main outputs. To biologically interpret such region-sets, researchers often use colocalization analysis, where the statistical significance of colocalization (overlap, spatial proximity) between two or more region-sets is tested. Existing colocalization analysis tools vary in the statistical methodology and analysis approaches, thus potentially providing different conclusions for the same research question. As the findings of colocalization analysis are often the basis for follow-up experiments, it is helpful to use several tools in parallel and to compare the results. We developed the Coloc-stats web service to facilitate such analyses. Coloc-stats provides a unified interface to perform colocalization analysis across various analytical methods and method-specific options (e.g. colocalization measures, resolution, null models). Coloc-stats helps the user to find a method that supports their experimental requirements and allows for a straightforward comparison across methods. Coloc-stats is implemented as a web server with a graphical user interface that assists users with configuring their colocalization analyses. Coloc-stats is freely available at https://hyperbrowser.uio.no/coloc-stats/.

Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome

Mehran Karimzadeh-1,2,3, Michael M. Hoffman-1,2,3,4

1-Department of Medical Biophysics, University of Toronto, Canada; 2-Princess Margaret Cancer Centre, Toronto, Canada; 3-Vector Institute, Toronto, Canada; 4-Department of Computer Science, University of Toronto, Toronto, Canada

Transcription factors (TFs) bind DNA and control expression of genes. Identifying TF binding sites is the first step in pinpointing mutations that disrupt the gene regulatory network and promote disease. ChIP-seq is the most common method for identifying TF binding sites, but performing it on patient samples is hampered by the amount of available biological material and cost of the experiment. Existing methods for computational prediction of regulatory elements primarily predict binding in genomic regions with sequence similarity to known TF sequence preferences. This has limited efficacy since most binding sites do not resemble known TF sequence motifs, and many TFs are not even sequence-specific.
We developed Virtual ChIP-seq, which predicts binding of individual TFs in new cell types using an artificial neural network that integrates ChIP-seq results from other cell types and chromatin accessibility data in the new cell type. Virtual ChIP-seq also uses learned associations between gene expression and TF binding at specific genomic regions. We train Virtual ChIP-seq on a concatenated matrix of genomic regions and predictive features from 13 training cell types and evaluate the performance on 5 different validation cell types. Virtual ChIP-seq correctly predicted TF binding genome wide, even in regions without chromatin accessibility, similarity to TF's sequence preference, or previous report of TF binding. Virtual ChIP-seq outperforms methods that use TF sequence preferences in the form of position weight matrices, predicting binding for 31 TFs (area under receiver operating characteristic curve > 0.97; area under precision-recall curve > 0.3). We predicted binding of these 31 TFs in 34 Roadmap Consortium tissues with matched transcriptome and chromatin accessibility data. This publicly available resource allows us to better study disease epigenomics.
We further improved performance of the model using matrix factorization to learn TF binding from the transcriptome data. We also found that a convolutional neural network which models dependency among genomic regions in 5 kbp windows boosts the performance of Virtual ChIP-seq.

Barcode correction for linked-read sequencing data

Birte Kehr,

Berlin Institute of Health, Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany

The linked-read sequencing protocol by 10X Genomics enables genomic analyses that require long-range information with accurate short reads, such as haplotype phasing, scaffolding of de novo assemblies and structural variant discovery. The protocol augments Illumina short reads with long-range information using a barcoding strategy. Barcodes label read pairs originating from the same long (~50 kb) DNA molecules and are incorporated as the first 16 bases of the first read in each pair. However, as barcodes are part of the sequence reads, they are subject to Illumina sequencing errors. Some errors can be corrected when a whitelist of barcodes present in a sample is given - a crucial preprocessing step for recovering erroneous barcodes and rescuing valuable long-range information.

Here, a new toolbox for barcode correction, bctools, is introduced. It can infer a whitelist of barcodes from linked-read data, correct barcodes using a custom index data structure, and compute basic barcode statistics. The whitelist is inferred from a barcode occurrence histogram and filtered for sequence entropy. Correction is based on an index that allows efficient retrieval of whitelisted barcodes at Hamming distance 1 from a given query barcode using bit vector rank operations. The index can store possible alternative corrections for a single query barcode, needs less than 1 GB of space and can be constructed in less than a minute of time for a typical barcode whitelist.

The toolbox was tested on linked-read sequencing data from NA12878 (Germline Genome v2) downloaded from the 10X Genomics website. The inferred whitelist for this data set includes 1,492,110 barcodes, of which 99.7% can be found on the whitelist shipped with the Long Ranger software by 10X Genomics. Of the 8% of read pairs labeled with barcodes not on the whitelist, bctools corrects a third of which more than 80% can be confirmed with a read cloud. In comparison to the Long Ranger basic pipeline, bctools corrects 1.08 times as many barcodes and 6.5% more barcode corrections can be confirmed with a read cloud. Furthermore, bctools is at least twice as fast and offers more flexibility in its output format for downstream analyses. Therefore, bctools will be a valuable asset for the analysis of linked-read sequencing data.

The toolbox is implemented in C++, GPL licensed and the source code will be available at https://github.com/kehrlab/bctools.

Serverless cloud technologies for variant discovery and interpretation of human genetic disease

Ben Kelly1, Patrick Brennan1, Harkness Kuck1, David Gordon1, Grant Lammi1, Gregory Wheeler1, Jeffrey Gaither1, James Fitch1 and Peter White1,2

1The Institute for Genomic Medicine, Nationwide Children's Hospital
2Department of Pediatrics, The Ohio State University

Next Generation Sequencing technologies continue to show significant increases in throughput, further escalating the critical need for scalable and cost-effective bioinformatics applications. To address this need, we developed Churchill, a secondary analysis pipeline, and Varhouse, a system for tertiary analysis and data warehousing, using serverless cloud technologies.

Churchill employs an innovative parallelization strategy to implement a best practice workflow for variant discovery of single nucleotide variants, indels and structural variants. The pipeline has been completely written to operate in a serverless cloud environment, and is fully automated to utilize AWS Lambda, ECS, Batch and S3. Churchill has been containerized, enabling scalable, large cohort analyses. Taking raw sequencer output as input, Churchill produces a high quality VCF ready for tertiary analysis.

Varhouse is a cloud-powered, serverless application that joins dozens of databases to sample genotypes to enable rapid variant annotation and experimental interpretation. The entire process is automated from VCF through data annotation via technologies such as AWS Lambda, Athena, EMR and S3 along with Apache Spark. The sample data is stored in the open, columnar Apache Parquet format, allowing for straightforward exploratory research. Varhouse's use of Apache Spark allows us to efficiently leverage distributed machine learning algorithms through its built-in ML Pipeline framework, as well as integrations with other libraries such as scikit-learn and TensorFlow. Embarrassingly parallel learning algorithms such as Random Forest are extremely well suited to this framework and we have used this technique to perform unbiased classification and association of phenotypes with large cohort exome data.

The serverless architecture of Churchill and Varhouse enables all compute resources to be utilized on-demand, eliminating the need for dedicated server resources and reducing costs dramatically versus a more traditional HPC cluster. Through use of many of the services available from AWS, Churchill and Varhouse reduce compute costs by 75% and storage costs by 90% without an increase in compute time. Most importantly, the stateless nature enables near limitless horizontal scalability to easily handle the increased amount of data produced from even the latest NGS sequencers.

Dynamic Transcriptome Changes in Porcine Endometrium through a Female Estrous Cycle

Jun-Mo Kim, none

Department of Animal Science and Technology, Chung-Ang University, Anseong, Gyeonggi-do, 17546, Republic of Korea

Estrous cycle is initially related to figure out the female reproductive system, thereby could support to manipulate mating and resolve the infertility or reproductive problems. Moreover, it has great commercial interest in the pigs whose increase of reproductive rate and litter size. However, the complicated differing signs on hormonal secretions and physiological or clinical changes appear during the estrous cycle and those of biological alterations can be closely cooperated with the multiple molecular mechanisms in regulations. In this study, we studied transcriptome changes in endometrium through an estrous cycle. Therefore, we took the endometrium tissue from the seven time series samples to embrace the estrous cycle: 0 day (Estrus onset time which is marked by the preovulatory surge of luteinizing hormone), 3 day, 6 day, 9 day, 12 day, 15 day, 18 days. Total RNA and their libraries in the 21 groups (n = 4 per group) were prepared for produce the transcriptome data by RNA-sequencing analysis. Gene co-expression network (GCN) for seven time points in the endometrium tissue was constructed by the PCIT algorithm. The network illustrated by multiple substantial core groups which were clearly separated by each time points under the stringent cut-off (absolute log 2 fold-change ≥ 3.0) for use only highly significant differentially expressed genes (DEGs). The overall GCN contained 338 genes connected by 581 edges. Gene clustering analysis by the k-means clustering algorithm revealed five distinct expression profiles across estrous stages. Enrichment analyses of biological meaning for the network were conducted by using the DAVID database. The first three significantly enriched KEGG pathway terms were Alsosterone-regulated sodium reabsorption, Pantothenate and CoA biosynthesis and Riboflavin metabolism ($P < 0.05$). To use the ClueGO app in the Cytoscape, we tried to visualize their networks and connectivities among the counted genes. Finally, we suggested candidate genes which can be important genes in endometrial regulations for estrous cycle. According to this study, we could provide dynamical transcriptome changes in the endometrial estrous environment and summarize the essential regulative genes from the constructed GCN model.

Automated analysis and result reporting for targeted sequencing data of a hemophilia A & B patient cohort

Philip Kleinert, Beth Martin, Martin Kircher

Berlin Institute of Health (BIH), Berlin, Germany; University of Washington, Department of Genome Sciences, Seattle, WA, USA

Targeted sequencing of genomic regions associated with Mendelian disease or somatic cancer variation is a cost- and time-efficient approach for screening patient cohorts in modern medicine. Here, we introduce a fast and efficient pipeline to analyze highly imbalanced, targeted next-generation sequencing (NGS) data sets generated using enrichment by molecular inversion probes (MIPs). MIPs are single-stranded DNA molecules that allow targeted amplification of genomic regions and enrichment using Capture by Circularization (Turner EH et al., Annu Rev Genom Hum Genet., 2009). By pooling multiple barcoded MIP reactions, cost-efficient multiplex sequencing can be performed and samples computationally separated using sequence barcodes. The proposed pipeline processes (MIP-arm trimming and overlap read merging), aligns and sorts MIP reads via burrows wheeler transform alignment (BWA), handles coverage imbalance and calls variants using either GATK v3 UnifiedGenotyper in combination with IndelRealigner or GATK v4 HaplotypeCaller in genomic VCF (gVCF) mode. Further, the pipeline supports the analysis of MIPs specifically designed to capture certain structural variants and determines the genetic sex of patients using Y-chromosome-unique probes. In a user-friendly report, we summarize coverage information (incl. incompletely covered regions) as well as variant effect predictions (based on Ensembl VEP) and variant call qualities for SNVs and InDels of each patient and each targeted region. We developed and tested the pipeline using data for a MIP design of >450 probes targeting the Factor VIII (F8) and Factor IV (F9) genes in a hemophilia A & B cohort from the "My Life, Our Future" initiative (Johnson et al., Blood Advances, 2017). Our setup enables the screening of 384 patients on a single Illumina NextSeq run. The pipeline is available as a Ssnakemake implementation on GitHub (https://github.com/kircherlab/hemoMIPs).

Smart Variant Filtering

Vladimir Kovacevic, Jack Digiovana

Seven Bridges Genomics

Variant filtering consists of preserving highly confident variants and removing falsely called variants. Secondary genomic DNA analysis is mainly oriented toward alignment and variant calling because these two processes strongly influence overall quality. Previously, the variant filtering step was mostly overlooked or analyzed only in deeper testing. However, variant filtering can boost precision of variant calls substantially.

Here, we created a Smart Variant Filtering (SVF) framework. Conceptually, the SVF framework has three phases: (i) selecting a locally-optimal machine learning algorithm configuration for the Genome In A Bottle variant-called samples (HG001-HG005); (ii) learning parameters for that configuration with a training set; (iii) using learned parameters to perform variant filtering on novel datasets. SVF is available on Github (https://github.com/sbg/smart-variant-filtering) and also as a Public project (https://igor.sbgenomics.com/u/sevenbridges/smart-variant-filtering) on the Seven Bridges Platform. It is open-sourced and free to use by any party (BSD-3 license).

Phase (i) included brute-force testing across 372 different algorithm and parameter configurations. It included 10-fold, automatitized cross-validation using 123,000 variants. Based on these results we selected a locally optimal classifier and configuration (Multi Layer Perceptron with 250 nodes in the hidden layer). Phase (ii) trained the network selected in the prior phase with 25 million variants to learn the network weights (model) to be applied in Phase (iii).

We will show results from deep, 3-stage testing to demonstrate that SVF outperforms standard variant filtering solutions currently used within most secondary DNA analyses. Smart Variant Filtering increases the precision of called SNVs (removes false positives) for up to 0.2% while keeping the overall f-score higher by 0.12-0.27% than in existing solutions. Indel precision is increased by up to 7.8%, while the f-score increase is in range of 0.1 to 3.2%.

Population-scale detection of non-reference sequence insertions using colored de Bruijn graphs

Thomas Krannich, Birte Kehr

Berlin Institute of Health (BIH), Berlin, Germany

Non-reference sequence insertions are a less frequently investigated class of genomic structural variation where DNA sequence is found within an individual that is novel with respect to a given reference. These sequences occur primarily due to the fact that a reference genome misses some ancestral sequence since it is mostly derived from few or a single individual. Therefore, newly sequenced individuals can yield genomic sequence which is absent from a reference genome. In recent years, only a few methods and tools have been developed to detect these non-reference sequence insertions and even fewer to perform the detection on large cohorts because of the difficulties to handle large-scale data. To find such insertions in short-read sequencing data, it is indispensable to perform a de novo assembly which is algorithmically challenging and computationally expensive. Especially repetitive genomic regions can complicate the assembly.
We develop PopIns2, a successor of the insertion caller PopIns that exploits the simultaneous analysis of multiple samples. Similar to PopIns, PopIns2 extracts unmapped reads from all individuals, assembles a non-redundant set of contig sequences across all individuals, anchors these sequences to the reference and finally genotypes against each individual. The novelty of PopIns2 is in the assembly step, where it uses a highly specialized data structure from the Bifrost API, a compacted colored de Bruijn graph. In this colored de Bruijn graph, k-mers originating from unmapped reads of a single individual are labeled with the same color. By making use of the color encoding, we aim to generate a more complete set of non-reference sequence insertions compared to PopIns, which assembles contigs per individual using Velvet prior to a multiple sequence alignment (MSA) for merging the contigs across individuals.
With the new joint assembly approach, we recover more sequence than with the MSA strategy of PopIns. Furthermore, the highly compressed data structure of the graph has a small memory footprint and is independent of alignment parameters and scoring schemes. We anticipate PopIns2 will considerably improve the detection of non-reference sequence insertions and, therefore, offer more comprehensive variant call sets for population and disease studies.

Translational regulation in aggressive B-cell lymphomas

Joanna A. Krupka, Jie Gao, Daniel Hodson, Shamith Samarajiwa

Department of Haematology, University of Cambridge, Cambridge, UK; MRC Cancer Unit, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK

High grade aggressive lymphoma, such as Diffuse Large B-cell Lymphoma (DLBCL), is one of the most genetically and phenotypically heterogeneous tumours. Despite these advances, combination chemotherapy with anti-CD20 monoclonal antibody (R-CHOP regimen) is still the first-line therapy in most cases. A fundamental challenge in the implementation of molecular diagnostic and subtype-specific treatment is to understand reciprocal dependencies of identified molecular alterations in tumour phenotype and clinical presentation.

Many studies on gene expression regulation focus primary on the transcriptome dynamics. However, because of complex post-transcriptional control of protein abundance, mRNA quantification is not always the most optimal approximation for gene expression. Recently developed techniques, such as Ribosome Profiling (Ribo-Seq), allow for high-throughput measurement of translation rate and provide deep insight into translation dynamic at sub-codon resolution.

We assessed the utility of recent, open-source bioinformatic tools for Ribo-Seq analysis, compared their performance in simulated and experimental Ribo-Seq data and optimized protocol to study translational regulation during early stages of lymphomagenesis. Our protocol achieved high reproducibility between biological replicates ($R2 < 0.97$) and allowed to capture main features of translational dynamics, such as accumulation of ribosomal footprints in CDS region, 3 nt periodicity and translation events outside annotated open reading frames.

In order to investigate if any translational alterations occur during malignant B-cell transformation, we isolated GC B-cells from fresh human tonsil tissue discarded after tonsillectomy and transduced them with 3 constructs: BCL2, BCL2 + BCL6 and BCL2 + MYC. Primary GC B- cells were co-cultured in a feeder system with irradiated follicular dendritic cells in presence of IL-21. We observed that overexpression of oncogenic transcription factors (MYC and BCL6) caused independent changes in transcriptome and translatome. It suggests that translational regulation contribute to lymphoma phenotype and may help to identify functional significance and redundancy of genomic and transcriptomic alterations.

Disruptions in SACS Gene and Likely Mitochondrial Dysfunction may lead to Autosomal Recessive Spastic Ataxia of Charlevoix- Saguenay in Consanguineous Family from Tribal J&K, India

RAJA AMIR HASSAN KUCHAY, YASIR RAFIQUE MIR

DEPARTMENT OF BIOTECHNOLOGY, BABA GHULAM SHAH BADSHAH UNIVERSITY, J&K, INDIA

Autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) is a neurodegenerative disorder characterized by late infantile onset spastic ataxia and other neurological features. Here, we present the case of a 28-year-old male from consanguineous tribal region of J&K, India, who presented with progressive ataxia, spasticity, and peripheral neuropathy with imaging features and genetic testing suggestive of SACS gene-related ARSACS. No previous medical examination was done due to lack of medical facilities in this area. Patient had cerebellar signs in the form of bilateral finger-to-nose and heel-knee incoordination, dysdiadokokinesis, and past pointing. Linear T2 hypointense striations in the pons and atrophy of superior cerebellar vermis were noted on magnetic resonance imaging (MRI) of the brain. Methylation specific PCR was carried out rule out Fragile-X Syndrome. Karyotyping was done to rule out any chromosomal aberrations. Whole exome sequencing and subsequent variant filtration releveled following mutation in exon 8 of SACS gene: c.8164delT:p.C2722Vfs*15 (NM_001278055). To further understand the consequences of SACS disruption, gene network analysis was performed in-silico. This identified alterations in genes for oxidative phosphorylation and oxidative stress with subsequent mitochondrial dysfunction. Mutations in SACS gene can be used in diagnosis of rare ARSACS. SACS might plays an important role in maintaining mitochondrial health and its mutations may lead to abnormality in oxidative phosphorylation and mitochondrial fission. Further research is needed to elucidate the pathway through which SACS disrupts mitochondrial function.

Selective single molecule sequencing and assembly of a human Y chromosome of African origin

Lukas F.K. Kuderna 1,*, Esther Lizano 1,*, Eva Julia Arteaga 2,3, Jessica Gomez-Garrido 4, Aitor Serres 1, Martin Kuhlwilm 1, Regina Antoni Arandes 4, Marina Alvarez Estapé 1, Tyler Aliotto 4, Marta Gut 4, Ivo Gut 4, Mikkel Heide Schierup 5,6, Oscar Fornas 2,3*, Tomas Marques-Bonet 1,4,7 *

1Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona, Catalonia 08003, Spain
2Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology (BIST), Carrer del Doctor Aiguader 88, PRBB Building, Barcelona 08003, Spain
3Universitat Pompeu Fabra (UPF), Carrer del Doctor Aiguader 88, PRBB Building, Barcelona 08003, Spain
4CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028, Barcelona, Spain
5Bioinformatics Research Center, Aarhus University, C.F. Moellers Alle 8, Aarhus C, Denmark
6Department of Bioscience, Aarhus University, Ny Munkegade 116, Aarhus C, Denmark
7Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Catalonia 08010, Spain

Mammalian Y chromosomes are often neglected from genomic analysis. Due to their inherent assembly difficulties, high repeat content, and large ampliconic regions, only a handful of species have their Y chromosome properly characterized. To date, just a single human reference quality Y chromosome, of European ancestry, is available due to a lack of accessible methodology. To facilitate the assembly of such complicated genomic territory, we developed a novel strategy to sequence native, unamplified flow sorted DNA on a MinION nanopore sequencing device. Our approach yields a highly continuous and complete assembly of the first human Y chromosome of African origin. It constitutes a significant improvement over comparable previous methods, increasing continuity by more than 800%, thus allowing a chromosome scale analysis of human Y chromosomes. Sequencing native DNA also allows to take advantage of the nanopore signal data to detect epigenetic modifications in situ. This approach is in theory generalizable to any species simplifying the assembly of extremely large and repetitive genomes.

Transposon Insertion Sequencing (TnSeq) in Galaxy

Delphine Larivière1, Anton Nekrutenko1, and the Galaxy Team 2

1 Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania, USA; 2 https://www.galaxyproject.org/

Identifying genes influencing bacteria growth and fitness to a medium containing antibiotics is a key factor in identifying the mechanisms of resistance. Transposon Insertion Sequencing allows the modification of gene expression across all genome by inserting promoter sequences randomly. Analysing how the transposon insertion impacts the survival and growth of bacteria provides precious data on the genes influencing the resistance of bacteria on selected media. A wide range of tools and methods can be used to analyze TnSeq data and among the variety of options available it is often difficult to identify which one provides a reliable and reproducible analysis addressing their peculiarities :
- The need to identify and curate data based on various barcodes, promoter sequences and primers on several locations of the sequenced read call for a powerful and reliable pre-processing.
- The genomic sequences left after the removal of all primers and barcodes is very short and poses alignment challenges.
In this talk, we are going to present how to perform a reproducible and reliable analysis of TnSeq data in Galaxy, including data de-multiplexing, promoter sequence identification, alignment of transposon flanking region to a reference genome and read count repartition across the genome.
We are also going to discuss the quality control in the process: how to determine the best parameters for the analysis, filtrate non-relevant data, and check for errors.
We use a set of antibiotic resistance TnSeq analysis to show how to identify growth advantage and growth inhibition from transposon insertion, and how to use this information to link genes to antibiotic resistance.

A graph-based framework for unified identification of short and structural genetic variants in whole-genome sequencing data

Dillon Lee, Yi Qiao, Andrew Miller, Alistair Ward, Gabor Marth

University of Utah

Several state-of-the-art, easy to use tools are available both for short-variant detection (e.g. GATK, FREEBAYES), and structural variant (SV) detection (e.g. LUMPY, MANTRA, DELLY), but these tools often produce divergent variant calls, especially INDELs, and it is very difficult to reconcile such variants into a single, accurate set. Furthermore, while it would be highly desirable to also detect larger, structural variants (SV), existing SV detector packages are typically difficult to integrate, highly resource-intensive to run, and result in call sets that require expert manual review to reduce false positive detection rate.

Our algorithm, GRAPHITE (https://github.com/dillonl/graphite) requires as input a collection of variant calls, made by one or more short-variant or SV detection tools. Typically, this starting set is high sensitivity (i.e. inclusive), but low specificity (i.e. have a high false discovery rate). We then apply a novel "variant adjudication" procedure to discard false positives, while keeping true positive calls. This is accomplished by constructing a graph from these variants representing allelic variants as graph branches, in addition to the branches formed by the current, linear genome reference sequence. Using a graph mapping algorithm (GSSW, a graph extension of the Smith-Waterman alignment algorithm) we developed earlier, we re-map all reads from each of the samples contributing to the candidate calls. We retain candidate variants confirmed by mappings to those branches in the graph that represent the corresponding variant allele, and discard those candidates that were not confirmed by such mappings. This procedure results in a highly specific callset that also maintains the high sensitivity of the inclusive starting callset constructed by multiple primary variant calling methods. Because the graph construction and mapping approach works for most types of SVs in addition to all short variants, variants of all different types can be integrated in a single step.

Here we present the application of this method for validating somatic variants sampled from a patient at four time points in a longitudinal tumor dataset. GRAPHITE uses these four time point samples along with the normal sample to improve allele frequency measurements across all samples. Allele frequency improvements not only aid in identifying somatic variants, they also enhance the ability to detect low frequency variants at earlier time points allowing for more accurate reconstruction of the tumor phylogeny.

Functional analysis of polymorphic inversions in the human genome

Jon Lerga-Jaso1, Sergi Villatoro1, Marta Puig1, Alejandra Delprat1, Mario Cáceres1 2

1Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain; 2Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Advances in genomic techniques have generated an increasing interest in structural variants. However, there is limited information on their functional impact. This is particularly true for inversions, whose study is still challenging due to their balanced nature and the complex regions where they appear. To address this problem, we took advantage of a large genotyping project to carry out a complete bioinformatics analysis of the potential consequences of 45 human polymorphic inversions. First, we applied three different approaches to detect gene-expression changes associated to inversions in lymphoblastoid cell lines and maximized the concordance among the pipelines, finding eight inversions that may influence the transcription of neighboring genes. We also looked for additional inversion effects in other tissues through linked SNPs already reported as eQTLs in the GTEx project. This analysis confirmed our previous results and identified seven more candidate inversions that may affect gene regulation. In total, we could identify ten inversions as top markers of these associations, indicating that these could be indeed the cause of the expression changes. One inversion with a clear molecular impact is HsInv1051, which causes the disruption of the pseudogene CCDC114B and creates a novel fusion transcript with new 3' sequences after losing a premature stop codon. In addition, by analyzing alternative eQTL collections, we uncovered condition-specific expression changes as those associated to HsInv0201 in infection. Second, we found a marginally significant enrichment of GWAS signals in the inversion regions, supporting their possible implication in disease. Specifically, we discovered three inversions driving this result. An example is HsInv0030, which creates hybrid transcripts of CTRB1 and CTRB2 genes, and is consistently enriched in diabetes and pancreatic cancer signals. Moreover, seven inversions were in high LD ($r2 \geq 0.8$) with trait-associated hits. A good candidate is HsInv0031, which is associated with FAM92B expression in cerebellum and is linked to a SNP conferring risk for Alzheimer's disease. Thus, the extension of the genotyping effort to a higher number of inversions currently underway will help us get a more global idea of the potential functional role of inversions on the human genome and reveal previously missing variants related to phenotype variability.

Variant identification in whole exome sequencing of patients with adult onset hearing loss

Morag A. Lewis (1,2), Lisa S. Nolan (3), Barbara Cadge (3), Lois J. Matthews (4), Bradley A. Schulte (4), Judy R. Dubno (4), Sally Dawson (3), Karen P. Steel (1,2)

1) Wolfson Centre for Age-Related Diseases, King`s College London, SE1 1UL, UK
2) Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK
3) UCL Ear Institute, University College London, WC1X 8EE, UK
4) The Medical University of South Carolina, SC, USA

Hearing loss is one of the most common sensory deficits in the human population, and it has a strong genetic component. However, although to date more than 140 loci relating to human hearing loss have been mapped, and over 100 genes identified, the majority of genes involved in hearing remain unknown. It is also unclear whether adult-onset hearing loss results from rare Mendelian gene variants with large effect size or multiple variants each making a small contribution to hearing loss. Several genome-wide association studies have been carried out, but only five loci have been associated with hearing status at the genome-wide significance level. We have therefore chosen to carry out whole exome sequencing to explore the landscape of variation associated with adult-onset hearing loss. However, variant identification is a challenge in a common complex diseases. In particular, while the recommended minor allele frequency against which to filter is limited to 0.5% (for a recessive gene), some of the variants known to cause hearing loss are more common than that. We have carried out a pilot test on 30 exomes from patients with hearing loss, filtering them based on quality, minor allele frequency, predicted consequence and predicted severity of impact. We compared the number of genes with predicted pathogenic variants to a list of genes known to be associated with deafness in mice and humans. We found multiple variants in known deafness genes in both our patient population and the 2504 individuals from the 1000 Genomes study, and the distribution of deafness genes remained similar between the two populations across a range of minor allele frequency, consequence and pathogenicity filters. This was an unexpected finding, and has significant implications for current diagnostic sequencing in deafness as well as for gene discovery research. We are proceeding with exome sequencing analysis on a further 532 exomes, including 78 from older people with normal hearing, which are a more appropriate control population.

Assessment of the predictive accuracy of prediction tools on missense variants in CFTR

Lonishin LR, Serebryakova EA, Nasykhova YA, Glotov AS

Peter the Great St.Petersburg Polytechnic University, St.Petersburg, Russian Federation; D.O.Ott Research Institute of Obstetrics, Gynecology, and Reproductology, St.Petersburg, Russian Federation

Missense mutations of evolutionarily conserved amino acids are not necessarily deleterious so there is a difficulty for the interpretation of rare missense variants even in such a studied gene as CFTR. Functional studies in relevant model systems can determine the significance, but it is not always possible to implement them in a short time. As a result, the use of prediction tools is currently very important and should be reasonably accurate.
Three pathogenicity prediction programmes freely available on the web were used to determine their ability to correctly predict the impact of a missense variant on CFTR protein function. We selected all variants with certain clinical significance from ClinVar database. To evaluate the performances of the programmes, seven measures (sensitivity, specificity, accuracy, precision, positive predictive value (PPV), negative predictive value (NPV), and Matthews correlation coefficient (MCC)) were calculated by comparing the results of all programmes with previously generated functional data. After that, all mutations with inaccurate prediction were located at exons and modeling structure programme.
Three in silico prediction programmes (SIFT, PolyPhen2, PROVEAN) freely available on the web were used to determine their ability to correctly predict the impact of a missense variant on CFTR protein function. We selected all variant with certain clinical significance from ClinVar database. To evaluate the performances of the programmes, sensitivity, specificity, accuracy, and MCC were calculated by comparing the results of all programmes with previously generated functional data. After that, all mutations with inaccurate prediction were located at exons and protein modeling structure programme (Phyre2) were performed on this mutations.
None of the prediction programmes were able to identify all of the variants tested correctly as either 'damaging' (CF-causative variants) or as 'benign' (common sequence variants). The overall sensitivity of predictions ranged from 66% to 89% depending on the programme used, with specificity from 39% to 66%. The MCCs were between 0.2 and 0.4, which is described as weak and moderate positive relationship.
According to this research, Provean has shown the better result in prediction neutral mutations than other programmes. There were no significant differences between CFTR domains structure. We suggest that the usage of these prediction programmes should be done in a combination with other predictors, nevertheless the accurate prediction of missense mutation can't be guaranteed.

Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project

Ernesto Lowy-Gallego, Susan Fairley, Holly Zheng-Bradley, Laura Clarke, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, United Kingdom

The latest build of the human genome (GRCh38) constitutes the best representation of the human genome to date. GRCh38 has been adopted as the reference assembly and, to enable the research community to take full advantage of the improved reference, it is necessary for genomic resources to make their data available on this assembly.

The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and one of the main objectives is to move the data from phase three of the 1000 Genomes Project to GRCh38, including the development of the computational pipelines needed to achieve this and directly recall variants on GRCh38.

As a part of this work we have already mapped the 1000 Genomes sequence data to GRCh38 using the alt-aware BWA alignment algorithm. These alignments provide the foundation for generating variant calls directly on GRCh38.

We are using three established methods (BCFtools, Freebayes and GATK UnifiedGenotyper) to identify biallelic SNPs in the autosomes (chr1-22) and in the PAR regions for chrX. A consensus callset is being generated by the union of the sites from each callset and by the calculation of the genotype likelihoods for each site using GATK UnifiedGenotyper. We are then filtering the spurious variants using Variant Quality Score Recalibration (VQSR). The consensus approach enables us to take advantage of the strengths of each method and thus increase the sensitivity in the final callset.

The calculated genotype likelihoods are being integrated with microarray genotype data available on the same samples. These data are used to create a highly accurate haplotype scaffold by leveraging family information and then the sites in our consensus callset are phased onto this scaffold. This same strategy was used in the phase three of the 1000 Genomes Project and has been shown to produce low error rates for genotype calls.

We will release the biallelic SNP call set as soon as it is complete and plan thereafter to build on this work, increasing the range of samples in IGSR with variation calls generated directly against GRCh38.

A systematic interrogation of technological artefacts in whole exome sequence data

Juliet Luft, Robert S. Young, Martin S. Taylor

Medical Research Council Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, Scotland, UK

In recent years, multiple large-scale genomic analyses have been performed using whole exome sequencing (WES). Like all sequencing technologies WES is vulnerable to various technological artefacts, resulting in recurrent sequencing errors. Whilst these issues have been documented, the downstream errors are not always accounted for during data processing and often persist in variant repositories and databases. Accordingly, we present a systematic interrogation of WES data from The Cancer Genome Atlas (TCGA).

During preliminary analysis we identified multiple samples with evidence of non-self contamination. Contaminated samples were characterised by high numbers of heterozygous variants with extreme B-allele frequencies (BAF) in both the tumour and matched normal sample, and a prominent left-shift of rare variants within the BAF distribution.

Analysis of common heterozygous loci across patients found a subset of variants with high mean relative read depth, and low mean BAF - furthermore, these variants tended to appear heterozygous within our population more often than expected based upon population allele frequency estimates. These variants are likely to represent common sequencing errors and mismapped reads, often appearing in mono-nucleotide runs and loci with high sequence identity to other regions of the genome.

Whilst subsequently investigating batch effects, we found heterozygous variants that were enriched within sequence data generated using the same exome target capture kit. Using a linear regression model, we separated out kit specific biases from cancer sub-type specific batch effects. Consequent analysis of batch-specific variants revealed an enrichment of T>G mutations downstream of a 7bp motif, likely to be an artefact generated during library preparation or read alignment. Interrogation of motif-linked variants in public databases found them to have higher population allele frequencies in exome data, compared to genome data, indicating that this is a pervasive artefact within exome sequencing.

Our analysis revealed many artefacts in WES data, as well as potential contamination of 1.4% (n = 140) of tumour-normal pairs in TCGA. Comparison of our results revealed persistence of these artefacts in online variant repositories. This work hence demonstrates a requirement for more vigorous downstream analysis of WES data.

NIH Data Commons Pilot Phase Consortium (DCPPC) Crosscut Metadata Model: A generalized model to represent genomic datasets in the Commons

Anup Mahurkar, Jonathan Crabtree1, Suvarna Nadendla1, Alejandra Gonzalez-Beltran4, Philippe Rocca-Serra4, Victor Felix1, Olukemi Ifeonu1, Michelle Giglio1, Carl Kesselman2, Ian Foster3, Susanna-Assunta Sansone4, Owen White1

1 Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD
2 Keck School of Medicine, University of Southern California, CA
3 Department of Computer Science, University of Chicago, IL
4 Oxford e-Research Center, Department of Engineering Science, University of Oxford

The NIH Data Commons (https://commonfund.nih.gov/commons) will accelerate biomedical discovery by providing a cloud-based platform where investigators can store, share, access, and compute on digital objects including data, software, workflows, and more. Particular attention will be paid to ensure that the resources produced by the project will adhere to FAIR principles. The initial implementation will be organized around a set of targeted high-value data sources that will serve as test cases. The test datasets will come from the Genotype-Tissue Expression (GTEx) Project, the Trans-Omics for Precision Medicine (TOPMed) Program, and several model organisms from the Alliance of Genome Resources (AGR). In addition to seeding the Commons, these test datasets will allow us to understand the challenges in bringing such diverse datasets together and will inform the inclusion of other datasets in the Commons in the future. This pilot project also serves the scientific goal of enabling data discovery and hypothesis generation by permitting queries across these diverse datasets.

These data, particularly their metadata, are stored natively in different formats, and often using different vocabularies for the same concept. Thus, any team wanting to enable search across multiple datasets must undertake transformation and harmonization tasks. To address this, we are developing a Crosscut Metadata Model (C2M2) based on the Data Tag Suite (DATS) (https://github.com/datatagsuite) to define a standardized representation of diverse metadata, and Stage 1 Metadata Instances, products that use the C2M2 to render these diverse metadata in a unified, and progressively harmonized form. We package the Metadata Instance as a BDBag (http://bd2k.ini.usc.edu/tools/bdbag/), a common exchange format. DCPPC Full Stack teams (cloud service providers), and other data consumers, can then import the contents of this Metadata Instance into their indexing and search infrastructure, avoiding duplication of effort across the DCPPC and encouraging a consistent basis for interoperation. The DATS-based metadata instances and scripts to generate the instances from various sources can be found at https://github.com/dcppc/crosscut-metadata.

Tissue specific gene expression patterns and chromosome organization in chicken

David Martín-Gálvez (1,2), Duncan Odom (3) and Paul Flicek (2)

(1) Complutense University of Madrid, 28040, Spain
(2) European Molecular Biology Laboratory, European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
(3) University of Cambridge, Cancer Research UK - Cambridge Institute, Li Ka Shing Centre, Cambridge, CB2 0RE, UK.

The causes behind chromosomal organization and their effects on the evolutionary process are not fully understood yet. Birds are an interesting group to address such questions as they have a typical karyotype that is fairly conserved across the evolution of the group. The typical avian karyotype consists of 6-7 pairs of macrochromosomes, a pair of sex chromosomes and 30-32 pairs of microchromosomes. The existence of mechanisms preventing abrupt changes in the karyotype of birds have previously been suggested. One possibility is that the different types of chromosomes in birds provide the specific genomic environment to ensure an optimal level of expression necessary for the type and function of the genes that they contain. This could be especially relevant for those genes requiring similar regulation, such as tissue-specific genes and house-keeping genes. We address this question in chicken by comparing at the chromosomal level the tissue specificity (using Shannon entropy) of existing gene expression (RNA-seq) data from 21 tissues. Chicken macrochromosomes had smaller values of gene entropies (i.e. variable gene expression among tissues) and less tissues-specific genes (genes within decile 1) than expected. In contrast, chicken microchromosomes show greater gene entropies (i.e. uniform gene expression among tissues) and more house-keeping genes than expected. Sex chromosomes had smaller values of gene entropies and were positively enriched with tissues-specific genes. Enrichment analysis for Gene Ontology (GO) terms established a set of significant genes, mainly housekeeping genes in microchromosomes. Our results suggest a link between the morphology of chromosomes and regulation of the expression of their genes. We are now investigating possible mechanisms underpinning this link.

Accurate and complete genome assembly for the Vertebrate Genomes Project

Shane A McCarthy1,2, Arang Rhie3, Martin Pippel4, Olivier Fedrigo5, Sergey Koren3, Maria Simbirsky6, William Chow2, Jo Wood2, Iliana Bista2, Zemin Ning2, Harris Lewin7, Kerstin Howe2, Erich D. Jarvis5, Gene Myers4, Richard Durbin1,2, Adam Phillippy3 on behalf of the VGP Assembly Group

1University of Cambridge, Cambridge, UK; 2Wellcome Sanger Institute, Hinxton, UK; 3NIH, NHGRI, Bethesda, MD, USA; 4Max Planck Institute, Dresden, Germany; 4The Rockefeller University and HHMI, New York, NY, USA; 6DNAnexus, Mountain View, CA, USA; 7UC Davis, Davis, CA, USA

The Genome10K Vertebrate Genomes Project (VGP) is an international effort to create at least one high-quality, near-gapless, phased and annotated chromosomal-level assembly of all extant vertebrate species. Phase 1 of this project is focused on finishing one species from each vertebrate order, totaling 260 individual species, to a quality standard of >1 Mb N50 contig size, >10 Mb N50 scaffold size, average base quality >QV40, and 90% of the sequence assigned to chromosomes. This has been enabled by the maturation of long-read sequencing and long-range scaffolding technologies. The VGP has been collecting and sequencing ordinal samples using four such emerging technologies: PacBio long reads, 10X Genomics linked-reads, Bionano optical maps, and Arima Genomics Hi-C libraries.

The VGP assembly working group has been comparing and evaluating sequencing and assembly strategies using an initial set of 16 species including mammals, birds, fishes, a reptile and an amphibian. The current assembly process involves contig generation using PacBio, followed by scaffolding using 10X Genomics, Bionano, and Hi-C. Draft assemblies are then evaluated for correctness using the gEVAL platform and genome-to-genome alignments.

An improved, comprehensive assembly strategy is under continued development, including new methods aimed at better separation of haplotypes by integrating additional information from the scaffolding technologies and pedigree data. In parallel to the ordinal level project aimed at creating a resource with phylogenetic breadth, some deeper, clade-specific projects using subsets of the technologies are also underway aimed at more specific scientific questions within these groups.

All initial assemblies are being submitted to the public sequence archives, and all raw data will be released as it is generated in coordination with DNAnexus and Amazon Web Services (http://genomeark.s3.amazonaws.com). As of July 2018, we have submitted 5 vertebrate genome assemblies, with a further initial batch of ~10 new ordinal genomes to be completed over the summer.

Mapping Genes to Proteins in UniProtKB

Alistair MacDougall1, UniProt Consortium1-4, Ensembl1

1 EMBL-EBI, Cambridge, UK; 2 Swiss Institute of Bioinformatics, Centre Medical Universitaire, Geneva, Switzerland; 3 Protein Information Resource, Georgetown University Medical Center, Washington, USA; 4 Protein Information Resource, University of Delaware, Newark, USA.

Connecting the genome and proteome worlds is critical to enable the mining of genomes, for instance to identify functional variants which have disease consequences.

UniProt provides proteome sets for species with completely sequenced genomes. To make effective use of these data for genome studies, it is essential to have accurate mapping from gene to protein sequence, and in the reverse direction from protein to gene.

Many gene:protein mappings are already established, arising from the existing close relationship between Ensembl and UniProtKB, where UniProt proteins are included in the gene building pipeline in Ensembl, and newly reported protein sequences from Ensembl genes are routinely incorporated into UniProtKB as predicted sequences. However, for many genes and proteins, even in well studied organisms such as human, the mapping is not 1:1 and there are discrepancies between genome and proteome.

UniProt and Ensembl are establishing a detailed mapping between nucleotide and protein data, initially starting with the human genome, and extending to other organisms. This will lead to the public availability of reference gene/transcript/protein sets from Ensembl and UniProtKB of whole genomes that remain synchronized through changes in the sequence model or annotation.

Data flow within the Multiscale Genomics Virtual Research Environment for 3D Genome Analysis

Mark McDowall, Pablo Acera, Reham Fatima, Andrew Yates

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

The Multiscale Genomics Virtual Research Environment (VRE) is a resource to investigate the structure of chromosomes by integrating multiple sources of biological knowledge concerning DNA proximity, locations of genes, expression, nucleosomes, promoter regions and methylation sites. The VRE is open to the scientific community and supports experimental data upload, analysis and visualisation using tools, such as TADbit, nucleR and pyDock. APIs facilitate tool integration and data management. For example, the Tools API wraps individual tools in a common format and enables a common interface to the VRE via a web interface that matches relevant input data and appropriate tools.

Files can be held in different locations within the VRE (currently the Barcelona Supercomputing Centre and EMBL-EBI). Tracking file locations is crucial so that pipeline execution is performed on compute clusters close to the location of the data for computational efficiency and to minimise transfer. The Data Management API handles the tracking of files, data types and matching metadata to describe the information within the file, including how it was generated and the source files that were required for its creation. Clear descriptions about how experimental data was generated and analysed facilitate publication reproducibility and helps when submitting to data archives.

RESTful APIs allow access from multiple servers for analysis and visualisation, while indexing and storage optimisations create a seamless experience when searching through the results.

The VRE is accessible at http://vre.multiscalegenomics.eu with limited default storage or, when logging in using a Google or LinkedIn account, with 20GB of space. From the VRE homepage there are tutorials and help documents for all of the tools and pipelines. Code is available from https://github.com/Multiscale-Genomics under an Apache 2.0 license. Documentation is also available to help developers integrate their own tools within the VRE framework (http://multiscale-genomics.readthedocs.io).

SeeGEM: Lightweight Interactive Visualization of DNA Sequence Variation Prioritization

David McGaughey, David McGaughey

National Eye Institute (NIH)

Identification of genetic variation causing human disease in a proband is a complex process; raw DNA sequence must be aligned to a reference genome, genotypes called from the aligned reads, annotations (e.g. gnomAD, CADD) appended, and family structure (proband, mother, father) delineated. The most powerful and popular program for doing the filtering and inheritance pattern testing is GEMINI. However even after running GEMINI to filter for DNA variants passing quality control metrics, match proper inheritance patterns (e.g recessive, dominant), and have sufficiently low allele frequency in healthy population(s), a proband will still have dozens to hundreds of possible variants that an analyst must hand assess. An analyst must balance several different metrics (protein consequence, in silico pathogenicity, etc.) as well as assess whether the sequencing for the trio or cohort has passed technical quality controls. As GEMINI is a command line tool, the output is a dense display of tab separated values filling the computer screen. To more efficiently aid this manual curation, I have written the R package SeeGEM, which wraps the output from GEMINI into a reactive html document. These portable (no internet required) and compact (usually <2 megabytes) files can be used in any device running a modern web browser and can also incorporate crucial sequencing quality control metrics from the peddy tool. The reactive tables can be labeled with color to delineate pathogenicity, the columns and rows can be re-ordered and searched within, and several columns data types can be turned into URLs to facilitate searches of variant information in web resources like OMIM, gnomAD, dbSNP, and Google Scholar. Shortcuts in the header allow for user-specific sets of annotations, which is crucial given that hundreds of annotations are possible for each DNA variant. The combination of GEMINI, peddy, and SeeGEM provide similar functionality to tools like Seave and VarApp without requiring complicated installation procedures and maintenance of web servers/apps. SeeGEM can be installed and run with only a few commands in R. It has been engineered to be easily be integrated into existing DNA analysis pipelines.

Distill, a random forest and deep neural network based ensemble learner, provides class leading DNA variant pathogenicity prediction

Dr David McGaughey, David McGaughey

National Eye Institute (NIH)

Identification of pathogenic variant(s) from human exome or genome sequencing (WGS) is a crucial and difficult task. With the dramatic reduction in sequencing costs, it is an increasingly common task. Even after filtering for rare coding and splicing variants, numerous prioritized variants remain for most patients. Current prioritization strategies use a combination of in silico predictions and knowledge of likely deleterious genes for the condition, along with large variety of scoring metrics. Popular metrics include conservation and predicted functionality-based scores like GERP, SIFT, GERP, CADD, and REVEL. To better guide variant analysis, we use a novel and broad dataset of rare and richly annotated variants to train a mendelian disease DNA variant pathogenicity model. We curated a machine learning input dataset including a high quality ClinVar pathogenic/benign dataset, 425 solved retinal degeneration WGS cases, and rare variants from gnomAD. The hundreds of thousands of variants were richly annotated with hundreds of variant metrics including constrained coding regions (ccr), ENCODE epigenetic data, GTEx gene expression data, and other pathogenicity metrics (e.g. REVEL, FATHMM, MetaSVM). The advantage of this strategy is the use of a broad, curated set of rare benign variants and a richer set of annotations. Our dataset, when used with a random forest model to predict DNA variant pathogenicity, highlights ccr, the ExAC missense and loss of function Z-scores, PhyloP, CADD, and gnomAD population metrics as the most crucial scores. We then trained a deep LSTM neural network model and merged it with the random forest model to create an ensemble learner, Distill, with class leading pathogenicity prediction.

Choose your k-mers wisely: a new approach to indexing a large collection of genomes

Páll Melsted 1, Guillaume Holley 1, Lior Pachter 2

1. Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland. 2. Departments of Biology and Computing & Mathematical Sciences, California Institute of Technology

Sketching k-mers from large collections of genomes using Mash,was a breakthrough that allowed researchers to quickly compute relatedness from a large collection of genomes. One of the interesting upsides was being able to compute this information from Short Read High Throughput Sequencing data in order without the need for a separate assembly.

Rather than working with distance, i.e. answering the question "How far is my readset from each genome", we look at the releated containment question "Which genomes are present in the readset". Recently Mash and sourmash have tackled this question using the MinHash index developed for Mash. In this work we develop a new indexing based structure we call mindex, for selecting a small subset of k-mers for indexing a large collection of genomes. Unlike the MinHash based data structure, we require the entire set of genomes to be available in order to perform the indexing.

The selection of k-mers is reduced to a bipartite graph problem and the resulting index is computed solely from this graph. The detection of containment is performed by matching k-mers from the index from the readset and solving a related graph problem to detect genomes present.

The index construction runs in 10 minutes for 8726 complete bacterial genomes and detects genomes present in less than 3 minutes for a simulated dataset of 26M reads. Additionally we benchmark the performance on related strains of 400 E. coli genomes showing that the mindex can distinguish readsets better than Mash.

Deep Hi-C Multi task model for detecting A/B compartments, TADs, hubs and long range interactions

Tanmoy Mukherjee, Oisin Faust, Dóra Bihary, Shamith Samarajiwa

MRC Cancer Unit, University of Cambridge, Hutchison-MRC Research Centre, Cambridge Biomedical Campus, Cambridge, UK.

Gene and genome regulation is modulated not only by DNA binding transcriptional regulatory proteins but also by complex changes in the epigenome and the chromatin landscape. The advent of chromosomal conformation capture methods such as Hi-C has enabled the study of chromatin higher order structure and long range interactions at a previously unimaginable resolution. While Hi-C can generate genome-wide contact frequency maps that facilitate the detection of A/B compartment, Topologically Associated Domains (TADs) and Chromatin Loops, the resulting large data-sets are computationally challenging and novel methods and approaches are needed for extracting biologically important features.

Recently Deep Neural Network methods have achieved great success in several disciplines ranging from Image Analysis, Natural Language Processing to Computer Vision. In particular, Deep Learning models such as Convolution Neural Networks show great promise applied to certain types of problems in both computational biology and genomics. While most Deep Learning solutions are targeted to solve a single task, some related tasks often share common characteristics and can be jointly solved. In this work we show some exciting and promising directions of extending Convolution Neural Networks in a multitask learning framework towards detecting A/B compartments, TADs and loops. In particular our model is a joint model which is able to learn on all three tasks together and provide a means of extracting biologically important epigenomic and chromatin features.

Reappraising the protein-coding gene count in GENCODE

Jonathan M. Mudge1, Toby Hunt1, Michael Tress2, Laura Martinez2, Irwin Jungreis3, Paul Flicek1, Adam Frankish1

1 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom; 2 Bioinformatics Unit, Spanish National Cancer Research Centre, Madrid, Spain; 3 MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA

While the question 'how many protein-coding genes are there in the human genome?' may seem like a legacy of the early 2000s, it remains unanswered - and indeed hotly debated - to this day. There are two reasons why current gene annotation catalogs may contain imprecise counts: firstly, additional protein-coding genes could remain to be discovered; secondly, existing coding annotations may be incorrect.

We produce the Ensembl / GENCODE reference genesets for human and mouse as part of the Ensembl project, and a key focus of our current efforts is to provide more accurate annotation of coding sequences (CDS). Here, we will discuss the creation of a multi-faceted workflow that allows us to identify prospective novel CDS as well as potentially dubious existing annotations, and its implementation in human. This workflow incorporates experimental data from transcriptomics and proteomics assays, evolutionary analyses based on PhyloCSF and comparative annotation, and a consideration of the spectrum of genetic variation in human populations.

In recent months we have added 144 high-confidence protein-coding genes to the human GENCODE geneset, as well as additional coding annotations to over 300 existing loci. The majority of these annotations are truly novel. Conversely, our in-progress survey of dubious protein-coding genes has found nearly 200 loci that appear unlikely to produce genuine, functional protein products. While these changes may appear modest in number, we demonstrate that they could have dramatic consequences for genome interpretation. Most importantly, we are able to present new interpretations for dozens of variants linked to diseases or traits.

Inclusion of pseudones in the Ensembl Comparative Genomics resources

Matthieu Muffato, Guillaume Giroussens, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Pseudogenes are segments of DNA that are related to functional genes but have lost functionality, often from the accumulation of multiple mutations. Pseudogenes can thus be found and annotated using sequence similarity, and linked to a parent or source gene, giving us insights into the history of this gene. However, very few resources consider pseudogenes when running comparative genomics analyses, and pseudogenes are largely missing from the major orthology databases.
Ensembl is a platform that provides integrated genomics resources of more than 100 vertebrate species and a comprehensive comparative genomics database. Currently, this includes phylogenetic trees and orthology calls across all functional genes. In this work, we extend the Ensembl resources to include pseudogenes at multiple levels to help understanding their evolution.
First, we link the pseudogenes to their closest functional homologue with PseudoPipe (Zhang et al., Bioinformatics, 2006), an existing homology-based pseudogene identification pipeline. Then we update the multiple-sequence alignments and phylogenetic trees of their functional counterparts by constraining the original alignment and topology. We are thus able to supplement our orthology predictions with pseudogenes-to-functional orthologues (such as unitary pseudogenes), and between-pseudogenes orthologues. Finally, we run our quality-assessment analyses based on conservation of local gene order and congruence with whole-genome alignments.
We will present the results and some statistics of this approach on a test dataset comprising human and rodents, giving insights into the recent evolution of pseudogenes. We will also present prototype Ensembl comparative genomics displays (phylogenetic tree, orthologue and paralogue lists) that include pseudogenes, and we are seeking feedback before releasing the new data in a future version of Ensembl.

Applying Apache Spark to Genomics at Industrial Scale

Frank Austin Nothaft, Ram Sriharsha, Henry Davidge

Databricks

As the size and scope of genomics datasets continue to grow, many researchers are building large genomics analyses using the Apache Spark distributed computing framework. This is reflected in the adoption of Apache Spark by large open source projects like the GATK4, Hail, and BDG ADAM. While Apache Spark reduces the effort of implementing parallel genomic analyses, it is still complex to use Apache Spark across large cohorts.

In this talk, we will provide a brief survey of genomics libraries that exist on Apache Spark, and will present lessons from deploying Apache Spark in a cloud framework across industrial scale datasets including variation data across hundreds of thousands of exomes, short read data across tens of thousands of whole genomes, and large scale transcriptomics datasets. From these large scale case studies, we will focus on opportunities for improving the performance and productivity of these cohort scale analyses, while also illustrating where machine learning techniques can be applied to large cohorts of genomic data.

Building on the lessons from these large scale analyses, we have identified a number of optimizations for manipulating genomic data in the cloud using Apache Spark. These optimizations span across cloud data access, as well as optimizing the Spark SQL engine to better support genomics-specific query patterns. With these optimizations, we have been able to improve performance by >50x over the state-of-the-art, enabling significant reductions in end-to-end analysis time.

Annotation of Cats in Ensembl

Denye Ogeh, Konstantinos Billis, Carlos García Girón, Leanne Haggerty, Thibaut Hourlier, Osagie Izuogu, Fergal J. Martin, Paul Flicek

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

The recently redesigned Ensembl gene annotation system is a major part of our ambition to transform vertebrate biology by providing genome resources for thousands of species. With processing times for gene sets reduced to a matter of days, our methods support the rapid inclusion of new species in Ensembl while maintaining our standard for high-quality annotations with a large number of secondary analysis tracks.

Ensembl is based on a combination of RNA-seq alignments, annotation projection via whole genome alignments and protein-to-genome alignments using selected UniProt proteins. This approach allows us to perform clade-based annotations with consistency and efficiency.

Using our methods, we carried out a simultaneous annotation of domestic cat (Felis_catus_9.0), tiger (PanTig1.0), and leopard (PanPar1.0). After an initial annotation of repeat features, CpG and other simple genomic features, we created both comprehensive protein coding and non-coding gene sets. These robust annotations should help to understand the evolution of the cat family.

Where available, annotated genomes include RNA-seq data, which can be viewed on the Ensembl genome browser. Data and tools to facilitate research on feline genomes and other species are accessible through our website (www.ensembl.org), REST API (http://www.ensembl.org), REST API (http://rest.ensembl.org), the Ensembl Variant Effect Predictor (www.ensembl.org/Tools/VEP), BioMart (http://www.ensembl.org/biomart) and our public MySQL server (ensembldb.ensembl.org).

The gEAR portal v2: now integrating single cell RNA-seq, epigenomics and data analysis tools

Joshua Orvis[1], Dustin Olley[1], Jayaram Kancherla[2], Beatrice Milon[1], Yang Song[1], Hector C. Bravo[2], Anup A. Mahurkar[1] and Ronna Hertzano[1]

Affiliations: [1]University of Maryland School of Medicine-Baltimore; [2]University of Maryland-College Park

The gEAR portal (gene Expression Analysis Resource, umgear.org) is an online tool for multi-omic and multi-species data visualization, sharing, and analysis. Originally designed for auditory and vestibular researches, the gEAR portal has now been expanded for general use. The gEAR is unique in its ability to allow users to upload, view and analyze their own data in the context of previously published datasets, as well as confidentially share their data with collaborators prior to publication. It is also unique in combining not only multiple species but multiple data types including bulk RNA-seq, sorted cell RNA-seq, single cell RNA-seq (scRNA-seq) and epigenomics in a one page, user-friendly, browseable format. Individual expression datasets can be displayed in a variety of ways alongside each other, including interactive bar, line or violin plots, and colorized anatomical SVGs.

Most recently, scRNA-seq has matured to a commonly used technique for measuring gene expression across tissues. To provide researchers access to scRNA-seq data regardless of their programming knowledge, we have integrated a scRNA-seq workbench into the gEAR. The gEAR scRNA-seq workbench provides access to both the raw data of scRNA-seq datasets, as well as to saved expert analyses where cell types have already been assigned – giving researchers rapid insight into gene expression of their cell type of interest. This presentation functions as a step-by-step introduction to the gEAR portal, now a mainstream multi-omic data source for the ear research community.

Multi-omics approaches towards personalised medicine for rare diseases.

Georg W. Otto, Daniel Kelberman, Xueting Wang, Andrey Gagunashvili, Hamzah Syed, Rosalind Davies, Janna Kenny, Louise Ocaka, Lamia Boukhibar, Hywel Williams, Jochen Kammermeier, Christopher J. Piper, Meredyth G. Wilkinson, Claire T. Deakin, Lucy R. Wedderburn, Chiara Bacchelli and Philip L. Beales

National Institute for Health Research Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust
University College London Institute of Child Health, London, UK

Research in Mendelian diseases is primarily concerned with the discovery of pathogenic genetic variants and the mechanisms of disease causation, providing entry points for therapeutic interventions. However, patients affected by monogenetic diseases, including those carrying the same primary genetic defect, can markedly vary in their expression of clinical phenotypes or response to therapies. This suggests an important role of genetic variation and non-genetic factors in causing a considerable phenotypic diversity. To understand the underlying molecular mechanisms of this is crucial for disease classification and clinical management.

Functional genomics provides genome-wide assessments of molecular phenotypes, and the analysis of multiple classes of molecules (multi-omics) targets the complex interactions between the different levels of gene regulation in its cellular context. Hence, a multi-omics based investigation of disease and control cohorts has the potential to yield detailed insights into disease mechanisms and their variation between individuals.

The GOSgene initiative at the UCL Institute of Child Health is investing in performing multi-omics studies of rare pediatric diseases. In addition to utilising genome sequencing for the identification of disease causing variants, we are developing functional genomics approaches and combine these with the use of clinical information. As part of initial pilot studies in a select number of rare diseases we have used multiple technologies including whole genome and RNA sequencing, mass spectrometry and targeted proteomic assays. The resulting data provide in-depth molecular phenotypes in conjunction with detailed clinical characterisations and genetic variation.

Initial investigations of individual datasets in specific disease cohorts using differential transcript and protein expression and various pathway analyses have highlighted a number of cellular processes involved in disease pathology. In Juvenile Dermatomyositis, comparing pre-treatment patients, on-treatment patients and healthy controls identified a strong IFNα signature in isolated B-cells, with molecules downstream of IFNα receptor signalling being highly expressed in pre-treatment patients versus controls. Likewise, in Very-Early-Onset Inflammatory Bowel Disease, signatures of inflammation were identified on the level of mRNA- and protein-abundance. Integrating these data into multi-omics analyses will enable a better understanding of disease mechanisms.

Population and allelic variation of A-to-I RNA editing in human transcriptomes

Eddie Park1, Jiguang Guo2, Shihao Shen1, Levon Demirdjian3, Ying Nian Wu3, Lan Lin1 and Yi Xing1

1 Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA.
2 Department of Microbiology & Parasitology, Medical School of Hebei University, Baoding, Hebei Province 071002, China.
3 Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095, USA.

A-to-I RNA editing is an important step in RNA processing in which specific adenosines in some RNA molecules are post-transcriptionally modified to inosines. RNA editing has emerged as a widespread mechanism for generating transcriptome diversity. However, there remain significant knowledge gaps about the variation and function of RNA editing. In order to determine the influence of genetic variation on A-to-I RNA editing, we integrate genomic and transcriptomic data by combining an RNA editing QTL (edQTL) analysis with an allele-specific RNA editing (ASED) analysis. We identify widespread associations between RNA editing events and cis genetic polymorphisms. Additionally, we find that a subset of these polymorphisms is linked to genome-wide association study signals of complex traits or diseases. Finally, compared to random cis polymorphisms, polymorphisms associated with RNA editing variation are located closer spatially to their respective editing sites and have a more pronounced impact on RNA secondary structure. Our study reveals widespread cis variation in RNA editing among genetically distinct individuals and sheds light on possible phenotypic consequences of such variation on complex traits and diseases.a

The Ensembl Variant Effect Predictor, an extensible variant annotation toolset

Andrew Parton, William McLaren, Irina Armean, Laurent Gil, Helen Schuilenburg, Anja Thormann, Sarah E. Hunt, Fiona Cunningham

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Accurate prediction of variant effects, which may cause transcriptional alterations and influence protein function, is vital for biological research and clinical studies. The Ensembl Variant Effect Predictor (VEP) is a powerful toolset for the interpretation of genomic variants. VEP utilises the extensive variant, regulatory and transcript data held in Ensembl to provide comprehensive variant annotation. To ensure maximum accessibility, VEP can be accessed through our online web tool, REST service or downloadable command line tool.

We provide reference data for all species held in Ensembl to enable rapid local analysis of genome-scale variant data. For human, these currently include both Ensembl/GENCODE and RefSeq transcript sets; data from the Ensembl Regulatory Build; and allele frequency data from the 1000 Genomes Project and Genome Aggregation Database (gnomAD). These data can be replaced or supplemented by gene annotation or variant frequency data in standard formats. This feature enables analysis of gene sets, assemblies or species that are not currently supported within Ensembl.

Here we present ways to leverage VEP's highly configurable analysis options and powerful result filtering. We will also demonstrate how VEP functionality can be further extended utilising its plugin infrastructure. Plugins facilitate construction of custom analysis pipelines often required in novel study designs and can be easily created with basic programming knowledge. We have recently extended the range of tools supported via plugins to include additional protein impact scores such as REVEL and MTR, as well as improved splice site annotation. Alongside VEP's core capabilities, its extensible plugin architecture can simplify the use of powerful variant interpretation tools to suit studies across the genomic spectrum.

Genome-wide associations between host genotypes and M. tuberculosis

Stephanie Pitts, M. Möller, E. Hoal, H. Schurz, E. Streicher, R. Warren, G. v.d.Spuy, G.C. Tromp , C. Kinnear

DST/NRF Centre of Excellence for Biomedical Tuberculosis Research; South African Tuberculosis Bioinformatics Initiative (SATBBI); South African Medical Research Council Centre for Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town

TB is a complex disease caused by infection with Mycobacterium tuberculosis (M. tb). While the majority of infected, immune-competent individuals remain asymptomatic, approximately 10% will develop active disease. Numerous studies have investigated the association of candidate genes with TB, and with different M. tb clades, with one recent study investigating genome-wide associations in a Thai cohort. This study aimed to investigate genome-wide association(s) between the host genotype and the genome of the infecting M. tb pathogen.

Sputum and blood samples were collected from TB patients residing in a suburb in Cape Town. Genotyping was performed using the Affymetrix 500k SNP array, and M. tb clades were identified using spoligotyping and IS6110 RFLP. Genotypes passing strict quality control (QC) filters were phased followed by imputation. Multinomial logistic regression (MLR) was performed using SNPTest and the standard genome-wide significance cut-off of alpha = 5 x 10-8 was used.

The cohort was dominated by LAM, followed by the Haarlem/LCC, Beijing/CAS1, Other, and Quebec superclades. MLR was performed using ~7 million SNPs for 445 samples, and the five M. tb superclades. The strongest association was with SNP rs9389610 (g.139039029G>A) on chromosome 6, at a p-value of 1.6x10-7. Individuals with the A allele of this SNP were twice as likely to be infected with a member of the Beijing/CAS1 superclade, as compared to the Haarlem/LCC (OR: 0.49) or LAM (OR: 0.46) superclades.

Although the association did not reach genome-wide significance, the results suggest replication of this approach in a larger cohort of this population may provide significant associations with the infecting M. tb clade, thereby improving our understanding of TB pathobiology. This is the first study to analyse associations between gene variants and M. tb superclades at a genome level, and this working method will now be used to investigate a Ghanaian cohort.

Degenerative differentiation between the social chromosome supergene variants of the fire ant

Rodrigo Pracana, Eckart Stolle, Yannick Wurm

School of Biological Chemical Sciences, Queen Mary University of London, London, UK; Institut für Biologie, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany

Antagonistic selection can favour the linkage between beneficial combinations of alleles via the suppression of recombination, forming genomic structures such as supergenes and sex chromosomes. However, recombination suppression can also reduce the efficacy of purifying selection and can therefore lead to chromosomal degeneration. A limitation in the study of degenerated chromosomes (e.g. Y chromosomes) is the difficulty of performing assemblies and variant identification in regions with a high repeat content. Species with diploid genomes pose the further problem of assigning haplotypes to either supergene variant or to either sex chromosome. We study the fire ant social chromosome supergene system, which controls dimorphism in the social organisation of colonies. The young age of this system (~400,000 years) provides the opportunity to study the early effects of suppressed recombination, with haploid males offering particularly tractable samples to perform this research. We use whole-genome sequencing and optical mapping of haploid fire ant males to show that the two variants of the social chromosome supergene are differentiated, and that the variant with most limited recombination is undergoing 'degenerative expansion'. Furthermore, we find evidence that differentiation at the coding sequence level is due to the degeneration of this chromosome rather than the result of antagonistic selection. Our results suggest that the functional differentiation between the supergene variants may thus result either from changes in regulatory elements or from changes with strong phenotypic effects at a few coding loci. We discuss how comparable supergene regions could be similarly affected by the interplay between antagonistic selection and degeneration, particularly in young sex chromosomes.

Long molecule sequencing improves genome assembly of the red fire ant

Anurag Priyam, Eckart Stolle, Yannick Wurm

School of Biological and Chemical Sciences, Queen Mary University of London, London, UK

Ants live in colonies organised by reproductive division of labour: queens specialise in production of brood, while workers participate in rearing the brood, foraging, and nest maintenance. In the red fire ant, Solenopsis invicta, a colony may contain exactly one, or multiple reproductive queens. The mode of colony organisation is associated with several morphological, physiological, and behavioural differences. Draft assembly of the fire ant genome created in 2011 comprises 69,511 sequences and represents about three quarters of fire ant's genome. Although the draft assembly has been instrumental in studying genetic basis of differences in social organisation, its fragmented and incomplete nature limits the scope of questions that can be answered.

Algorithmic advances in handling high error rate of long molecule sequencing have enabled more contiguous and complete genome assembly for several species. We used long reads from Pacific Biosciences' Sequel platform (~40x coverage) to obtain an improved de novo assembly of the fire ant genome. Our approach comprised of three steps. First, using Canu assembler (version 1.6) we obtained forty-five sets of corrected reads by varying three input parameters. Second, we evaluated the accuracy of corrected reads and selected fifteen (out of forty-five) sets of corrected reads for assembly. Finally, we selected the best assembly using five independent metrics derived from mapping Illumina reads to the assemblies and phylogenetic signals. The resulting contigs assembly is forty-seven-fold more contiguous than the previous assembly.

The novelty of our work extends beyond obtaining an improved reference assembly for an important model system for studying social evolution. First is our extensive parameter space exploration of the popular Canu assembler, requiring months of compute time. Second, we find that compared to default parameters, a more stringent error threshold for detecting overlaps between raw reads, and not trimming or splitting reads during correction stage produce more accurate assemblies. Finally, we show that increasing the error rate during assembly step to smash haplotypes together for obtaining a haploid reference assembly is a potentially poor strategy.

CADD v1.4 – variant effect scoring on GRCh37 and GRCh38

Philipp Rentzsch1, Daniela Witten2, Gregory M Cooper3, Jay Shendure4, Martin Kircher1,4

1 Berlin Institute of Health, Berlin, Germany
2 Department of Biostatistics, University of Washington, Seattle, WA, USA
3 HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA
4 Department of Genome Sciences, University of Washington, Seattle, WA, USA

Recognizing disease causing genetic variants is one of the main challenges of personalized medicine. While modern sequencing technologies enable the rapid identification of variants in patient genomes, interpreting thousands of new or very rare variants remains an unsolved problem. Computational prioritization can support variant interpretation, specifically with genome-wide scores available across variant types. Such scores integrate diverse types of data including functional annotations, sequence conservation, and biochemical activity read-outs to a measure of variant effect.
Combined Annotation Dependent Depletion (CADD) is a method for genome-wide deleteriousness scoring of single nucleotide variants (SNVs) and short insertion/deletion events (InDels). CADD uses machine learning techniques to separate simulated de novo variants (proxy-deleterious) from sequence changes since the common ancestor of human and chimpanzee (proxy-benign). Since CADD's publication on human genome build GRCh37 in 2014, some studies suggested that novel annotations are available and that tweaks in model training may improve performance of variant scoring. Additionally, the latest human genome build, GRCh38, became more broadly adopted by the community.
Here, we revised CADD's source code to allow easy integration and preprocessing of new annotations as well as model training by the open source, machine learning library scikit-learn. Using the new code base, we integrated additional annotations (e.g. a splice effect score), updated genomic annotations and trained new models for GRCh37 and GRCh38. We show that CADD v1.4 has similar performance on both genome builds, outperforms previous versions in separating sets of known benign and pathogenic variants, and better predicts the effect size of multiplexed molecular assays. With this new version, we are able to offer CADD scoring for the latest genome build, providing an important tool for applications in personalized medicine.

LncRNA atlas of vascular cells

Julie Rodor, Andrew H. Baker

BHF/University Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, UK

Long non-coding RNAs, defined as a heterogeneous class of transcripts longer than 200 nucleotides, have emerged as key regulators in biology and disease. While the functionality of most lncRNAs still needs to be established, an increased number of studies have shown the role of lncRNAs in the regulation of gene expression through varied mechanism. Due to their high spatio-temporal specificity of expression, lncRNAs are potential critical regulatory molecule and, consequently, ideal therapeutics targets. Many lncRNAs have been involved in vascular biology and disease but the complete repertoire of lncRNAs in vascular cells has never been described. The two main components of vessels are endothelial cells (ECs) and vascular smooth muscle cells (vSMCs). These two cell types, which acquire vascular bed (arterial, venal or lymphatic) specific differences, are disrupted in diseases.
Here, we used deep RNA-seq from 16 ECs and vSMCs human cell lines as well as other 15 relevant samples, obtained by the ENCODE consortium, to characterise the lncRNA profile of vascular cells. We quantified gene expression based on recent GENCODE annotation of the human genome and this analysis confirmed the clustering of samples by cell types as well as a clear separation of EC samples depending on their vascular bed of origin. Genes were clustered into groups depending on their expression profile specificity providing a clear catalog of ECs- and vSMCs enriched lncRNAs. We also carried out a new transcript discovery analysis of the ECs and vSMCs RNA-seq, giving us a comprehensive transcriptome annotation of the different cell lines. Novel lncRNAs were obtained after robust coding-potential filtering using PLAR pipeline1.
Future work will focus on predicting the function of the annotated and novel lncRNAs, in particular in ECs, based on a guilt by association approach and thanks to localisation and chromatin interaction high throughput data. LncRNA implication in vascular disease will be investigated by screening Genome-wide association study (GWAS) associated SNP and assessing the lncRNA expression level in RNAseq of pathological conditions.
1. Hezroni et al. Cell Reports 2015

Promiscuous expression of lincRNAs in medullary thymic epithelial cells

Kévin Rue-Albrecht1, Kathrin Jansen1,2, Adam Handel3, Chris Ponting4, George A. Holländer2, Stephen N. Sansom1

1The Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK; 2Department of Paediatrics and the Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK; 3Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, UK; 4Department of Biomedicine, University of Basel, Basel, Switzerland; 5MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

Medullary thymic epithelial cells (mTEC) are important both for the negative selection of self-reactive thymocytes and for the positive selection of regulatory T-cells, processes that are essential for the avoidance of autoimmunity. The identification of thymocytes with affinity for self-antigens involves the promiscuous expression and presentation of almost all protein-coding genes by the mTEC population. Previous work has shown that promiscuous gene expression (PGE) in mTEC is in part controlled by the Autoimmune regulator (Aire), a transcriptional facilitator that is understood to recognise epigenetically silenced loci via interactions with various cofactors. We investigated whether the mechanisms of PGE in mTEC might extend to (or involve) long intergenic non-coding RNAs (lincRNAs), a class of RNA transcripts increasingly associated with gene regulatory functions in the immune system. Using a transcriptome assembly of RNA-seq data from 22 peripheral tissues (from the mouse ENCODE project) and FACS-isolated TEC populations, we identified a set of 6,572 robustly expressed lincRNA transcripts (from 4,715 lincRNA genes). We found mTEC to express 79.5% of this set of lincRNA genes, a proportion comparable to that transcribed by the testis (80.5%) but markedly larger than that observed in other peripheral tissues (in which we could detect 28-60.7% of these lincRNAs). In mature mTEC, we found Aire expression to up-regulate the expression of some 526 lincRNA genes, a fraction (18.8%) comparable to that observed for protein-coding genes (20.4%) in this cell type. Outside of the thymus, we found both Aire-regulated and Aire-independent lincRNA genes to be highly tissue-restricted in expression. Our results suggest that Aire-dependent and Aire-independent mechanisms of PGE in mTEC do not discriminate between protein coding and non-coding loci.

Robustness of metrics for assessing similarity of genomic datasets

Stefania Salvatore 1, Knut Dagestad Rand 2, Ivar Grytten 1, Egil Ferkingstad 1,3, Diana Domanska 1, Lars Holden 4, Marius Gheorghe 5, Anthony Mathelier 5, Ingrid Glad 2 and Geir Kjetil Sandve 1

1 Department of Informatics, University of Oslo, Oslo, Norway, 2 Department of Mathematics, University of Oslo, Oslo, Norway 3 Institute, University of Iceland, Reykjavik, Iceland, 4 Statistics For Innovation, Norwegian Computing Center, Oslo, Norway 5 Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, Oslo, Norway.

The generation and systematic collection of genome-wide data is ever increasing. Consortia such as the Encyclopedia of DNA Elements (ENCODE) and the International Human Epigenome Consortium (IHEC) systematically sequence and collect genome-wide data on DNA methylation, histone modifications, chromatin accessibility and transcription factor binding. This has enabled researchers to study the interplay between biological processes and their relation to traits and diseases.
We here consider the setting where relations are investigated by comparatively assessing genomic co-occurrence of experimental datasets, for example determining the most relevant cell types for trait-associated genetic variation or determining highly co-occurring transcription factors. Technically, this corresponds to ranking a set of genome-wide binary vectors based on similarity to a separate query vector, a problem that has long roots in other fields like ecology.
We show that on both simulated and real genomic data, commonly used similarity metrics have very different properties and can lead to very different rankings of similarity for the same input data. In particular, the metrics are affected very differently by variation in dataset size. We show that basic modelling assumptions for the datasets can be used to guide the choice of an appropriate similarity metric, and provide some general recommendations for the setting of genomic data. Specifically, we show that arguments could be made for using either fold change (Forbes coefficient) or tetrachoric correlation, while the commonly used Jaccard index should be avoided as it is strongly affected by the number of elements in genomic datasets.

Handling dependencies and heterogeneities in genomic colocalization analysis

Geir Kjetil Sandve 1, Chakravarthi Kanduri 1,2, Diana Domanska 1, Stefania Salvatore 1, Ivar Grytten 1, Knut Dagestad Rand 1, Boris Simovski 1, Sveinung Gundersen 1,3, Eivind Hovig 1,3,4

1 Department of Informatics, University of Oslo, Oslo, Norway, 2 K. G. Jebsen Coeliac Disease Research Centre, University of Oslo, Oslo, Norway, 3 Elixir Norway - Oslo node, Department of Informatics, University of Oslo, Oslo, Norway 4 Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway

Reference genomes allow DNA sequencing-based data for a variety of omics features to be represented as coordinates on a genome assembly. This uniform representation opens for generic analysis methodology, where the analysis of feature colocalization is one of the well established methodologies. When genome-wide base pair-resolution data started appearing a decade ago, it was convincingly shown that occurrences for features related to e.g. chromatin and gene regulation followed complex structures along the genome. This included frequency heterogeneity at the broad scale and clumping at the fine scale, which may lead to underestimation of the variance of colocalization. If ignored in statistical analysis, this will lead to overoptimistic significance and potentially to false findings.

While a variety of approaches have been proposed for colocalization analysis, both at the generic level and tailored to particular features, each approach is focused on certain challenges while ignoring others. We will present published and recently submitted work on the variety of challenges involved in colocalization analysis, the extent of significance overestimation if particular aspects are ignored, how the detailed choice of colocalization metric may affect biological interpretations and how conclusions may vary substantially depending on chosen methodology. We will also present ongoing work on how dependency structures between features may delude biological interpretations when performing statistical analyses across sets of omics features, an important issue that is unaddressed in existing literature. Finally, we propose that the colocalization community should come together to establish biologically relevant benchmarks that can guide the evaluation and development of future colocalization methodology that comprehensively tackles the full spectrum of challenges established by the field.

Give a Dog a Genome: generating a stake-holder funded bank of whole-genome sequences with which to elucidate benign and disease-associated variation within the canine genome

Ellen Schofield(1), Louise Burmeister(1), Rebekkah Hitti(1), Chris Jenkins(1), Bryan McLaughlin(1), Louise Pettit(1), Sally Ricketts(1) and Cathryn Mellersh(1)

(1)Kennel Club Genetics Centre, Animal Health Trust, Newmarket, Suffolk, UK

The advent of whole genome sequencing (WGS) has promised to revolutionise genetic research, and the rapid fall in per-sample costs in recent years has made the revolution an affordable reality for geneticists. The technology is especially useful for the study of simple Mendelian conditions where disease-causing mutations have the potential to be identified from the WGS of a single case. However, when comparing a typical canine genome with the reference sequence (CanFam3.1) or a control genome, at least 2-3 million variants will typically be identified. Many of these variants are likely common polymorphisms which could be excluded by comparing with multiple control genomes. We devised the Give a Dog a Genome (GDG) project to build a resource of canine genetic variants across the genome using WGS; currently projected to contain 90 genomes from 78 breeds, and investigate genetic diseases in at least 69 breeds. We used a crowd-funding approach, with the costs of the project being shared between multiple stakeholders. To date (two years after GDG was launched), 74 samples from 69 breeds have been sequenced comprising 62 dogs affected with a suspected genetic condition (27 conditions in total) and 12 apparently healthy older dogs. The GDG variant bank has been used to validate several disease-associated mutations and DNA tests have been developed to improve the health and wellbeing of dogs. All of the WGS data generated through GDG will be shared with the Dog Biomedical Variant Database Consortium (DBVDC), and specific sequences will be shared with at least 20 scientists from Europe and the USA to contribute to their research.

Genome analysis of Pantoea ananatis strain MHSD5 on web-based platform

Mahloro Hope Serepa-Dlamini, SG Mahlangu

Department of Biotechnology and Food Technology, Faculty of Science, University of Johannesburg, Doornfontein Campus, PO Box 17011 Doornfontein 2028, Johannesburg, South Africa

The genus Pantoea has diverse species, which have been isolated from several environments such as aquatic and terrestrial. Species in this genus have associations with humans, plants, insects, and animals, which can be either parasitic, mutual or commensal. Although most species have been reported to have pathogenic associations with humans, animals and plants, few have been reported to be symbiotically associated with plants. As plant endophytes, Pantoea bacteria promote plant growth via a variety of mechanisms and produce bioactive compounds with antibiotic activities. Since Pantoea species have associations with different hosts in different environments, there is need for understanding genetic factors that allow this group of bacteria to successfully colonize various hosts. In addition, the availability of various genomes of Pantoea genus will promote whole genome comparison within this group and further our understanding of genetic factors that contribute to Pantoea species thriving in different environments and thus delineating their biology and evolution. We present here the draft genome sequence and annotation of P. ananatis strain MHSD5, which is a bacterial endophyte isolated from the surface sterilized leaves of Pellaea calomelanos, a medicinal plant obtained in Limpopo province of South Africa. All the pre-annotation analysis were performed on Galaxy web platform (https://usegalaxy.org). The de novo genome assembly with Unicycler version 0.4.1.1 and assessed with Quast version 4.6.3 was 4.6 Mb with an N50 of 550,557 bp. 4,350 putative protein coding sequence genes were predicted with PGAAP. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession PUEK00000000. This draft genome/study is one of the many studies that illustrates the benefit of open source Galaxy web platform for genome data analysis.

Direct identification of fusion genes from whole genome sequencing data

Riccha Sethi, Barbara Schrörs, David Weber, Martin Löwer, Martin Suchan, Jos de Graaf, Ugur Sahin

TRON gGmbH – Translational Oncology at Johannes Gutenberg-University Medical Center gGmbH, Mainz, Germany

Cancer is largely driven by accumulation of somatic mutations that can be subdivided into small mutations (e.g. point mutations, small insertions and deletions) and large structural variants (e.g. translocations, inversions, deletions, duplications and insertions). While point mutations affect few bases, structural variants (SV) can affect large stretches of DNA. The simple and complex genomic rearrangements can often give rise to fusion genes. The prediction of fusion genes from only RNA-sequencing data is masked by high false positive rate. In principle, fusion genes expressed should be explained by genomic structural rearrangements except the ones produced by read-through transcription. Thus, fusion genes predicted from whole genome sequencing and expression checked through RNA-sequencing data would reduce number of false predictions made from whole genome and RNA sequencing data individually.
Here, we show that FuseSV is a novel pipeline that merges calls from multiple SV-calling algorithms from whole genome sequencing data (WGS) with high sensitivity and specificity. We tested our approach in the MCF7 breast cancer and in a patient derived melanoma cell line by validating predicted SVs with PCR and amplicon sequencing. Furthermore, we developed a tool, FUDGE (Fusion of DNA Genes) that predicts exon structure of classical fusion genes between annotated genes A and B based on structural rearrangement. The advantage of this approach is the discrimination and elimination of read-through fusions, which are potentially not somatic events and thus have uncertain tumor specificity. This prediction pipeline takes SV events as input, builds a genomic rearrangement graph, uses depth-first search to identify classical fusion genes and predict the respective exons fused together. The combined prediction of fusion genes from SVs and their corresponding expression through RNA-sequencing data allow us to confirm the predicted fusion gene. Using this approach, we identify novel fusion genes in MCF7 and the patient derived melanoma cell line, which were missed by predictions utilizing only RNA-sequencing data. In summary, integrative analysis of genomic and transcriptomic data can predict fusion genes directly from genomic SV. To further improve this approach, we use the validation data in a machine learning approach to train a classifier that will further improve specificity.

Comparative analysis of cfDNA between plasma and pleural effusion samples from lung cancer patients.

Seung-Ho Shin1,2, Yeon Jeong Kim1, Ku Bo Mi3, Hyo-Jeong Jeon1, , Danbi Lee1, Dae-Soon Son1, Woong-Yang Park1,2,4, Myung-Ju Ahn3, Donghyun Park1

1Samsung Genome Institute, Samsung Medical Center, Seoul 06351, South Korea,
2Department of Health Sciences and Technology, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University, Seoul 06351, South Korea,
3Division of Hematology-Oncology, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea
4Department of Molecular Cell Biology, Sungkyunkwan University School of Medicine, Suwon 16419, South Korea

As Cell-free DNAs (cfDNAs) released from various cell types to the bloodstream, circulating cfDNAs from tumor cells (i.e., circulating tumor DNAs (ctDNAs)) have been used for detecting tumor variants. While plasma has been frequently analyzed to profile tumor mutations in cfDNA, pleural effusion (PE) has been also proposed as a resource for the purpose in lung cancer patients. In fact, malignant pleural effusion has been routinely examined in clinics for diagnosis, treatment and prognosis of lung cancer patients. However, there have been few systematic comparison studies between two types of specimens as a resource for tumor variant detection. In this study, using a targeted deep sequencing method, we analyzed cfDNAs of pleural effusion (PE) and plasma from lung cancer patients to compare the allele frequency levels of tumor variants. Variants from a patient displayed significantly higher allele frequency in PE cfDNA than plasma cfDNA. When we examined tumor DNA fraction in cell pellet obtained from PE, the variant allele frequency (VAF) was frequently lower than those in plasma and PE cfDNAs. Our results suggest that PE is a better resource for tumor variant detection with higher sensitivity than plasma if lung cancer patients accumulate PE.
Next, because the correlation between VAFs in PE cfDNA and PE cell pellet was low, we wondered what fraction of ctDNA in PE was derived from tumors resided in pleural space or lung tissue. Thus, to estimate tissue-of-origin of PE cfDNA, we analyzed DNA methylation patterns. Both plasma and PE specimens carrying higher VAF mutations displayed the significantly higher fraction of lung cancer origin supporting the validity of the method. These results will help us understand the origin of PE cfDNA and thus will provide a foundation for expanding the clinical applications utilizing PE.

Streamlining bisulfite sequence analysis workflow using CyVerse

Jawon Song, Joe Stubbs, James Carson, Matthew Vaughn

Texas Advanced Computing Center

Recent advances in sequencing techniques have improved our understanding of epigenetic modifications in many of ways. One specific example is sensing methylated cytosines through normal sequencing methods after unmethylated cytosines are transformed into uracil with bisulfite conversion. Through this, it has been possible to capture DNA methylation at the base-pair resolution.

Bioinformatics tools that are readily used for DNA methylation studies aligns the sequence reads and calls methylated cytosines and their methylation ratios. While this is extremely useful, the processes are discrete and involves handling of intermediate datasets.

Using tools that are available publicly and developed in-house, we have created a workflow in which users can start with raw sequencing reads and get differential methylation information between two samples. This way, time can be saved in waiting for prior task to be finished for submission of the next analysis, as well as prevent any chances of data loss or corruption.

UniProtKB: providing protein sequence and functional information to facilitate scientific discovery

Elena Speretta1, UniProt Consortium1-4

1 EMBL-EBI, Cambridge, UK; 2 Swiss Institute of Bioinformatics, Centre Medical Universitaire, Geneva, Switzerland; 3 Protein Information Resource, Georgetown University Medical Center, Washington, USA; 4 Protein Information Resource, University of Delaware, Newark, USA.

The data curation of the UniProt Knowledgebase (UniProtKB) is one of main activities of the UniProt Consortium. It involves the evaluation and integration of information from multiple resources so as to represent biological knowledge in the most accurate and comprehensive way. UniProtKB consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The former contains manually reviewed records while the latter contains computationally generated records enhanced by automatic classification and annotation.
UniProtKB/Swiss-Prot manual curation includes a critical review of experimental and predicted data for each protein. Sequences from the same gene are merged while discrepancies and variants are evaluated case by case. The curation protocol includes manual extraction and structuring of information from the literature, performance and evaluation of computational analyses, and mining and integration of large-scale data sets. Information is updated continuously as new data become available.
UniProtKB has also developed UniRule and the Statistical Automatic Annotation System (SAAS), two complementary approaches to automatically annotate protein records in UniProtKB/TrEMBL. Both use InterPro and the manually curated data in UniProtKB/Swiss-Prot as templates to provide information about the many uncharacterised proteins for which there is no experimental data available.
Here, we will present the current annotation process in UniProtKB and show how literature-based manual curation coupled with large-scale automatic annotation provides high-quality data across the proteomes provided by UniProtKB, enabling researchers working on comparative and evolutionary genomics to analyse proteomes, and thus genomes, across broad taxonomic ranges.
UniProt is updated every four weeks and can be freely accessed or downloaded from https://www.uniprot.org. Programmatic access is also offered through a dedicated API.

AViDE - An interactive data visualization application for gene expression

Peter Taschner, Stephen Pieterman1, Aldo Jongejan2, Floyd Wittink1

1University of Applied Sciences Leiden, Leiden, Nederland; 2Bioinformatics Laboratory, Academic Medical Center, Amsterdam, Nederland

AViDE - An interactive data visualization application for gene expression
Stephen Pieterman1, Peter Taschner1, Aldo Jongejan2, Floyd Wittink1
1University of Applied Sciences Leiden, Leiden, Nederland; 2Bioinformatics Laboratory, Academic Medical Center, Amsterdam, Nederland
Taschner@generade.nl

Gene expression analysis using RNA sequencing (RNA-Seq) has become popular giving rise to a large amount of data, which poses challenges for effective visualization of results. Many tools provide static figures and have to be re-run with new parameter settings when different thresholds or cutoff values are required. Adding interactivity to these tools would increase flexibility, simplify the analysis and interpretation of the data and aid in the recognition of biological patterns or mechanisms.
Therefore, we developed AViDE. an interactive visualization application for gene expression data. AViDE is written in R with Shiny1 and can be run as server on any platform with R installed. The input for AViDE is a count file containing the RNA-Seq data and a sample sheet. The sample sheet contains the information about the samples and their respective groups. After uploading input data, users can switch quickly between different time points (contrasts) or samples, select different thresholds and cutoff values, and visualize plots in the most informative way.
AViDE offers different plot visualizations: Volcano plot, MA-plots, Heatmaps, log-plots, Beeswarm plots, 3DPCA, Circle of Correlations and Sample Distance Heatmaps, PCA, and MDS-plots. Data points can be clicked or hovered over to display more information. This type of interactivity is common across all plots. Changing settings for one plot will also change other plots. The ability to visualize data in a convenient and intuitive way and on-the-fly analysis from RNA-Seq files speeds up down-stream gene expression data interpretation.

1) Beeley, C. 2016. Web application development with R using Shiny. Packt Publishing Ltd.

Funded in part by SIA (Grant RP-2015-02-69P).

Germline variant calling using the Seven Bridges Graph Toolkit

Huseyin Serhat Tetikol, Vladimir Kovacevic, Ozem Kalay, Devin Locke

Seven Bridges Genomics Inc., Cambridge, MA 02142, USA

The human reference genome lays the basis for genomic analyses by enabling alignment of sequenced reads. However, the linear human reference genome released by the Genome Reference Consortium only represents a single consensus haplotype, and therefore constitutes a suboptimal representation of human genetic variation. Directed acyclic graph reference genome representations that incorporate information on genetic variation have been shown to improve the accuracy of read alignment, variant calling and other subsequent genomic analyses. A set of bioinformatics tools utilizing graph genomes (Seven Bridges Graph Toolkit; SBGT) for the analysis of next-generation sequencing (NGS) data has recently been published by Seven Bridges in preprint*. In this study, we benchmark the SBGT pipeline on germline variant calling and compare it against other state-of-the-art whole genome and whole exome pipelines, namely BWA-MEM + GATK4 by the Broad Institute, Sentieon DNAseq by Sentieon Inc., BWA-MEM + Strelka2 by Illumina Inc., Dragen by Edico Genome and BWA-MEM + DeepVariant by Google. The analysis includes five whole-genome samples (HG001-HG005) with a total of 8 library preparations (from 30x to 150x) with the truth data established by the Genome in a Bottle Consortium. The results show that the SBGT pipeline has the highest accuracy in small (<50bp) INDEL calling with an F1-score of 0.25% above the average, while scoring 0.06% above the average in SNP calling. Moreover, we show that, unlike the other pipelines, the SBGT pipeline is capable of finding INDELs as long as several kilobases without any additional processing steps. The Mendelian inheritance discordance measurements on CEPH/CEU and Ashkenazim trios demonstrate that the SBGT pipeline offers the most consistent germline variant calling with a score that is 1-2 percentage points better than the other pipelines. Finally, we present an optimized version of the SBGT pipeline that can process a whole-genome sample with 30x coverage on the cloud in 2 hours and under $5.

* Rakocevic, Goran, et al. "Fast and Accurate Genomic Analyses using Genome Graphs." bioRxiv (2018): 194530.

Characterising and visualising gene families with GeneSeqToFamily and Aequatus

Anil S. Thanki1, Nicola Soranzo1, Wilfried Haerty1, Javier Herrero2, Robert P. Davey1

1 Earlham Institute, Norwich Research Park, Norwich NR4 7UH, UK
2 Bill Lyons Informatics Centre, UCL Cancer Institute, London WC1E 6DD, UK

The phylogenetic information inferred from the study of homologous genes helps us to understand the evolution of gene families and plays a vital role in finding ancestral gene duplication events as well as identifying genes that are under positive selection within species. Various tools exist to identify gene families and provide an overview of syntenic region evolution at the family level but they typically do not provide information about structural changes within a gene. Similarly, collating and configuring the myriad software to discover gene families often requires many dependencies to be manually fulfilled, and results in bespoke pipelines that are tailored to a single computing environment. Here, we present a complete Galaxy workflow for finding gene families using GeneSeqToFamily (Thanki et al. GigaScience 2018) and visualising their relationships using Aequatus (Thanki et al. bioRxiv 2018).

GeneSeqToFamily is a Galaxy workflow based on Ensembl GeneTrees pipeline, and generates gene families based on coding sequences, providing details about exon conservation. It helps users to run potentially large-scale gene family analyses without requiring command-line usage while still allowing parameter configuration to be modified, and tools themselves to be replaced by other compatible software as appropriate.

Aequatus is a standalone web-based tool that provides an in-depth view of gene structure across gene families, including gene order and protein domains. It relies on pre-calculated alignment and gene feature information held in, typically, the Ensembl Compara and Core databases, or generated by the GeneSeqToFamily workflow. To aid reuse, Aequatus.js, an open source JavaScript module, is available as a plugin within the Galaxy web platform to visualise gene trees generated by GeneSeqToFamily.

All source code is available on GitHub. All tools used in the GeneSeqToFamily workflow and the full workflow itself are available within the Galaxy ToolShed. GeneSeqToFamily is also available for use within the usegalaxy.eu public instance.

Working with Linkage Disequilibrium data in Ensembl

Anja Thormann, Irina Armean
Jyothish Bhai
Laurent Gil
Will McLaren
Andrew Parton
Helen Schuilenburg
Stephen Trevanion
Sarah Hunt
Fiona Cunningham

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK

An understanding of linkage disequilibrium (LD), the non-random association between alleles at different loci, is important when interpreting disease association data and performing population genetics studies. Since the completion of the HapMap project, the first large scale project to study genomic variation between different human populations, Ensembl has provided tools for the calculation and visualisation of LD patterns. Our LD calculation is implemented in C and follows an expectation-maximization algorithm which does not require phased data. The script calculates both r2 and D' values. The Ensembl Perl API provides a wrapper for the C script and implements different functionality for retrieving LD results. The API supports population-specific calculations for a given genomic region, a list of variants and for a variant and all surrounding variants in a given window size. The Ensembl browser shows both Manhattan and characteristic LD triangle plots alongside our rich gene annotation. The 1000 Genomes Project propelled the number of sequenced individuals to a new level, triggering changes in how Ensembl stores and makes genotype data available. Here, we present the changes that have been introduced to how we calculate and disseminate LD data. We moved from storing genotypes in a MySQL database to retrieving genotypes from VCF and using the htslib library for fast data retrieval. Our tools are not limited to supporting human calculations and can be used for any species with appropriate genotype data.

In addition to the updated infrastructure, we will also describe the various interfaces we provide to retrieve LD results, which have been tailored to different use cases. Our REST endpoints support the same functionality as our Perl API and can be conveniently integrated into analysis pipelines and websites. REST queries are however time limited, so if the region of interest is large or variant rich we recommend using our standalone script or web tool. Our web tool presents a simple interface for configuring the LD analysis type and inputting data. Results are presented as a downloadable table with r2 and D' values alongside variant annotation, including predicted functional consequence and phenotype association. Lastly, we provide a Perl standalone script which hides the complexity of the Perl API and allows for easy selection of the requested LD analysis and configuration of input data.

IsoformSwitchAnalyzeR: Enabling genome wide analysis of alternative splicing and isoform switches

Kristoffer Vitting-Seerup [1,2], Malte Thodberg [1], Albin Sandelin [1]

[1] Sandelin Lab, Bioinformatics Centre, Department of Biology | Biotech Research & Innovation Center (BRIC), University of Copenhagen, Denmark.
[2] Brain Tumor Biology Lab, The Danish Cancer Society, Denmark.

Alternative splicing, transcript start- and termination-sites play essential roles in development, homeostasis and diseases by giving rise to proteins with different properties. Due to the increased recognition of the importance of alternative splicing in amongst other many cancer types, the analysis of alternative splicing is a rapidly growing field. Due to its biological interpretability, the approach which focus on identifying switches in full length isoforms or transcripts have great potential. However, from a statistical perspective, devising tests for differential isoform usage is non-trivial and current methods suffers from low scalability and high false discovery rates. Furthermore, since such isoform switches are often identified in hundreds of genes, there is an acute need for tools which can help elucidate the genome-wide patterns of isoform switches and their consequences.

Here we extend IsoformSwitchAnalyzeR to address both these problems, with a new isoform switch test that can analyze thousands of libraries in a matter of minutes. The speed and robustness, combined with the ability to correct for confounding effects additionally enables the new test to scale from bulk RNA-seq to single cell RNA-seq datasets. The second module we developed enables genome-wide analysis of both alternative splicing as well as the functional consequences identified in individual isoform switches. Combined, these updates enable users to investigate both gene- and genome-level effects of changes in alternative splicing in both bulk and single cell RNA-seq data.

To illustrate the new capabilities of IsoformSwitchAnalyzeR, we applied the updated version of IsoformSwitchAnalyzeR to RNA-seq data from TCGA. Using the genome-wide analysis tool we found a pan-cancer tendency to use isoform switches resulting in loss of protein domains. Surprisingly, we find that the biological mechanisms behind these switches are different in the individual cancer types - with some cancer types relying on exon skipping while other cancer types rely on alternative transcription start- and termination-sites.

Comparative genomics of Ebolavirus species reveals determinants of pathogenicity

Mark Wass, Morena Pappalardo, Henry J Martell, Stuart G Masterson, Martin Michaelis

Industrial Biotechnology Centre and School of Biosciences, University of Kent, Canterbury, Kent, UK

The West African Ebola virus outbreak killed thousands of people and demonstrated the scale on which the virus threatens human life. Using extensive sequencing data obtained during the outbreak, we compared Ebolavirus genomes to identify potential determinants of Ebolavirus pathogenicity. Of the five Ebolavirus species, only Reston viruses are not pathogenic in humans. We compared Reston virus genomes with those from the four human pathogenic species to identify specificity determining positions (SDPs) that are differentially conserved and may therefore act as determinants of pathogenicity.

We identified 160 SDPs using 1406 Ebolavirus genome sequences. Protein structural analysis was performed to identify SDPs that are likely to alter protein structure and function and could be associated with pathogenicity. The most striking findings were in Ebolavirus proteins VP24 and VP40. Particularly SDPs present in VP24 are likely to alter binding to human karyopherin alpha proteins and therefore prevent inhibition of interferon signaling in response to viral infection.

Further, Ebola viruses do not normally cause disease in rodents. To establish in vivo models, Ebola virus has been adapted to rodents through serial passaging experiments. We analysed the mutations that occurred during four independent rodent adaptation experiments. Our analysis showed that only three Ebolavirus genes consistently acquired mutations, VP24, GP (glycoprotein) and NP (nucleoprotein). The functional effects of the GP and NP mutations were unclear, but structural analysis demonstrated that the VP24 mutations either cause conformational changes of the protein or are located at the interface site with human karyopherin alpha. Thus, we propose that VP24 is critical for Ebolavirus adaptation to novel hosts.

Taken together, our findings demonstrate that VP24 is vital to determining species-specific pathogenicity. Since only a few SDPs distinguish Reston virus VP24 from VP24 of other Ebolaviruses, it is possible that human pathogenic Reston viruses may emerge. This is of particular concern as Reston viruses circulate in domestic pigs in The Philippines and China.

Diagnosing the undiagnosed: expanding the genetic etiology and phenotypic spectrum of rare pediatric conditions

Peter White, Benjamin Kelly, Patrick Brennan, Theresa Mihalic Mosher, Scott Hickey, Kim McBride, Daniel Koboldt and Richard Wilson

The Institute for Genomic Medicine, Nationwide Children's Hospital and The Ohio State University Department of Pediatrics, Columbus, Ohio, USA.

While a condition is considered "rare" if it affects fewer than 200,000 persons, collectively the total number of people with a rare disease is large (~350 million people worldwide), representing a significant public health burden. There are thought to be as many as 7,000 different rare diseases, of which 80% may have a genetic cause. While advances in genome sequencing technologies have accelerated rates of discovery of new functional variants in syndromic and rare monogenic disease, many more disease-causing genes and novel genetic etiologies remain to be discovered.

Accurate molecular genetic diagnosis of a rare disease is essential for patient care. However, our current best efforts leave 60-75% of patients undiagnosed. For these individuals with a suspected rare or as yet to be diagnosed disease we have been evaluating the use of whole genome sequencing and alternative analysis approaches to uncover novel genetic etiologies. Our institute has enrolled 50 families suffering rare inherited conditions into a research genomics rare disease protocol that utilizes whole genome sequencing of the proband and available family members.

Through whole genome sequencing and implementation of novel computational methods, we are developing approaches to identify pathogenic variants that do not directly impact the protein coding sequence, such as intronic or synonymous variants, and structural variants that are not possible to detect with traditional molecular methods. For example, whole-genome sequencing of a male with myoclonic dystonia uncovered a de novo 182-kb inversion disrupting SGCE. The copy-neutral status and intronic breakpoints of the inversion make it refactory to detection by conventional assays (array or clinical exome sequencing). Further, this represents a new class of disease-causing alterations that are refractory to detection by conventional array and exome sequencing.

While the majority of families enrolled in our study have negative findings from clinical whole exome sequencing, through application of whole genome sequencing we have identified likely causal variants in 30% of cases and strong candidate variants in another 20%. Our study demonstrates the power of research genomics to uncover novel genetic etiologies or rare or as yet to be diagnosed genetic disease, and provide long-sought molecular diagnoses for patients and their families.

Linking gene expression with genetic variation at single-cell resolution in human Acute Myeloid Leukemia

Stephen R. Williams, Allegra Petti, Christopher Miller, Ian Fiddes, Sridhar Srivatsan, Catrina Fronick, Robert Fulton, Deanna M. Church, and Timothy J. Ley

10x Genomics, Inc., Pleasanton, CA, USA; Department of Genetics, Washington University School of Medicine, St Louis, MO, USA; Department of Medicine, Division of Oncology, Washington University School of Medicine, St Louis, MO, USA; McDonnell Genome Institute,
Washington University School of Medicine, St Louis, MO, USA

Acute myeloid leukemia (AML) is the most common form of acute leukemia in adults, with ~20,000 new cases in the US each year. The 5 year survival rate is ~27%. Because of the clonally heterogeneous nature of virtually all AML tumors, detection of genetically distinct and clinically relevant subclones is essential for understanding AML biology. We developed a methodology combining Enhanced Whole Genome Sequencing (eWGS) with 10x Genomics Chromium Single Cell 5' Gene Expression workflow for Single Cell RNA Sequencing (scRNA-Seq) that allows for assignment of expressed mutations to single AML cells at unprecedented resolution.

Bone marrow collected at presentation from five AML patients were assessed with eWGS, bulk RNA-seq, and scRNAseq. We used eWGS to identify germline and somatic variants in paired tumor/normal DNA from each patient. scRNAseq from the malignant samples had a median of 20,474 cells/sample, with an average read coverage of 192,427 reads/cell. We developed bioinformatic methods to identify individual cells harboring the eWGS mutations, assign those cells to mutationally-defined subclones, and identify mutation- or subclone-specific expression profiles within the context of scRNAseq expression clustering.

Using these methods, we recovered 22%-46% of expressed somatic mutations in at least one cell (including SNVs, indels, and a gene fusion event). The mutation detection rate was highly correlated with expression levels and variant allele frequencies determined by eWGS and bulk RNA-seq data. We identified the AML cells within every sample, major subclones within a subset of samples, and mutation-specific and subclone-specific expression profiles.

In summary, this methodology builds upon existing technologies to help better understand the clonal heterogeneity of AML at the single-cell level, and directly link mutations that drive the outgrowth of subclones to distinct expression signatures. By doing so, we are gaining mechanistic insight into how specific subclonal mutations affect transcriptional profiles, and ultimately drive differences in growth, differentiation, and responses to therapy.

Evaluation of Constrained Coding Regions (CCRs) for rare disease gene identification in 100000 Genomes Project pilot dataset

Hywel J Williams1, Chris Odhams2, Dimitris Polychronopoulos2 and Genomics England2

1 GOSgene, UCL Great Ormond Street Institute of Child Health, Genomics and Systems Medicine, London WC1N 1EH. 2 Genomics England, Queen Mary University of London, Dawson Hall, London, EC1M 6BQ.

The identification of pathogenic mutations causing rare diseases (RD) is the first step in reducing the diagnostic odyssey experienced by many RD patients. Through the application of genomic sequencing techniques, we are currently able to derive a genetic diagnosis for around 40% of patients from across the clinical spectrum. This however leaves the majority of patients undiagnosed. We know that approximately 80% of RD patients have a genetic cause and that mutations affecting the coding sequence form the largest category of pathogenic variants. Therefore, if we can improve the ways we annotate and interpret coding variants we may be able to detect novel pathogenic variants in the remaining undiagnosed patients.

Recent work harnessing the power of public repositories of genomic data such as gnomAD have allowed researchers to characterise specific regions within the coding sequence of genes that are devoid of known functional variants. By modelling the likelihood of seeing a variant within these regions researchers have been able to rank these regions into percentile bins and show that those regions in the highest percentiles (>95) are enriched for known pathogenic variants.

We have sought to evaluate the utility of this resource by analysing the rare disease pilot dataset from the 100000 Genomes Project (100KGP). This dataset comprises a total of 4125 individuals that have undergone genomic sequencing, with the vast majority (82%) in the form of parent offspring trios. For our analysis we have derived a list of high quality de novo calls and extracted those that intersect CCRs with a percentile score of ≥95.

We identified 32 de novo variants that intersected a CCR ≥95. Of these 12 (38%) are deemed pathogenic with a further 5 (16%) being likely pathogenic candidates. The remaining 15 variants were considered unlikely to be pathogenic.

The conclusion to this analysis is that the use of novel ways to annotate and interpret genomic sequence data from RD patients such as CCR analysis has the potential to increase the diagnostic yield and improve the outcome for RD patients.

Citizen Science charts the oral microbiome of Spanish adolescents and reveals links with diet, hygiene, and the composition of tap water.

Jesse R. Willis1,2, Pedro González-Torres1,2, Alexandros A. Pittis1,2, Luis A. Bejarano1,2, Luca Cozzuto1,2, Nuria Andreu-Somavilla1,2, Miriam Alloza-Trabado1,2, Antonia Valentín3, Ewa Ksiezopolska1,2, Carlos Company1,2, Harris Onywera1,2,4, Maria M. Montfort1,2, Antonio Hermoso1,2, Susana Iraola-Guzmán1,2, Ester Saus1,2, Annick Labeeuw1,2, Carlo Carolis1,2, Jochen Hecht1,2, Julia Ponomarenko1,2, and Toni Gabaldón1,2,5,*.

1) Bioinformatics and Genomics Programme. Centre for Genomic Regulation (CRG). Dr. Aiguader, 88. 08003 Barcelona, Spain
2) Universitat Pompeu Fabra (UPF). 08003 Barcelona, Spain
3) ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
4) Institute of Infectious Disease and Molecular Medicine (IDM), University of Cape Town (UCT), Anzio Road, Observatory 7925, Cape Town, South Africa
5) Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.
* corresponding author: tgabaldon@crg.es

The oral cavity comprises a rich and diverse microbiome, which plays important roles in health and disease. Previous studies have mostly focused on adult populations or in very young children, whereas the adolescent oral microbiome remains poorly studied. Here we used a citizen-science approach and 16S profiling to assess the oral microbiome of 1500 adolescents around Spain and its relationships with life-style, diet, hygiene, and socioeconomic and environmental parameters. Our results provide a detailed snapshot of the adolescent oral microbiome and how it varies with life-style and other factors. In addition to hygiene and dietary habits, we found that the composition of tap water was related to important changes in the abundance of several bacterial genera, pointing to an important role of drinking water in shaping the oral microbiota. Overall, the microbiome samples of our study can be clustered into two broad compositional patterns, which show striking similarities with those found in unrelated populations. We hypothesize that these two stomatotypes represent two possible global equilibria in the oral microbiome that reflect underlying constraints of the human oral niche. As such they should be found across a variety of geographical regions, lifestyles, and ages.

Signatures of socially antagonistic selection in the fire ant social chromosome

Yannick Wurm, Carlos Martinez-Ruiz, Rodrigo Pracana, Richard Nichols

Organismal Biology Department, Queen Mary University of London

Colonies of the red fire ant Solenopsis invicta have either one or multiple queens. These differences in queen number are associated with a suite of other phenotypic traits including independent colony founding. Colony type is determined by a "social chromosome" region including >400 genes on chromosome 16. This system therefore provides a unique opportunity to understand the molecular processes underlying social evolution.

The social chromosome has two variants, SB and Sb, which are inverted with respect to each other. Recombination between SB and Sb is severely repressed and, since the Sb homozygotes seem to be lethal recessives, there is no effective recombination within Sb. This genetic architecture is very similar to that of sex chromosomes, which also have variants that are essentially non-recombining (the Y in an XY system).

Here, we tested whether the evolution of the social chromosome system has been shaped by social antagonism between the two social forms, in a similar manner to how the evolution of sex chromosomes has been shaped by sexual antagonism between males and females. Specifically, we test whether expression patterns in the social chromosome match those expected by evolutionary conflict. Additionally, we model the gene flow between the two social forms to determine the potential importance of social antagonism in this system. Our results are consistent with socially antagonistic selection being responsible for the divergence of the social chromosome. These findings shed light on the phenotypic effects of genome architecture in the context of complex social behaviours.

Annotation and prediction of human polyadenylation site by deep learning

Chen Yang (1, 2), Chenkai Li (1, 2), Ka Ming Nip (1, 2), René Warren (2), Inanç Birol (1, 2, 3)

1 University of British Columbia, Vancouver, BC, Canada; 2 Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada; 3 Simon Fraser University, Vancouver, BC, Canada

As a major mechanism of gene regulation in eukaryotic cells, alternative 3' UTR cleavage and polyadenylation (APA) of mRNA precursors has been found to be tissue-specific. With our evolving understanding of the important functional role of APA in several human diseases, it is desirable to characterize this phenomenon in large cohort studies. Sequencing technologies have been proved to be useful in studying such cohorts at the genome and transcriptome levels, and there are vast amounts of publicly available datasets. While there are sequencing protocols specifically designed to interrogate APA, most of these large-scale studies use more general-purpose data types, such as whole genome (WGS) or transcriptome (RNA-seq) sequencing reads.

Recent studies showed the value of RNA-seq data for detailed characterization of polyadenylation sites in transcriptomes. Approaches proposed fall into two major groups: alignment based analysis, and machine learning. These approaches, however, may either lose sensitivity in low coverage sequencing libraries, or potentially miss critical sequence features because of their heavy reliance on manually selected features. To address these problems, we propose a new deep learning based polyadenylation site prediction tool called DeepA.

We trained DeepA on a dataset composed of 160,000 transcript sequences represented as consecutive k-mers, and evaluated its performance via 5-fold cross validation. We explored various deep neural network architectures, and various hyper-parameters, including the optimal k-mer and sequence lengths. The best performing model resulted in an average of 95.5% testing accuracy. We also benchmarked our method against competing tools, and observed DeepA to perform favourably in its accuracy as well as its runtime.

DeepA essentially has two use cases: (1) it can be applied on draft genome assemblies to annotate the 3' end of genes, or on transcriptome assemblies to determine whether they are complete or truncated; and (2) it can characterize APA using RNA-seq reads. For the latter, we will introduce a characterization pipeline built on DeepA, and demonstrate how it can be used for retrospective studies on existing large cohorts, such as TCGA and GTEx. We expect DeepA to find wide application in the field, also benefiting prospective studies.

Porcine subcutaneous fat tissue transcriptome analysis, differential gene expression, and target analysis on backfat thickness divergent samples

PaoloZambonelli, Martina Zappaterra, Roberta Davoli

Department of Agricultural and Food Sciences (DISTAL), University of Bologna, Viale Fanin 46, 40127, Bologna, Italy

The identification of the molecular mechanisms regulating pathways associated to fat deposition aptitude in pigs is essential to detect key genes to utilize in selection plans for the genetic improvement of fat traits. Short noncoding RNAs (sRNAs) modulate the expression of messenger RNAs (mRNAs). In particular, microRNAs (miRNAs) interactions with target mRNAs regulate gene expression and modulate pathway activation. In pigs, miRNA discovery is far from saturation and the knowledge of miRNA expression in backfat tissue is still fragmentary. Moreover, the identification of the porcine backfat mRNAs transcription profile is still to be completely defined because gene expression changes between breeds, developmental stages and rearing condition. This work describes the backfat tissue transcription profiles of sRNAs, and mRNAs, in Italian Large White pigs and reports genes differentially expressed between fat and lean animals obtained with RNA-seq. The backfat transcription profile was characterized by the expression of 23,483 genes of which 54.1% were known genes. Of 63,418 expressed transcripts about 80% were non previously annotated isoforms. By comparing the expression level of fat vs. lean pigs we detected 86 differentially expressed (DE) mRNAs, and 31 DE sRNAs. The main functional categories enriched in DE mRNAs were immune system process, response to stimulus, cell activation, skeletal system development, for the overexpressed genes, unfolded protein binding and stress response, for the under-expressed ones. Adipose tissue alterations and changes in stress response are linked to inflammation and, in turn, to adipose tissue secretory activity similarly to what is observed in human obesity. To understand the biological impact of the observed miRNA expression variations, we analyzed the co-expression of DE miRNA and their target transcripts in the same samples with the aim to define a regulatory network of interactions between DE miRNAs and DE target transcripts showing opposite expression profiles. The knowledge of relationships between the two categories of transcripts opens new possibilities to understand the patterns of gene regulation in pigs with different aptitudes for backfat deposition. These results highlight some of the gene networks involved in fat traits and suggest genes that can be further investigated for the identification of new biomarkers for the implementation of innovative strategies in pig genetic selection.

Functional annotation of GWAS regulatory variants powered by Ensembl

Daniel R. Zerbino 1,2, Myrto Kostadima 1,2, Thomas Juettemann 1, Michael Nuhn 1,2, Ilias Lavidas 1,2, Jose Carlos Marugan 1,2, Verena Zuber 1,2, William Jones 1,2,4, William Newell 2,3, Gareth Peat 1,2, Andrew Hercules 1,2, Miguel Carmona 1,2, Maya Ghoussaini 2,4, Sarah Spain 2,4, Gautier Koscielny 2,3, Ian Dunham 1,2, Oliver Stegle 1,2

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
2. Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
3. GSK, Medicines Research Center, Gunnels Wood Road, Stevenage, SG1 2NY, UK
4. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

Ensembl is one of the world's leading sources of information on the structure and function of the genome. It brings together genome sequences, genes, non-coding RNAs, known variants, and other data to create an up-to-date, comprehensive and consistent resource. The Ensembl annotations are widely used for the analysis and interpretation of genome data using tools such as the Ensembl Variant Effect Predictor (VEP), which can quickly annotate the known variants of an individual and report on the potential effects of each.
However, the analysis of individual variants remains challenging when analysing the results of genome wide association studies (GWAS). In a majority of cases, the associated variants are not likely causal coding variants, rather potential regulatory variants with weak phenotypic associations. Here, we present how weak associations with causal genes can be detected through a genome-wide and multi-layered integrative analysis.
Ensembl's Regulatory Build synthesises epigenomic datasets produced by large-scale projects such as ENCODE, Roadmap Epigenomics or BLUEPRINT. The resulting regulatory annotation defines biochemically active regions across 68 human cell types and 79 mouse cell types, assigning them a function wherever possible. To support the Regulatory Build, Ensembl maintains the International Human Epigenome Consortium's (IHEC) Epigenome Reference Registry (EpiRR), where large epigenomic consortia collect the metadata describing their datasets.
Regulatory elements are chiefly of interest because of their effect on gene expression. To gain further insight, Ensembl is developing a database of cis-regulatory interactions that attach them to their target genes; as a first step all of the GTEx summary eQTL data is incorporated and can be accessed and viewed.
Having brought all this data together, it is possible to develop advanced functional analysis methods without being constrained by the scale of the data, as we demonstrate with our post-GWAS analysis platform, Postgap. This algorithm compares human GWAS results, as stored in public archives or in a private individual study, to a collection of genomic annotations, many of which are stored in Ensembl, producing a list of putative causal genes along with linked evidence. It optionally integrates a Bayesian colocalisation algorithm that stochastically tests possible sets of causal variants at each GWAS peak. Postgap can be run locally or precomputed results can be explored on the Open Targets Genetics Portal.

Notes

Notes

Notes

Delegate list

Naila Adam
New York University Abu Dhabi
nsa313@nyu.edu

Jessica Adams
Royal Devon and Exeter NHS
j.adams8@nhs.net

Stuart Aitken
MRC HGU, University of Edinburgh
stuart.aitken@igmm.ed.ac.uk

Craig Anderson
The University of Edinburgh
craig.anderson@igmm.ed.ac.uk

Federico Ansaloni
SISSA
fansalon@sissa.it

Christian Arnold
EMBL
christian.arnold@embl.de

Endre Bakken Stovner
NTNU
endrebak85@gmail.com

Tracy Ballinger
IGMM
tracy.ballinger@igmm.ed.ac.uk

Gal Barel
Max Planck Institute for Molecular Genetics
barel@molgen.mpg.de

Jemma Beard
WGCSC
jb36@sanger.ac.uk

Jordana Bell
King's College London
Jordana.Bell@kcl.ac.uk

Jonathan Belyeu
University of Utah
jrbelyeu@gmail.com

Ewa Bergmann
Illumina
ebergmann@illumina.com

Dora Bihary
University of Cambridge
db679@mrc-cu.cam.ac.uk

Simon Boutry
de Duve Institute (UCL)
simon.boutry@uclouvain.be

Bryony Braschi
EBI
bbraschi@ebi.ac.uk

Patrick Brennan
Nationwide Children's Hospital
patrick.brennan@nationwidechildrens.org

Benedikt Brink
Ludwig-Maximilians-Universität München
b.brink@lmu.de

Lucile Broseus
CNRS
lucile.broseus@igh.cnrs.fr

Elspeth Bruford
EMBL-EBI
elspeth@ebi.ac.uk

Carlos Fernando Buen Abad Najar
University of California, Berkeley
cfbuenabadn@berkeley.edu

Laura Buggiotti
The Royal Veterinary College
lbuggiotti@rvc.ac.uk

Kathryn Burdon
University of Tasmania
kathryn.burdon@utas.edu.au

Shane Burgess
The University of Arizona
sburgess@cals.arizona.edu

Mario Caccamo
NIAB EMR
mario.caccamo@niab.com

Sarah Carey
University of Florida
sarah.carey@ufl.edu

Alex Caulton
University of Otago
alex.caulton@agresearch.co.nz

Dineika Chandrananda
University of Cambridge
dineika.chandrananda@cruk.cam.ac.uk

Keira Cheetham
Illumina
kcheetham@illumina.com

Yibu Chen
University of Southern California
yibuchen@usc.edu

Nadia Chuzhanova
Nottingham Trent University
nadia.chuzhanova@ntu.ac.uk

Lieven Clement
Ghent University
lieven.clement@ugent.be

Dorothy Clyde
Springer Nature
d.clyde@nature.com

Cory Colaneri
Indigo Agriculture
ccolaneri@indigoag.com

Camilla Colombo
Illumina
ccolombo@illumina.com

Rachel Colquhoun
University of Oxford
rmnorris@well.ox.ac.uk

Michael Cormier
University of Utah
cormiermichaelj@gmail.com

Carolina Correia
University College Dublin
carolina.correia@ucdconnect.ie

Raúl Oscar Cosentino
Ludwig-Maximilians-Universität München
Raul.Cosentino@para.vetmed.uni-muenchen.de

Andrew Cosgrove
Springer Nature
andrew.cosgrove@genomebiology.com

Anthony Cox
Illumina Cambridge Ltd.
acox@illumina.com

Ottavio Croci
IIT
ottavio.croci@iit.it

Lucy Crooks
Sheffield Hallam University
L.Crooks@shu.ac.uk

Danang Crysnanto
ETH Zürich
danang.crysnanto@usys.ethz.ch

Carla Cummins
EMBL-EBI
carlac@ebi.ac.uk

Nishadi De Silva
EBI
nishadi@ebi.ac.uk

Deniz Demircioglu
Genome Institute of Singapore
ddemircioglu@gis.a-star.edu.sg

Joe Dennis
University of Cambridge
jgd29@cam.ac.uk

Fiona Dick
University of Bergen
fiona.dick91@gmail.com

Diana Domanska
University of Oslo
dianadom@ifi.uio.no

Alessia Donato
University of Siena
alessia.donato@student.unisi.it

Kevin Donnelly
University of Edinburgh
kevin.donnelly@igmm.ed.ac.uk

Ruth Dunn
Goura Victoria Consultants
rdunn_81@yahoo.co.uk

Hannes Eggertsson
University of Iceland
hannese@decode.is

Barbara Engelhardt
Department of Computer Science
bee@princeton.edu

Ailith Ewing
University of Edinburgh
Ailith.Ewing@igmm.ed.ac.uk

Marta Farre Belmonte
Royal Veterinary College
mfarrebelmonte@gmail.com

Andrew Farrell
University of Utah
afarrell@genetics.utah.edu

Fred Farrell
Illumina
ffarrell@illumina.com

Ian Fiddes
10x Genomics
ian.fiddes@10xgenomics.com

Daniel Fischer
Natural Resources Center Finland (Luke)
daniel.fischer@luke.fi

Paul Flicek
EMBL-EBI
kwalsh@ebi.ac.uk

Anna Fowler
University of Liverpool
a.fowler@liv.ac.uk

Richard Francis
Telethon Kids Institute
rfrancis@ichr.uwa.edu.au

Adam Frankish
EMBL-EBI
frankish@ebi.ac.uk

Timothy Freeman
University of Sheffield
tmfreeman1@sheffield.ac.uk

Mike Furness
TheFirstNuomics
mike@thefirstnuomics.com

Daniel Gaffney
Wellcome Sanger Institute
dg13@sanger.ac.uk

Jeff Gaither
Nationwide Children's Hospital
jeffrey.gaither@nationwidechildrens.org

Carlos Garcia Giron
EMBL-EBI
Carlos@ebi.ac.uk

Raquel Garcia Perez
Institute of Ev Biology
raquel.garcia@upf.edu

Philippe Gautier
MRC IGMM University of Edinburgh
philippe.gautier@igmm.ed.ac.uk

Guillaume GAUTREAU
Genoscope
guillaume.gautreau@free.fr

Gregory Gimenez
University of Otago
gregory.gimenez@otago.ac.nz

Madalina Giurgiu
Medizinisch Genetisches Zentrum
Madalina.Giurgiu@mgz-muenchen.de

Jonathan Goeke
Genome Institute of Singapore, A*STAR
gokej@gis.a-star.edu.sg

Sung Sam Gong
University of Cambridge
ssg29@cam.ac.uk

Giorgio Gonnella
University of Hamburg
gonnella@zbh.uni-hamburg.de

Cristina Yenyxe Gonzalez Garcia
EMBL-EBI
cyenyxe@ebi.ac.uk

Mar Gonzalez Porta
Illumina
mgonzalez@illumina.com

Asier Gonzalez Uriarte
Rothamsted Research
asier.gonzalez@rothamsted.ac.uk

Casey Greene
University of Pennsylvania
greenescientist@gmail.com

Graeme Grimes
IGMM, University of Edinburgh
graeme.grimes@gmail.com

Ivar Grytten
University of Oslo
ivar.grytten@gmail.com

Xiaolian Gu
Umeå University
xiaolian.gu@umu.se

Dengfeng Guan
University of Cambridge
dg539@gen.cam.ac.uk

Romain Guitton
University of Bergen
r.guitton@uib.no

Toby Gurran
University of Edinburgh
s1582371@sms.ed.ac.uk

Saber HafezQorani
Genome Sciences Center
shafezqorani@bcgsc.ca

Leanne Haggerty
EMBL-EBI
leanne@ebi.ac.uk

Tom Aharon Hait
Tel-Aviv university
sthait@gmail.com

Mihail Halachev
University of Edinburgh
mihail.halachev@igmm.ed.ac.uk

Michael Hall
EMBL-EBI
michael.hall@ebi.ac.uk

Daniel Halligan
Fios Genomics
dan.halligan@fiosgenomics.com

John Hamilton
Michigan State University
jham@msu.edu

Jennifer Harrow
ELIXIR
jen.harrow@elixir-europe.org

Jim Havrilla
University of Utah
semjaavria@gmail.com

Haynes Heaton
Sanger
whheaton@gmail.com

Raphael Helaers
de Duve Institute (UCL)
raphael.helaers@uclouvain.be

Javier Herrero
UCL Cancer Institute
javier.herrero@ucl.ac.uk

Benjamin Hitz
Stanford University
hitz@stanford.edu

Phuc Hoang
The Institute of Cancer Research
phuc.hoang@icr.ac.uk

Guillaume Holley
University of Iceland
guillaumeholley@gmail.com

Ian Holmes
UC Berkeley
ihholmes@gmail.com

Laura Huerta
EMBL-EBI
lauhuema@ebi.ac.uk

Jack Humphrey
University College London
jack.humphrey@ucl.ac.uk

Ho Wan Ip
Queen Mary Hospital
iphowan@gmail.com

Zamin Iqbal
EMBL-EBI
zi@ebi.ac.uk

Rafael Irizarry
Dana-Farber Cancer Institute
chair@jimmy.harvard.edu

Kathryn Jackson Jones
MRC Institute of Genetics and Molecular Medicine
s1668455@sms.ed.ac.uk

Emma Jones
University College London
e.jones.17@ucl.ac.uk

Chakravarthi Kanduri
University of Oslo
skanduri@ifi.uio.no

Efstathios Kanterakis
Illumina
ekanterakis@illumina.com

Mehran Karimzadeh
University of Toronto
mehran.karimzadehreghbati@mail.utoronto.ca

Ann Marie Keane
Quadram Institute Bioscience
adminfoodandhealth@quadram.ac.uk

Birte Kehr
Berlin Institute of Health
birte.kehr@bihealth.de

Nikka Keivanfar
10x Genomics
nikka.keivanfar@10xgenomics.com

Ben Kelly
Nationwide Children's Hospital
ben.kelly@nationwidechildrens.org

Hossein Khiabanian
Rutgers University
h.khiabanian@rutgers.edu

Jeffrey Kidd
University of Michigan
jmkidd@umich.edu

Jinho Kim
Samsung Medical Center
jinho80@gmail.com

Jun-Mo Kim
Chung-Ang Univeristy
junmokim@cau.ac.kr

Philip Kleinert
Berlin Institute of Health
philipkleinert@hotmail.com

Sriram Kosuri
UCLA
skosuri@gmail.com

Vladimir Kovacevic
Seven Bridges
vladimir.kovacevic@sbgenomics.com

David Kovalic
Webster University
davidkovalic76@webster.edu

Thomas Krannich
Berlin Institute of Health
thomas.krannich@bihealth.de

Joanna Krupka
University of Cambridge
jak75@mrc-cu.cam.ac.uk

RAJA AMIR HASSAN KUCHAY
BGSB UNIVERSITY
kuchayamir@gmail.com

Lukas Kuderna
Institut de Biologia Evolutiva (UPF-CSIC)
lukas.kuderna@upf.edu

Richard Kuo
University of Edinburgh
richard.kuo@roslin.ed.ac.uk

Delphine Lariviere
Pennsylvania State University
lariviere.delphine@gmail.com

Dillon Lee
University of Utah
dlee123@gmail.com

Ellen Leffler
University of Oxford
leffler@well.ox.ac.uk

Richard Leggett
Earlham Institute
business.support@earlham.ac.uk

Hui Sun Leong
CRUK Manchester Institute
HuiSun.Leong@cruk.manchester.ac.uk

Ion Lerga Jaso
Universitat Autònoma Barcelona
jlerga@alumni.unav.es

Morag Lewis
King's College London
morag.lewis@kcl.ac.uk

Jianjun Liu
Genome Institute of Singapore
liuj3@gis.a-star.edu.sg

Liubov Lonishin
D.O.Ott Research Institute
liubov.lonishin@gmail.com

Ernesto Lowy Gallego
EMBL EBI
ernesto@ebi.ac.uk

Juliet Luft
MRC Human Genetics Unit
juliet.luft@ed.ac.uk

Nina Luhmann
University of Warwick
N.Luhmann@warwick.ac.uk

Rachel Lyne
InterMine
rachel@intermine.org

Shamoni Maheshwari
10x Genomics
shamoni.maheshwari@10xgenomics.com

Joseph Mahon
Leeds Teaching Hospitals Trust
joseph.mahon@nhs.net

Anup Mahurkar
Institute for Genome Sciences
amahurkar@som.umaryland.edu

Klaus Maisinger
Illumina
kmaisinger@illumina.com

Andreas Maos
MedImmune Ltd
slidelt@medimmune.com

David Martín-Gálvez
Complutense University of Madrid
dmartingalvez@ucm.es

Shane McCarthy
University of Cambridge
sam68@gen.cam.ac.uk

Alistair Mcdougall
EMBL-EBI
amcdouga@ebi.ac.uk

Mark McDowall
EMBL EBI
mcdowall@ebi.ac.uk

David McGaughey
National Eye Institute (NIH)
mcgaugheyd@mail.nih.gov

Páll Melsted
University of Iceland
pmelsted@gmail.com

Nana Mensah
Guy's and St Thomas' NHS Trust
Nana.mensah1@nhs.net

Wouter Meuleman
Altius Institute for Biomedical Sciences
meuleman@gmail.com

Alison Meynert
MRC Human Genetics Unit
alison.meynert@igmm.ed.ac.uk

Younes Mokrab
Sidra Medicine
ymokrab@sidra.org

James Morris
University of Cambridge
james.morris@cruk.cam.ac.uk

Tanmoy Muckherjee
MRC Cancer Unit
tm646@mrc-cu.cam.ac.uk

Jonathan Mudge
EMBL EBI
jmudge@ebi.ac.uk

Matthieu Muffato
EMBL EBI
muffato@ebi.ac.uk

Zemin Ning
Wellcome Sanger Institute
zn1@sanger.ac.uk

Frank Nothaft
Databricks
frank.nothaft@databricks.com

Denye Ogeh
EMBL-EBI
do1@ebi.ac.uk

Jacob Oppenheim
Indigo Ag
jnoppenheim@gmail.com

Joshua Orvis
Institute for Genome Sciences
jorvis@gmail.com

Georg Otto
UCL Institute of Child Health
g.otto@ucl.ac.uk

Alicia Oshlack
MCRI
alicia.oshlack@mcri.edu.au

Donghyun Park
Samsung Medical Center
eastwise37@gmail.com

Eddie Park
CHOP
eddiep@uci.edu

Andrew Parton
EMBL EBI
aparton@ebi.ac.uk

Liam Paul
Illumina Inc.
lpaul@illumina.com

Luca Penso Dolfin
Earlham Institute
luca.penso-dolfin@earlham.ac.uk

Stephanie Pitts
Stellenbosch University
steph15@sun.ac.za

Katherine Pollard
Gladstone Inst of Data Science & Biotechnology
kpollard@gladstone.ucsf.edu

Claudia Pommerenke
Leibniz-Institute DSMZ
claudia.pommerenke@dsmz.de

Thomas Sean Powell
IMBA-Institut für Molekulare Biotechnologie GmbH
sean.powell@imba.oeaw.ac.at

Rodrigo Pracana
Quen Mary University of London
rodrigopracana@gmail.com

Anurag Priyam
Queen Mary University of London
anurag.priyam@qmul.ac.uk

Christopher Pyatt
NCYC - Quadram Institute
christopher.pyatt@quadram.ac.uk

Aaron Quinlan
University of Utah
aaronquinlan@gmail.com

Knut Dagestad Rand
University of Oslo
knutdrand@gmail.com

Philipp Rentzsch
Berlin Institute of Health
philipp.rentzsch@bihealth.de

Alessandro Riccombeni
DNAnexus
ariccombeni@dnanexus.com

Maria Rigau
Barcelona Supercomputing Center
maria.rigau@bsc.es

Mark Robinson
University of Zurich
mark.robinson@imls.uzh.ch

Julie Rodor
University of Edinburgh
Julie.Rodor@ed.ac.uk

Maša Roller
EMBL-EBI
roller@ebi.ac.uk

Kevin Rue Albrecht
University of Oxford
kevin.rue-albrecht@kennedy.ox.ac.uk

Stefania Salvatore
University of Oslo
stefasal@ifi.uio.no

Shamith Samarajiwa
University of Cambridge
ss861@mrc-cu.cam.ac.uk

Kaitlin Samocha
Wellcome Sanger Institute
ks20@sanger.ac.uk

Geir Kjetil Sandve
University of Oslo
geirksa@ifi.uio.no

Anna Saukkonen
King's College London
anna.saukkonen.17@ucl.ac.uk

Christoph Schlaffner
WT Sanger Instsitute
christoph.schlaffner@childrens.harvard.edu

Sebastian Schoenherr
Medical University of Innsbruck
sebastian.schoenherr@i-med.ac.at

Ellen Schofield
Animal Health Trust
ellen.schofield@aht.org.uk

Dominik Seelow
Berlin Institute of Health
dominik.seelow@charite.de

Stefan Seemann
University of Copenhagen
seemann@rth.dk

Colin Semple
MRC Human Genetics Unit
colin.semple@igmm.ed.ac.uk

Mahloro Hope Serepa
University of the Witwatersrand
hopeserepa@gmail.com

Riccha Sethi
TRON (Translation Oncology)
riccha.sethi@tron-mainz.de

Seungho Shin
Samsung Medical Center
sin12ho@gmail.com

Tim Slidel
MedImmune Ltd
slidelt@medimmune.com

Graeme Smith
The University of Manchester
graeme.smith-2@postgrad.manchester.ac.uk

Guillaume Smits
Medical Genetics - IB2 - ULB
guillaume.smits@erasme.ulb.ac.be

Dae Soon Son
Samsung Medical Center
ds3.son@samsung.com

Jawon Song
University of Texas at Austin
jawon@tacc.utexas.edu

Lucia Spangenberg
Institut Pasteur de Montevideo
lucia@pasteur.edu.uy

Mirjam Spengeler
Qualitas AG
mirjam.spengeler@qualitasag.ch

Elena Speretta
EMBL-EBI
esperett@ebi.ac.uk

Prithika Sritharan
Quadram Institute Bioscience
julie.buckenham@quadram.ac.uk

Alexander Suh
Uppsala University
alexander.suh@ebc.uu.se

Najeeb Ashraf Syed
Sidra Medicine
nsyed@sidra.org

Lin Tang
Nature Communications
lin.tang@nature.com

Peter Taschner
University of Applied Sciences
taschner.p@hsleiden.nl

Martin Taylor
University of Edinburgh
martin.taylor@igmm.ed.ac.uk

Philip Tedder
illumina
ptedder@illumina.com

Sarah Teichmann
Wellcome Trust Sanger Institute
st9@sanger.ac.uk

Christian Tendeng
Fios Genomics
christian.tendeng@fiosgenomics.com

Huseyin Serhat Tetikol
Seven Bridges Genomics
serhat.tetikol@sbgenomics.com

Anil Thanki
Earlham Institute
anil.thanki@earlham.ac.uk

Anja Thormann
EMBL-EBI
anja@ebi.ac.uk

Hagen Tilgner
Weill Cornell Medicine
hut2006@med.cornell.edu

Minerva Susana Trejo Arellano
Swedish University of Agricultural Sciences
minerva.trejo@slu.se

Ian Tully
Cardiff University
tullyij1@cardiff.ac.uk

Remco Ursem
HZPC
remco.ursem@hzpc.com

Koen Van den Berge
Ghent University
koen.vandenberge@ugent.be

Imke van Ettinger
University of Edinburgh
s0790250@sms.ed.ac.uk

Kristoffer Vitting-Seerup
University of Copenhagen
kristoffer.vittingseerup@bio.ku.dk

Massimiliano Volpe
Stazione Zoologica Anton Dohrn
mas.volpe@gmail.com

Lixiao Wang
Umeå University
lixiao.wang@umu.se

Dennis Wang
NIHR Sheffield BRC
dennis.wang@sheffield.ac.uk

Helen Warren
NHS
helen.warren4@nhs.net

Mark Wass
University of Kent
m.n.wass@kent.ac.uk

Peter White
Nationwide Children's Hospital
peter.white@nationwidechildrens.org

Stephen Williams
10x Genomics
stephen.williams@10xgenomics.com

Hywel Williams
UCL Institute of Child Health
hywel.williams@ucl.ac.uk

Jesse Willis
Center for Genomic Regulation (CRG)
jesser.willis@crg.eu

Melissa Wilson Sayres
Arizona State University
melissa.wilsonsayres@asu.edu

Yannick Wurm
Queen Mary U London
y.wurm@qmul.ac.uk

Sergei Yakneen
EMBL
llevar@gmail.com

Chen Yang
Genome Sciences Center
cheny@bcgsc.ca

Rob Young
MRC Human Genetics Unit
robert.young@igmm.ed.ac.uk

Paolo Zambonelli
Bologna University
paolo.zambonelli@unibo.it

Luke Zappia
Murdoch Children's Research
Institute/University of Melbourne
luke.zappia@mcri.edu.au

Daniel Zerbino
EMBL-EBI
zerbino@ebi.ac.uk

Martine Zilversmit
American Mus. of Nat. History
mzilversmit@amnh.org

Antonino Zito
King's College London
antonino.zito@kcl.ac.uk

Gregory Zynda
TACC
gzynda@tacc.utexas.edu

Index

Anderson	P1
Ansaloni	P2
Arnold	P3
Bakken Stovner	P4
Ballinger	S31
Barel	P5
Bell	S69
Belyeu	S97
Bergmann	P6
Bihary	P7
Boutry	P8
Braschi	P9
Brennan	S21
Brink	P10
Broseus	S63
Buen Abad Najar	P11
Buggiotti	P12
Caccamo	S43
Carey	S89
Caulton	P13
Clement	P14
Colquhoun	S95
Cormier	P15
Cosentino	P16
Cox	P17
Croci	P18
Crooks	P19
Cummins	S49
De Silva	S11
Demircioglu	P20
Dennis	P21
Dick	S67
Domanska	P22
Donato	P23
Donnelly	P24
Eggertsson	P25
Englehardt	S61
Farre Belmonte	S41
Farrell	P26
Fiddes	S99
Fowler	P27
Frankish	P28
Freeman	P29
Gaffney	P30
Gaither	S65
Garcia Giron	P31
Garcia Perez	S83
Gautreau	S37
Gimenez	P32
Giurgiu	P33
Gong	P34
Gonnella	S3
Gonzalez Garcia	P35
Gonzalez Porta	P36
Greene	S9
Gurran	P37
HafezQorani	P38
Haggerty	P39
Hait	P40
Hamilton	P41
Harrow	P42
Havrilla	P43
Helaers	P44
Holley	P45
Holmes	P46
Huerta	S7
Humphrey	P47
Irizarry	S51
Jackson Jones	P48

Kanduri P49
Karimzadeh P50
Kehr P51
Keivanfar S57
Kelly P52
Khiabanian S29
Kidd S93
Kim P53
Kleinert P54
Kosuri S19
Kovacevic P55
Krannich P56
Krupka P57
Kuchay P58
Kuderna P59

Lariviere P60
Lee P61
Leffler S35
Lerga Jaso P62
Lewis P63
Lonishin P64
Lowy Gallego P65
Luft P66
Luhmann S45
Lyne S15

Maheshwari S25
Mahurkar P67
Martín-Gálvez P68
McCarthy P69
Mcdougall P70
McDowall P71
McGaughey P72
McGaughey P73
Melsted P74
Meuleman S71
Muckherjee P75
Mudge P76
Muffato P77

Ning S91
Nothaft P78

Ogeh P79
Orvis P80
Otto P81

Park P82
Parton P83
Penso Dolfin S39
Pitts P84
Pollard S17
Pracana P85
Priyam P86
Pyatt S47

Rentzsch P87
Robinson S53
Rodor P88
Roller S73
Rue Albrecht P89

Salvatore P90
Samocha S27
Sandve P91
Schofield P92
Serepa P93
Sethi P94
Shin P95
Song P96
Spangenberg S33
Speretta P97
Sritharan S87
Suh S77

Taschner	P98
Teichmann	S1
Tetikol	P99
Thanki	P100
Thormann	P101
Tilgner	S55
Trejo Arellano	S81
Van den Berge	S59
Vitting-Seerup	P102
Volpe	S75
Wang	S23
Wass	P103
White	P104
Williams, S	P105
Williams, H	P106
Willis	P107
Wilson Sayres	S85
Wurm	P108
Yakneen	S13
Yang	P109
Zambonelli	P110
Zappia	S5
Zerbino	P111
Zito	S79