CSHL ‘Genome Informatics’ conference report – September 2018

Report by Tom Hait, October 1, 2018; Reviewed by Ron October 2, 2018

This meeting was held on 17-20 September 2018 at Wellcome Genome Campus (WGC), Hinxton, South Cambridgeshire, England. Hinxton is a village and a home to the WGC, which includes the Wellcome Trust Sanger Institute, and the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI).

The purpose of Genome Informatics meeting is to bring together researchers and PhD students from different bioinformatics fields for sharing new research directions. The conference was divided into 6 sessions from different research fields and two keynote talks. The number of participants was 252 (mostly Europeans, few Americans, and a single Israeli). This meeting is held once a year either at CSHL or at Cambridge. Abstract book is available in the lab and in the link. Below I tried to cover talks that were relevant to our group members from different sessions.

A personal note:

This is my second time in a row participating in this conference. In the first time I was really impressed by the high number of tweets in Twitter made on the talks (hashtag #GI2017; see statistics of tweets), which convinced me to start following scientists in Twitter in order to get ’out of the oven’ new updates about discoveries and publications, and to understand the scientific language in my field. This year was not different and the number of tweets was even higher (hashtag #GI2018; see statistics of tweets), allowing me to easily follow and summarize the interesting talks. So my message to you – If you haven’t opened a twitter account yet then please do so, and start following your relevant scientists.

In addition, this year the organizers added 'lightning talks' – 1 minute talks to present posters as done in our Safra retreats. Only a subset of poster presenters was selected (due to time limit) and each presenter was allowed to present a single slide without animations, which saved time. The delegates were very positive to these lightning talks and said that this session was one of the interesting parts of the conference. The organizers also added a poster/slides channel to upload the posters and slides presented in the conference. In this channel the posters/slides are given a DOI and are citable.

The next meeting will be held at CSHL on November 6-9, 2019, and I hope more people from our lab and from Israel in general will attend it.

Posters: Some poster presenters uploaded their posters to the poster channel. I took pictures of other posters, and they are available at: https://www.dropbox.com/sh/3c6z9gdo4w07zdl/AABjOd5l2b9aUeyQeUWt viwa?dl=0.

There are relevant posters for most of you covering CNVs/SNPs, De Bruijn Graph coloring, k-mer indexing of large genomes, and more.
Keynote talk: A population genetic view of human chromatin organization

Katherine S. Pollard – Pollard’s group, Gladstone Institute of Data Science & Biotechnology and UCSF, San Francisco, CA, USA

It is well known that most of the human disease-associated variants (via GWAS) and human accelerated regions (HARs) are non-coding. In terms of interpretation, one can think about disruption of transcription factor binding motifs (TFBMs) but also disruption of 3D enhancer-promoter folding. Pollard raised three hypotheses that were tested in three preprints from her lab during her talk in which I briefly cover them below.

The first hypothesis is: if the genome structure is functional then mutations that change the structure and function would be more deleterious. It is well known that disruption of topological associated domain (TAD) boundary elements (BEs), e.g., TAD duplication/inversion/deletion, results with changes of enhancer’s target genes/promoters leading to diseases. The effect of structural variants on chromatin features like TAD BEs has received relatively little systematic attention outside of cancer until the past year (2017). Pollard showed by systematic evolutionary analysis that structural variants disrupting TAD BEs are under strong negative selection (i.e., depleted/vanished) between apes, healthy and sick human genomes, which allow identification of rare variants in human healthy/sick populations (Fig1). This signature of negative selection is strongly absent in patients with autism and developmental delay, where deletions occur uniformly across the genome (Fig1; Fudenberg & Pollard, bioRxiv, 2018).
The main message here is that variant effect predictors should model TAD BEs explicitly (e.g., noncoding mutation scoring tools).

The second hypothesis is: mapping non-coding variants to the closest gene promoter or to promoters within the same linkage disequilibrium (LD) block is not a good practice. Past examples have shown that regulatory variants can affect phenotypes by changing the expression of target genes up to several megabases (mb) away, far beyond their LD block. To show that the gene’s regulatory program is not related to local LD haplotype structure, Pollard screened 22 cell types and compared the block structure of the LD (computed using plink haplotype block estimation on each of the 5 super populations in 1000 Genomes; see algorithm in Gabriel et al., Science, 2002) and the chromatin (using HiC). Fig2 shows an example of such comparison between NHEK and HUVEC cells (Whalen & Pollard, biorxiv, 2018). The results show that chromatin interactions are more enriched for expression quantitative trait loci (eQTL) and gene ontologies (GOs; for further details see ‘Gene Ontology Enrichment’ section in Whalen & Pollard, biorxiv, 2018) than does the closest gene within the SNP's LD block. Comparison of nested blocks (overlapping LD and chromatin
blocks) shows that these blocks are uncorrelated in distances > 5kb. Most distal variants (>5kb) do not map to their nearest gene and are in low LD with SNPs closer to their target genes.

Figure 2. LD blocks (green squares), HiC contact domains (dashed purple squares), and significant HiC loops (yellow box). It can be seen that in (b) for NHEK cells the SNP (rs17032996) in LD block with the closest gene MANBA, however, according to eQTL and HiC this SNP interacts with SLC9B2. This is not the case in (c) for HUVEC cells where this interaction is missing.

The third hypothesis is: transcription factors (TFs) can recognize DNA shape even without nucleotide sequence recognition. Focusing on the set of DNA binding proteins (DBPs), Pollard showed that the top 30% of 2000 ChIP-seq peaks have no sequence motifs consistent across cell types (Samee et al, bioxiv, 2017). Pollard presented ShapeMF, a method to infer de-novo TF shape-motif, which is a significantly over-represented pattern in the profiles of DNA shape features at the TF’s binding sites as compared to non-bound regions, without taking TF-motif sequence information into account. The shape features are computed based on 5-tuple statistics from GBshape database (Chiu and Yang et al., Nucleic Acid Res, 2015). For instance, we would say that a TF has a minor groove width shape-motif if its binding sites are enriched for windows with a particular sequence of minor groove values (e.g., low, high, low) compared to flanking non-peaks. The results show that most DBPs have a shape-motif and 17% of the tested DBPs only had DNA shape without sequence motif.

Pollard finished with some of her insights on wrong usage of ML approaches in genomics. (1) Using balanced training data for unbalanced problems (e.g., equal numbers of enhancers and non-enhancers when most of the genome is not an enhancer). (2) Failing to realize that good performance is due to "always no" predictor. (3) Selecting features from whole data and then performing cross-validation (feature selection should be done within the cross validation). (4) Using biologically inappropriate negative examples or controls. (5) Utilization of performance statistics that may not reflect the biological goal (e.g., ROC AUC when we care about the precision or power at low false discovery rates). One of the biggest offenders in ML is that most ML models assume observations are independent and identical distributed (IID). Observations are almost never independent and we should strive to make
them IID as much as possible by understanding the correlation structure. We should start with simple methods (e.g., logistic regression before jumping to the black box deep learning) to understand the coefficients and how they behave under different settings of the parameters. Once we understood what are the confounders and limitations in our data we can proceed to more complex methods.

**Session:** Data Curation, Integration, and Visualization

**Immunogenomics one cell at a time**

**Sarah Teichmann – Teichmann's group, Welcome Sanger Institute, UK**

Single cell (SC) next generation sequencing provides great promises in dissecting the immune responses in different cell states, including antigen receptor diversity, cell-cell interactions in tissues, and how cellular phenotypes change between lymphoid and non-lymphoid tissues.

Teichmann started by presenting the Human Cell Atlas, which is the "google maps" of the human body. We wish to develop google-like methods for the human cells (e.g., Google street view). There is an exponential growth in the number of cells assayed in a single experiment of SC as the technology advances ([Fig.1](#); Svensson et al, Nat Prot, 2018). To compare different technologies in terms of power analysis, one need to use spike-in molecules added manually to the experiment instead of using the endogenous mRNA molecules since they differ by cell type and condition ([Fig. 2a; Svensson et al, Nat Prot, 2017](#)). Sequencing depth below <1M reads can affect the sensitivity (spike-in detection) but less the accuracy (spike-in prediction levels; [Fig. 2b-c](#)) across technologies (see Figure 2 a-b in Svensson et al, Nat Prot, 2017).

![Figure 1. The number of cells assayed in each single cell technology](#)

**Figure 1.** The number of cells assayed in each single cell technology
Next, Teichmann presented spatialDE (Svensson et al, *Nat Methods*, 2018), a method to find spatially variable (SV) genes in scRNA-seq. Each gene may have different expression levels across cells within the same tissue. SpatialDE builds on Gaussian process regression, a class of models used in geostatistics. Briefly, for each gene, SpatialDE decomposes expression variability into spatial and nonspatial components (Fig. 3a,b), using two random effect terms: a spatial variance term that parametrizes gene expression covariance by pairwise distances of samples, and a noise term that models nonspatial variability. The ratio of the variance explained by these components quantifies the fraction of spatial variance. One can identify significant SV genes by comparing this full model to a model without the spatial variance component. SpatialDE was applied on mouse olfactory bulb data (see organ image on the right). SpatialDE identified 67 SV genes while some of them had periodic (distance b/w the centers and hemispheres; orange bar) or general spatial (not periodic; blue bar) trends. More examples in the paper.
Figure 2. Power analysis of single cell sequencing. (a) Endogenous mRNA levels vary by cell type and condition and cannot be used to compare protocols applied to different cell types. By contrast, protocols can be compared, regardless of cell type, by measuring the same spike-in RNA standards added at known concentrations to all experiments. (b,c) We define two global technical performance metrics for spike-ins: sensitivity, the number of input spike-in molecules at the point at which the probability of detection reaches 50% (b), and accuracy, the Pearson product-moment correlation ($R$) between estimated expression levels and actual input RNA-molecule concentration (ground truth) (c). TPM, transcripts per million.
Figure 3. (a) In spatial gene expression studies, expression levels are measured as a function of spatial coordinates of cells or samples. SpatialDE defines spatial dependence for a given gene by using a nonparametric regression model, testing whether gene expression levels at different locations covary in a manner that depends on their relative location, and thus are spatially variable. (b) SpatialDE partitions expression variation into a spatial component (using functional dependencies $f(x_1, x_2)$), characterized by spatial covariance, and independent observation noise ($\psi$). Representative simulated expression patterns are plotted below the corresponding covariance matrices for the null model (None) and the alternative model (Spatial covariance) with different length scales.

Flexible and interactive visualization of GFA sequence graphs

Giorgio Gonnella – Stefan Kurtz’s group, University of Hamburg, Germany

Gonnella presented a graphical fragment assembly (GFA) tool for visualization and saving sequence graphs. Graph assemblers need a way to communicate and share their graphs in an acceptable and common file format. The first representation was FastG (2011) and it was replaced by GFA2 (nowadays). Here is a nice link to Paull Melsted blog on this issue. The mentioned GFA tools were missing a good a way to visualize the graphs. Gonnella developed GfaVis, which will be published later this year, a tool that can visualize scaffolding between contigs, can visualize read alignments, and has pipelines that transform minimap2 PAF (Pairwise mApping Format) output to GFA2 formats and reads to contigs (sam format to GFA2). GfaVis is inspired by Bandage graph assembly visualization tool. It’s worth following Gonnella's github page for further updates.

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

Laura Huerta – Irene Papatheodorou's group, EMBL-EBI, UK

Huerta presented Expression atlas, an open-access resource for gene expression data that has >3,300 curated datasets (from GTEx, ENCODE, and FANTOM5) across 45 species and various conditions dominated by humans (1,189 datasets). It also includes a large number of RNA-seq experiments. Most of the datasets (3,200) contain differential expression analysis. The resource also performed cancer variation analysis using information taken, among
others, from PCAWG consortium, allows searching genes with similar expression patterns to
the query gene, and provides nice visualizations including heatmaps.

Next, Huerta presented the Single Cell Expression Atlas, an open-access resource of
\( \sim 53,000 \) cells, 43 SC experiments, and 9 species including human and mouse. Similar
analyses are done as in the regular Expression Atlas. In the future they will design an
interface for zooming/combining both bulk and SC RNA-seq data in one place, and meta-
analyses for batch correction to derive gene expression signals and differentially expressed
genes.

Can "big data" help us tackle rare diseases? - slides

Casey Greene – Greene lab, University of Pennsylvania, USA

Greene decided to switch his planned talk and to discuss about how we can incorporate
pathway information and \( \sim 3.7 \)M gene expression assays (worth \( \sim \$3.8 \) billion) to uncover
latent variables by genes and samples using matrix decomposition. Greene mentioned in
twitter that pathway information can be replaced with het.io knowledgebase, which is an
integration of biomedical data (e.g., drugs, pathways, genes, diseases) into a single resource.

Greene presented PLIER (Pathway Level Information Extractor; Mao et al, bioRxiv,
2017), a method that tries to find \( k \) latent variables (LVs) giving for each gene a loading
measure in a set of pathways/genesets across samples/conditions (see Fig.1.a). PLIER was
applied on 35 human blood samples assayed by RNAseq and direct CyTOF measurement
of cell type proportions. 605 pathways, which included 60 cell-type markers and 555 canonical
pathways, were taken from MSigDB. 14 LVs were annotated with high confidence (AUC>0.7,
FDR<0.05, see methods in Mao et al, bioRxiv, 2017) to one or more genesets, of which 9
represented cell types also measured by CyTOF. The correlation between the cell types
PLIER LVs and CyTOF measurement achieved a mean of 0.76 across 35 samples, which was a
bit higher than other methods (Fig.1.b).

Next, Greene questioned whether we can infer biological meaningful patterns in
large gene expression data and then transfer them to small datasets of interest. ML methods
are suitable for analyzing/interpreting large datasets. However, usually datasets created on
a small scale must be combined in order to fully benefit from ML methods applied on them
(see our ADEPTUS paper, for more details). Greene presented MultiPLIER (Taroni et al,
bioRxiv, 2018), a method that applies PLIER on large datasets and project the small datasets
onto the learnt LVs’ space. PLIER was applied on recount2 large compendium, which is
comprised of approximately 37,000 samples. The large sample size and breadth of molecular
processes represented increases the number of gene sets that the model captures and the
specificity of the model—the learned LVs are more interpretable due to a “separation” of
signals. Once the individual datasets are projected to the same latent space learnt from
recount2, the LVs can serve as input into numerous downstream applications, including
supervised tasks such as predicting response to treatment or testing for differential
expression. The comparison between MultiPLIER and PLIER shows a slightly higher
prediction power to the former (see Figure 3 in Taroni et al, bioRxiv, 2018).
**Figure 1.** PLIER overview. PLIER is a matrix factorization approach that decomposes gene expression data into a product of a small number of latent variables and their corresponding gene associations or loadings, while constraining the loadings to align with the most relevant automatically selected subset of prior knowledge. Given two inputs, the gene expression matrix $Y$ and the prior knowledge (represented as binary geneset membership in matrix $C$), the method returns the latent variables ($B$), their loadings ($Z$), and an additional sparse matrix ($U$) that specifies which (if any) prior information genesets and pathways are used for each latent variable. The light gray area of $U$ indicates the large number of zero elements of the matrix. We apply our method to a whole blood human gene expression dataset. The positive entries of the resulting $U$ matrix are visualized as a heatmap, facilitating the identification of the correspondence between specific latent variables deconvolved from the data and prior biological knowledge-based geneset categories. We validate the latent variables mapped to specific cell-types by comparing PLIER estimated relative cell-type proportions with direct measurements by Mass Cytometry. We find that the PLIER estimates are highly accurate, outperforming other matrix decomposition methods as well as the state-of-the-art dedicated blood mixture deconvolution method, Cibersort [Newman et al., 2015]. (*) indicates that the Cibersort estimate had near zero negative correlation with the directly measured proportion for these cell-types. Taken from Mao et al, *bioRxiv*, 2017.
Session: Personal and Medical Genomics

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

**Dennis Wang** – NIHR Sheffield Biomedical Research Centre, University of Sheffield, UK

Wang questioned why two similar persons with the same lung adenocarcinoma disease, stage, non-smokers, and EGFR mutation have different recurrence profiles when given the EGFR inhibitor drug ‘gefitinib’? He assumed that there are non-driver somatic alterations that confer good prognosis in non-small cell lung cancer (NSCLC). Wang used NSCLC-patient-derived xenograft (PDX; Tentler et al. Nat Rev. Clin Onc 2012) genomes, which are mouse models with transfected patients’ removed tumors and are used for drug testing and identifying patient-specific biomarkers. Wang applied a penalized regression (probably Cox proportional hazards model) to identify 865 genes for which high burden of somatic CNVs (also known as: SCNAs) and point mutations are associated with longer disease-free survival (D Wang et al, biorxiv, 2018; top right figure; NAG – number of altered genes). 95% of somatic alterations in these genes were not previously associated with cancer. When comparing SNV and CNA frequency in primary, cell line and PDX, PDXs seem to have a lower number of SCNA, but sample size (n=36) was too small to draw a conclusion (mid right figure – SNV & CAN frequency). However, PDXs and patient tumors were more similar in protein expression than that to cell-lines (mid right figure – heatmap). As a diagnostic tool for treatment, PDX models are not useful for cancer patients as the patients pass away before the models are ready. However, available PDX models can be compared with similar patients in combination with TCGA/GEO datasets to predict the mutation burden impact on survival rate (see pipeline in bottom right figure).

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

**Hossein Khibanian** – Khibanian Lab, Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ, USA

In a research setting we call somatic mutations by comparing cases vs. normal. However, in a clinical setting, normal patients are not always available. Khibanian suggested modeling the somatic and germline mutation (i.e., variant allele frequencies; VAFs) separately according to the cell ploidy (total number of chromatids in all cells) and tumor purity (fraction of tumor cells out of all cells in the solid tumor, including normal ones). The method suggested, LOHGIC (Loss Of Heterozygosity-Germline Inference Calculator), finds the most consistent model for mutations’ germline-versus-somatic status, and infers LOH, mutant allele’s copy-number, and cancer cell fraction.
Given a high-depth clinical DNA sequencing we measure the tumor purity (defined as $p$) and the heterozygous somatic/germline VAF (right Figure). We wish to estimate the tumor purities, $p$, as follows (the denominator is the tumor ploidy – weighted average of normal/cancer chromosome duplication cells, the numerator is the number of mutant alleles in normal/cancer cells while in germline we expect that mutant allele to be inherited in normal cells; For further details see this paper):

The likelihood of observing a variant with frequency $f$ at sequencing depth $D$ ($f$ and $D$ are known) is:

$$L = \text{Bin}(f \times D | D, (VAF))$$

Model selection is performed using the AIC statistic:

$$AIC = -2 \log(L) - 2, \text{ and } \Delta AIC = AIC - \min(AIC)$$

AIC weights are defined as:

$$W = \frac{\exp(-\frac{1}{2}\Delta AIC)}{\sum \exp(-\frac{1}{2}\Delta AIC)}$$

Now since there were uncertainties in measuring $f$ (dependent on $D$) as well as estimating $p$, Khibanian suggested using confidence intervals in model selection:

$$W_{\text{Mut. Model}} = \sum_{p_i \in \text{Purity CI}} \sum_{f_i \in \text{VAF CI}} W(f_i | (VAF)(p_i))$$

On the right, we have an example of confidence intervals for different models with number of mutant alleles in cancer cells (CN or $C_{\text{mut}}$) with $Y=4$ duplicated chromosomes and tumor purity $p=0.6$. Simulations show that the lower bound value for unambiguous inference calls for $W$ or sum of Ws can be set at 0.7 when mutational ploidy is less than 4. Applying LOHGIC on 1,636 solid tumors to access the BRCA1/2 alterations and comparing the results with testing data yielded 93% accuracy, 100% precision, and 96% recall.
(sound too good to be true...). Here is a link to a website where you can estimate tumor purity and VAF.

Session: Comparative, Evolutionary, Metagenomics

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

Nina Luhmann – Achtman Lab, Warwick Medical School, University of Warwick, Coventry, UK

Luhmann described a joint data structure for thousands of bacterial genomes, using colored De-Brujin Graphs (colored dBG; see description in 'Colored de Bruijn graph algorithms' from Iqbal et al., NG, 2012). Luhmann first presented GrapeTree (Zhou et al., Genome Res, 2018; github), a visualization tool of minimum spanning trees of assemblies including support for allelic profiles. Then, Luhman showed how to use Bifrost (github), a tool for constructing and indexing colored and compacted dBGs, on 160K salmonella genomes (available in EnteroBase DB). The tool took 4 running days and 150G memory, and produced 30M unitigs. Bifrost is a C++ api that allows you to work with the graphs generated, and can associate data to unitigs (i.e., set of reads that have been linked together based on overlaps and have no ambiguities).

Luhmann's goal is to query for specific sequences in the constructed colored dBG that are highly similar across assemblies. She first suggested finding overlapping k-mers in the graph (seeds), and then estimating alignment score (extend; similar to what is done in blast seed-and-extend; similar approach to BIGSI by Bradley et al, bioRxiv, 2017). However, for k=31 the results were too specific to few strains. As an alternative, Luhmann used hamming-neighboring k-mers. This solution seems to work for k=31 and 2 mismatches. Such query of sequences could be useful in finding the distribution of pathogenicity islands over different subspecies or within single genomes, 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.

Session: Transcriptomics, Alternative Splicing and Gene Predictions

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

Nikka Keivanfar – 10x Genomics, CA, USA

Nikka described a pipeline for de-novo genome assembly from 10x linked reads taken from long single molecules. The pipeline is as follows: (1) use Supernova V2 genome-assembly (Weisenfeld et al., Genome Res, 2017) on the reads, and (2) use Cactus (Paten et al., Genome Res, 2011), a multi-alignment tool, followed by the Comparative Annotation Toolkit (CAT; Fiddes et al., Genome Res, 2018), to generate an annotated draft reference. Supernova novelty is that it allows identification of diploid genome assemblies. Nikka applied the pipeline on peripheral blood cells from a male donkey. Supernova yielded superior scaffold N50 [Supernova: 49.6 Mb, while aligning to the reference genome gave 3.8 Mb] and contig N50 [Supernova: 494 Kb vs. 66.7 Kb]. CAT generated high-quality annotation on the donkey using the known horse annotation, keeping the track of orthologous relationships (>91% of horse genes have had an identified donkey orthologue gene). scRNA-Seq was applied on
peripheral blood lymphocytes (7,000 sequenced cells clustered based on gene expression profiles) from the same donkey. CAT enabled the annotation of clusters (of cells) representing all expected major cell types, including subtypes present at less than 1%.

Next, Nikka showed how the pipeline can be used on 8 human assemblies (7 had molecule length of approx' 100 kb and one of 10 Kb taken from peripheral blood mononuclear cells) for creating a reference for subsequent gene expression analysis. All 8 assemblies had a performance comparable to aligning to the human reference genome GRCh38 (using Cell Ranger, an automated SC pipeline) in identifying the subpopulations/cell-types.

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

Koen Van den Berge – Clement Lab, Ghent University, Ghent, Belgium

Berge presented a method for finding differentially expressed (DE) genes in scRNA-Seq data. Finding DE genes in scRNAseq is not trivial due to zero-inflation (ZI) problem. There two types of ZI: (1) biological zeros, when a gene is simply not expressed in the cell, and (2) technical zeros (dropouts), when a gene is expressed in the cell but not detected.

Berge introduced the zero inflated negative binomial-based wanted variation extraction (ZINB-WaVE; Risso et al, Nat Com, 2018) to fit ZINB models for every gene. NB distribution is known to model bulk RNAseq count data (edgeR, deseq2) accounting for both Poisson distribution of number of aligned reads to a reference genome of length L and for the different between the sample variance and mean count data (dispersion). ZINB allows down-weighting excess of zeros in subsequent analysis using the posterior probability that a count belongs to the NB count component. Berge proposed a method to combine both ZINB-WaVE method, designed specifically to scRNA-Seq data, and the deseq2/edgeR tools, designed for bulk RNAseq data, for boosting performance for scRNAseq (Van den Begre et al, Genome Biology, 2018).

The ZINB distribution is a two-component mixture between a point mass at zero and a NB distribution. Specifically, the density function for the ZINB-WaVE model is:

\[
(1)f_{ZINB}(y_{ij}; \mu_{ij}, \theta_j, \pi_{ij}) = \pi_{ij} \delta_0(y_{ij}) + (1 - \pi_{ij})f_{NB}(y_{ij}; \mu_{ij}, \theta_j)
\]

where \(y_{ij}\) denotes the read count for cell \(i\) and gene \(j\), \(\pi_{ij}\) the mixture probability for zero inflation, \(f_{NB}(\cdot; \mu_{ij}, \theta_j)\) the NB probability mass function with mean \(\mu_{ij}\) and dispersion \(\theta_j\), and \(\delta_0\) the Dirac delta function. From the ZINB-WaVE density of Eq. 1, one can readily derive the posterior probability that a count \(y_{ij}\) was generated from the NB count component:

\[
(2)w_{ij} = \frac{(1 - \pi_{ij})f_{NB}(y_{ij}; \mu_{ij}, \theta_j)}{f_{ZINB}(y_{ij}; \mu_{ij}, \theta_j, \pi_{ij})}
\]

Berge proposed using \(w_{ij}\) probabilities as weights in bulk RNAseq DE analysis methods (see further details in the Methods). The results show high performance for Berge’s method (tied with EDGER) in identifying the true DE genes from 10x Genomics PBMC single-cell RNA-seq dataset (n=1200) [ZINB-WaVE_DESeq2_common; ZINB-WaVE_edgeR_common].
Comparison of differential expression methods on simulated scRNA-seq datasets. Differential expression methods are compared based on FDP-TPR curves for data simulated from a 10x Genomics PBMC single-cell RNA-seq dataset (n=1200). Zoomed versions of the FDP-TPR curves are shown here and full curves are in Additional file 1: Figure S12. Circles represent working points on a nominal 5% FDR level and are filled if the empirical FDR (i.e., FDP) is below the nominal FDR. 10x Genomics sequencing typically involves high-throughput and massive multiplexing, resulting in very shallow sequencing depths and thus, low counts, making it extremely difficult to identify excess zeros. Unweighted and ZINB-WaVE-weighted EDGER are tied for best performance, followed by ZINB-WaVE-weighted DESEQ2. In general, bulk RNA-seq methods perform well in this simulation, probably because the extremely high zero abundance in combination with low counts can be reasonably accommodated by the negative binomial distribution. The behavior in the lower half of the curve for NODES is due to a smooth increase in true positives with an identical number of false positives over a range of low FDR cut-offs. FDP false discovery proportion, FDR false discovery rate, PBMC peripheral blood mononuclear cell, TPR true positive rate. Source: Van den Begre et al, Genome Biology, 2018

Next, Berge briefly showed a preliminary method for identifying DE genes within and between trajectories. Smooth functions, based on generalized additive model with NB (NB-GAM), of gene expression along developmental pseudotime allow identification of genes that are DE between branching trajectories, resulting in more informed results as compared to discrete DE (data taken from Paul et al, Cell, 2015). He did not have the time to go over the results.
A generative model for single-cell RNA-sequencing

Barbara Englehardt (BE) – Englehardt Lab, Princeton University, NJ, USA

BE introduced a low-dimensional generative manifold model for scRNA-seq to address confounders in the data. Gaussian process latent variable model (GPLVM), is a nonlinear manifold that can be used to describe the generative process of each cell in a scRNAseq data set. GPLVM can address genomic challenges such as normalization and RNA quantification, batch correction, impute drop out, pseudotime estimation, visualization, differential expression, and more. BE incorporates GPLVM with student’s t (tGPLVM) with Q latent variables: $Y \in \mathbb{R}^{n \times p}$ observations, $X \in \mathbb{R}^{n \times Q}$ latent variables, $x_i \sim N_Q(0, I_Q)$ latent variable prior, $f_p(x) \sim N_Q(0, K_{mm})$ noiseless observations, $K(x, x') = \sum_{j=1}^{K} k_j(x, x')$ composite kernel, $y_i \sim \text{StudentT}(f_{i,p}, v, \tau_p^2)$ heavy-tailed residual. Each single cell is an observation in 5 dimensional space (Q=5 latent variables) and we assume that the observations follows a Gaussian process. Clustering of the cells on the 5-latent space (discovered by tGPLVM) was made using k-NN clustering.

BE compared tGPLVM clustering with PCA, t-SNE, ZIFA (zero-inflated factor analysis). The data was taken from 301 human cerebral cortex cells (Pollen et al., Nat Biotech, 2014) and clustering comparison was made using the cell type labels based on adjusted rand index statistic. tGPLVM separates cell types better than PCA and ZIFA, and scales better with higher latent dimensions. t-SNE is sensitive to perplexity parameter, and, therefore, should be regarded as clustering method (It’s great for visualization). BE showed how 3D structured kernels (right figure – representing Cell Cycle) can identify gene’s specific signals. Unsynchronized cells are oscillating in cell state (axes x and y). BE overlaid specific gene expression on the manifold and found that the genes LTB/TYROBP (longitudinal data from Leng et al, Nat Methods, 2015) have an oscillating gradient signals on the manifold (yellow->green->light-blue->blue->purple). Next, BE applied GPfates (Lönnberg et al., Science Immunology, 2017), a method for trajectory construction, with tGPLVM on the same data to identify developmental trajectories and added them to the clustering methods. In this case, PCA was better in clustering. BE hypothesized that maybe they should filter cells that don’t proliferate under treatment for achieving a better clustering on the trajectories.

Session: Epigenetics and non-coding genome

Interpreting variation in the human methylome

Jordana Bell - Jordana Bell Lab, King’s College London, UK

Bell studied methylation variability in 400 twins from the twinsUK cohort in the context of ageing and in age-related diseases. Bell examined 772k CpGs after quality control (QC) and showed high correlation in twin methylation profiles consistent with heritability. 10% of the CpGs show >40% heritability. Bell identified DNA methylation QTL (mQTLs; SNPs associated with methylation level changes) and about 26% of the CpGs tested were under a local genetic control. Repeating in a larger cohort (~2k samples) results with 33% of CpGs tested identified as being under a local genetic control. Bell presented GoDMC (Genetics of DNA
Methylation Consortium), a meta-analysis database that analyzed 450k GWASs from 38 cohorts (n>100 European ancestry samples), i.e., Epigenome-wide association testing. This is computationally challenging, therefore, Bell presented a two stage approach using linear regression models. The regression model treated the methylation at CpGs (i.e., beta values) as the dependent variable (response vector) and the GWAS SNP's genotypes and age, sex, and white-blood cell counts (measured or imputed) as covariates for removing biases.

In the first stage 22 cohorts were analyzed and separated to positive and negative cohorts (i.e., with/without GWAS), and identified significant mQTL had p<10^{-5} (resulting with 120M SNP-CpG pairs). The resulting significant mQTL were later tested in the second stage on all cohorts (n=38). The findings from the fully adjusted model were considered as the primary output. In a cohort size of ~28k (36 chorots), Bell found 190k CpGs with at least 1 mQTL (45%) after Bonferroni correction, the percentage variance explained in all 420k CpGs was 1.3%, and 45% of the methylome was under genetic influence. The majority have cis-SNPs (170k;90%), only a few have trans-SNPs (7k;3.7%), but largest effects on TFBSs are seen for those with both (12k;6.3%). Bell’s conclusions from her study are: (1) SNPs influencing CpG sites are most likely GWAS signals - 19 CpG sites with > 5 mQTLs and GWAS signals from 118 traits, (2) SNPs that are methylation QTLs are more likely to impact human health and disease, and (3) insulators and enhancers are under strong genetic control while promoters are with the least control.

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

**Wouter Meuleman - Altius Institute for Biomedical Sciences, Seattle, USA**

DNase I hypersensitive sites sequencing (DNase-seq or DHS) could give us the idea which region of the genome is accessible. The goal is to find methods for a better mapping of the regulatory regions in the genome and annotation of them using a vocabulary of regulatory regions, which characterizes motifs, genes and pathways. Meuleman’s pipeline is as follows: (1) Create an index of human DHS accessible sites (733 DHS samples; 439 unique cell types; github) using majority vote of samples within each datasets to what region is accessible (21% of genome was covered by DHS sites; median width 200 bp), (2) Decompose the cell types to tissues using non-negative matrix factorization (NMF):

![Diagram](image.png)
Here we decompose the >3M DHS sites from 100 cell types using a combination of unknown k=16 terms (the number 16 was selected after several simulations; a 'k-term' was later referred to as 'regulatory term'), we assign colors to each k-term to aid interpretation and each DHS site is described by a mixture of colors (see right Figure). To interpret the colors, Meuleman looked on NMF factor loadings of the cell types (W in the NMF above) with respect to the tissues (see Figure below). It can be seen that cell types belonging to the same tissue are classified to the same k-term (or, regulatory term) having the largest NMF loading compared to other k-terms (by color).

Meuleman also showed that they can classify genes by the dominant regulatory term in the gene's locus. ~20% of genes get assigned to a regulatory term, e.g., GATA6 to cardiac; FOXG1 to neuronal; GATA1 to HSC/myeloid/erythroid with many “red” DHSs GATA1 locus. The same principle can be applied to other genes, TFBSs, pathways, and used to associate between GWAS variants and DHS region using a particular regulatory term. These examples highlight that regulatory term is more informative than its associated cell type.

Session: Variant Discovery and Genome Assembly

Encoding yeast genomic diversity using variation graphs

Prithika Sritharan (PS) – Joanna Dicks Group, Quadram Institute Bioscience, University of East Anglia, Norwich, UK

Yeast is one of the earliest species to be domesticated and are important in many biindustrial applications. The National Collection of Yeast Cultures (NCYC) has 4K strains from >530 species of yeast. PS mentioned that the regular variant discovery involves
alignment to reference genome resulting in a poor representation of the entire species. This approach is suboptimal for species with greater degree of sequence diversity, and introduces reference allele bias. All of these produce false variant calling. To accommodate common variants, Variation graphs (VG) can be used via VG toolkit ([Garrison et al., Nat Biot, 2018; github]).

VGs provide an enriched reference structure in which the genomes of many individuals within a species population can be incorporated as variants (including those with insertions, deletions, and duplications) forming embedded paths within a bi-directed sequence graph (see example below). 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.

PS applied a VG on the S288c reference genome + 19 S. cerevisiae strains ([Bergstrom et al., Molecular Biology and Evolution, 2014]) and compared read alignment between linear and graph-based references. The graph based reference alignments led to 8-41% of the reads with increased read alignment quality scores. While trimming reads improved read alignment scores when using the linear reference alignment, it is not necessarily good for variant calling. PS finished with saying that a good VG is not necessarily obtained by adding many genomes, but by selecting genomes that capture enough variance (and relevant to your sample).
ScaffHiC Genome Scaffolding by Modelling Distributions of Hi-C Paired-end Reads

Zemin Ning – Zemin Ning Group, Wellcome Sanger Institute, Hinxton, Cambridge, UK

Chromosome-scale scaffolding methods aim to de-novo assemble the chromosomes’ 3D structure using Hi-C data. 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. Hi-C mapping pairs have a wide distribution (right figure) – 86% in pairs, 14% on a different chromosomes, and <0.3% have no hits. 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. For example, in the figure below we can see different distributions of the left/right ends of a contig across the genome - while the left (red) is more condensed the right (blue) is more spread. Therefore, Ning suggested taking into account the distribution of left/right ends in two contigs \( i, j \) in order to decide whether to join them or not.

Ning presented 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. He defined a contig distance index (CDI) to quantify the likelihood for each pair of contigs with significant mapped ends (see formulation in the right Figure). The idea is the find two contigs to join with similar paired end reads distribution such that the distance index between them is maximal. The results show the new metrics gives better results than just the number of paired reads, and they were able to distinguish between noisy junctions and true junctions of contigs. The algorithm improves scaffold-N50 and max length when applied to scaffolding.
Pandora variation inference for pangenomes from Nanopore or Illumina data

Rachel Colquhoun (RC) – Zamin Iqbal Group, University of Oxford, Oxford, UK

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. E.coli pan-genome has ~3k core genes (normal E.coli has ~5k genes), and 90k genes (shared or unique) in the full pan-genome (i.e., a collection of all genes in the clade). Therefore, the ability to accurately detect genetic variation throughout the pan-genome and compare many genomes remains a difficult problem.

RC developed a novel reference graph structure tool that allows us to look for variation within genes across species/strains, and on addition of genes. The tool, named Pandora (github), allows approximation of a sequenced genome as a recombinant of genomes in the reference panel. RC uses population reference graph (PRG) of the pan-genome, and builds local graphs of regions with variation that make-up parts of the pan-genome PRG. RC first builds a dictionary order of local graphs by indexing the pan-genome PRG, maps the target genome to collection of local graphs, generates local graph consensus sequence, and forms the genotype. Indexing is done by using (w,k)-minimizers (w - window size;k – kmer size), which define the smallest k-mer of a substring of size w in a string, rather than using k-minimizer. Sketch reads (seed) are compared to the index to find representative k-mers in local graphs in order to pick paths between hits using maximum likelihood. Genotyping is done by using a Poisson model.

RC constructed a pan-genome reference graph for E. coli from 23,000 genes and 15,000 intergenic sequence clusters, and demonstrated high quality variant calls and sequence inference using Nanopore (on E. coli K12) or Illumina sequence data from E. coli. RC demonstrated that they 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. 960X Nanopore coverage can genotype 90% of the sites with 99.9% accuracy.