Workshop Report: Challenges and Synergies in the Analysis of Large-Scale Population-Based Biomedical Data

Overview

Increasingly, biomedical research depends on the analysis of population-based data from a large numbers of cells, individuals, organisms, or combinations thereof. A number of subfields are independently developing techniques to characterize, quantify and understand the effects of structure and evolution of populations represented in these data. While at first glance these subfields are disparate; a broader viewpoint suggests that their problems have a similar mathematical structure and shared computational challenges.

Our workshop provided a forum for researchers from several of these seemingly disparate subfields to identify commonalities so as to leverage discipline-specific insights across disciplines.

We organized the workshop around four theme days: functional genomics and evolution, cancer evolution, population and statistical genetics, and new statistical and machine learning methodology that was generally applicable in these areas. The biological sub-fields represented all made use of large-scale machine learning and probabilistic modelling, thereby providing a shared formalism to communicate, identify and solve shared, fundamental problems.

Important themes

Before reporting on some of the specific developments presented on in these areas during our workshop, we first here give a brief overview of the main theme areas:

Tumour evolution: Tumours are comprised of a heterogeneous populations of cells that can develop at different rates and can respond differently to treatment but that nonetheless evolved from a common ancestor. Currently, individual subpopulations are reconstructed based on allelic frequencies of tumour mutations without any reference to how these populations might have evolved under the selective environment of the tumour. Population genetics provides a series of methods for such reconstruction methods are emerging within the cancer genomics community. At the same time, tumor evolution represents an interesting bridge to population genetic analysis. Whereas one mainly examines somatic mutations, and the other mostly germline, many tasks and problems are shared.

Microbial genetics: As in tumour evolution, microbial genetics applied to metagenomic data seeks to characterize mixed populations of fragmented genomes and understand

their evolution and population dynamics in response to external stress. For example, one goal is to understand how populations of pathogenic microbes spread and evolve to evade treatment and immune response. Even non-pathogenic microbes, such as those of the gut, have been shown to be important to human health and response to drug treatment. Like tumour evolution, reconstructing individual microbial genomes from aggregate population samples remains an unsolved problem. Challenges faced in this field also bear resemblance to problems of sub-phenotyping in genetics for precision medicine, as well as dealing with heterogeneous populations of cell-types in epigenetics. In all these areas one needs to understand how to distinguish and characterize different entities so that each can be properly taken into account in the model and understood in the context of the problem at hand.

Precision medicine: Precision medicine based on, for example, genetic and epigenetic markers seeks to detect genetic traits that predispose individuals to disease. In particular, cross-sectional population analysis helps us to tease apart the genetic underpinnings of disease but important questions remain on how to detect and correct for population structure and latent subtypes within large-scale, heterogeneous genetic studies--problems also present in the other subfields in this workshop.

Population genetics: Human population genetics seeks to understand how our genome came to be, owing to evolutionary pressures and random processes. This field has developed methods for inferring selective pressures and population dynamics which can benefit other subfields mentioned previously. Furthermore, these subfields provide a fertile ground to test and extend the basic theoretical foundations of the population genetics.

Next we synthesize the presentation’s mathematical and statistical content into relevant summaries.

1. Scaling computational methods and data structures to extremely large data sets.

To fully leverage the power of modern genomic data sets, one must adapt the data structures and algorithms to handle millions of individuals each with measurements for millions of markers. Participants described how they used tree sequences to build linked ancestral recombination graphs genome-wide as motivation for fast new data structures, yielding massive (250x) compression, and quick access. Guiding principles for development of their approaches was that the approaches must scale to millions of genomes.

Scaling and efficiency were a focus in the analysis of RNA-seq data as well. Here we discussed using pseudo-alignment methods to map RNA reads to equivalence classes of transcripts. This equivalence class representation was deemed to be an information-rich but scalable representation of an RNA-seq assay.

2. Developing robust statistical inference methods for noisy and incomplete population genomic data with potentially non-standard noise models and incorrect modelling assumptions.

Single-cell sequence methods provide a variety of challenges. In addition to measurement noise that is so typical of genomic analyses, there is a high degree of dropout -- 90%-95% of measurements are missing. Strategies discussed to address these issues including Bayesian modelling of measurement covariances to fill in missing data. Also, the populations being measured are undergoing differentiation, and various talks discusses methods for modelling and recovering the underlying discrete structure. We discuss both tree-based methods and methods that learn intersecting manifold. Problems with established methods for measuring heritability were presented, as well as fixes to them. These problems centered on the priors used for SNP effect sizes and on how to properly handle the linkage disequilibrium (LD) among the SNPs in these estimates. In the first case, the prior on effect sizes often scales with the minor allele frequency, whereas empirical evidence was presented that this may not be advisable. Additionally, a method for how to account for the LD was presented and shown to perform well.

3. Detecting and correcting for interactions among subpopulations or other latent confounding factors including sample heterogeneity, and hidden functional sub-types in input or output data.

Several participants discussed generalizing standard (marginal and mean-based) genome-wide association studies to account for and leverage related traits, latent sub-phenotypes, and testing for differences in covariances rather than means. For example, a method for prediction of and clustering of genes by co-expression was developed, as well as a new Bayesian covariance test. Generalized linear mixed models were developed to find power in related traits and interaction with environmental exposures. Finally, methods for discovering more relevant, underlying traits from those recorded, were discussed.

4. Teasing apart sub-populations as confounders, or as interesting entities in and of themselves.

Whole genome sequencing of tumour DNA provides a summarized assessment of mutational diversity in an entire population of cancer cells. We discussed an international effort (Pan-Cancer Analysis of Whole Genomes) of 800 researchers to joint analyze WGS from 3,000 tumour samples. Work was presented on the overview of this project, showing that almost all tumour samples contained multiple cancer subpopulations. Critical to this project were new methods developed to combining clusterings together to identify consensus clusters -- three such consensus building methods were discussed. We also discussed various means by which tumours can change the chromosomal structure of human genomes and how to detect these catastrophic events, and methods to reconstruct the evolutionary history of an individual

tumour based on allele frequency data and a small number of unphased pairs of mutations linked by paired-end reads.

Methods for handling confounding factors in epigenetics studies have been emerging over the past few years. Some of these are based on linear mixed models, and others on PCA, or combinations of these. A new method for sparse PCA was presented and shown to have nice theoretical guarantees under some assumptions. The key idea is to compute a low rank approximation of the methylation data matrix by using only the most correlated methylations sites and discarding the rest.

Presentations also covered the problem of sub-phenotyping disease populations in order to identify homogeneous groups within a single, heterogeneous disease, such as type 1 diabetes.

5. Using evolutionary theory to help understand present-day populations

Several participants using evolution as a lens to understand present-day species. We had various presentation that discussed how to identify transcription factors with shared DNA sequence specificity. These and similar methods were used to reconstruct the ancestral binding specificities of these TF, and one participant described an effort to reconstruct transposable elements and model the arms race between TFs that evolved to silence these elements and the elements themselves. We discussed a modern-day arms race where by a parasitic worm silences gene expression in mouse cells through a RNAi mechanism previously thought to only exist in worms.

We also discussed the mechanisms of genome evolution and gene regulation based on an analysis of the sponge genome. Sponge genomes are one of the most compact, but intron-rich, genomes. We discussed mechanisms by which this could occur.

Evolution is surprising. One presentation discussed a microbial evolution experiment during which yeast populations reproducibly evolve very similar copy number changes in response to nutrient stress. Particularly surprising was that even within a single population, in a single evolution experiment, this copy number event appeared almost simultaneously across multiple cells.

6. Adapting deep learning and other machine learning methods for particular domains

As data sets increase in size in some biological application domains, it becomes useful to leverage the automatic feature-generation of neural networks, some of which are deep. But no one went off the deep end. A new method for using selective attention mechanisms for predicting whether a gene was “off” or “on” based on histone modification data. Leveraging a reinforcement-learning based CNN added additional power. Several presentations were centered on how to predict protein structure from sequence, using statistical physics and neural network approaches. One particularly unique idea that emerged was the idea of using reinforcement learning to fold a protein, even though on the face of it can be viewed as a straightforward prediction problem.

However, using intermediate information available by way of simulation and RL improved the results. Finally, generalizations of Naive Bayes to regression settings and with more complex feature spaces were presented for the purpose of CRISPR guide design for gene knockout. Guide design in this context has two main problems that machine learning can help with: the on-target problem of designing guides which will do their job well, and the off-target problem of designing guides which will not cause off-target activity. The latter is an especially difficult problem because the space of inputs is plagued by combinatorial explosion, while the amount of available training data are few. One workaround is to make assumptions of independence, to bootstrap up to a more general model.

Conclusion: Divisive Discussions

Two main discussion points emerged throughout the meeting where there was enough debate that resolution was not achieved, signaling important areas for further thought. The first had to do with the correct (or better) prior on SNP effects in linear models for predicting the trait. Several priors have been presented and argued for in the literature, but ultimately, evaluation comes down to synthetic data, and how one synthesizes the data, since answers on real data are unknown. Another decisive point was centered on how much of the raw data to keep for different tasks. For example, when developing data structure and algorithms to scale with massive genomic data sets, do we want to keep more than just variant calls, such as read counts, or quality scores? A consensus was not reached in either of these hotly contested areas, but the discussions were thought-provoking and pointing to areas in need of deeper thought/experimentation.

In statistical genetics, admixture populations present a particular challenge to correcting for genetic ancestry when mapping from genotype to phenotype. We briefly discussed a Mexican GWAS study: the Mexican population has a high degree of admixture of indigenous and European populations. Performing corrections on admixture data was determined to be a target for future methodology development.