Research Statement

Haynes Heaton

Research Vision

I seek to enable the infeasible, accelerate the intractable, and scale up the limited in life sciences research through algorithmic methods and professional grade software tools. In the past several years, single cell sequencing has allowed scientists to map the transcriptional landscape in a way previously unthinkable. But these studies are isolated from one another and must be brought into a common space from which comparative questions can be made. In the future, the same system will be used to understand the differential transcriptional and isoform landscape under different conditions such as disease state, pharmaceutical therapies, hormonal state, and underlying genetics. This is an effort to understand how both genetics and environment affect phenotype at the most basic level. Due to the high dimensionality of the condition space, this will be one of the most formidable biological efforts ever undertaken involving large scale data and requiring novel methods for data integration. Probabilistic machine learning models must be brought to bear in order to integrate data across different perturbation experiments into a holistic and interpretable understanding of how cells work. Also, different experiments incur technical differences which can outweigh and confound true biological differences. Software exists to correct such "batch effects," but each of these is flawed by the inherent ambiguity between technical variation and biological difference. I propose a combinatorial experimental design and deconvolution allowing for correction of these batch effects without diluting true biological differences as well as other artifacts.

Past and current research

Single Cell

Droplet based single cell sequencing (scRNAseq) uses reverse-emulsion droplet technology to isolate cells and attach cellular barcodes to the reads generated from that cell's RNA. This is crucial to understanding the biology of different cell types, especially minority cell types. In some circumstances, a sample may naturally have a mixture of two different individuals, such as a transplant patient or maternal/fetal sample. Additionally, there are many advantages to an experimental design in which multiple individuals are artificially mixed into the same experiment. When comparing different experiments, technical variation can overshadow and confound true biological differences. By mixing multiple individuals' cells into the same experiment, these "batch" effects are overcome. But this mixture of individuals must be deconvolveed to know which cells came from which individual. By using genetic variants detected in the scRNAseq reads, it is possible to assign cells to their donor of origin. To do this, I developed **souporcell, a sparse mixture-model clustering algorithm to cluster cells by genotype**. Additionally, as the number of individuals grows, clustering becomes prone to local minima and erroneous clusters. I employ a deterministic annealing strategy to overcome these local minima[1]. This tool is being used widely in the field including multiple million-plus cell studies and by the single cell expression quantitative loci (sc-eQTL) consortium.

Phased Assembly

Efforts have begun on the Earth Biogenome Project (EBP), a global project to sequence the entire diversity of multicellular eukaryotic life. In the UK, the Sanger Institute at the University of Cambridge and partners have started to sequence 60,000 species from the British Isles in the Darwin Tree of Life (DToL) project. These projects aim to provide a scientific resource for the next generation of biological science, for environmental conservation, and to study evolution at a much broader and deeper scale than ever before. Long read assembly has progressed well toward producing highly contiguous and accurate genome assemblies. One of the primary remaining difficulties is high levels of heterozygosity such as found in many of the non-model organisms included in the EBG and DToL projects. When assembling a diploid or polyploid genome, inexact read matches must be determined to have arisen from sequencing errors, alternate haplotypes, or paralogous sequences. Current methods focus on assembling a primary haplotype and separating out as much as possible of the second haplotype after the fact. This can lead to errors when the incorrect inference is made about homology, or fragmented assemblies when the ambiguity is properly detected. Newer methods have used trios to haplotype bin reads prior to assembly, but often

trios are not available. To address this problem, I have developed **phasst (phased assembly tool), a method to phase heterozygous kmers, and thus the reads they occur on, from a single individual into haplotype specific bins prior to assembly.** Our method is applicable to PacBio HiFi data and can use the additional information from linked reads and Hi-C if available.

Industry R&D

Distance maps

I began my research career in R&D at sequencing start-up companies. I believed that to understand the full range of genetic variation, long range genetic information is needed in order to analyze the often overlooked structural variation, copy number variation, and to access the repeat regions of the genome. To this end, I joined Nabsys, a company creating nanopore distance maps. A long DNA molecule would be tagged at certain sequence motifs and translocated through a nanopore. By measuring the current across the pore, the time--and thus distances--could be measured between these sequence motifs on the molecule. At Nabsys I worked on distance map assembly algorithms and **created a novel graph theoretic distance map multiple alignment algorithm**[2]. In this graph, nodes represent instances of the sequence motif and edges between nodes represented matches in pairwise alignments between two molecules. Through the knowledge that two instances of the motif on a single molecule were distinct events, the graph could transitively imply that these distinct events are the same. These contradictions can be resolved through the minimum cut on the graph which would resolve all such instances. This is achieved in a single pass with a modified version of Karger's algorithm. It created far more accurate distance map assemblies than we had previously achieved.

Linked reads

Continuing with my interest in long range genetic information technology, I joined **10x Genomics**. There I worked on **unlocking the repeat regions of the genome**. In genomics, short read sequencing involves breaking DNA into pieces of roughly 100 bases in length and mapping those onto a reference genome. Then the differences between the reads and the reference genome are said to be the genetic variants. But if there are two or more regions in the genome with very similar or identical sequences longer than the length of the read, that read cannot be mapped accurately and thus the genetic variants in that region of the genome cannot be analyzed. Linked read technology uses a reverse-emulsion droplet system to attach the same barcode to every read originating from a long DNA molecule while different molecules get different barcodes with high probability. This information can then be used to differentiate which location a repeat sequence should map. I **developed Lariat, which uses a metropolis-hastings search to find the optimal read mappings** under the statistical model[3]. This method is able to **uncover more than half of the previously "dark" regions of the human genome** and contributed to the NIST genome in a bottle ground truth resource[4].

I also worked on haplotype phasing--the process of determining which sequences originated from the maternal vs paternal chromosomes. The molecule information, along with a statistical model and a particle filter search, was used to determine the optimal phasing[4]. One additional difference from other phasing algorithms is that phasing inconsistent variants are likely false positive variants. This algorithm allowed for an error state in the search which **allowed our phasing algorithm to correct the false genotype information**[5].

Future Goals

Single Cell

I am now continuing to shift my research toward single cell sequencing. During the decade I spent working on genomics, I believed it was the most important line of research. But in that decade, researchers and technologists fixed many of the problems genomics was facing. Going forward I believe we will learn more about biology by inspecting the effect genotype and environment have on phenotype at the cellular level.

In the past several years, single cell sequencing has allowed the transcriptional landscape to be mapped out in a way previously unthinkable. In the future, the same system will be used to understand the differential transcriptional and isoform landscape under different conditions such as disease state, pharmaceutical therapies, hormonal state, and underlying genetics. Currently the sc-eQTL consortium is in the process of mapping eQTLs across different cell types and genotypes in the population. Perturb-Seq, the combination of scRNAseq and

clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 allows for the simultaneous testing of multi-locus gene perturbation at the single cell level. And in the future we expect to have a growing set of condition perturbations across drug therapies, disease states, and hormonal conditions. To enable this research, problems of data integration, batch correction, and data interpretation must be solved.

I aim to solve the batch effect problem through a combinatorial experimental design in which multiple individuals' cells or perturbation conditions' cells are mixed together across multiple experiments in a pattern that will allow for the deconvolution of which cells belong to which individual or condition, which cluster corresponds to which individual or condition and batch correction across experiments. This will require interactive software for both experimental design and downstream deconvolution and batch correction. A simpler example of this is shown in table 1 of the souporcell paper[1].

For data integration and interpretation there are two major problems I would like to tackle. The first is integrating and analyzing the many different tissue specific cell atlas projects into a coherent full body cell atlas and enabling questions such as what is different between resident macrophages in the skin and mucosa. For this I aim to collaborate with wet lab biologists to create a minimal set of datasets which will develop a minimal set of tagged tissue mixture experiments which will allow for appropriate batch corrections of all of the previous tissue cell atlas projects into a comparable space. The next problem of integration and interpretation is with multi-individual and multi-condition datasets. For this, probabilistic machine learning will be used. This system would be designed to be used by researchers wanting to analyze multi-individual and/or multi-condition large single cell experiments with a model that takes into account the total covariation of variables instead of a single-input single-output model. The inputs to this system should be as close to the raw data as is feasible. Raw or normalized transcription count matrices should be used along with some annotations of the individual, their genotype, and the perturbation conditions. The model should be self-regularizing to avoid overfitting. Two possibilities of models would be probabilistic graphical models (PGM) or approximate gaussian processes (GP). Then for interpretability, an interactive graphical user interface needs to be made to inspect the model under different conditions. Some variables may be fixed or perturbed and other free variables will react. The largest changing (delta/standard deviation) free variables will be displayed to the user for further analysis.

Genomics

In addition to my main goals in single cell sequencing, I remain interested in several projects in genomics. Polyploid phasing continues to be a difficult problem with only limited solutions. Using a similar sparse mixture model clustering to the one used for my souporcell project, it is possible to phase diploid or polyploid genomes while identifying and being robust to phasing inconsistent variants. This algorithm (phuzzy phaser) will also seamlessly integrate long read, linked read, and HI-C data or any subset of those. Phuzzy phaser will scale linearly with ploidy whereas other polyploid phasing algorithms scale super linear as ploidy increases. I also believe that long read sequencing will continue to decrease in cost over time and see more use in germline and somatic mutations at all length scales and will eventually see uptake in clinical use to increase the diagnostic yield. I hope to build software to support this transition and work with fellow medical doctors to bring this into translational research and eventual clinical use.

Closing remarks

Our understanding of how both environment and genetics impact phenotype is at the very core of biological understanding, and answering these questions at the most basic level--that of the cell, transcript, and isoform--will build a foundation upon which the next generation of biological research will stand.

1. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
2. Goldstein, P., Heaton, W. & Preparata, F. Distance maps using multiple alignment consensus construction. *US Patent App. 14* (2014).
3. Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.*

29, 635–645 (2019).

4. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
5. Kyriazopoulou-Panagiotopoulou, S. & Marks, P. Systems and methods for determining structural variation and phasing using variant call data. *US Patent App. 15* (2016).