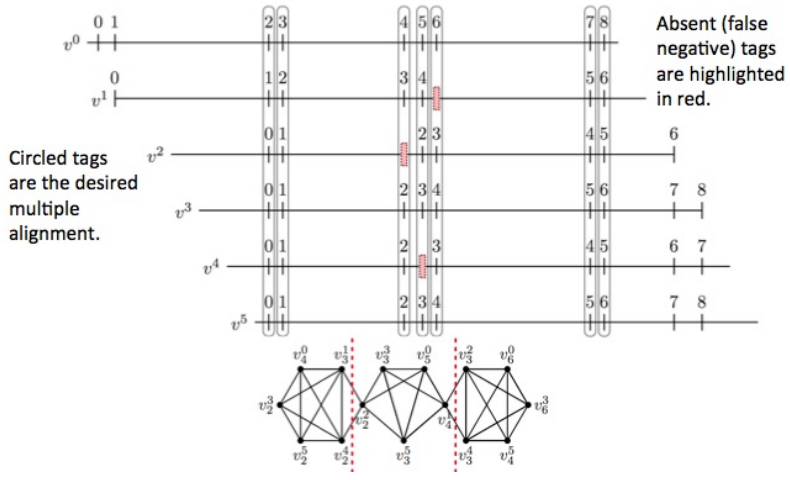
Algorithm Development for Single Cell Data and Genomics

Haynes Heaton

I aim to develop algorithms and professional grade software to enable life science researchers to answer previously infeasible questions in transcriptomics and genomics. In collaboration with wet lab scientists and fellow MDs, I hope to address these questions and take that knowledge from basic research into translational research and clinical use. In the past several years, single cell sequencing has allowed scientists to map the transcriptional landscape of developmental and disease systems at a resolution that was previously unthinkable. In the future, this approach will only be expanded to understand the cell state under varying 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 of individual cells. Due to the high dimensionality of the condition space, this will be one of the most formidable biological efforts ever undertaken. But these studies are isolated from one another and must be brought into a common alignment from which comparative questions can be made. Distinct experiments have technical artifacts that can outweigh and confound true biological differences. I propose a combinatorial experimental design and deconvolution system allowing for correction of these batch effects without diluting true biological differences. Also, 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. In addition to single cell research, I will continue to work on problems in genomics such as polyploid phasing methods and structural and copy number variant detection using cutting edge technologies and algorithms building on work I have done over the past decade in this field. My experience and expertise in computer science, working at 10x Genomics, and my MD give me unique insights and advantages in pursuing these research goals.

PAST AND CURRENT RESEARCH

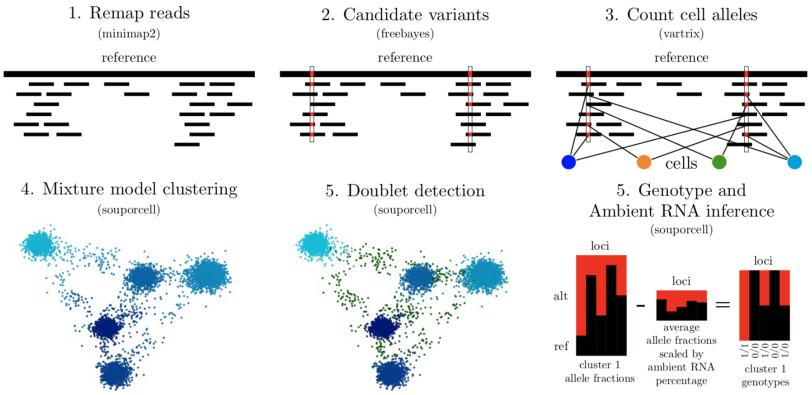
Nanopore Distance Maps - To understand the full range of genetic variation, long range genetic information is needed in order to analyze the often overlooked structural and copy number variations and to access the repeat regions of the genome. To this end, early in my career I joined Nabsys, a company creating nanopore distance maps. A long DNA molecule would be tagged at certain sequence motifs and translocated through a nanopore, measuring the distance between these motifs. **I developed distance map assembly algorithms and created a novel graph theoretic distance map multiple alignment algorithm**[1]. In this graph, nodes represent instances of the motif, and edges represent matches in pairwise alignments between two molecules. This graph can transitively imply that two distinct events on the same molecule are the same. These contradictions can be resolved through the minimum cut set on the graph which resolves all such contradictions. I achieved this with a modified version of Karger's algorithm. It created far more accurate distance map assemblies than we had previously achieved. These methods are now **patented**[1].



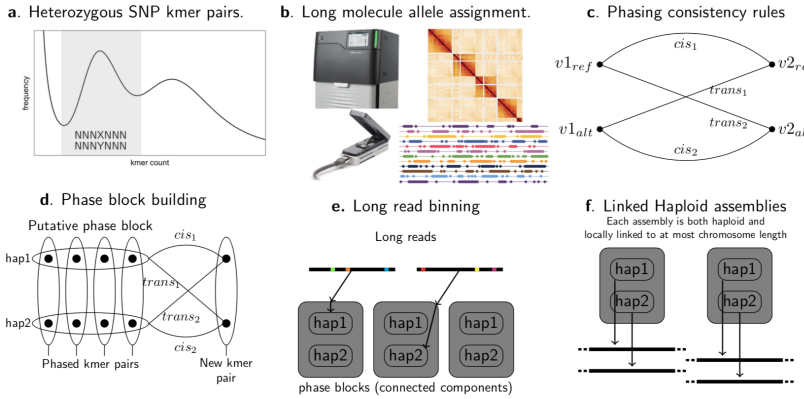
Linked reads - Continuing with my interest in long range genetic information technology, I joined 10x Genomics. There I worked on unlocking the repeat regions of the genome. Short read sequencing involves breaking DNA into pieces of 100 bases in length and mapping those onto a reference genome. Then the high quality and well supported differences between the reads and the reference genome are considered to be genetic variants. But if there are two or more regions in the genome with very similar 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 to which repeat locus a read should map. **I developed Lariat, which uses a Metropolis-Hastings search to find the optimal read mappings under a statistical model[2]. This method is able to uncover more than half of the previously "dark" regions of the human genome.** This work was published in **Genome Research** and **contributed repeat region variants and validation to the NIST (National Institute for Standards and Technology) Genome in a Bottle (GIAB)** ground truth resource which was published in **Nature Biotechnology**[3]. I also worked on haplotype phasing–the process of determining which sequences originated from the maternal vs paternal chromosomes. A statistical model of the data along with a particle filter algorithm was used to determine the best phasing. Phasing inconsistent variants are likely incorrectly genotyped and using this **signal allows the phasing algorithm to correct false genotype information**. This work was published in **Nature Biotechnology** and some aspects were **patented** [4][5].

Single Cell - Droplet based single cell sequencing (scRNAseq) uses reverse-emulsion microfluidic technology to isolate cells and attach cellular barcodes to the reads generated from that cell?s RNA. There are many advantages to an experimental design in which multiple individuals' cells 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 don't exist. But this mixture of individuals' cells must be demultiplexed 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. **I developed souporcell, a sparse mixture-model clustering algorithm** with deterministic annealing to assign cells to their donor of origin. I am the first author on this paper published in **Nature Methods**[6]. This tool is being used widely in the field including in multiple million-plus cell studies, by the **sc-eQTL (single cell expression Quantitative Trait Loci consortium)**, and it is being **incorporated into the 10x Genomics cellranger platform**.



Phased Assembly - Efforts have begun on the Earth Biogenome and Darwin Tree of Life projects, a global undertaking to sequence the entire diversity of multicellular eukaryotic life. These endeavors aim to provide a scientific resource for the next generation of biological research, for environmental conservation, and for the more extensive study of evolution than previously possible. To achieve these objectives, high quality genome assemblies must be created. Long read assembly has progressed well toward producing highly contiguous and accurate genomes. One of the primary remaining difficulties is high levels of heterozygosity found in many organisms. When assembling a diploid genome, inexact read matches must be determined to have arisen from sequencing errors, alternate haplotypes, or paralogous sequences. To address this problem, **I have developed phasst (phased assembly tool), a method to phase heterozygous kmers–sequences of length k, and thus the reads they occur on, from a**



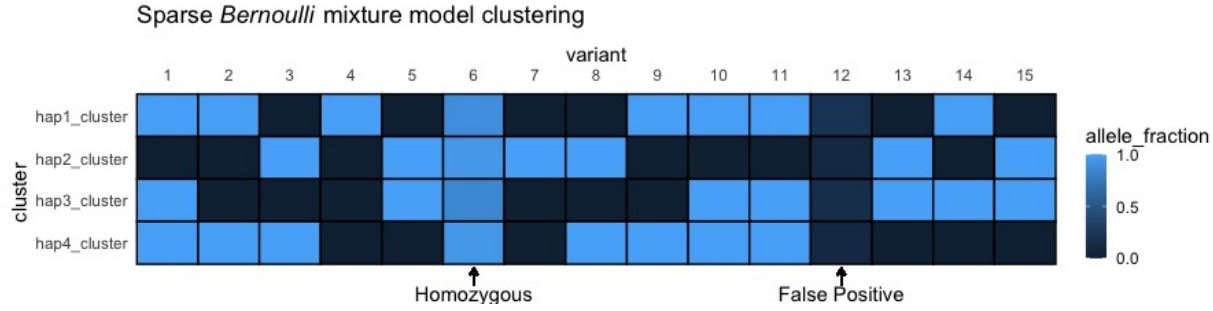
single individual into haplotype-specific bins prior to assembly (in prep). This method is applicable to PacBio HiFi data and can use the additional information from linked reads and Hi-C if available.

FUTURE RESEARCH

As an algorithm development lab, my lab will not be generating our own data. I will have a mixture of projects, some of which will be purely method development and rely on public data, and others will be part of collaborations with wet lab scientists. I will join the sc-eQTL consortium and am in the early phases of establishing collaborations in several different projects. For example, I will be working with Nicole Soranzo, faculty at the Sanger Institute, on massive single cell experiments using the UK biobank project, which offers data rich resources including genetics, disease state, and other phenotypic information to understand how these affect the transcriptional state. While I remain interested in genomics, I would like to shift my work to be 70% single cell and 30% genomics. Going forward I believe more will be learned about biology by inspecting the effect genotype and environment have on phenotype at the cellular level. I also believe that there are many opportunities to make major improvements in how we analyze single cell data that will keep pace with the technological improvements that are likely to occur (e.g. single cell multi-omics becoming more standard).

Single Cell - In the past several years, single cell sequencing has uncovered the diversity of cell transcriptional states. In the future, this system will be used to analyze cell activity under different conditions such as disease state, pharmaceutical therapies, hormonal state, and underlying genetics. **In my first grant proposal, I would seek to (i) develop a probabilistic machine learning framework for analyzing complex multi-genotype and perturbation experiments, (ii) create a system for combinatorial experimental design and deconvolution solving the batch effect problem without diluting true biological differences, and (iii) build methods for comparing cell types across different tissues and disparate cell atlases using these combinatorially multiplexed experiments to sync these previous datasets.**

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. I propose using a **sparse *Bernoulli* mixture model clustering** which will be able to phase diploid or polyploid genomes while identifying and correcting false genotypes. My 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 linearly as ploidy increases**. Other projects in genomics include **kmer and phasing based structural variant and copy number variant analysis building on my phased assembly methods.** I also believe that long read sequencing will continue to decrease in cost over time, see more use in germline and somatic mutations at all length scales, and will eventually see uptake in clinical use to improve diagnostics. I hope to develop methods and software to support this transition and work with fellow MDs to bring this into translational research and clinical use.



Closing remarks - Our understanding of how both environment and genetics impact phenotype is at the core of biological understanding, and answering questions at the most basic level–that of the cell, transcript, and isoform–will build a foundation upon which the next generation of life sciences research will stand. Computational methods to enable robust inference on these large, sparse, and difficult-to-integrate datasets will be vital to the success of this coming revolution.

REFERENCES

[1] Goldstein, P., Heaton, W., Preparata, F. & Upfal, E. Distance maps using multiple alignment consensus construction (2014). US Patent App. 14/212,458.

[2] Marks, P. *et al.* Resolving the full spectrum of human genome variation using linked-reads. *Genome research* **29**, 635–645 (2019).

[3] Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nature biotechnology* **37**, 561 (2019).

[4] Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology* **34**, 303–311 (2016).

[5] Kyriazopoulou-Panagiotopoulou, S. *et al.* Systems and methods for determining structural variation and phasing using variant call data (2016). US Patent App. 15/019,928.

[6] Heaton, H. *et al.* Souporcell: robust clustering of single-cell rna-seq data by genotype without reference genotypes. *Nature Methods* **17**, 615–620 (2020).