As part of the Simons Genomics semester that I am chairing in Berkeley, I organized an industry day where scientists from genomics companies and start-ups in the Bay Area spoke. Here is a brief summary about the companies and what they showed.

### Unzipping Diploid Genomes: From Contigs to “Haplotigs” - Jason Chin, Pacific Biosciences (PacBio)

**Company:** PacBio has the long-awaited technology of very long reads. Chin described their technology in brief, and reported that more than half their reads are now of length 12kb or longer. On the downside error rate in base reading is still very high (>10%).

**Problem and method:** The length of the reads allows them to separate haplotypes in a diploid genome, as every read covers some 10+ SNPs that all belong to the same haplotype. So even if few are lost due to errors, the ability to stitch together long haplotypes is good. In fact, there are many more structural variations (SVs) than what was believed before, which helps them in the phasing. They call the resulting contigs of haplotypes "haplotigs". They use the string graph idea of Gene Myers and a heuristic optimization approach from physics (Ising spin) due to the background of Chin.

**Results:** In an experiment that they did in mating two clean arabidopsis strains, the method gave N50 of 8M contigs for the second generation (F2). In human, they form haplotigs that cover 2Gb, with N50 of 330kb.

**Open problems:** many! rigorous solution to the phasing, using other long range info, overcoming mosaic-like repeats, and handling multi-ploid genomes (e.g. plants). Looks like Chin is eager for cooperation.

### De Novo Genome Assembly with “In Vitro” Proximity Ligation Libraries - Nicholas Putnam, Dovetail Genomics

**Company:** A 2-year old start-up from Santa Cruz.

**Problem and method:** They want to generate very long (ideally chromosome size) genome assemblies, by using a variation of Hi-C and illumina reads. The Hi-C variant (IVCA, due to Ed Green) creates DNA chromosomal segments of length ~100kb, incubates them in vitro with histones, and then continues like in HiC: cross-linking, restriction cutting, ligation and sequencing. The technique does not show TADs but only a clean signal near the diagonal.

**Results:** describes their pipeline to construct genomic scaffolds. They can create scaffolds of tens of Mbs, and also use the method to identify structural variations.

### Long Reads Sequencing Technology and Its Applications - Hayan Lee, Lawrence Berkeley National Laboratory

Hayan (a postdoc who takes part in the Simons semester, co-organized the day) described part of her phd work done at CSHL with Michael Shatz. Their work (published in Bioinformatics 2012) analyzed some 35 assembled genome to assess the effect of read length on various quality parameters of the assembly. They concluded that longer reads are more important for contiguity in the assembly than higher coverage (beyond 10-15x), and suggested what they call the Lee-Shatz model (!) as an
alternative for the Lander-Waterman model, which has been used to guide decisions on assembly process in genome projects. Most of the results are quite simple and expected, but there are a few interesting graphs and a lot of data is summarized.

**Novel methods for Long Range Information from Short-Read Sequencers - Chris Streck, 10X Genomics**

**Company:** A 1.5-year old Bay Area start-up for creating long assemblies from short reads. Their technology starts from a tiny amount of DNA (1ng), breaks the chromosomes into a million DNA fragments of 100kb each on average, inserts each fragment into a droplet (gel bead) that contains many copies of a unique barcode, and then does a duplication and fragmentation of the molecule, with the barcodes inserted in the ends. The droplets are then resolved and the mixture is Illumina sequenced (typically 30-40x). All the short reads that originate from the same bead have the same barcode and so the assembly problem is divided into many independent much smaller subproblems. (This simple divide and conquer idea was also the basis to the Moleculo company, but the scale here is bigger. Consider this as "Moleculo on steroids").

**Problem and method:** They address re-sequencing using a BWA aligner that they developed called **Lariat aligner**. They claim that the barcodes help overcome low complexity regions.

**Results:** In de novo assembly, they achieve N50 of 5Mb phased blocks, and 13Mb for unphased blocks. They are also working on single cell RNA-seq. No results were reported.

**A Few Outstanding Scientific and Computational Challenges and Opportunities in Genomics**

*Mark DePristo, Verily*

**Company:** Verily was previously called Google Life Sciences. They sit in the Google campus in the Silicon Valley. The company developed several gadgets in the medical field (e.g. a contact lens that measures glucose level for diabetic patients, a spoon that cancels tremors of Parkinson's disease patients and monitors them, and several medical monitoring devices that "talk" with the i-phone and the cloud).

DePristo gave a very high level talk. Their goal is medical treatment that is proactive and not reactive. They are building a study called **Baseline**, which constructs a cohort of 10k+ healthy people with clinical, omics, imaging and gadget data and will follow it over a long time.

One key message was that deep learning is the panacea to all problems in machine learning. ML algorithms are good, but the main problem in ML has been to design the features for the methods ("feature engineering"). Deep learning overcomes the problem by automatically selecting the features given sufficiently large data sets. He said that the **Google Brain** project (a deep learning project at Google) has reached the situation that deep learning software replaces many of the dedicated software previously developed for particular problems in Google given sufficient data. His message: "AI is eating software".

Another message (threat?) that he passed was that the big players in the industry are aiming to take control of the personalized medicine field. In his words, nobody in academia studies search any more, since Google "owns" it. Similarly, it is not inconceivable that in five years some of the personalized medicine problems will be "owned" by a big company, who has no problem spending a billion dollars on creating their cohort and data ("this is what Facebook spends in a couple of months"....)
Large Scale Genomics in the Cloud - Andrew Carroll, DNAnexus

**Company:** This Bay Area start-up enables analysis of genomic data on the cloud. It expects that within five years, the exomes of 2M persons will be available and we will face great challenges. A key difficulty is networking of information across different sites around the globe. For example, their platform was used after collecting samples of ebola in Africa, uploading them in the field to the DNAnexus cloud and analyzing them quickly at the Broad.

**Open problems:** Very large cohorts raise new problems of scalability. When the data is incremented over time, reanalysis from scratch after each new added sample ("the n+1 problem") is very inefficient. Storage and compression are hard problems.

**Results:** They analyzed 10K exomes to date. They observe new mutations all the time. In fact, the fraction of interesting mutations rises with increased sample size. They have a collaboration with Tina Graves (U Washington) that combines PacBio and BioNano (optical maps). Using this combination, they report human assemblies with N50 of 13Mb.

They mentioned a project of Regeneron to sequence 250K individuals and analyze their de-identified medical records. This seems the project with Geisinger of Pensylvania mentioned [here](#).

Liquid Biopsy: The New Gold Standard for Diagnostics - Raheleh Salari, Natera

**Company:** Natera was founded in 2009, and is a leader in non-invasive prenatal testing. The company has 700 employees of which more than 40 are PhDs. Their main product is a test for trisomy of Chromosomes 21,18 abd 13 by amplifying regions around SNPs from DNA found in the blood of the mother, and searching for differences between allele frequencies in the mother and the baby. They claim that this method is better than the competing methods, which are based on counting. They now plan to get into liquid biopsies too.

**Focus:** A liquid biopsy is a way to test for tumor presence from blood samples. The idea is to identify in the blood DNA of the somatic genome by sequencing DNA fragments from circulating tumor cells and from circulating tumor DNA (DNA pieces of 150bp or shorter from dying tumor cells). The difficulty is that the tumor DNA is only 0.5% of the circulating DNA in the blood.

**Results:** They use massively multiplexed PCR to extract select DNA segments, and look for copy number variations in cancer. They reported a preliminary study where they identified most of the CN changes in a small study (but they did not identify all, and the changes were in the largest and easiest to detect chromosomes). They also identified some but not all of 16 SNVs that they tested.

They have a collaboration with Cancer UK (Charlie Swanson) within a project called [Tracer-x](#), that will do a prospective lung cancer study of 850 patients over 10 years, following the cancer heterogeneity over time and using repeated liquid biopsies.

Engineering and Scientific Advancement in Genomics with Bina Technologies - Hugo Lam and Greg Gibeling, Bina, Part of Roche Sequencing Solutions

**Company:** Bina is a bioinformatics software start-up bought by Roche a year ago. They develop tools for the sequence analysis pipeline: mapping, mutation detection and structural variation detection. Their software did not impress me. What they described was integration of published algorithms from the literature, via [ensemble methods](#), with no new insights.
Company: Genia was formed in 2009 and was bought by Roche in 2014. It now has over 150 employees. Its goal is clinical sequencing.

Technology: This was by far the most exciting talk of the day. If focused on Genia's technology, which they (rightfully) called "Fourth generation sequencing". It combines integrated circuits (ICs) and nanopore technology. The current version has 128K wells on a single IC. In each 4.5 micron well a separate polymerase reaction takes place and is monitored. A nanopore is attached to a polymerase, and as each base is added by the polymerase, a chemically-tailored tail of the base is detached and passes through the pore. Each base has a different tail that induces a different electrical conductance readout. The rate is 1-3 bases per well per second, which translates to almost half Mb per second for a chip. This means 1Tb per hour per chip(!!). Some info on the technology and images can be found here. A cool video movie was shown in the talk, but I could not find it online.

Results: None presented. Time will tell if this will be the next killer technology or that the noise will overwhelm them.