**Long** genes: Can capture multiexon transcripts with PacBio RS sequencing platform – captures full-length RNA in hESCs.

Also identified ncRNA – from polyA tail based RNA-seq?? wtf?

Key aspects of de novo assembly: repeat resolution, read length, overlap mapping quality, assembly algorithm.

prior to 2007, clone-by-clone based sequencing, with 200 kb long overlapping frags cloned into BACs and assembled individually. BAC clones that were repetitive within the context of the entire genome were locally unique == gap free assembly.

Few modern genome assemblies exceed an N50 contig of 100 kb (avg = 41 kb). This translates into tens of hundreds of thousands of sequence gaps – corresponding to repeat and MPS biases in GC representation. (reduced coverage in AT- and GC- rich regions of the genome associated with the MPS platforms)

**When the result of a de novo assembly is a sequence per chromosome without gaps and with 99.99% base-pair accuracy, the assembly is considered complete; otherwise it is considered a draft.**

**There are a total of 261 alternative references in the current human reference genome (GRCh38), corresponding to regions of extreme genetic diversity often associated with segmental duplication.**

Retrotransposons that lack ORFs?

Optical mapping?

PacBio CCS reads?

**Class notes**

1953 – double helix structure of DNA.

Central dogma of molecular biology

Sanger sequencing – first milestone with DNA sequencing

HGP started in 1990, completed in 2003. 2001 was the draft genome. Budget was 3 billion dollar. A dollar a base assumption. Lot of money spent in the early phase, in technology development. Automated sanger sequencing was the crucial innovation (ABI technologies – company). **2nd phase of DNA sequencing technology.** Still sanger sequencing, but now automated.

After we got the human genome in 2003,

* SNP consortium (discovery of all SNPs), HapMap project
* In order to understand the genome sequence
* Compare sequence and function. GWAS studies. Relate polymorphisms with phenotypes.

Sequencing technologies that came around

* 454, solexa, ABI SoiD, iotorrent. Solexa got bought by illumine.

Genome assembly still relatively easier with a clean reference gneome, but de novo assembly of genomes is much more complicated. Enter long read sequencing – 3rd generation sequencing technologies. PacBio (2011) and Oxford Nanopore (2014).

Longer reads can assist assembly of highly complicated genomes. These emerging technologies can also read ‘additional marks’ on the DNA ***(ex. Methylation information).***

Find examples of why some genomes have not been fully assembled, and how PacBio can be used to assemble them.

**Small genomes != simple genomes**

Wheat genomes – 18 Gb, hexaploid genome! Lots of transposons, paralogs of genes.

A simple reference genome can be used to help detect genomic variation.

**Genomic variations**: SNVs, indels, tandem repeats

**Structural variations**: Translocations, CNVs

Different tools do well to reconstruct different structural variants. Ex. Tandem copies of genes (transcription factors, for ex.) are difficult to pick up on with short reads. CNV resolution of these tandem gene clusters may still sometimes have variations (genetic, or technical). Technical bias in sequencing prep can create bias and over-representation (higher coverage, just because those regions are amplified better in the protocol – gc bias, mappability bias, see <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0016685>).

Tandem duplications resolution – famous example is ribosomal DNA.

Longest Pacbio reads up to 80,000! However, most reads are much shorter. The bottleneck is the ability to perfect the library prep.

**PacBio and genome assembly (+ tractably representing genomic variations):**

(3rd point). Original purpose of PacBio is for genome assembly and De novo assembly. But there are other things you can apply it towards.

One of these things was to use the long reads to reconstruct full length transcripts. From pre-mRNA to mRNA, there is an intermediate process of splicing. RNA-seq can identify splice junctions quite well, but the problem comes when a gene has a large number of introns and can result in different isoforms (as intermediates or as final products). For ex., 38000 isoforms detected in the DISC-R? (dsx?) genes in Drosophila.

Through the direct sequencing of DNA, we can also read the methylation state of genomic regions.

So far efforts have been focused on small genomes – bacteria, archaea, and c. elegans (100 Mbp).

Metagenomics is another area – 16S rRNA (bacteria) and 18S rRNA (archaea) can be amplified with same primer, but represent different taxa due to Hyper Variable Regions. *Are there similar issues with inferring metagenomics studies as there are with isoform identification in RNA-Seq?* High error rate can also make it look like there are a lot more species than there really are – one way is CCS based correction. The number of cycles in CCS is influenced by the polymerase, the length of the DNA fragment you are sequencing (2k amplicons are sufficient for metagenomics). Oxford Nanopore has a similar adaptation too – can look into if interested, not discussed in class due to lack of time.

**Topics to explore**

1. How does PacBio work? (reference: Science paper)
   1. Advantages (CCS, read length) and disadvantages (high error rate etc.) of PacBio
   2. Relative drawbacks of the 1st, 2nd, and 3rd generation sequencing technologies (read length, throughput, cost perspectives)
   3. Strengths (field applications) of the respective technologies – what are the specific scenarios for which specific generation techs are best suited (ex. Things short read sequencing does better than long reads)
2. Genome assembly using PacBio, variant detection
   1. Small genome assembly (has been done quite extensively in bacteria, archaea, fungi)
   2. Large genome assembly
   3. Human genome specific projects with PacBio
   4. Structural variation detection in complex genomic regions

**welcome to compare why PacBio is better as compared to Illumina**

1. PacBio applications
   1. Full length transcripts construction (focus on PacBio, why previous methods were not able to do this effectively)
   2. Detection of DNA methylation
   3. Metagenomics

Group 1: Golnaz, Prasath

Group 2: Ben, Rachelle, Shams, Phil

Group 3: Sam, Matt, Jasleen, Tony