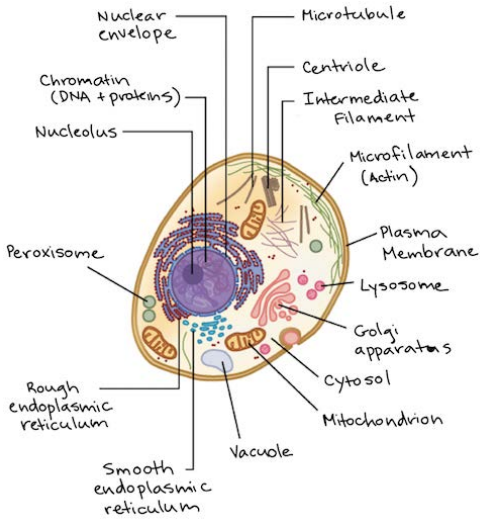
Biomedical Informatics
Week Of April 2ND: Part 1 Biology Primer

USC Institute Of
Translational Genomics
Keck Medicine of USC

Measuring Through Sequencing
Quick Primer



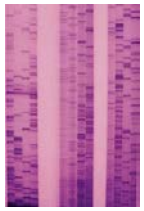
https://www.khanacademy.org/science/biology/structure-of-a-cell/prokaryotic-and-eukaryotic-cells/a/intro-to-eukaryotic-cells

With important exceptions…

- … you are diploid with a maternal and paternal copy
- … you have two copies of 22 chromosomes plus X and sometimes Y
- … there are four nucleotides (A, T, C, G), about 3 billion bases long (ATTATA..)
- … a copy of your genome is every cell.
- … there are 4 millions genetic variants between two people in their germline genome
  - Single nucleotide substitutions (SNVs), Insertions/Deletions (indels), Structural Variants (inversions, duplications, translocations)
- … changes occurring in a specific tissue or cell during our life are called somatic events
- … most genetic variants are not functional.
- … 1% of your genome is coded in genes, sometimes this is called your exome
- … in genes, DNA is transcribed to RNA, RNA is translated to proteins
- … genes are frequently transcribed as exons broken by introns, where the introns on spliced out of mRNA
- … a considerable number of modifications can occur to proteins (e.g. phosphorylation)
- … 99% of your genome we don't understand, but we all recognize its important.

The exceptions are often the most important aspects of understand and treating diseases.

Molecular Sequencing AKA Next-Generation Sequencing






Traditional Sanger Sequencing (1979 ->):

- Major improvements include capillaries, use of dyes, automated calling
- Consensus of billions of molecules
- P&E, AB, etc

Array-based sequencing (2002 ->)

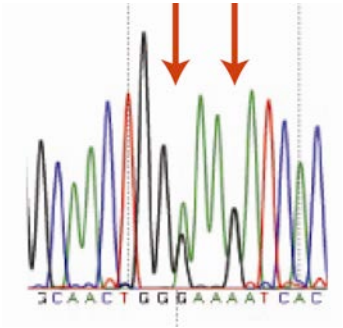
- Sequencing millions of pre-defined SNPs via hybridization or allele extension
- Affymetrix, Illumina

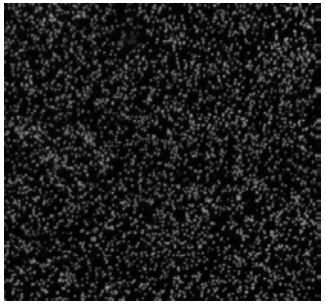
Pseudo single molecule sequencing (2006->)

- Each read derived from a single molecule, clonally amplified
- Millions of sequences sequenced base*base (lawn-sequencing)
- 454, Solexa, Agencourt, Life

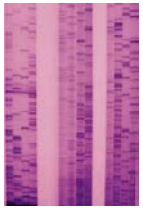
Real-time single molecule sequencing (2010->)

- Single molecule, fewer reads in realtime
- PacBio, Oxford, etc

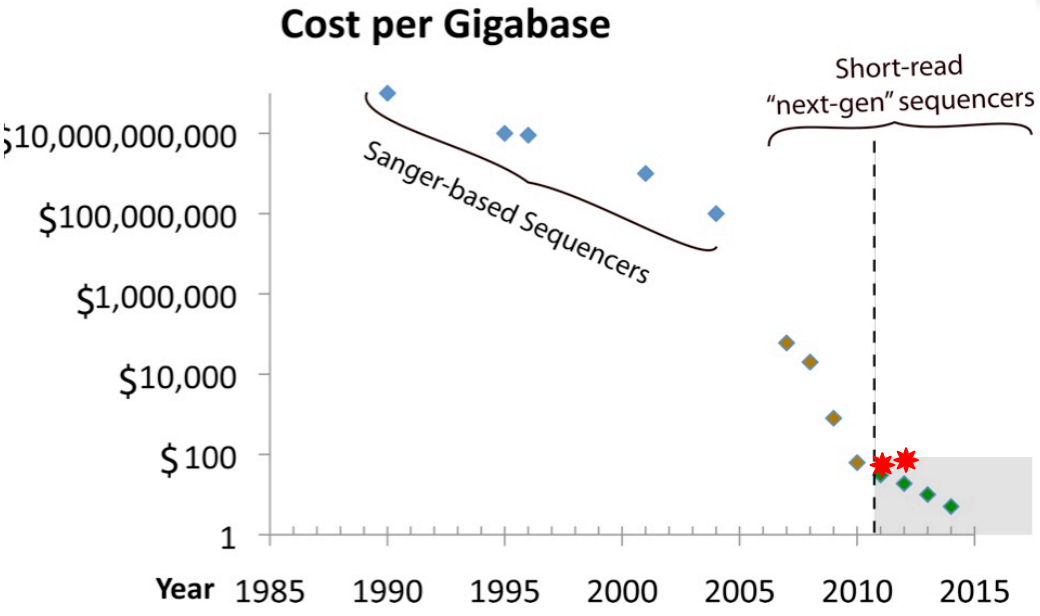




Molecular Sequencing

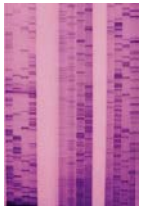




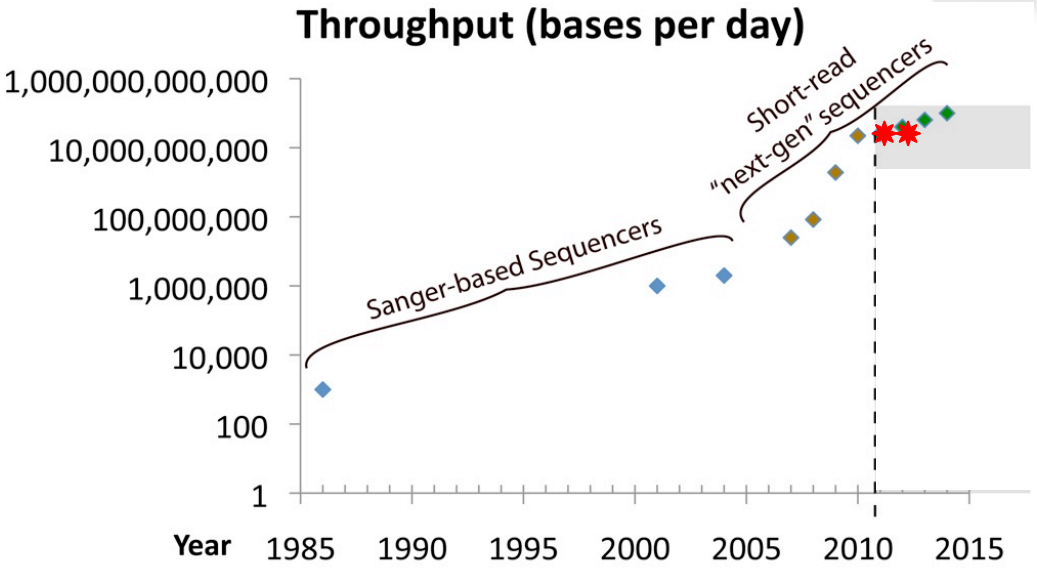



Molecular Sequencing

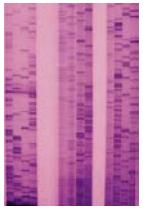
AKA Next-Generation Sequencing





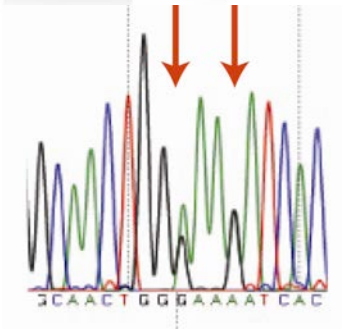
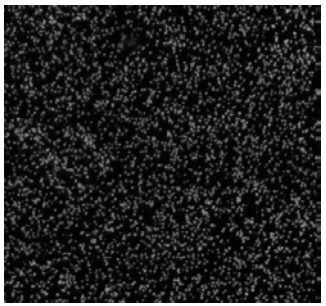



Molecular Sequencing



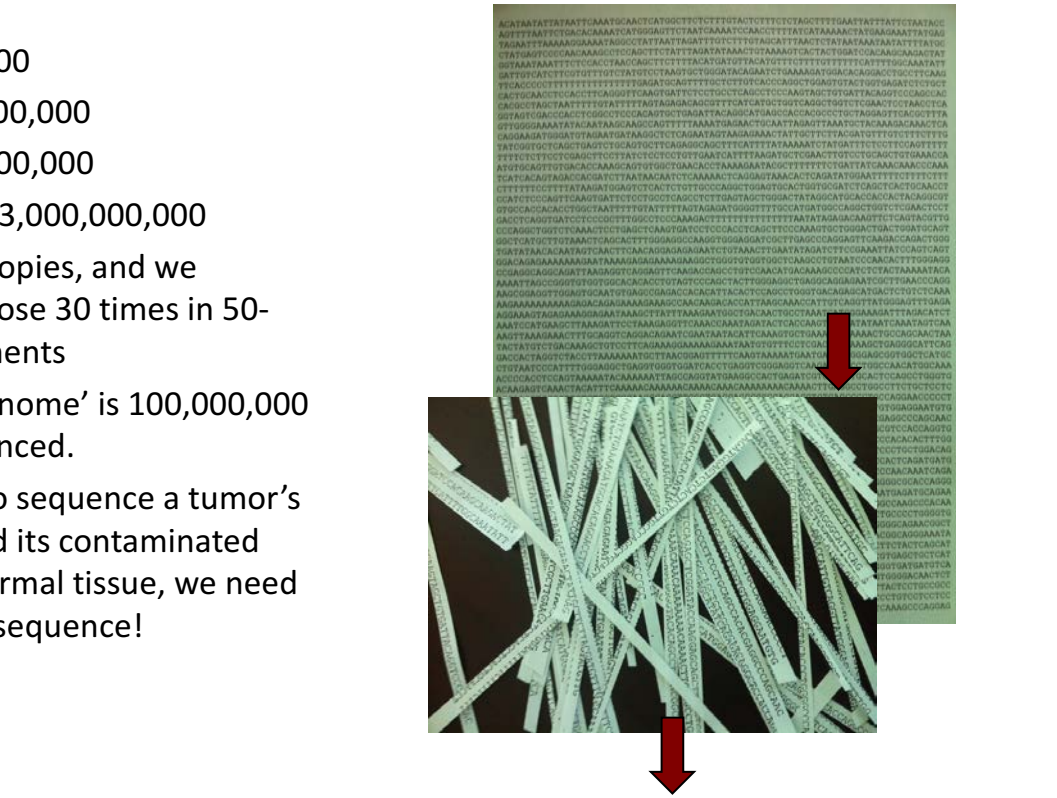



- Traditional Sanger Sequencing (1979 ->):
  - Major improvements include capillaries, use of dyes, automated calling
  - Consensus of billions of molecules
  - P&E, AB, etc
- Array-based sequencing (2002 ->)
  - Sequencing millions of pre-defined SNPs via hybridization or allele extension
  - Affymetrix, Illumina
- Pseudo single molecule sequencing (2006->)
  - Each read derived from a single molecule, clonally amplified
  - Millions of sequences sequenced base*base (lawn-sequencing)
  - 454, Solexa, Agencourt, Life
- Real-time single molecule sequencing (2010->)
  - Single molecule, fewer reads in realtime
  - PacBio, Oxford, etc

A genome is a lot of information

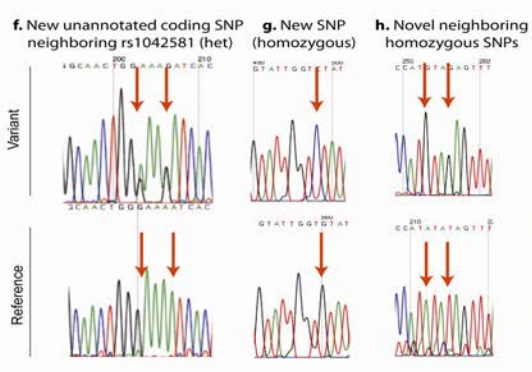
- 1 page = 5,000
- 1 ream =2,500,000
- 1 box = 25,000,000
- 120 boxes = 3,000,000,000
- You have 2 copies, and we sequence those 30 times in 50-100bp fragments
- A ‘decent genome’ is 100,000,000 bases sequenced.
- If we want to sequence a tumor’s genome, and its contaminated with 90% normal tissue, we need much more sequence!



KEY PRINCIPLES

Pseudo-single molecule reads

- A heterozygous SNP will give the paternal or maternal allele in a single read, not both

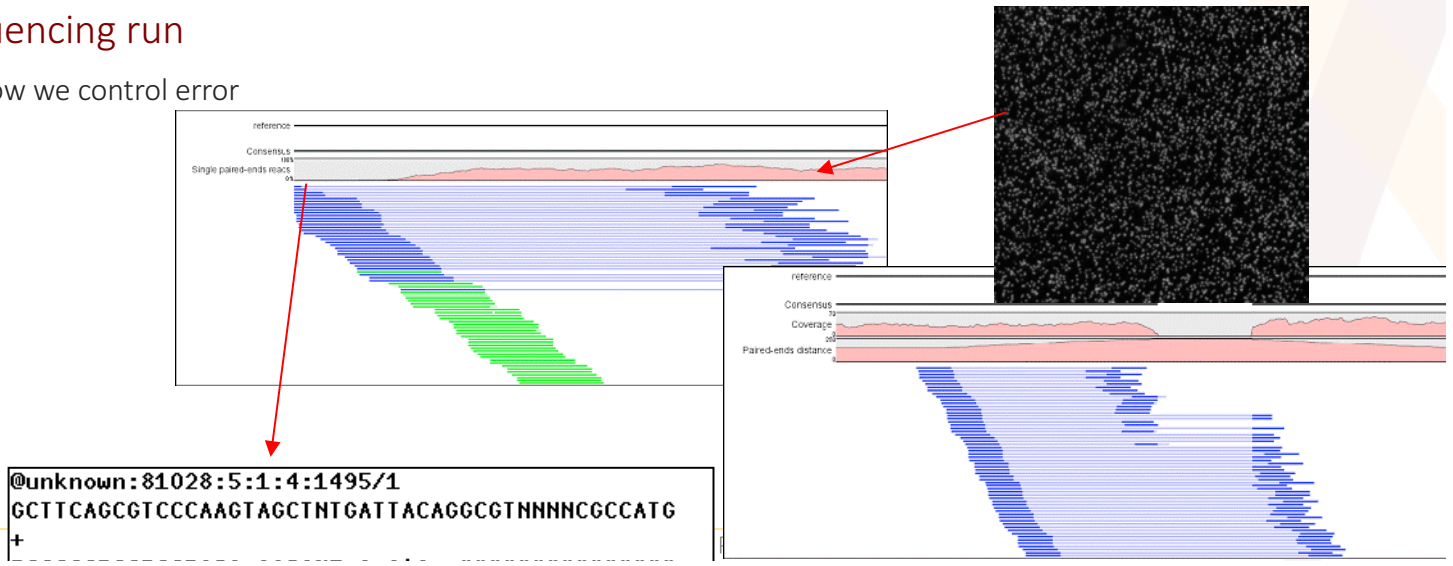


Paired-Reads

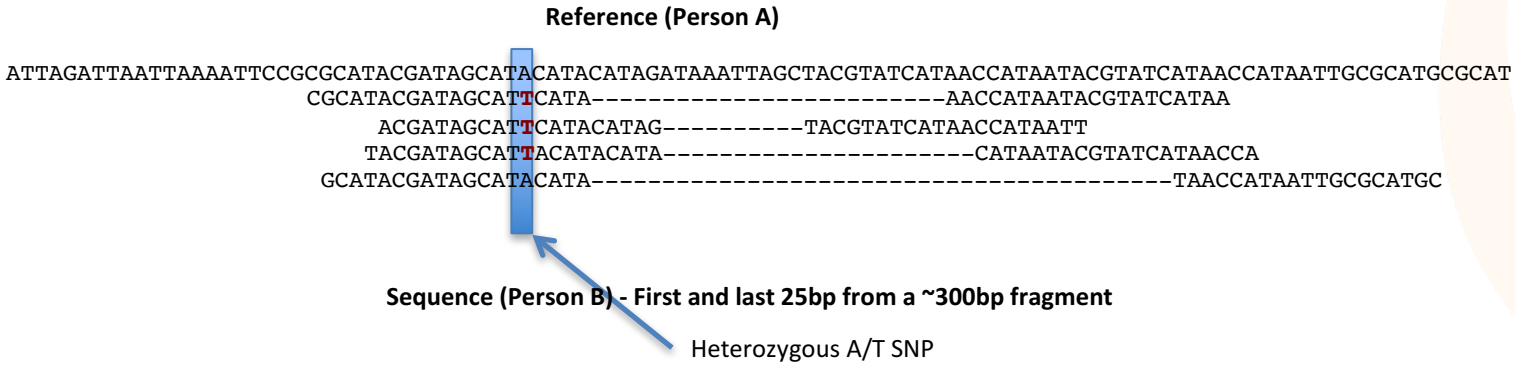
- First 100 bases and last 100 bases of a ~500bp DNA molecule

Billions of reads in a sequencing run

- Sampling matters and is how we control error



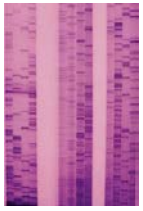
Concept of NGS Sequence Analysis



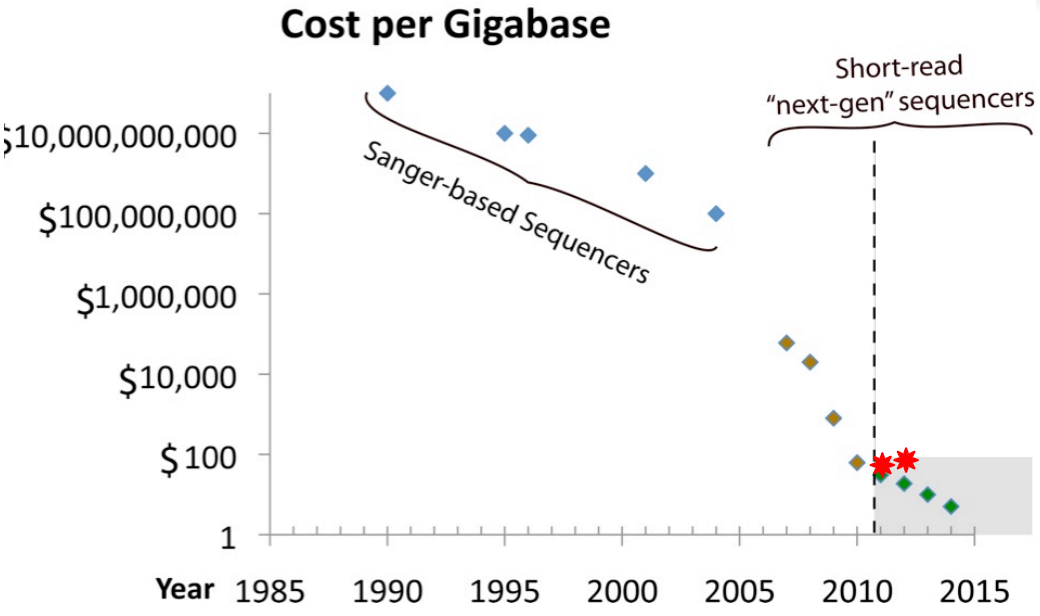
Steps to remember:

1. Alignment (produces BAM file)
2. Variant Calling (produces VCF file)
3. Interpretation (produces powerpoint)

Molecular Sequencing

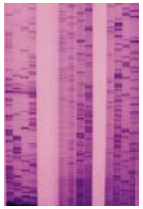




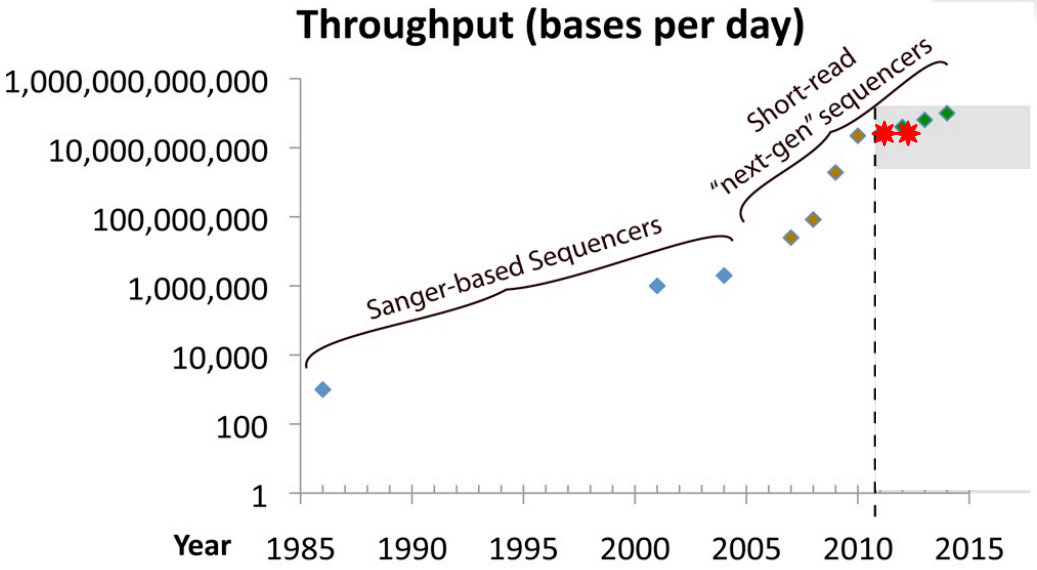



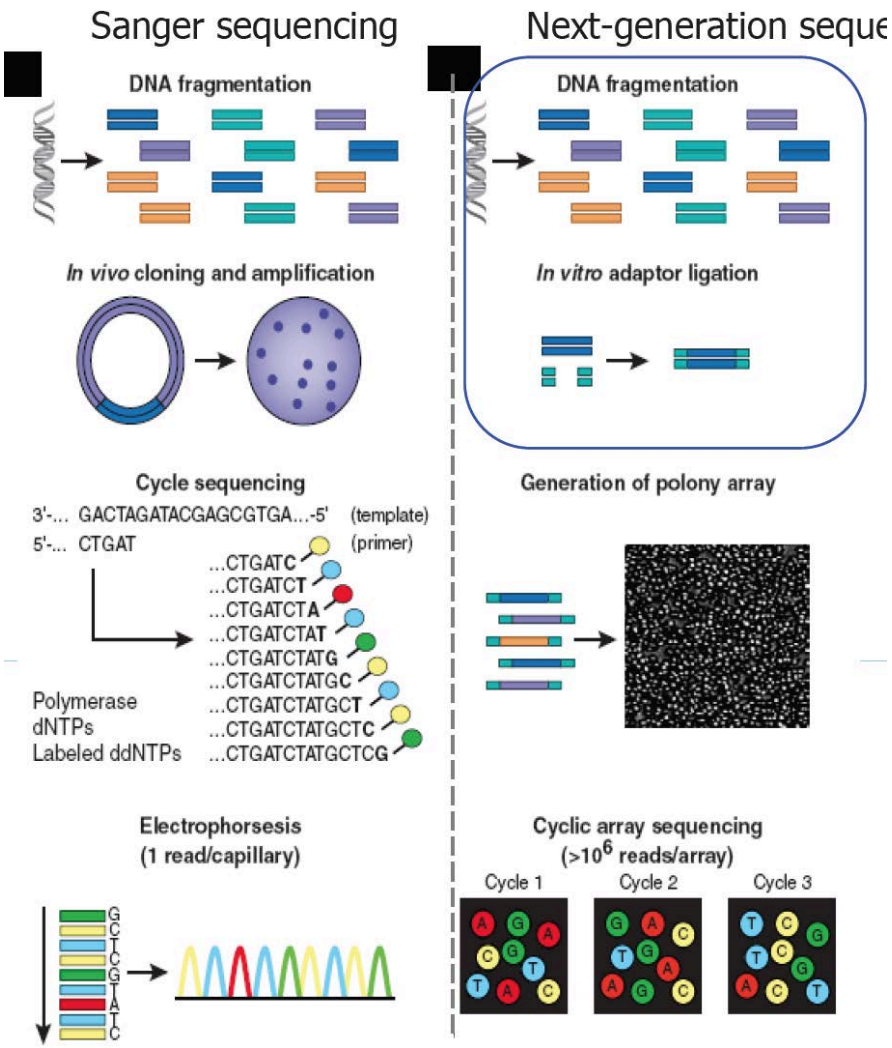
Molecular Sequencing

AKA Next-Generation Sequencing

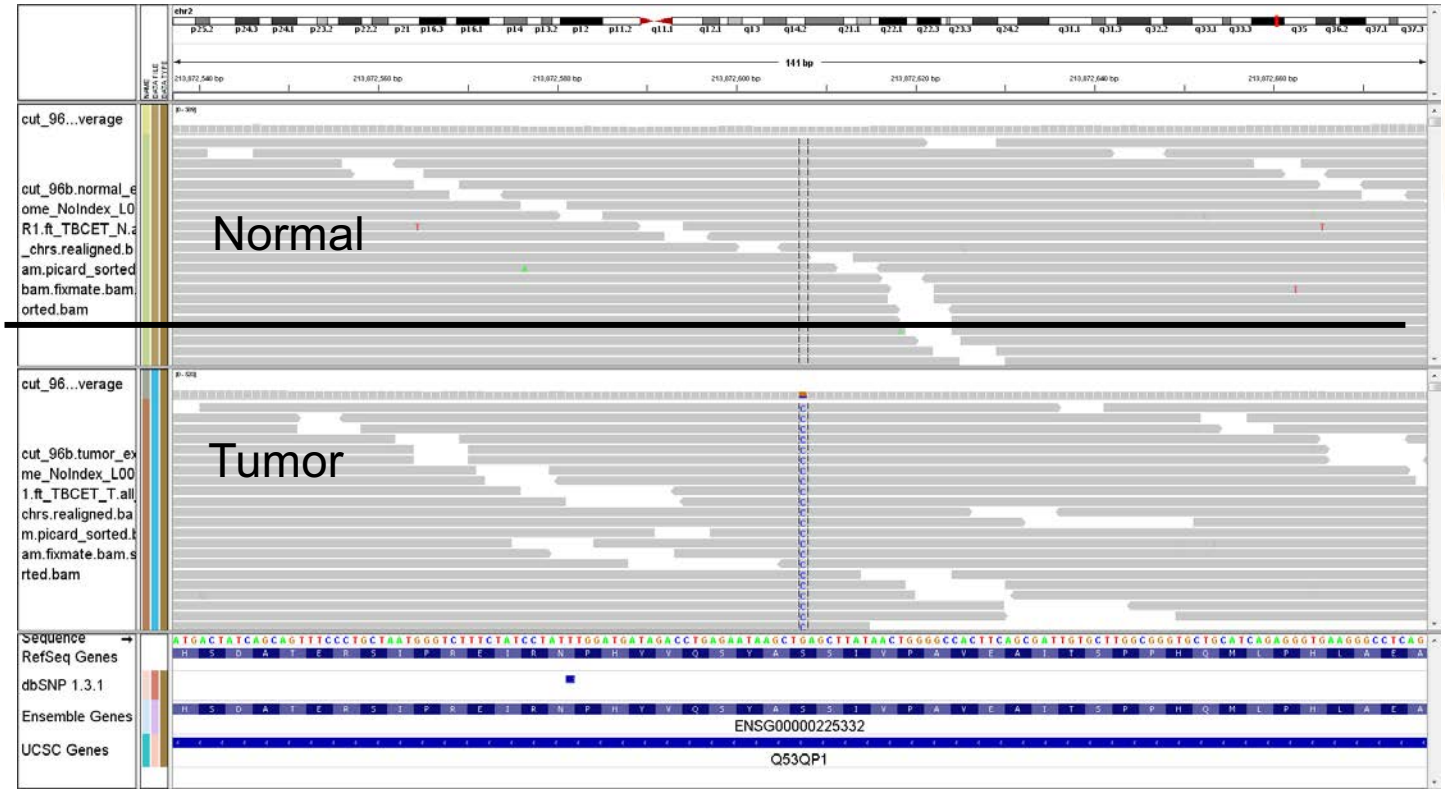




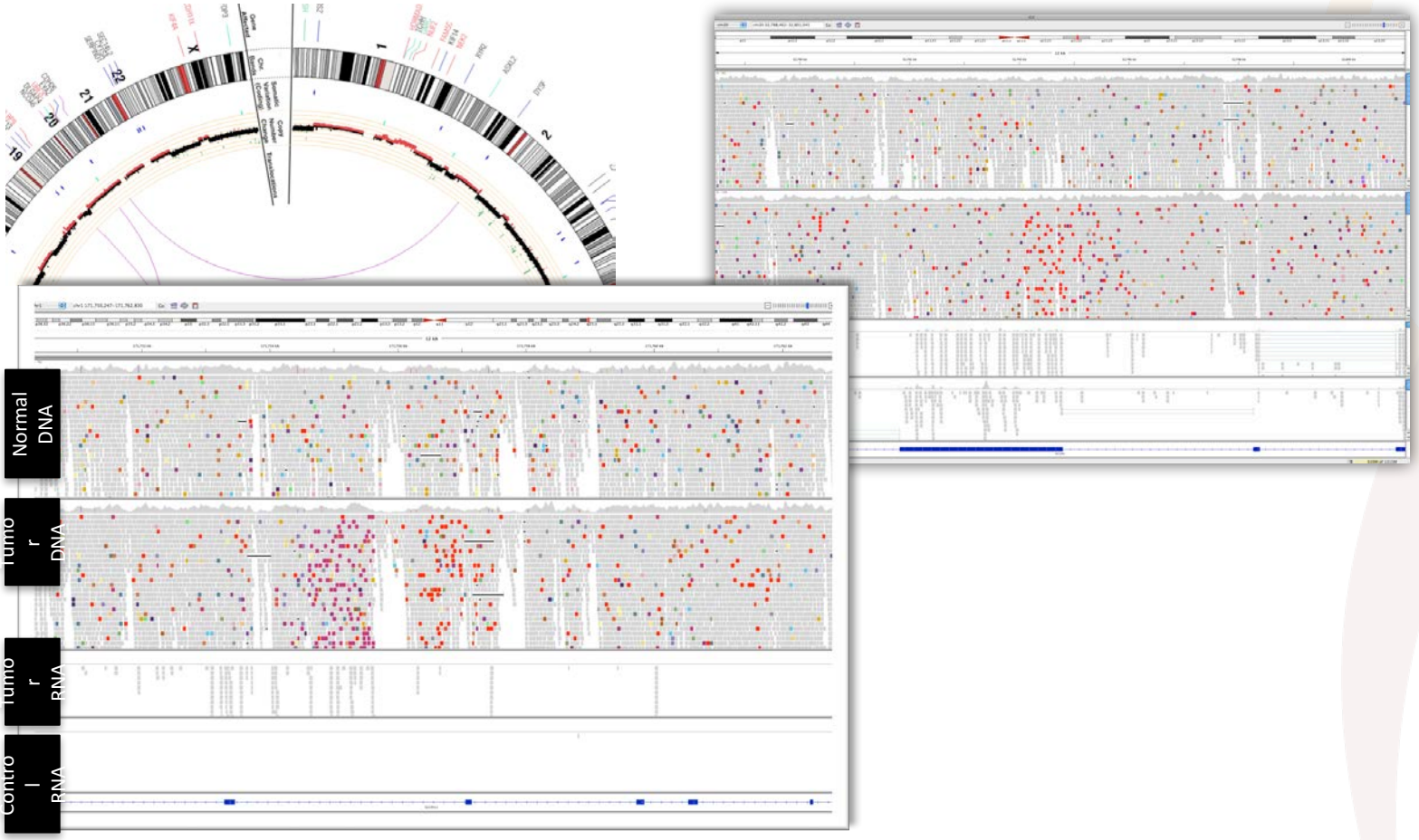




Sanger sequencing
Next-generation seque
DNA fragmentation
In vivo cloning and amplification
Cycle sequencing
3'-... GACTAGATACGAGCGTGA...-5' (template)
5'-... CTGAT (primer)
...CTGATC
...CTGATCT
...CTGATCTA
...CTGATCTAT
...CTGATCTATG
...CTGATCTATGC
...CTGATCTATGCT
...CTGATCTATGCTC
...CTGATCTATGCTCG
Polymerase
dNTPs
Labeled ddNTPs
Electrophorsesis
(1 read/capillary)
DNA fragmentation
In vitro adaptor ligation
Generation of polony array
Cyclic array sequencing
(>10^6 reads/array)
Cycle 1
Cycle 2
Cycle 3

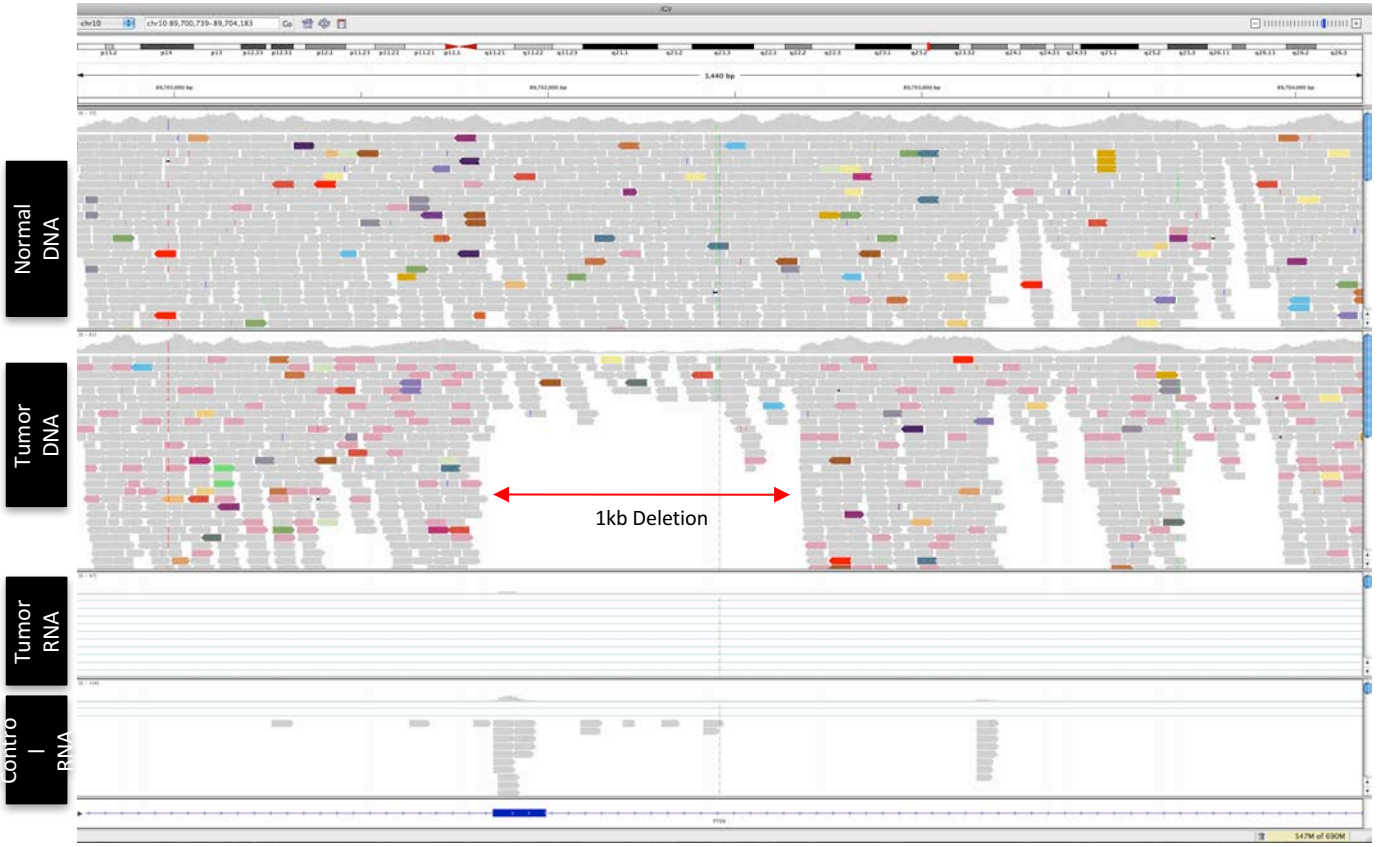
Variants: Example



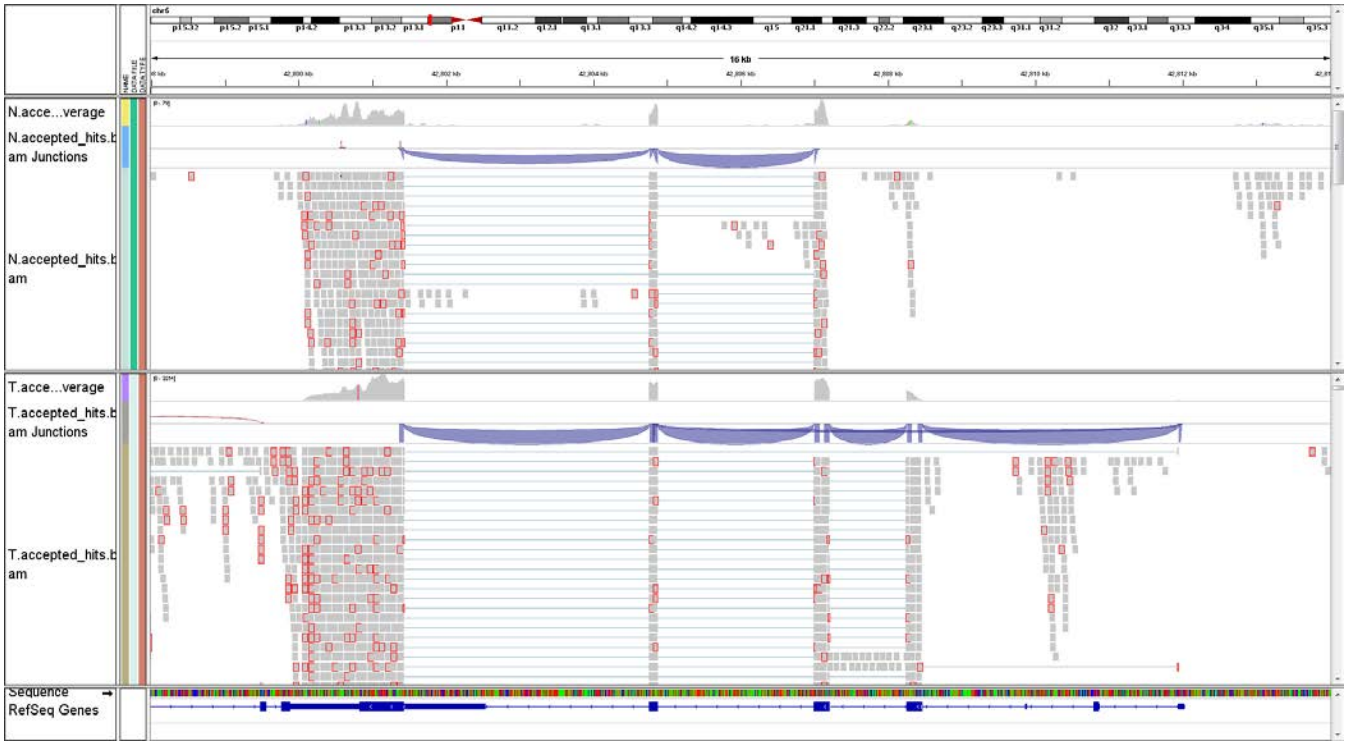
Translocations Leading to Fusion Events



Deletion of Exon 6 at *PTEN*



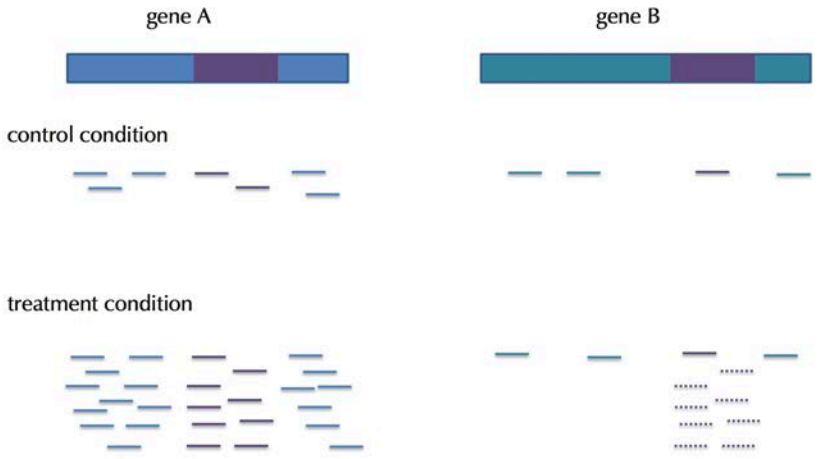
RNA


16 kb
N.acce...verage
N.accepted_hits.b
am Junctions
N.accepted_hits.b
am
T.acce...verage
T.accepted_hits.b
am Junctions
T.accepted_hits.b
am
Sequence
RefSeq Genes

COUNTS & NORMALIZATION

RNA Seq is at a basic level counting expressed genes

- Gene transcripts are broken into small fragments and counted by sequencing
  - Big genes will yield more transcripts – thus at some level we known we must correct or normalize for this.
- Different experiments generate different numbers of reads.
  - One sample may have millions of reads and another only yielded about 15% of the reads.
- Fragments Per Kilobase of transcript per Million mapped reads (FPKM) is one type of normalization for these effects building from these concepts.



Presentation Contact: David W. Craig, Ph.D. (davidwcr@usc.edu)

Normalization is Subjective – but core concepts remain

Normalization

- f sample A has been sampled deeper than sample B, we expect counts to be higher.
- Form a “virtual reference sample” by taking, for each gene, the geometric mean of counts over all samples – size factor approach

Fundamental rule:

- We may attribute a change in expression to a treatment only if this change is large compared to the expected noise.

Presentation Contact: David W. Craig, Ph.D. (davidwcr@usc.edu)

End of Part 1