The Elizabeth H.
and James S. McDonnell III
MCDONNELL GENOME INSTITUTE
at Washington University

Next-Generation Sequencing Technologies

Elaine R. Mardis, Ph.D.

Co-director, McDonnell Genome Institute

Robert E. and Louise F. Dunn Distinguished Professor of Medicine

Current Topics in Genome Analysis 2016

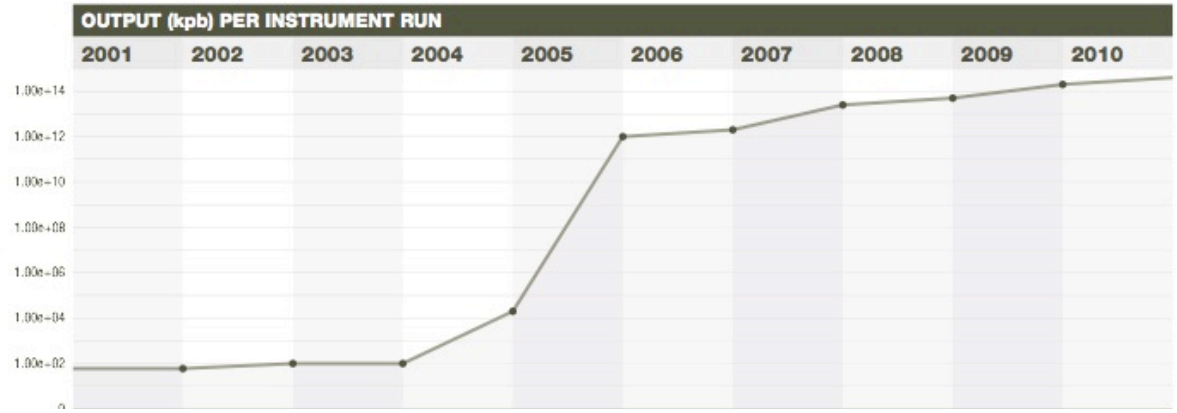
Elaine Mardis

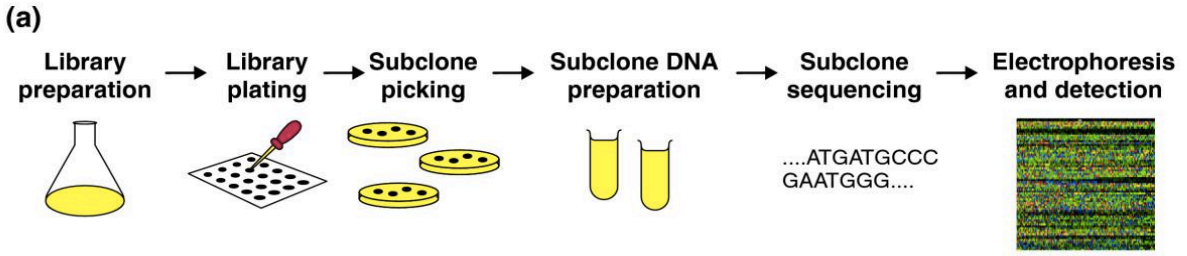
Qiagen NV
Paid Member of Board

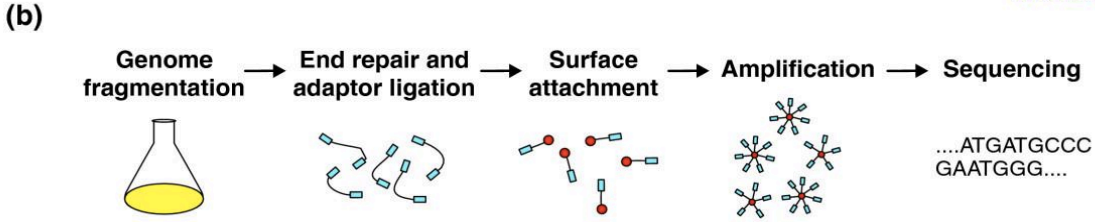
Massively Parallel Sequencing basics

How massively parallel sequencing works

NGS has transformed biomedical inquiry





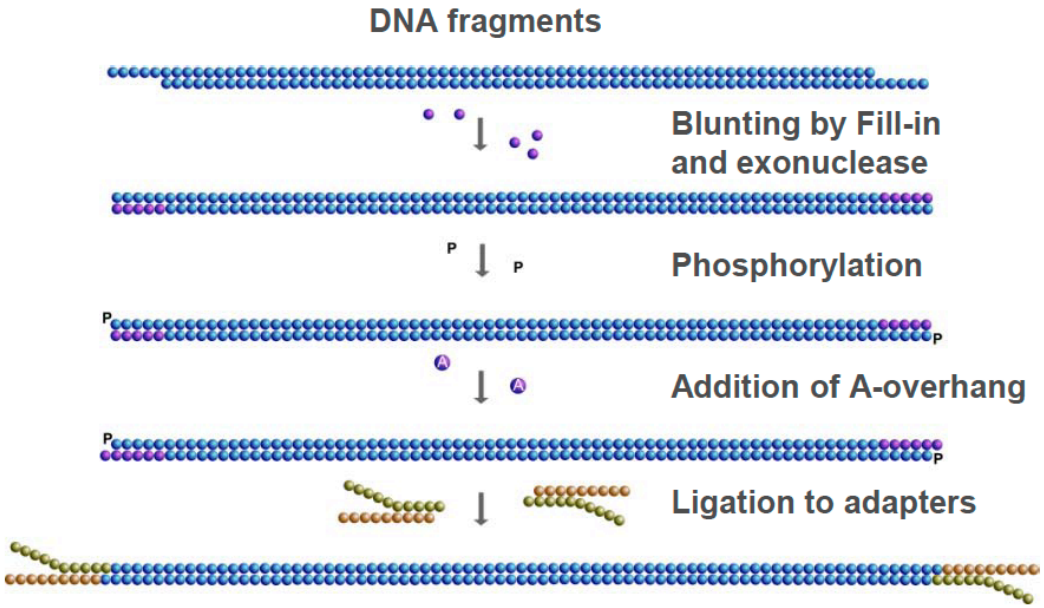


E.R. Mardis, Nature (2011) 470: 198-203, Ann. Rev. Analyt. Chem. (2013)

Massively Parallel DNA sequencing instruments

- All MPS platforms require a library obtained either by amplification or ligation with custom linkers (adapters)
- Each library fragment is amplified on a solid surface (either bead or flat Si-derived surface) with covalently attached adapters that hybridize the library adapters
- Direct step-by-step detection of the nucleotide base incorporated by each amplified library fragment set
- Hundreds of thousands to hundreds of millions of reactions detected per instrument run = "massively parallel sequencing"
- A "digital" read type that enables direct quantitative comparisons
- Shorter read lengths than capillary sequencers

Library Construction for MPS



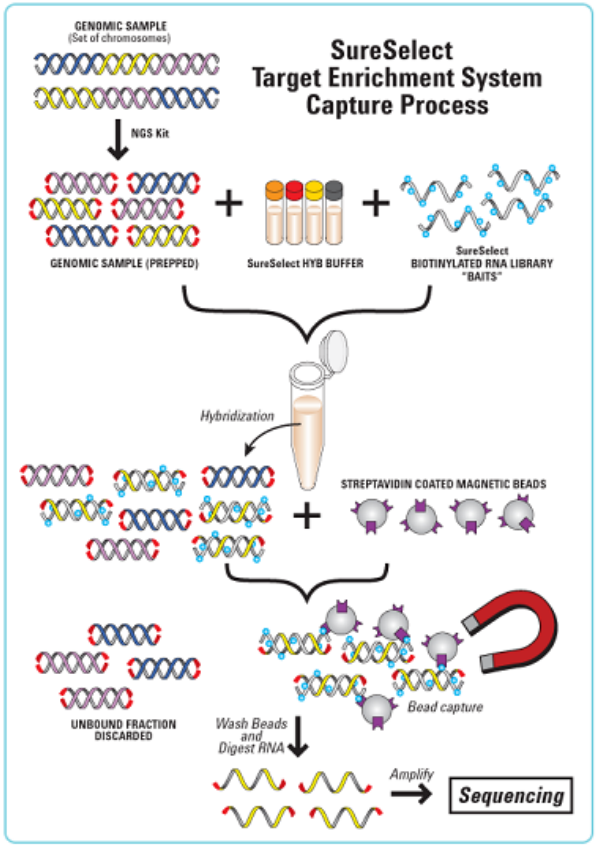
- Shear high molecular weight DNA with sonication
- Enzymatic treatments to blunt ends
- Ligate synthetic DNA adapters (each with a DNA barcode), PCR amplify
- Quantitate library
- Proceed to WGS, or perform exome or specific gene hybrid capture

PCR-related Problems in MPS

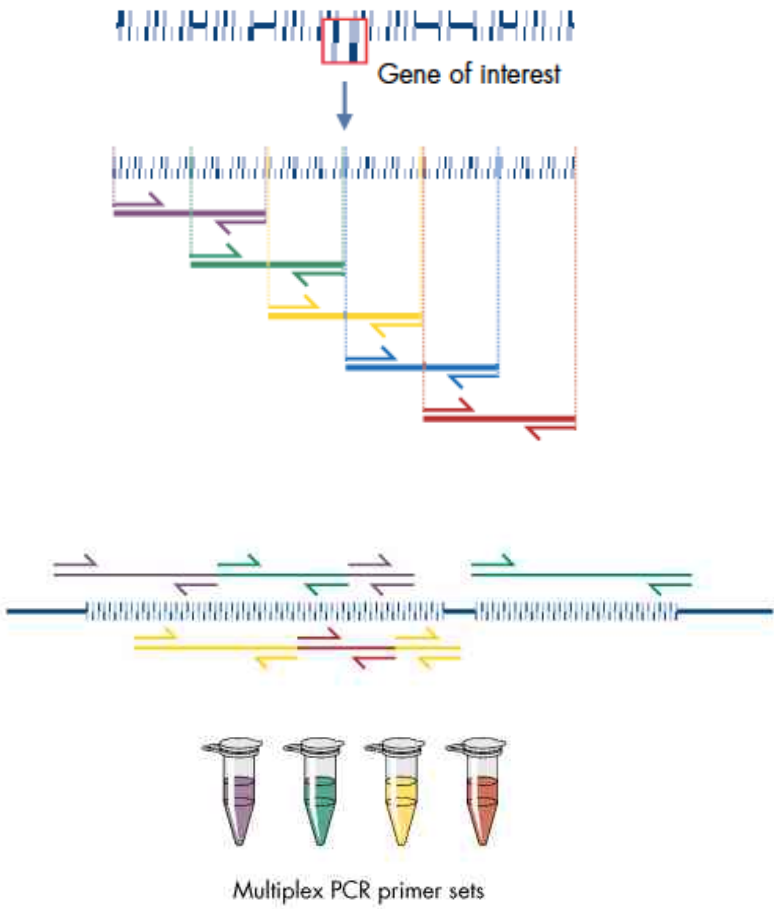
- PCR is an effective vehicle for amplifying DNA, however…
- In MPS library construction, PCR can introduce preferential amplification ("jackpotting") of certain fragments
  - Duplicate reads with exact start/stop alignments
  - Need to "de-duplicate" after alignment and keep only one pair
  - Low input DNA amounts favor jackpotting due to lack of complexity in the fragment population
- PCR also introduces false positive artifacts due to substitution errors by the polymerase
  - If substitution occurs in early PCR cycles, error appears as a true variant
  - If substitution occurs in later cycles, error typically is drowned out by correctly copied fragments in the cluster
- Cluster formation is a type of PCR ("bridge amplification")
  - Introduces bias in amplifying high and low G+C fragments
  - Reduced coverage at these loci is a result

Hybrid Capture

- **Hybrid capture** - fragments from a whole genome library are selected by combining with probes that correspond to most (not all) human exons or gene targets.
- The probe DNAs are biotinylated, making selection from solution with streptavidin magnetic beads an effective means of purification.
- An "**exome**" by definition, is the exons of all genes annotated in the reference genome.
- **Custom capture reagents** can be synthesized to target specific loci that may be of clinical interest.

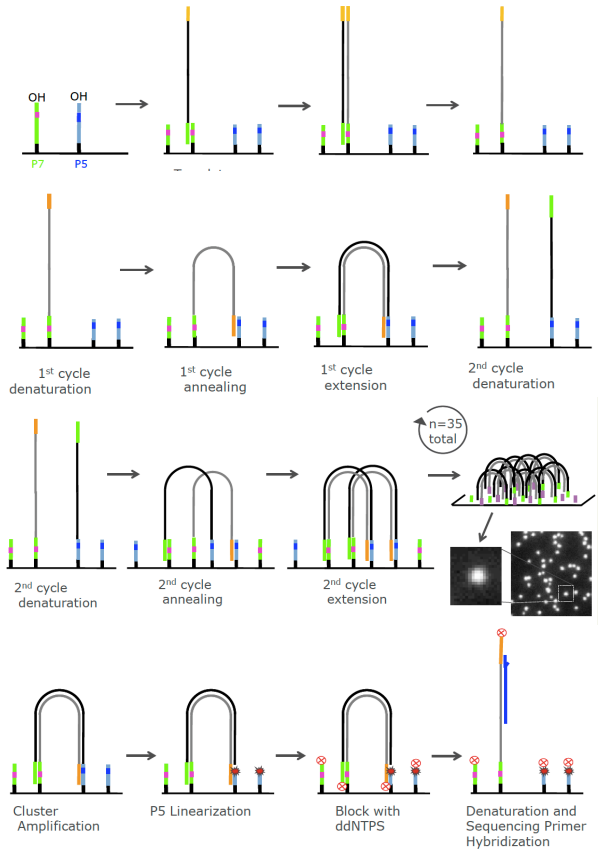


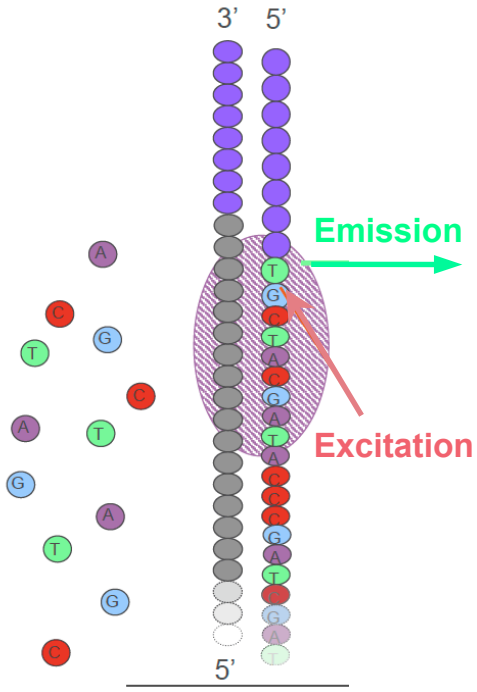
Multiplex PCR Amplification of Targets

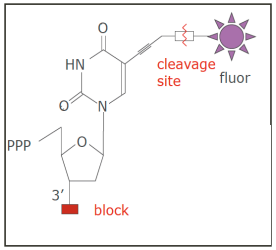


1. Design amplification primer pairs for exons of genes of interest; tile primers to overlap fragments in larger exons
2. Group primer pairs according to G+C content, Tm and reaction condition specifics
3. Amplify genomic DNA to generate multiple products from each primer set; pool products from each set
4. Create library by ligation or tail platform adaptors on the primer ends
5. Sequence

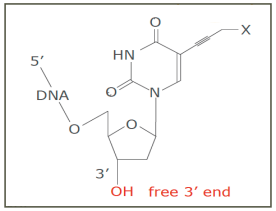
Massively Parallel Sequencing by Synthesis



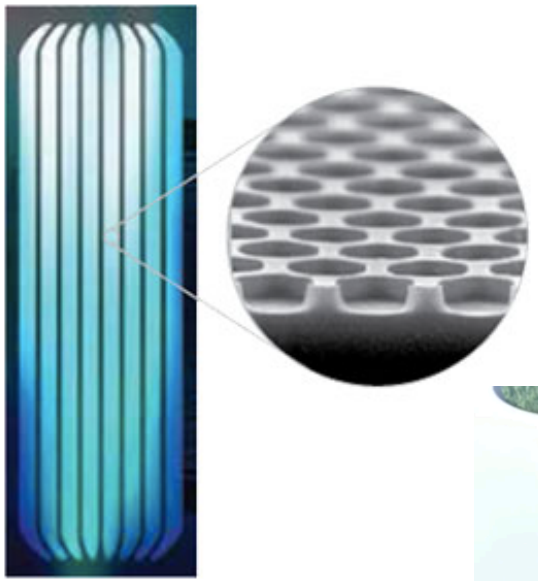


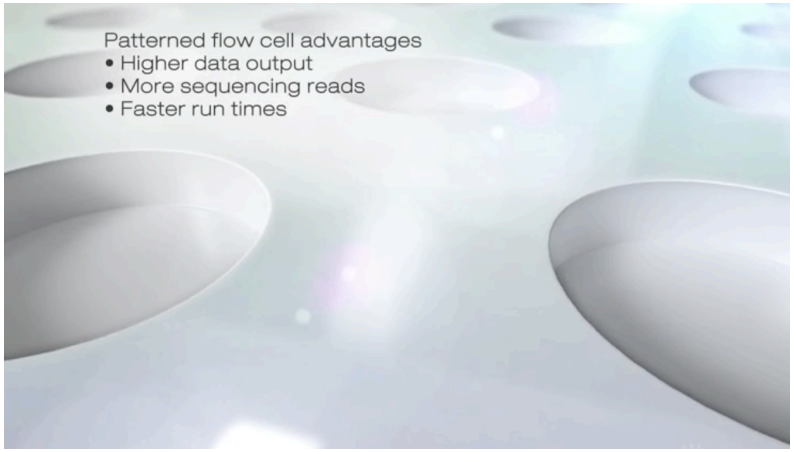


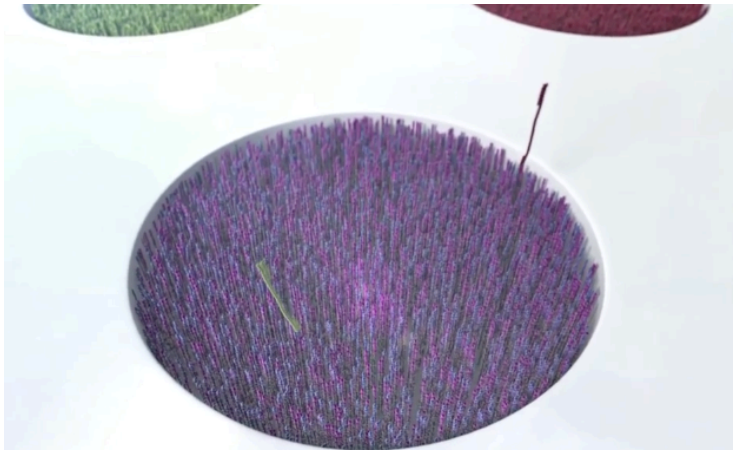
Incorporate
Detect
De-block
Cleave fluor



Illumina Patterned Flow Cell







Platforms: Illumina

MiSeq

NextSeq 500

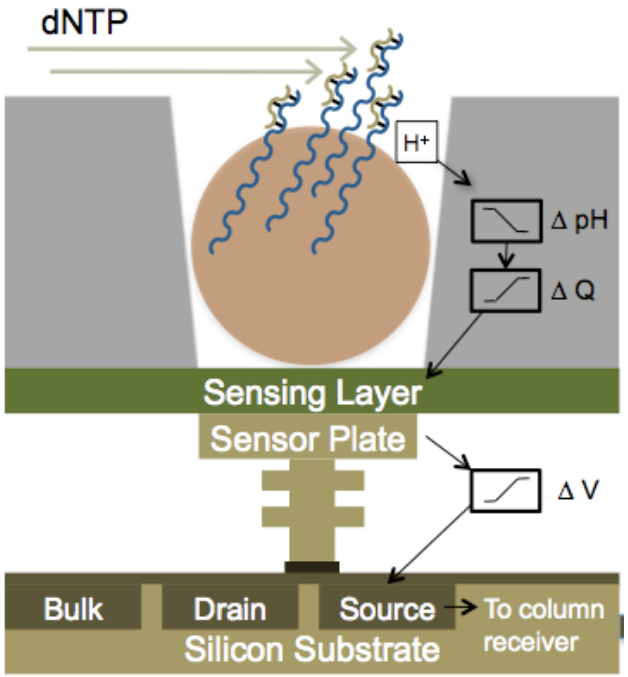
HiSeq 2500

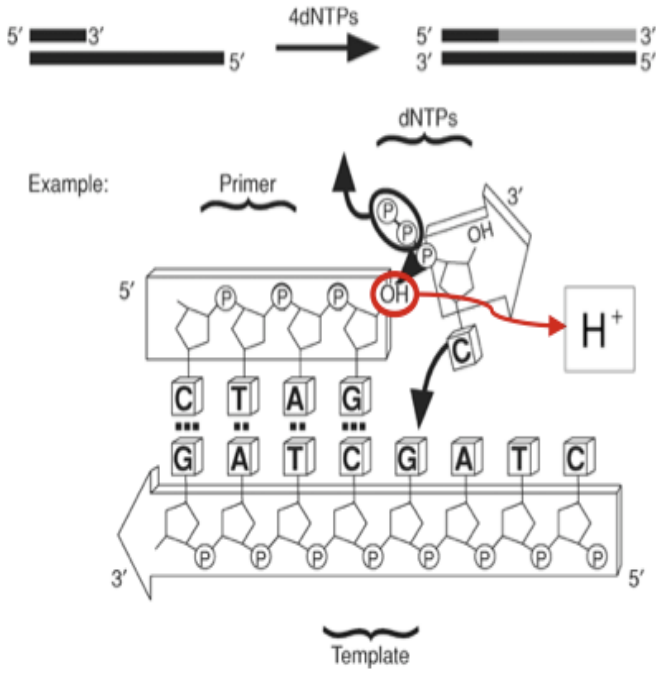
HiSeq X*

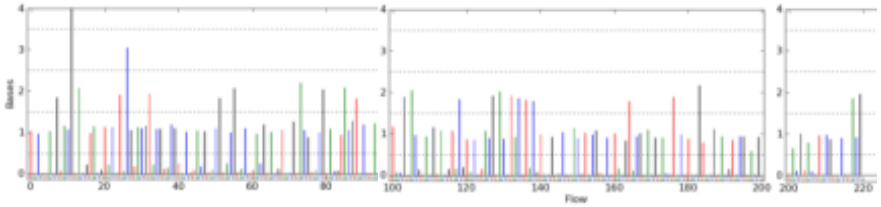
Key applications	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.		Production-scale genome, exome, transcriptome sequencing, and more.		Population-scale human whole-genome sequencing.
Run mode	N/A	Mid-Output	High-Output	Rapid Run	High-Output	N/A
Flow cells processed per run	1	1	1	1 or 2	1 or 2	1 or 2
Output range	0.3-15 Gb	20-39 Gb	30-120 Gb	10-180 Gb	50-1000 Gb	1.6-1.8 Tb
Run time	5-65 hours	15-26 hours	12-30 hours	7-40 hours	< 1 day - 6 days	< 3 days
Reads per flow cell†	25 Million‡	130 Million	400 Million	300 Million	2 Billion	3 Billion
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 125 bp	2 × 150 bp

- High accuracy, range of capacity and throughput
- Longer read lengths on some platforms (MiSeq)
- Improved kits, improved software pipeline and capabilities, cloud compute

ION Torrent-pH Sensing of Base Incorporation







Platforms: Ion Torrent

PGM

- Three sequencing chips available:
  - 314 = up to 100 Mb
  - 316 = up to 1 Gb
  - 318 = up to 2 Gb
- 2-7 hour/run
- up to 400 bp read length
- 400kreads up to 5 Mreads

Proton

- Two human exomes (Proton 1 chip) or one genome (@20X-Proton 2 chip) per run
- Ion One Touch or Ion Chef preparatory modules
- 2-4 hour/run
- ~200 bp average read length
- Proton 1 produces 60-80 Mreads ≥50 bp

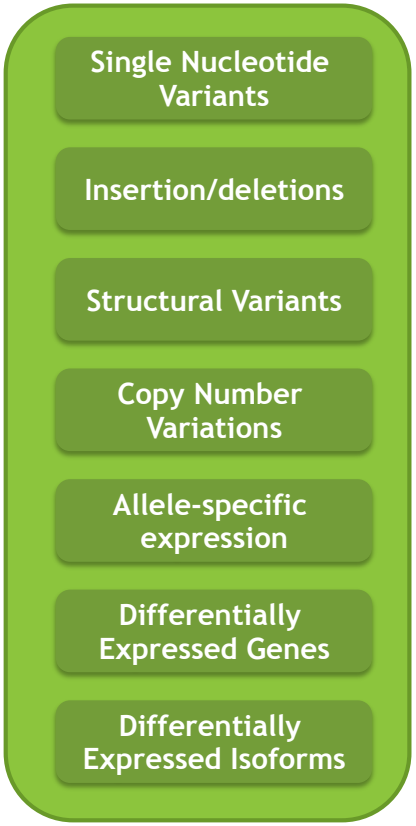
- Low substitution error rate, in/dels problematic, no paired end reads
- Inexpensive and fast turn-around for data production
- Improved computational workflows for analysis

Post Data Generation Analyses

Bioinformatic and computational approaches to NGS

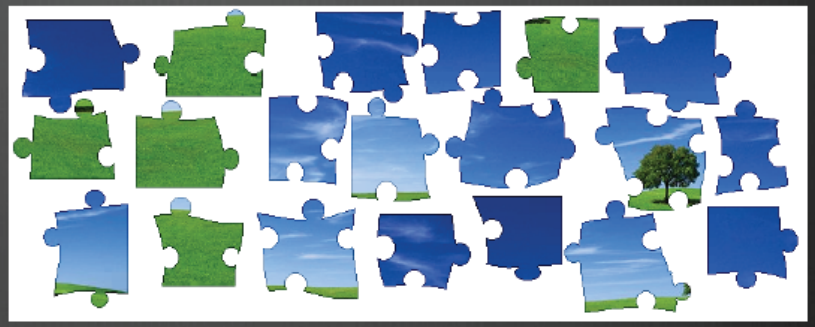
The Human Genome Reference enables MPS Genomics

- The human genome reference sequence is the keystone interpreting MPS sequencing read data
- **Alignment** of reads to the human reference sequence is the first step to identify variation of all types
- Mis-aligning sequences identify structural alterations
- Alignment and analysis of RNA sequence data provides information about gene expression changes



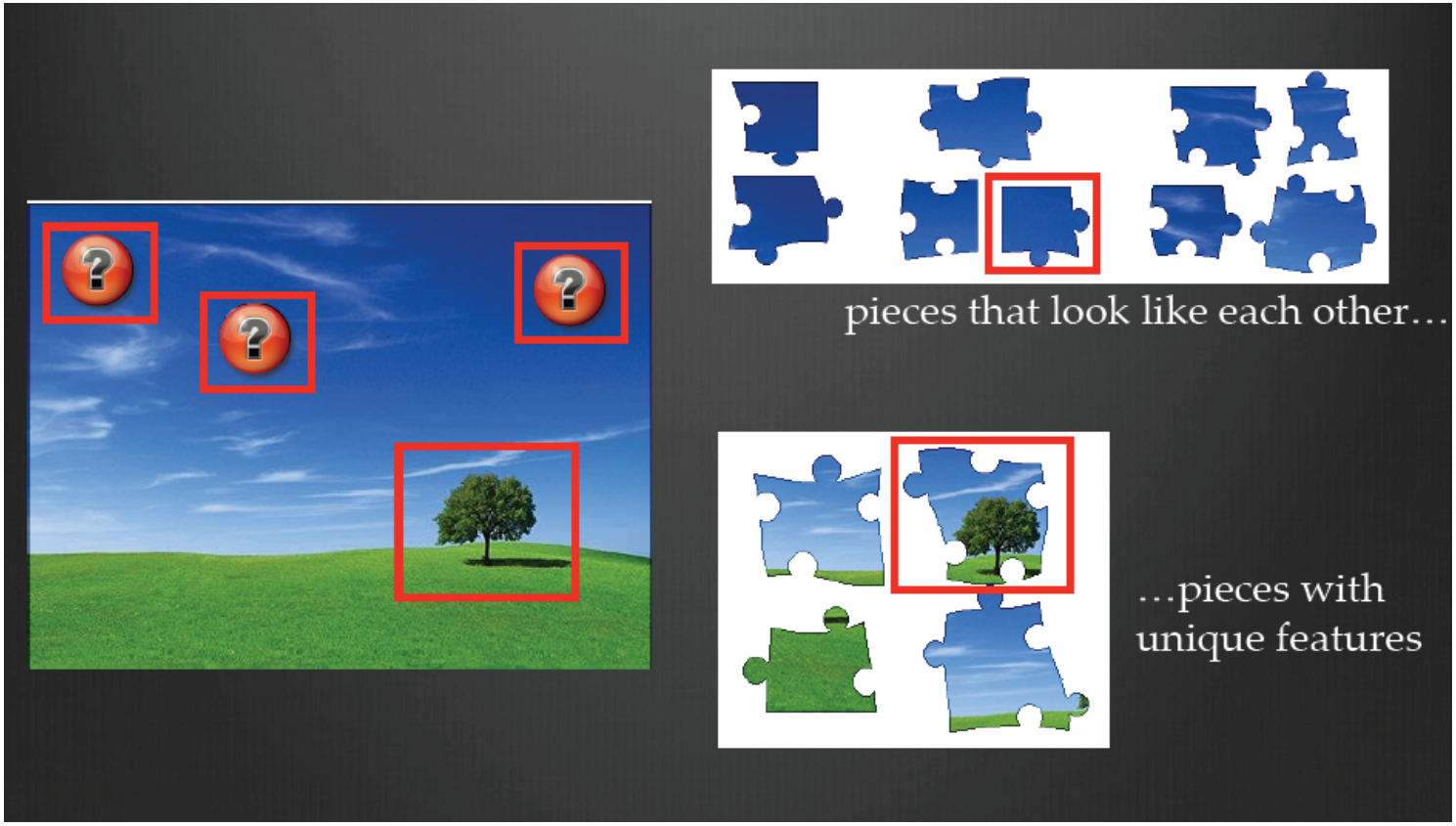
Short Read Alignment…

Is like a jigsaw puzzle…

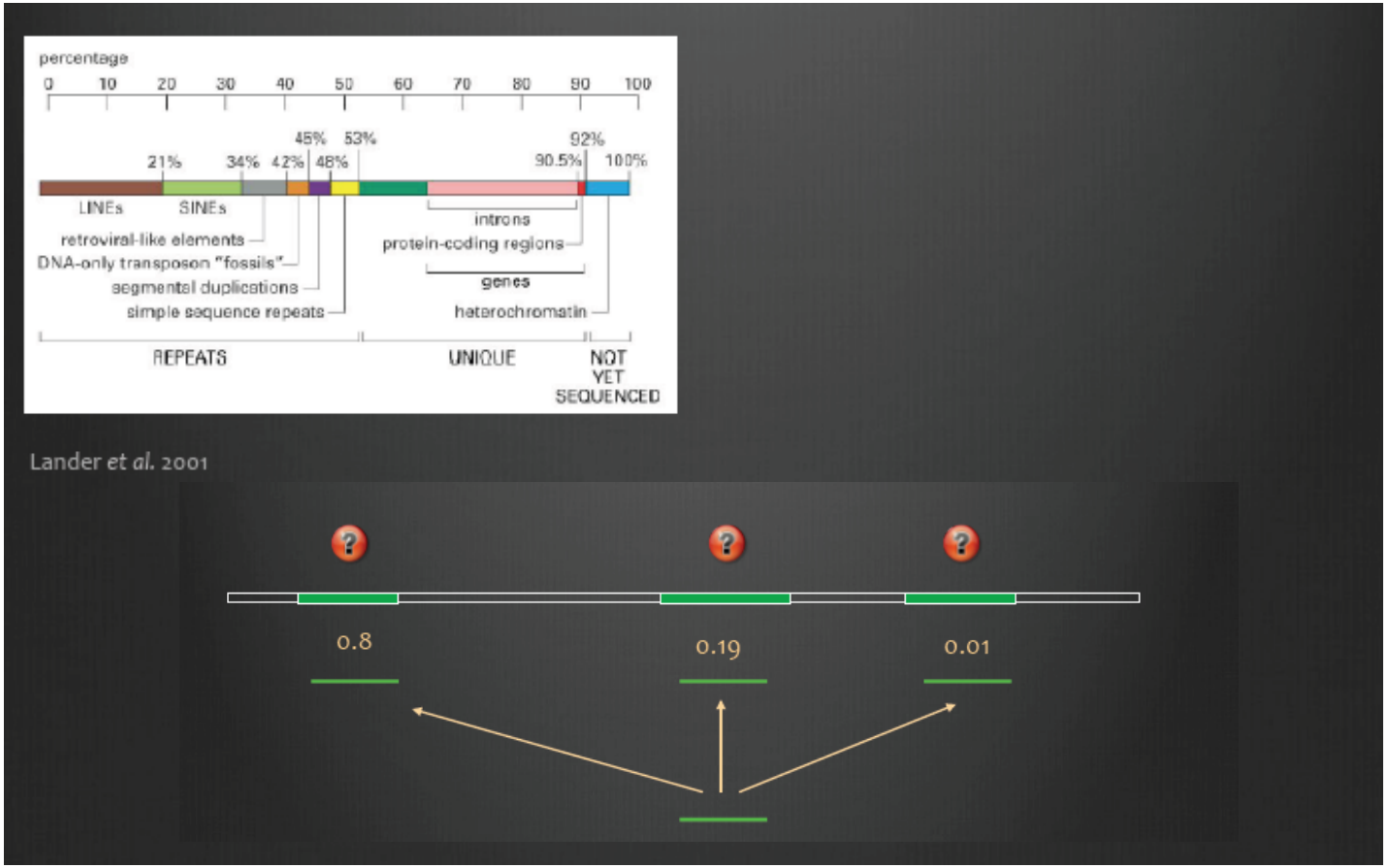


…where they give you the cover on the box

Some pieces are easier to place than others…



Repetitive Sequences Result in Multiple Read Alignments

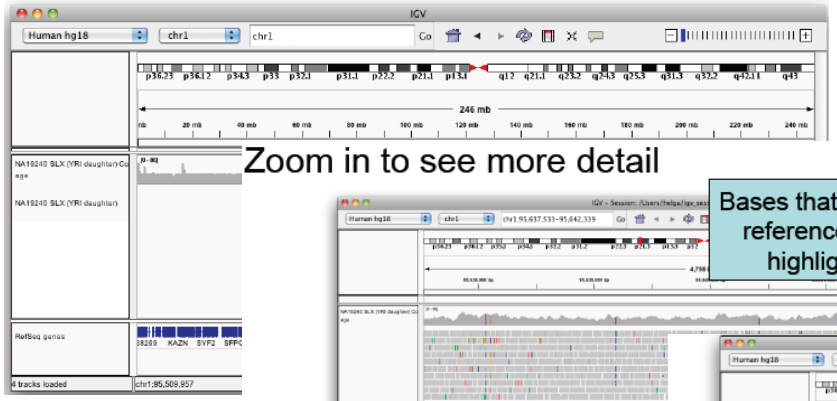


Reads are Aligned, Now What?

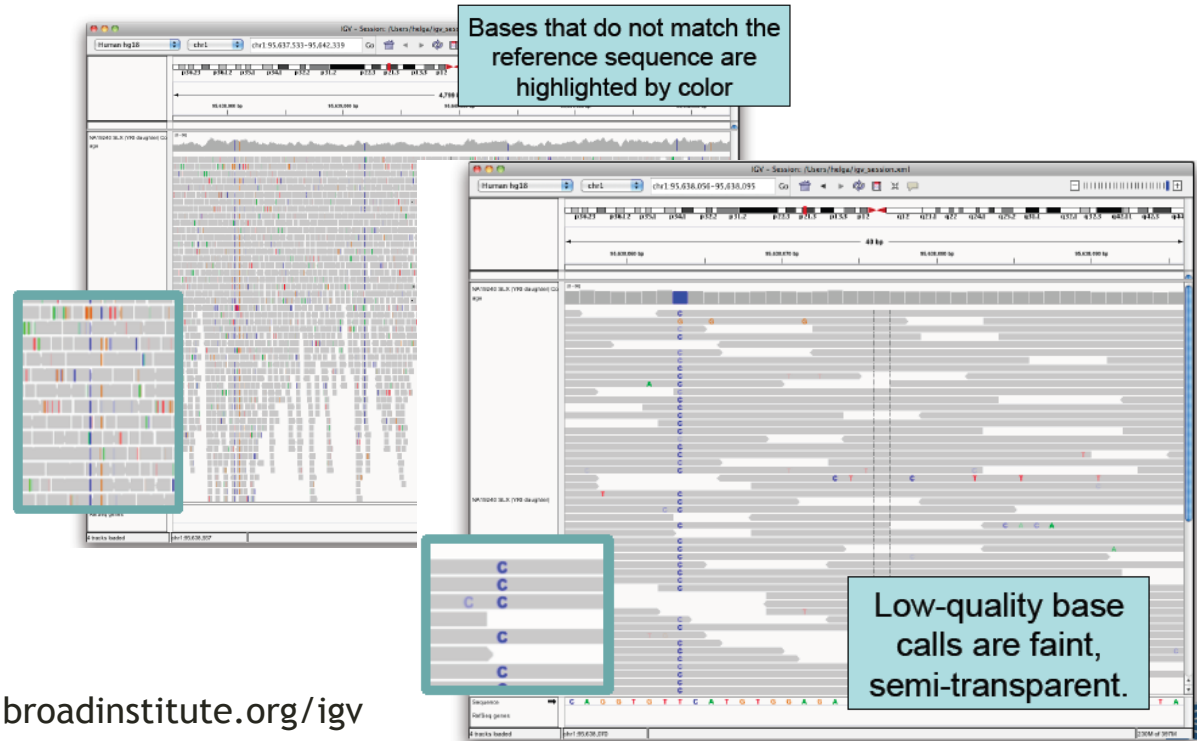
- Data calibration and cleanup:
  - Mark proper pairs (if applicable)
  - Mark duplicate reads!
  - Correct local misalignments
  - Recalculate quality scores
- Call SNPs
- Evaluate Coverage
  - Compare SNPs from NGS to SNPs from array data OR
  - Compare SNPs from tumor to normal (Which are shared? Number?)
  - Integrated Genome Viewer
  - RefCov and others
- Analyze the data

Integrated Genomics Viewer (IGV)

Whole chromosome view



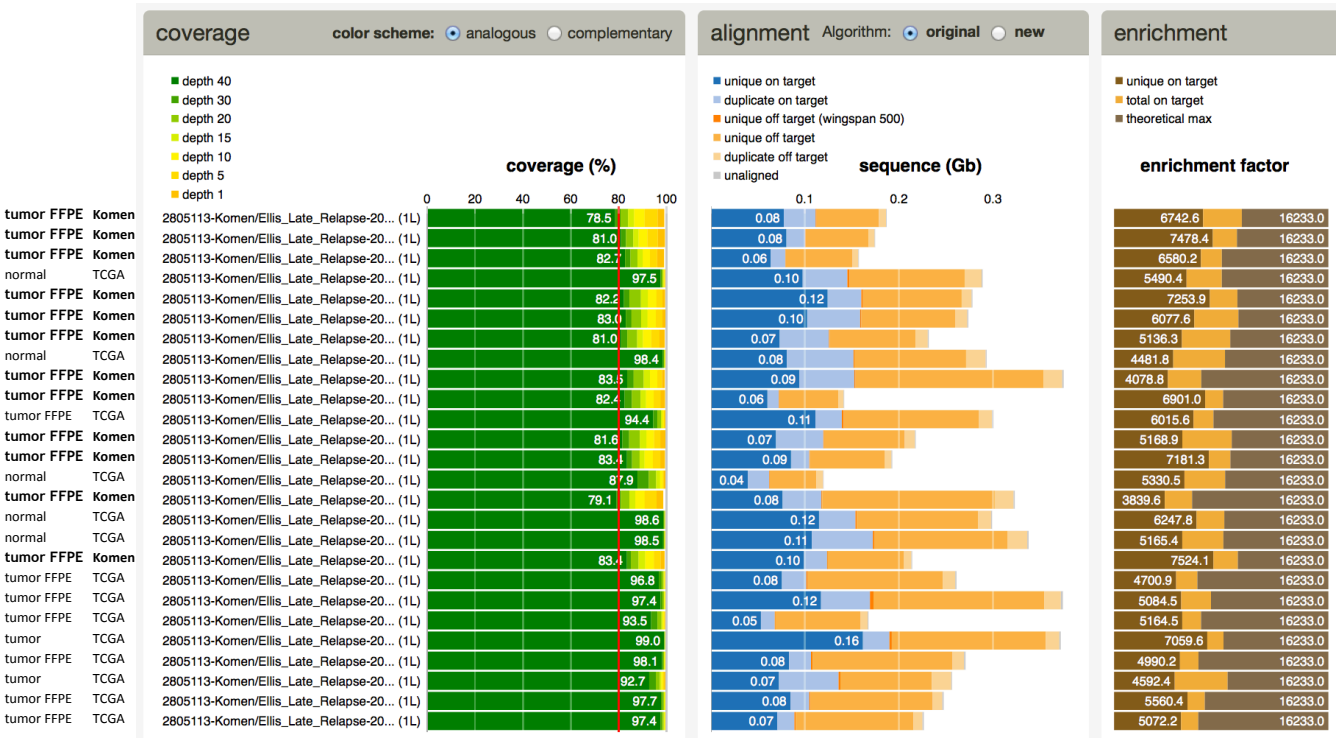
Zoom in to see more detail



http://www.broadinstitute.org/igv

IGV: Somatic Single Nucleotide Variant QC

RefCov: Coverage Depth and Breadth from Hybrid Capture



http://gmt.genome.wustl.edu/genome-shipit/gmt-refcov/0.3/index.html

Data and coverage characteristics

Sequencing Data		
	Exome	**Whole Genome**
Data generated	6 Gbp	118 Gbp
Target space	37 Mbp	3.2 Gbp
Typical map rate	98.5%	95.0%
Typical dup rate	8.8%	3.4%
Avg. CDS sequence depth	65x	30x
% of CDS covered >10x	90-95%	95-99%

Variant Detection		
	Exome	**Whole Genome**
SNV calling	Good	Good
Small indel calling	Good	Good
CNV calling	Poor	Good
SV calling	Poor	OK
Typical SNVs called	~50,000	~3 million

False Negativity/Positivity

- Most false negatives are due to lack of coverage
- False positives are due to multiple reasons, including:
  - Variant is only called on one strand
  - Variant is only called at the end of the read
  - Coverage of the matched normal at that locus is poor
  - Gene has a pseudogene/paralog and the reads are mis-mapped
  - High sensitivity variant calling algorithms have elevated false positive rates to achieve detection of subclonal variants and low false negative rates
- Data that verifies or refutes variant calls can help to define bioinformatic filters to remove them

Integrated Systems: Qiagen GeneReader

- Modules for Sample Prep, Library Construction/Amplification, SBS sequencing
- Onboard analysis and interpretation software for mutation detection and interpretation based on Ingenuity Variant and Pathway Analysis

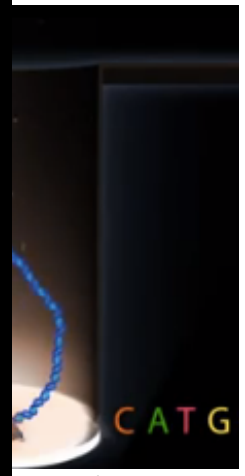
Third Generation Sequencers

Variations on a theme

Real Time Sequencing of Single DNA Molecules

DNA:polym
immobilize

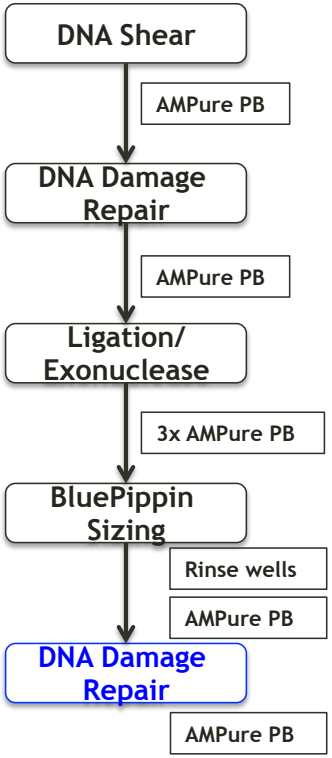
ulse is produced
fluorescent base
n the polymerase
te.
sphate is
during
ation, releasing
rophore



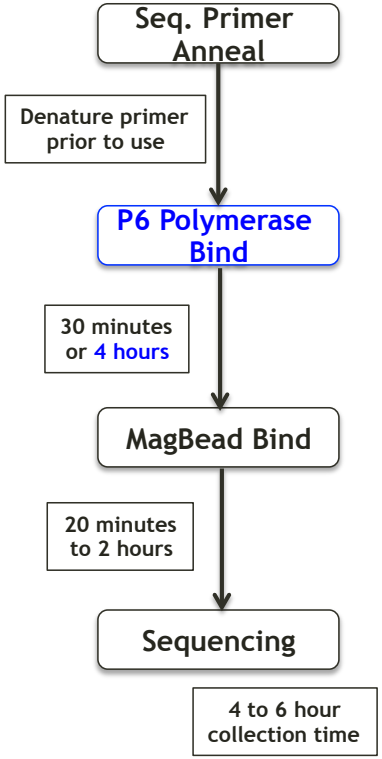
ccurs in
the loaded

Pacific Biosciences RS

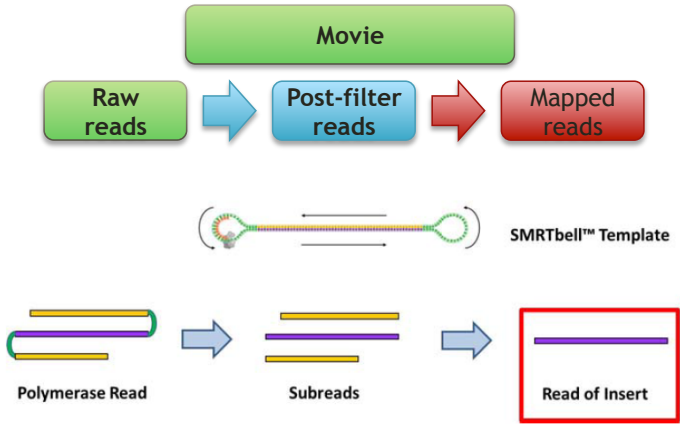
Sample Prep



Sequencing Prep



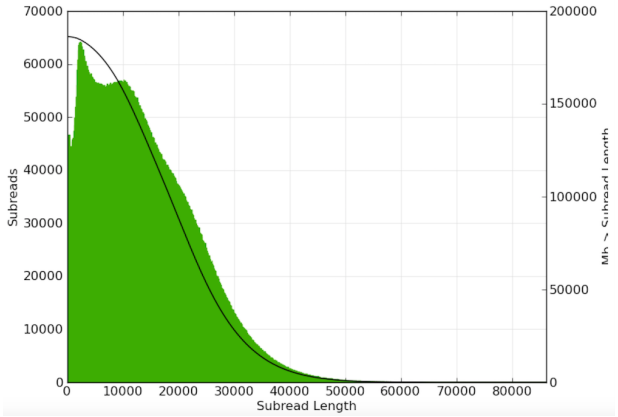
Sequencing



Considerations for PacBio WGS

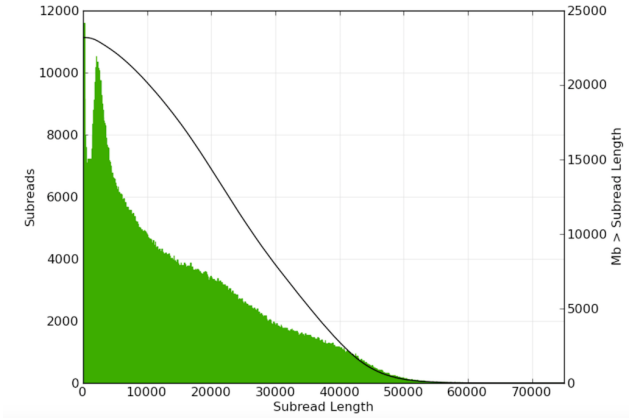
- High molecular weight genomic DNA
  - DNA must be of sufficient quality to allow for >30 kb shearing to produce PacBio Continuous Long Reads (CLR)
- Consistent shearing >30 kb
  - Shearing genomic DNA >30 kb is challenging and requires a consistent technology
  - Preferred method: Diagenode Megaruptor
  - Alternate method: Covaris g-Tube
- Sufficient DNA for PacBio sample prep
  - A single PacBio sample prep reaction requires 5 µg sheared DNA
  - One library is composed of 8-10 sample prep reactions
  - At least 2-4 libraries are required for 60x coverage

Read Length Comparisons



Library #1 18-50kb

- Mean Subread Length 13,720 bp
- N50 Length 19,411 bp



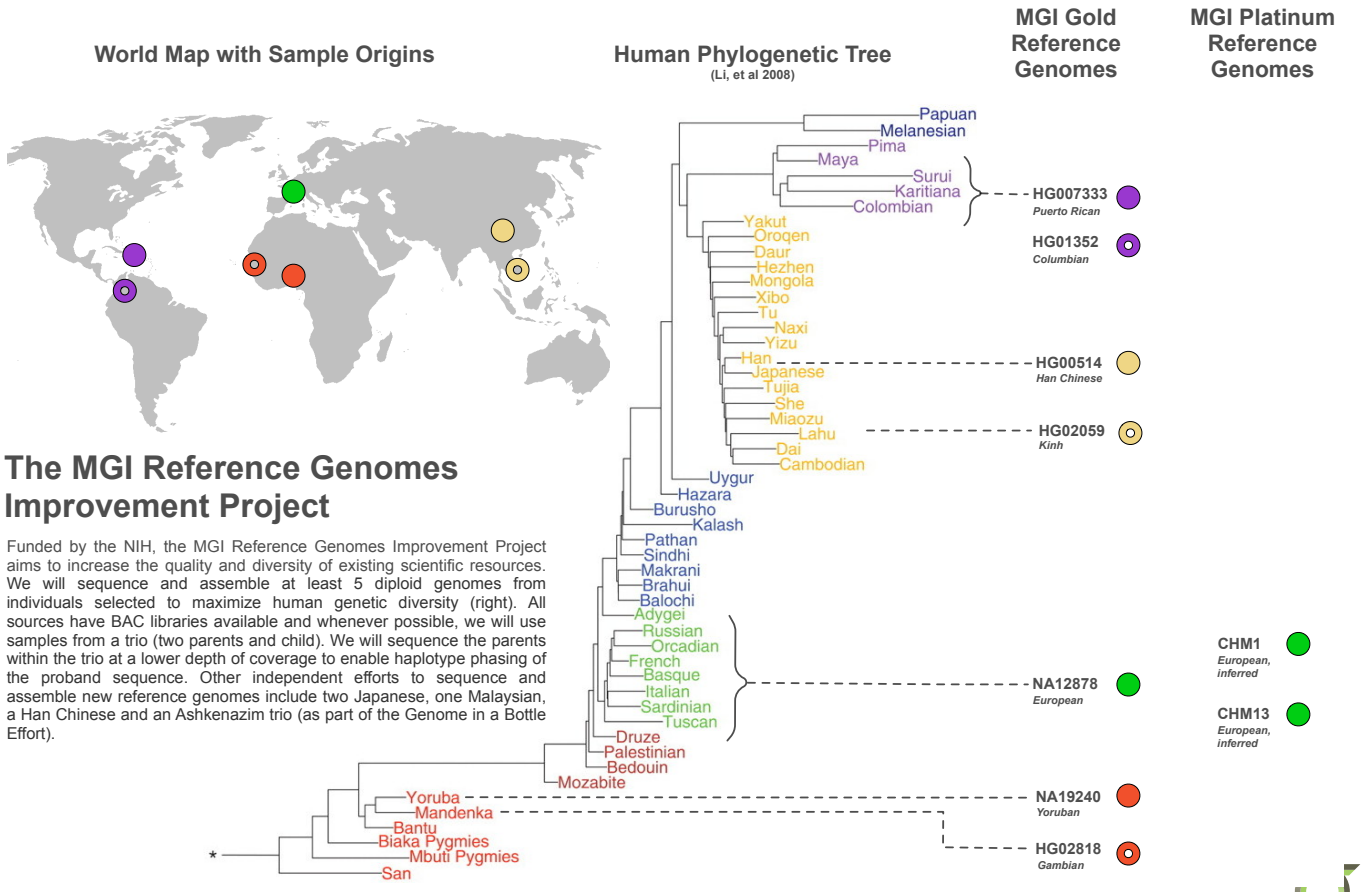
Library #3 30-80 kb

- Mean Subread Length 15,008 bp
- N50 Length 24,136 bp

The Human Reference is a Work in Progress!

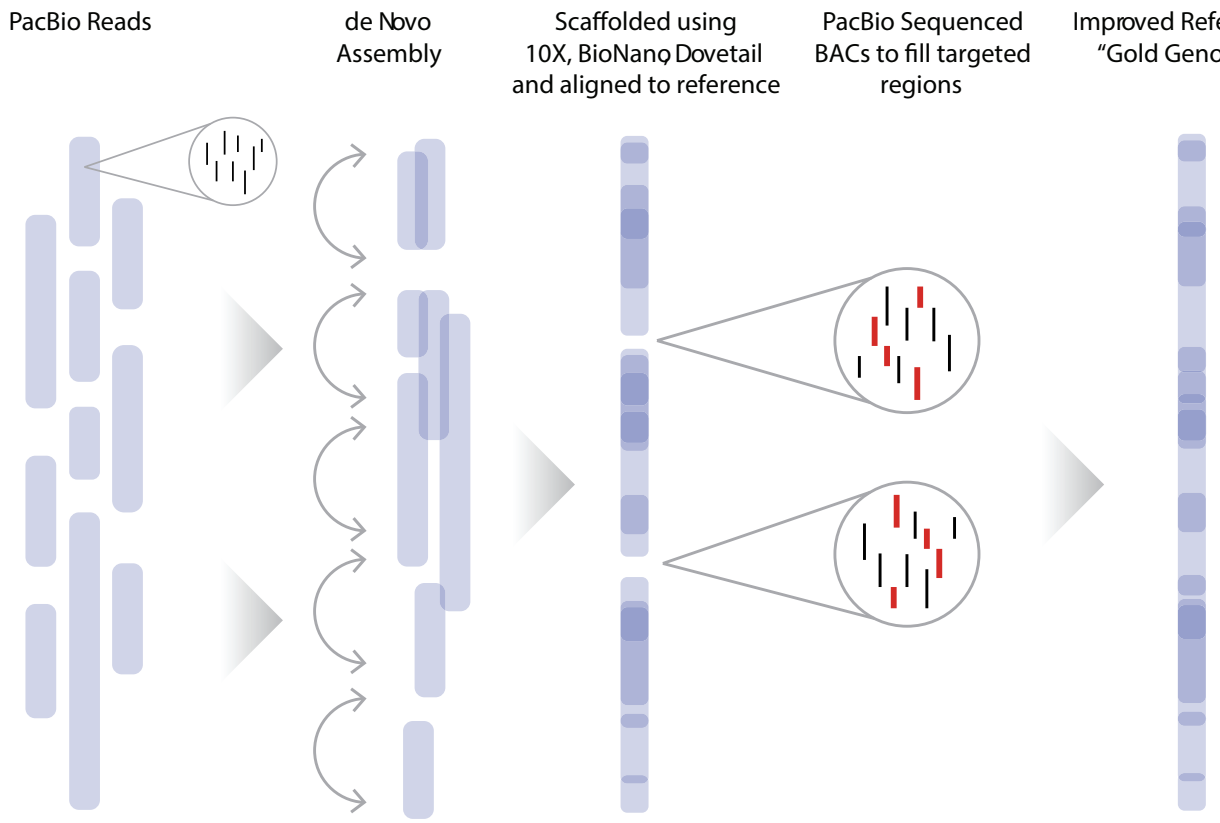
- The current reference – GRCh38 - is not optimal for some regions of the genome and/or some individuals/ancestries.
- GRCh38 is comprised of DNA from several individual humans.
- Allelic diversity and structural variation present major challenges when assembling a representative diploid genome.
- New technologies, methods, and resources since 2003 have allowed for substantial improvements in the reference genome.
- Additional high-quality reference sequences are needed to represent the full range of genetic diversity in humans

Improving the Human Reference Genome(s)



http://genome.wustl.edu/projects/detail/reference-genomes-improvement/

Sequencing Plan

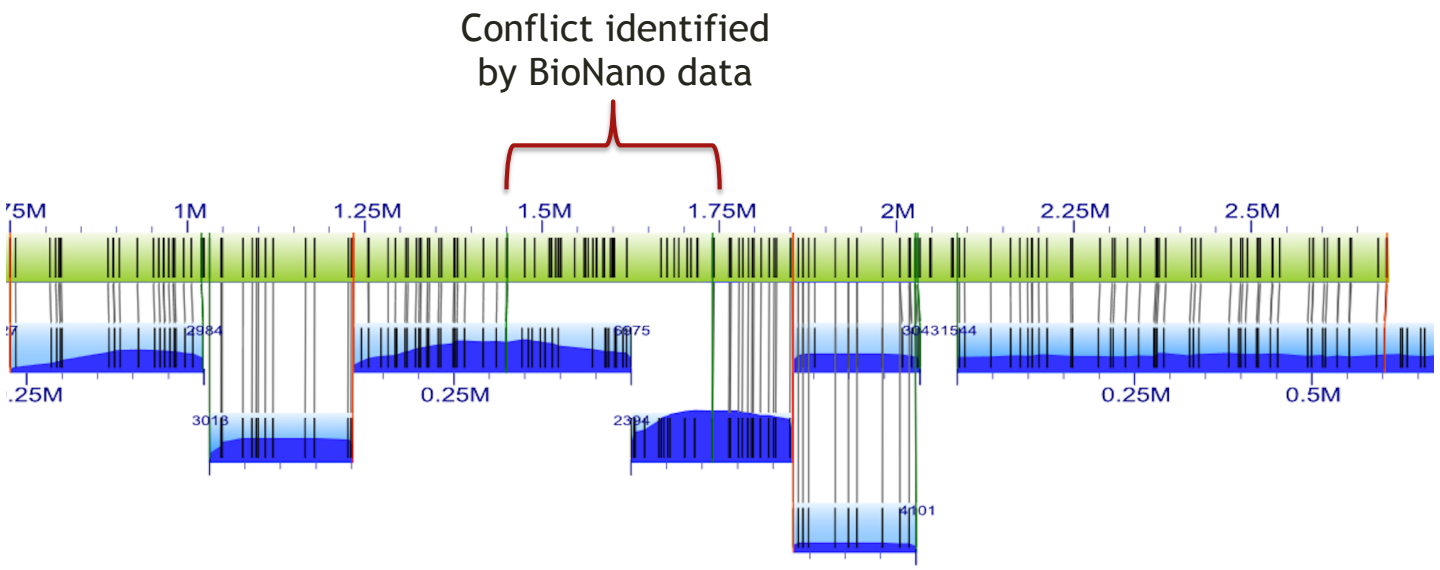


First Gold Genome - NA19240

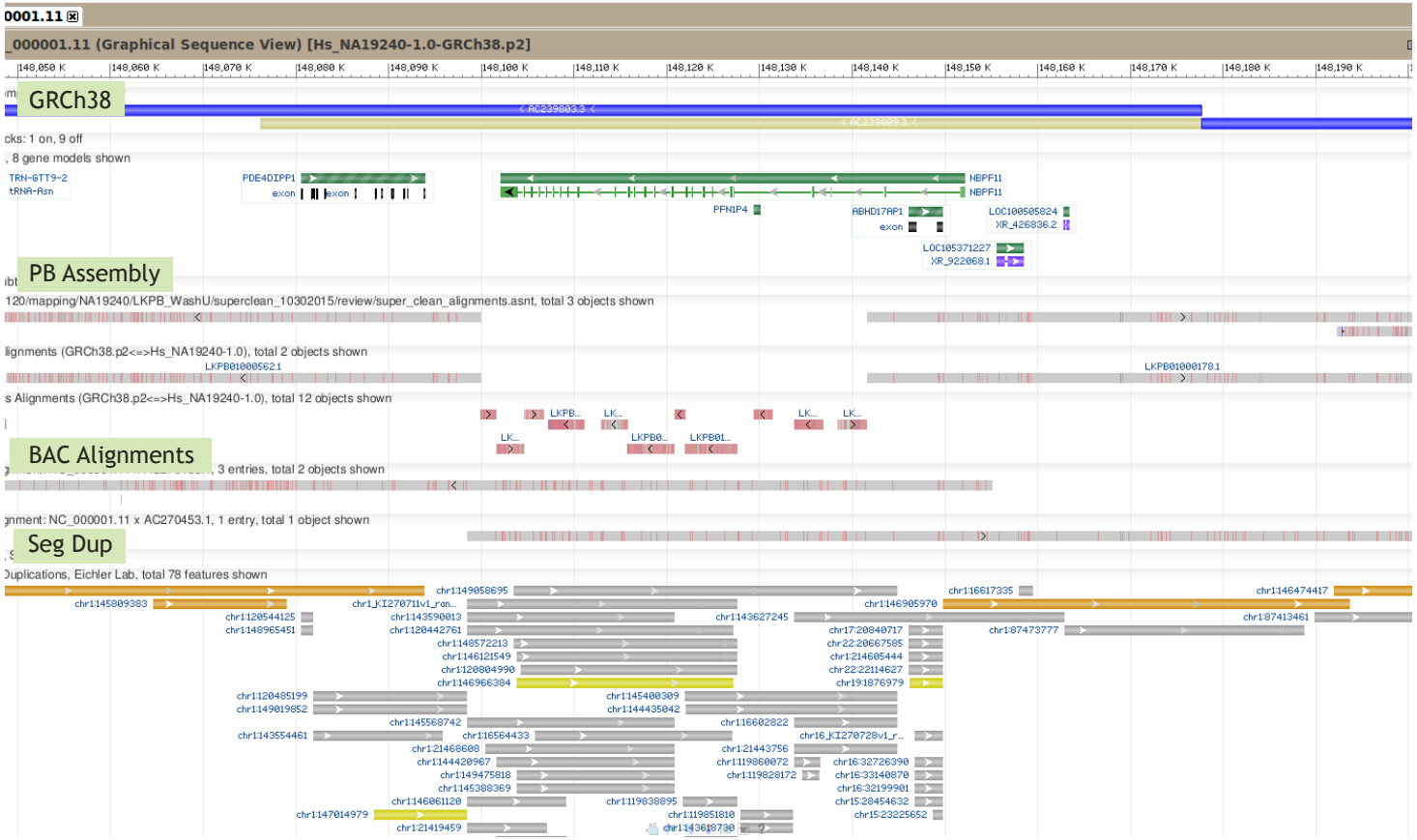
- NA19240 – Yoruban sample
- Generated >70X raw PacBio data

	Initial Assembly Stats
# Seq Contigs	3569
Max Contig Length	20,393,869 bp
Total Assembly Size	2,745,634,789 bp
N50	**6,003,115 bp**
N90	848,151 bp
N95	345,457 bp

Alignment of NA19240 to BioNano map



Finished BACs Resolve This Region

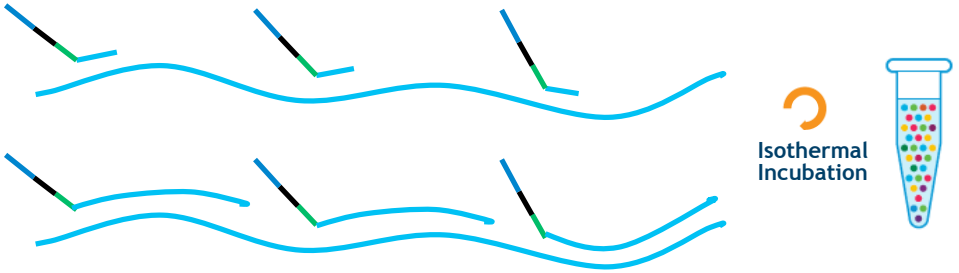


Platinum/Gold Genome Status

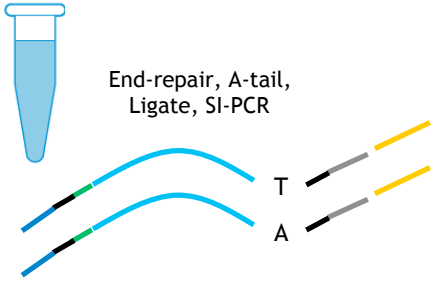
Data Source	Origin	Level of Coverage	Status
CHM1	NA	Platinum	Assembly Improvement
CHM13	NA	Platinum	Data Generation
NA19240	Yoruban	Gold	Analysis Underway
HG00733	Puerto Rican	Gold	Assembly Assessment
HG00514	Han Chinese	Gold	Assembly QC
NA12878	European	Gold	Assembly QC
HG01352	Columbian	Gold	Assembly Underway
HG02818	Gambian	Gold	Not Started
HG02059	Kinh Vietnamese	Gold	Not Started

10X Genomics: a novel twist

1. Molecular Barcoding in GEMs



2. Pool, Library Prep

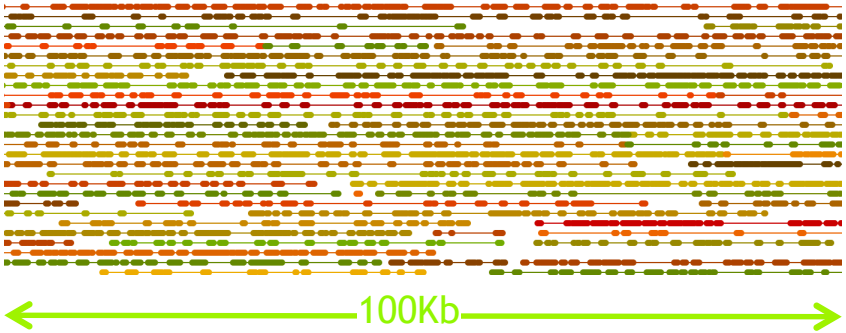


3. Sequence and Analyze

Final Library Construct

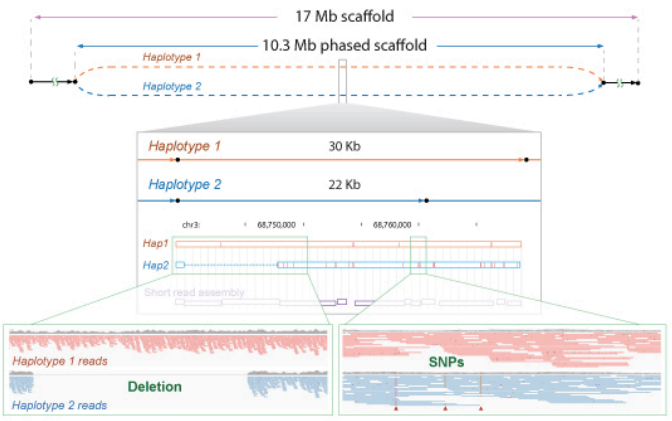
10X Genomics: Linked Reads

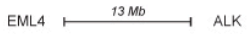
- Long-range information from short reads
  - Partition long input molecules into GEMs (***Gelbead-in-Emulsion***)
  - Gelbeads carry barcode oligos that are incorporated in sequencing library
  - Use barcodes to link short reads back to original long input molecules



- Resulting barcoded reads are called ***Linked-Reads***

10X Genomics: Power of Linked Reads

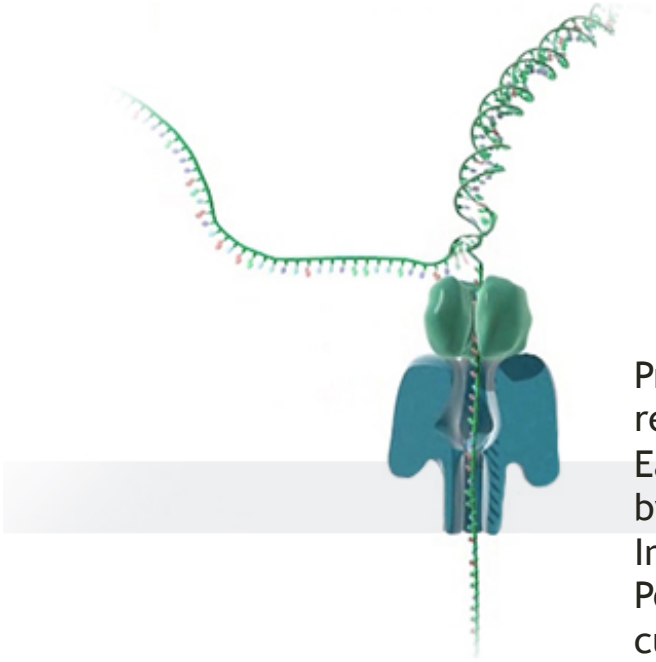




Multi-Megabase Diploid *De Novo* Assembly

Using the Supernova™ Assembler, reconstruct multi-megabase diploid assemblies, preserving phasing information for small variants, structural rearrangements and novel sequence without the need for a reference.

Oxford Nanopore Sequencing



Protein nanopore set in an electrically resistant polymer membrane
Each nanopore is controlled and measured by a matched ASIC (Application Specific Integrated Circuit)
Post-run data analysis compares pore current changes to a model of all possible multimers

- Variable read lengths
- Electrical current-based detection of multiplex nucleotides in pore
- Error rate is around 10-20% with newest pore/software

Nanopore Sequencing Devices

Genome-guided Immunotherapy Decision-making

Identifying “non-self” Neoantigen Landscapes

NGS to identify tumor-specific antigens

- Early work from the labs of Thierry Boon and Hans Schreiber (among others) suggested that tumor specific mutations can sometimes function as tumor specific antigens

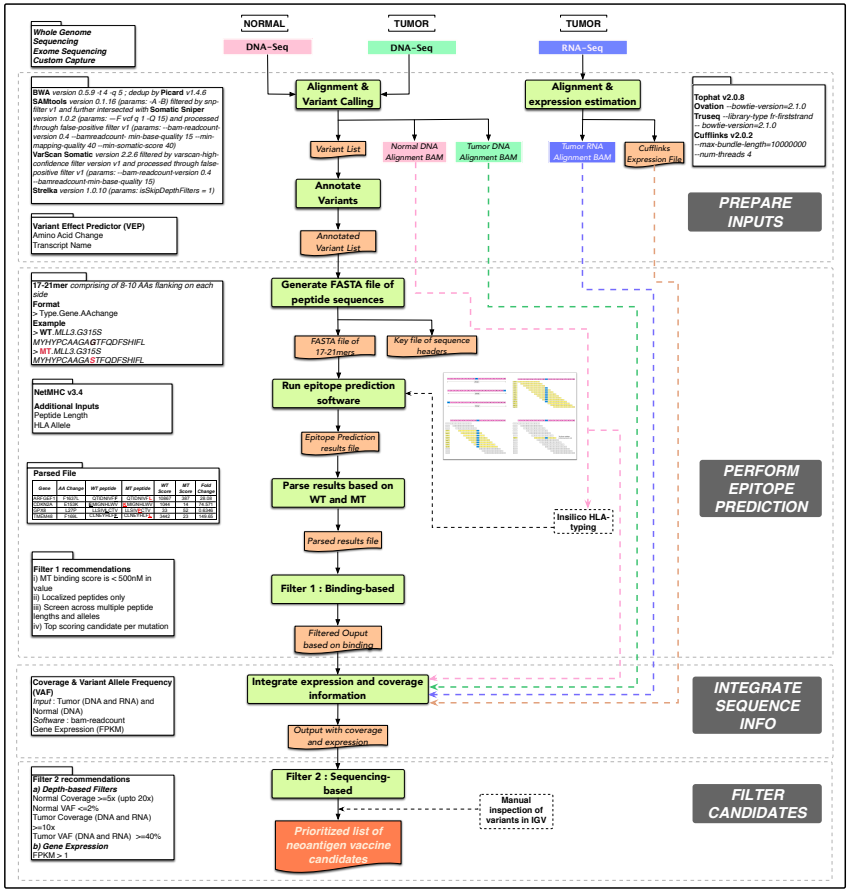
- James Allison and Bert Vogelstein predicted that many/most tumors should express mutational antigens based on their genomic repertoire of coding variants, and these might be the ideal tumor-specific targets for cancer immunotherapy

- In the past, identifying the tumor mutation landscape and its most immunogenic peptides has been hampered by technical obstacles, most of which have now been overcome by NGS and bioinformatic approaches to epitope prediction

Genome-guided Immunotherapy

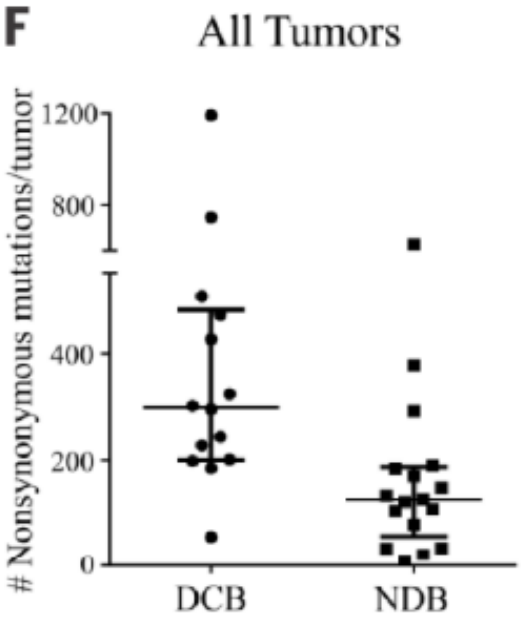
- In using genomic data to predict neoantigen load for a specific tumor, we utilize:
  - massively parallel sequencing and analysis to compare cancer and normal exomes and identify cancer-unique peptides
  - the HLA haplotypes of the individual patient
  - RNA sequencing data from the cancer cells to identify genes that are mutated **and** expressed
- These input data are considered by algorithms that model the binding of peptides to the MHC and calculate binding energies, producing a list of tumor specific mutated antigens (TSMAs) or neoantigens
- This information can describe the cancer's neoantigen load

pVac-Seq Pipeline: Open Source Neoantigen Prediction

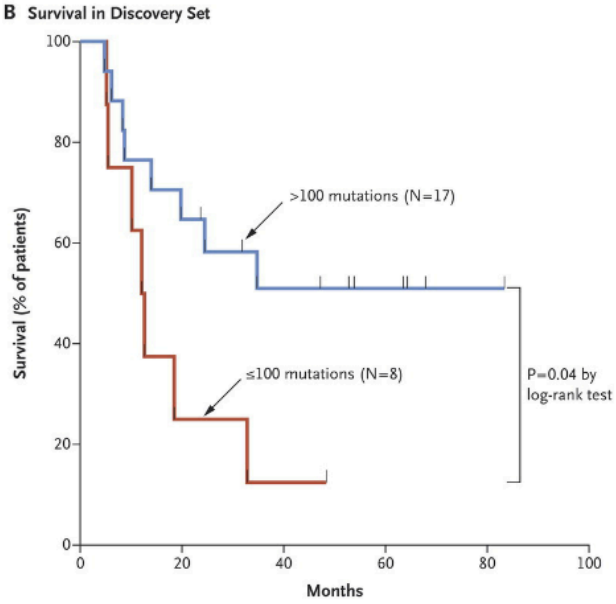


Hundal et al., Genome Medicine 2016

Mutation/Neoantigen Load and Checkpoint Blockade Response Potential



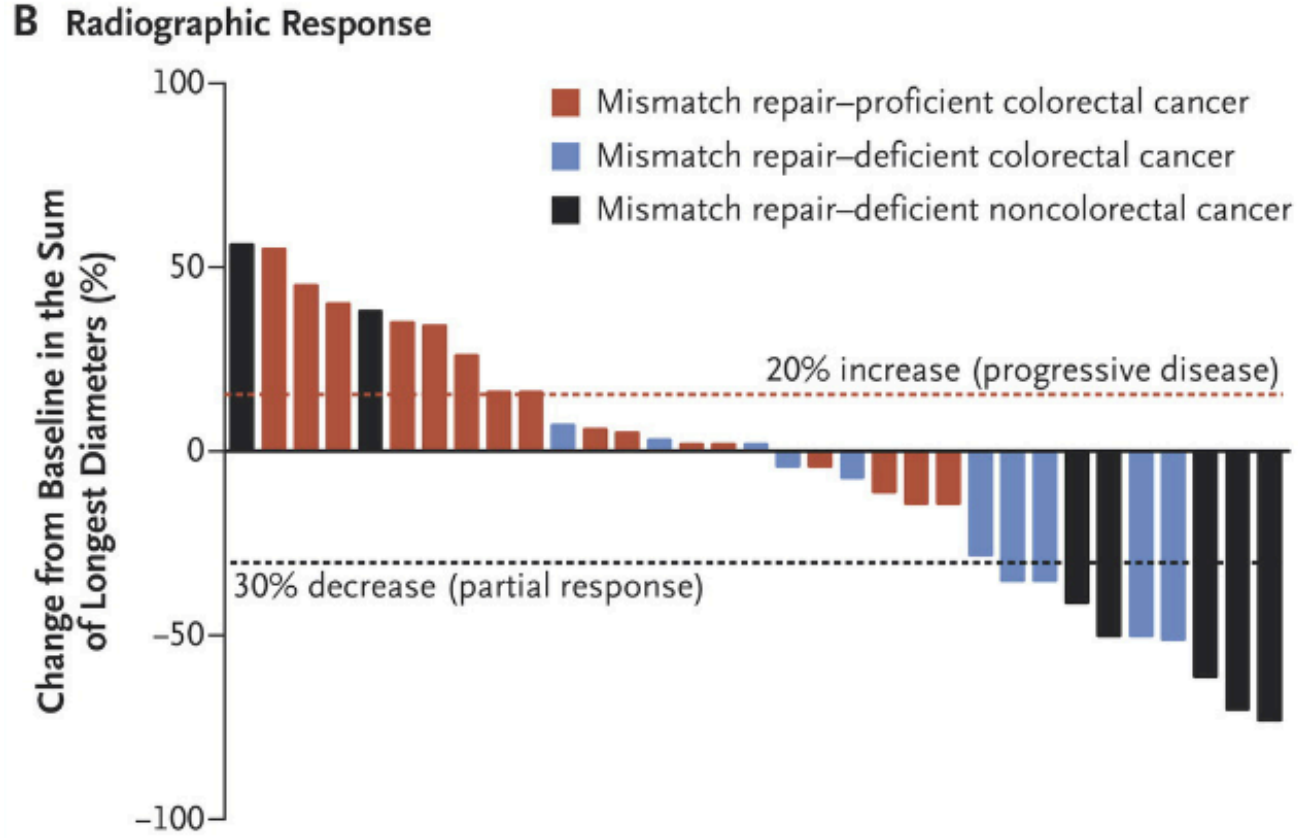
Rivzi et al., Science 2015



Snyder et al., NEJM 2014

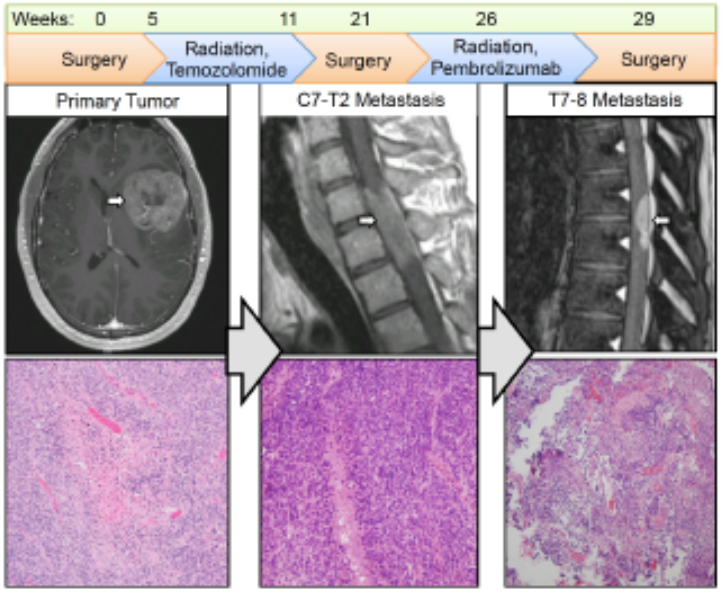
Tumors with a high mutational (neoantigen) load tend to respond to checkpoint blockade immunotherapy

Germline DNA Repair Defects and Checkpoint Blockade Response Potential



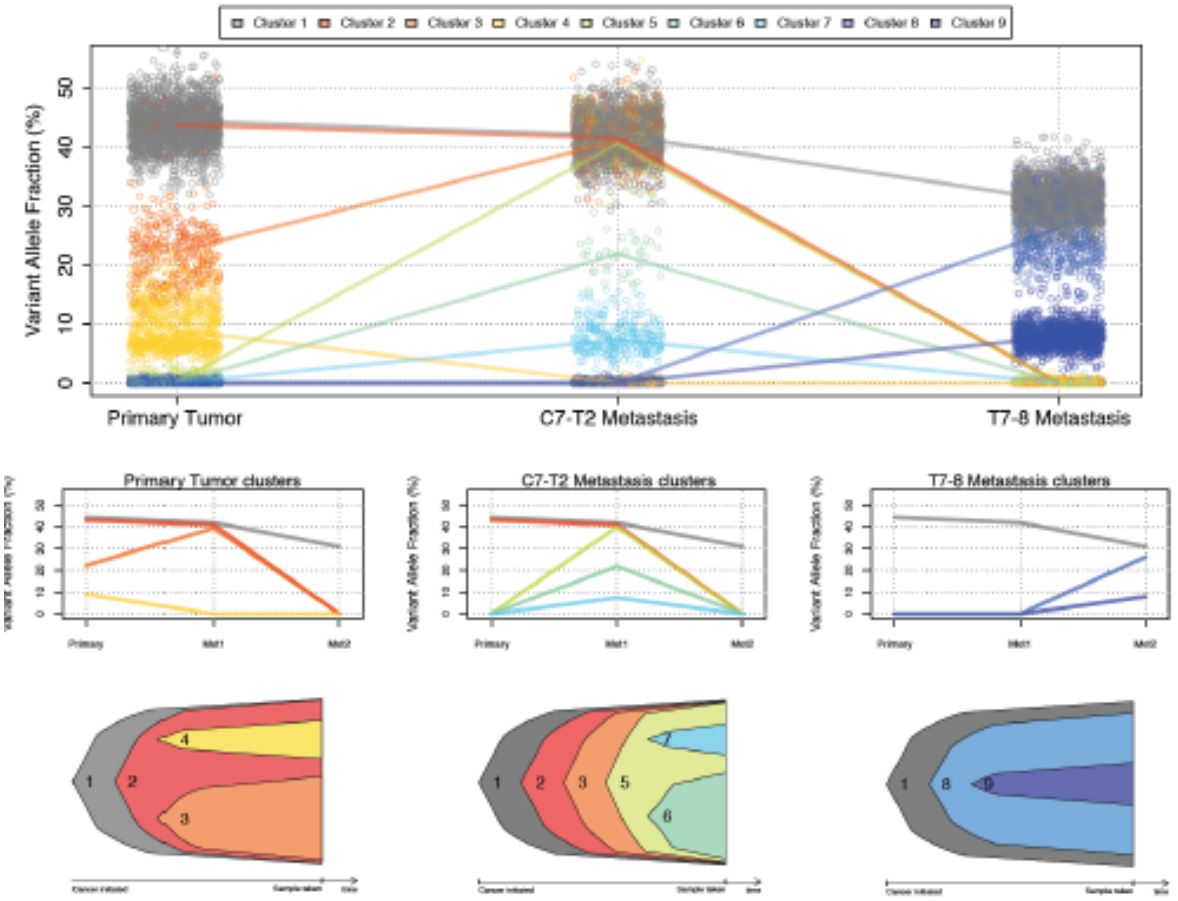
Dung et al., NEJM 2015

GBM27: Rapid Progression in CNS

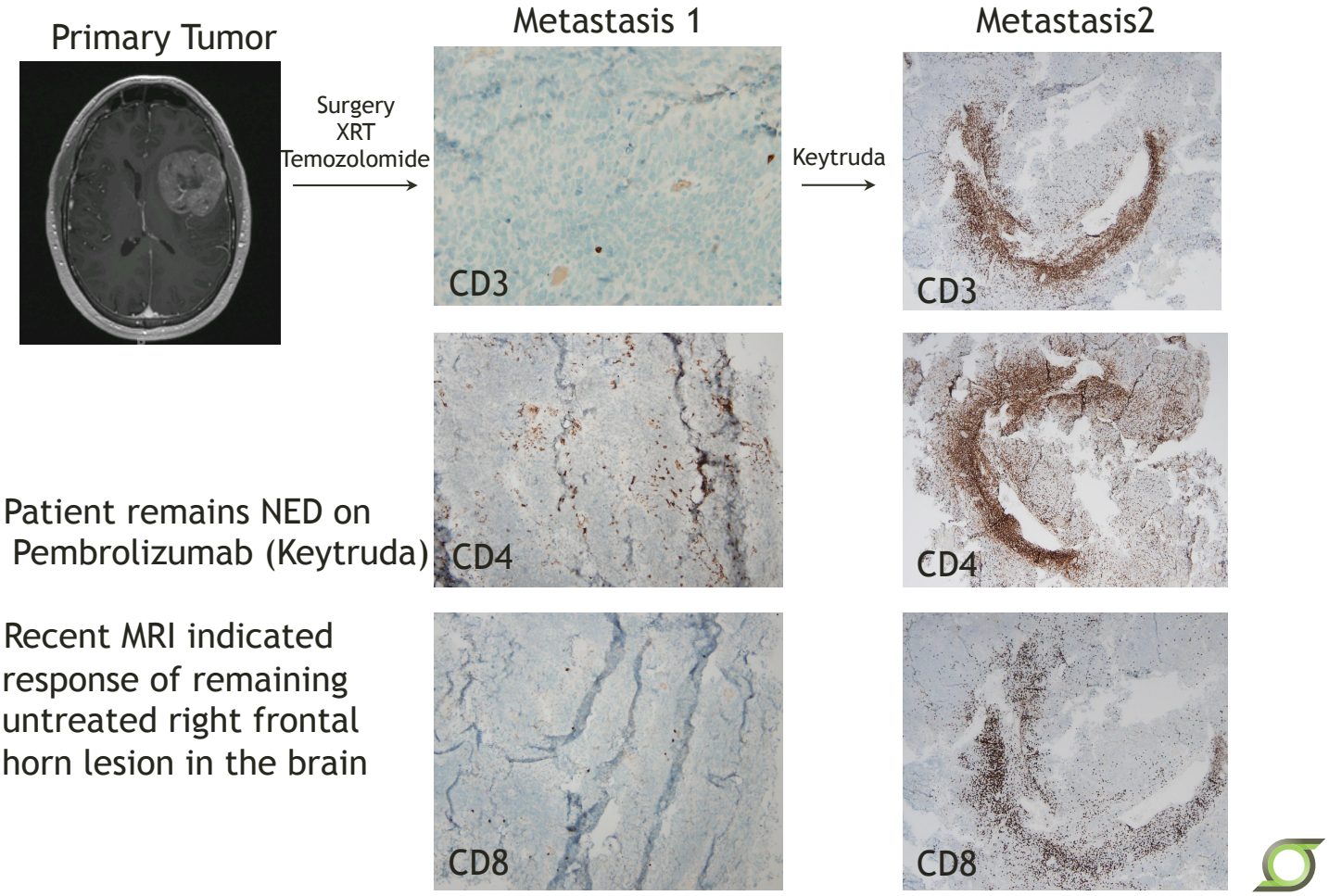


- Male patient, early 30's, prior history of colon polyps
- GBM removed by craniotomy 8 months ago
- Post surgery temozolomide (TMZ) therapy
- Spinal metastasis resected 3 months ago, FMI test indicated high mutation load/ pol E mutated germline status : treatment with Pebrolizumab
- Second spinal metastasis identified upon complications, removed 2 months ago
- All tumors studied by high coverage exome sequencing compared to PBMC normal, and by IHC

GBM27: Clonal Evolution



GBM27: Evolving Immune Response



- Patient remains NED on Pembrolizumab (Keytruda)
- Recent MRI indicated response of remaining untreated right frontal horn lesion in the brain

Acknowledgements

McDonnell Genome Institute
Malachi Griffith, PhD
Obi Griffith, PhD
Vincent Magrini, PhD
Sean McGrath
Ryan Demeter
Tina Graves-Lindsay
Bob Fulton
Chris Markovic

Richard K. Wilson, PhD

WUSM/Siteman Cancer Center
Gavin Dunn, MD, PhD
Joshua Rubin, MD, PhD
Robert Schreiber, PhD

Thanks also to:
Aaron Quinlan (Utah)
Gabor Marth (Utah)
Michael Zody (NYGC)
Evan Eichler (U Wash)
Pui Yan Kwok (UCSF)
Valerie Schneider (NCBI)
Jason Chin (Pac Bio)
Adam Phillippy (NHGRI)
Sergey Koren (NHGRI)

Our patients and their families