THE GENOME INSTITUTE at Washington University

Next-Generation Sequencing Technologies

Elaine R. Mardis, Ph.D.
Co-director, The Genome Institute
Robert E. and Louise F. Dunn
Distinguished Professor of Medicine

NHGRI Current Topics in Genome Analysis 2014

JOHNS HOPKINS MEDICINE
CONTINUING MEDICAL EDUCATION

Current Topics in Genome Analysis 2014

Elaine Mardis

No Relevant Financial Relationships with Commercial Interests

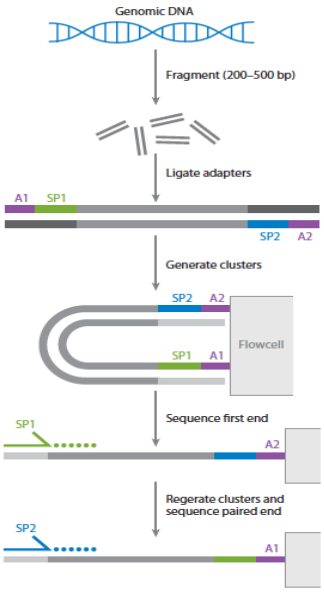
Next-generation Sequencer basics

How massively parallel sequencing works

Next-generation DNA sequencing instruments

- All NGS 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 and Amplification



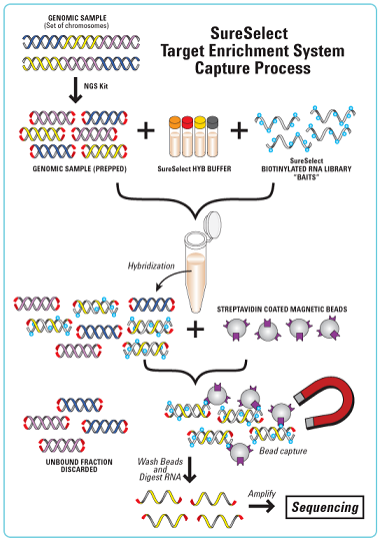
- Shear high molecular weight DNA with sonication
- Polish ends
- Ligate synthetic DNA adapters (PCR*)
- Produce size fractions (PCR*)
- Quantitate
- Amplify library fragments on flow cell surface (PCR*)
- Denature clusters to single-stranded
- Hybridize sequencing primer to linearized ss cluster DNAs
- Proceed to sequencing or hybrid capture

PCR-related Problems in NGS

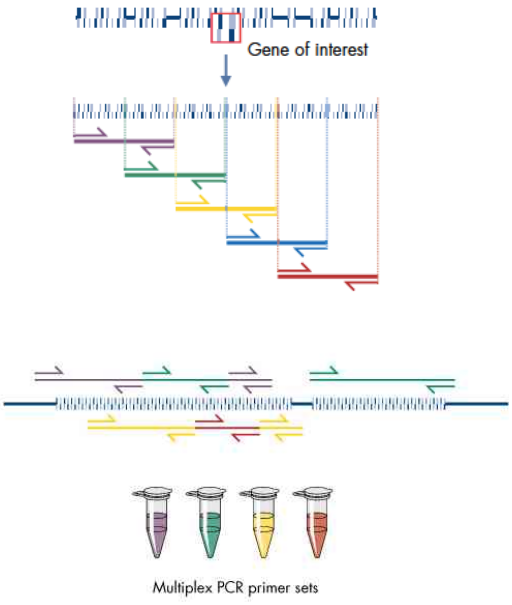
- PCR is an effective vehicle for amplifying DNA, however…
- In NGS library construction, PCR can introduce preferential amplification ("jackpotting") of certain fragments
  - Duplicate reads with exact start/stop alignments
  - Need to "de-duplicate" after alignment and keep only one pair
  - Low input 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.

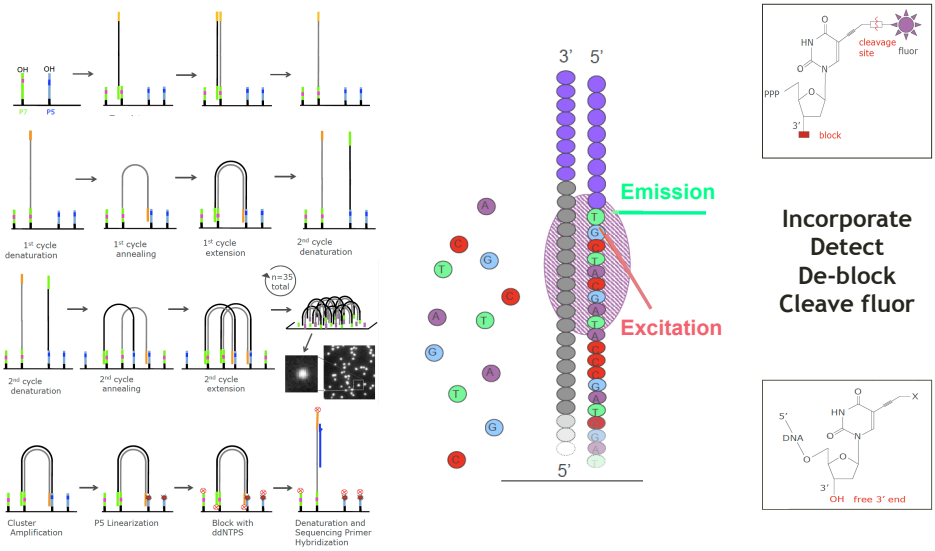


Multiplex PCR Amplification of Targets



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

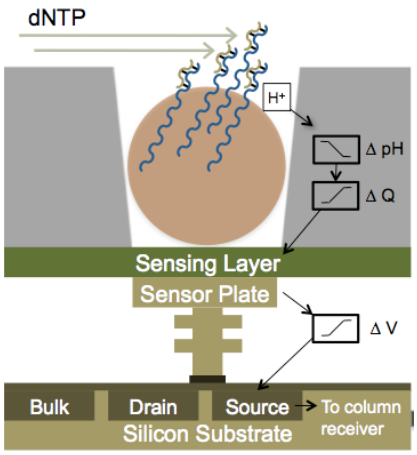


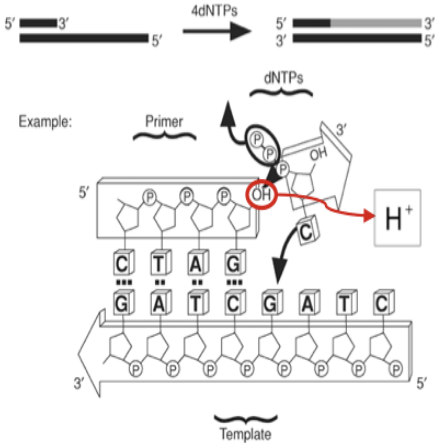
Platforms: Illumina

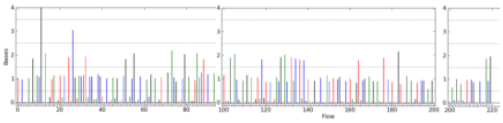
	MiSeq	NextSeq 500		HiSeq 2500		HiSeq X*
Key applications	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.		Production-scale genome, exome, transcriptome sequencing, and more.		Population-scale human whole-genome sequencing.
Run mode	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

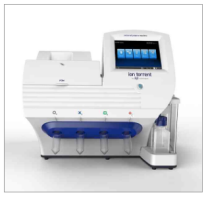
ION Torrent-pH Sensing of Base Incorporation







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

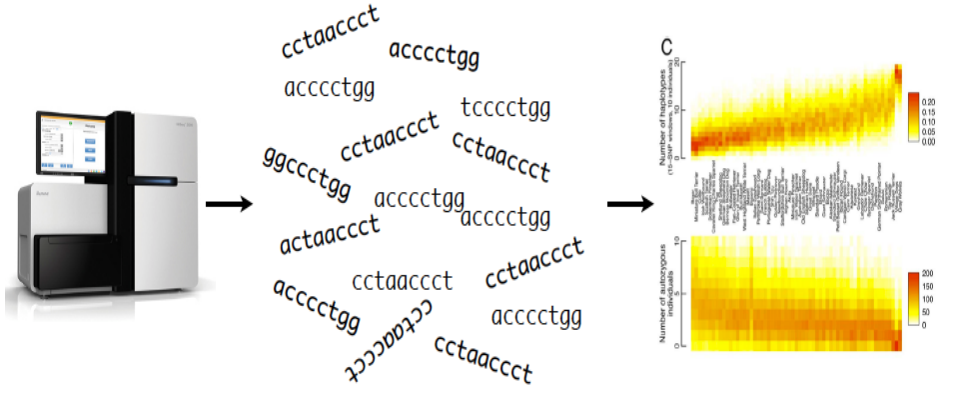
- Two human exomes (Proton 1 chip) or one genome (@20X-Proton 2 chip) per run
- Ion One Touch or Ion Chef preparatory modules
- 2-4 hour/run
- ~200 bp average read length
- Proton 1 produces 60-80 Mreads ≥50 bp

- 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

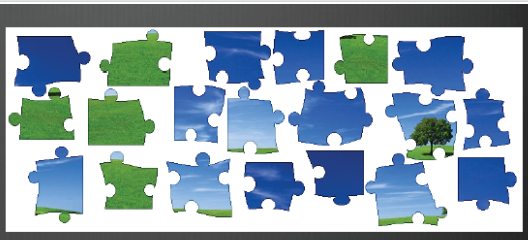
The Goal?



Sequence data alignment is the crucial first step!

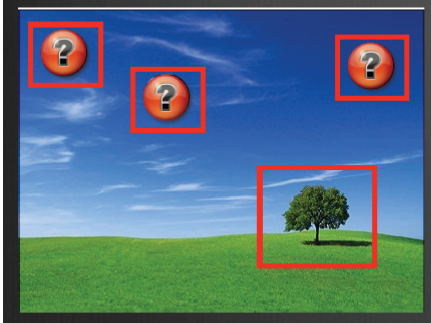
Short Read Alignment…

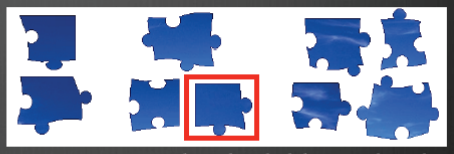
Is like a jigsaw puzzle…



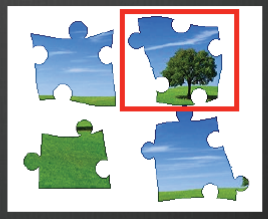
…where they give you the cover on the box

Some pieces are easier to place than others…



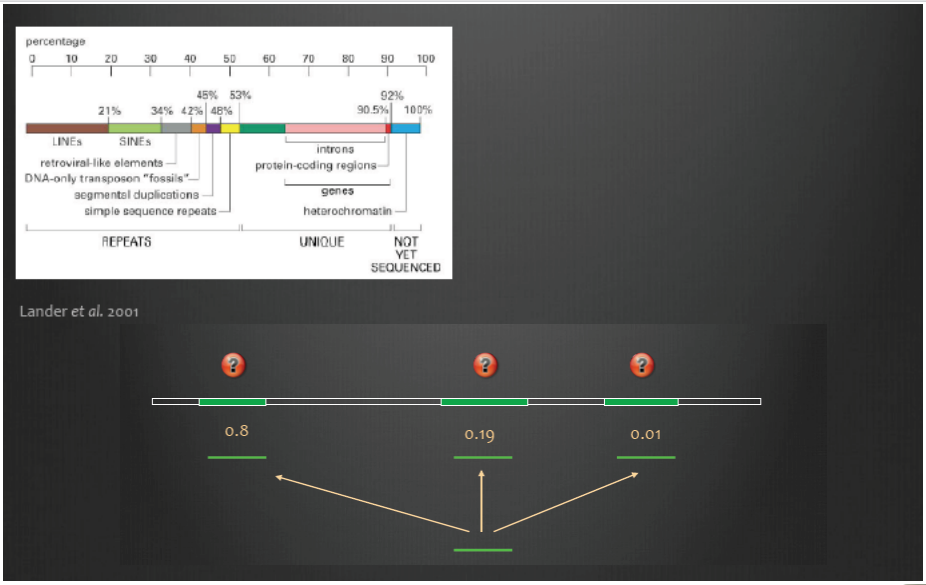


pieces that look like each other…



…pieces with unique features

Repetitive Sequences Result in Multiple Read Alignments



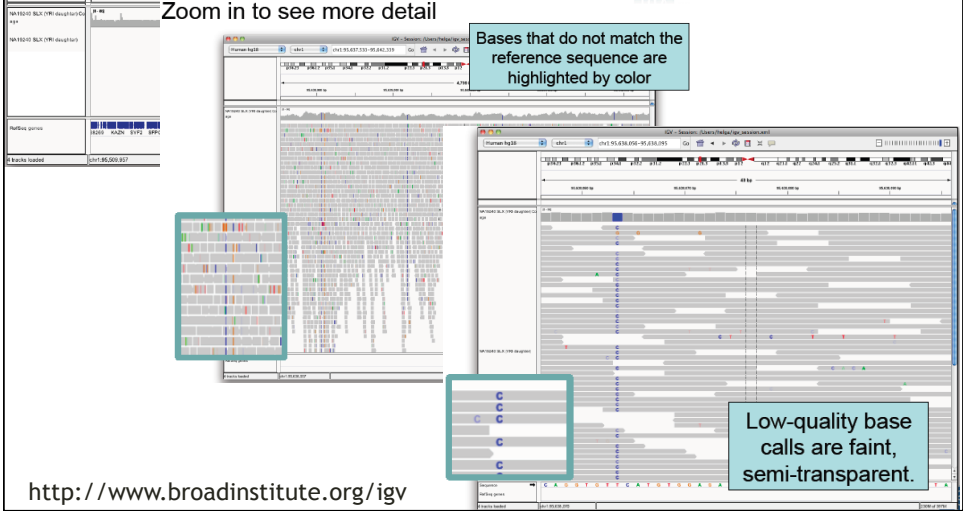
Reads are Aligned, Now What?

- 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
  - Integrated Genome Viewer
  - RefCov and others
- Analyze the data

Integrated Genomics Viewer (IGV)

Whole chromosome view

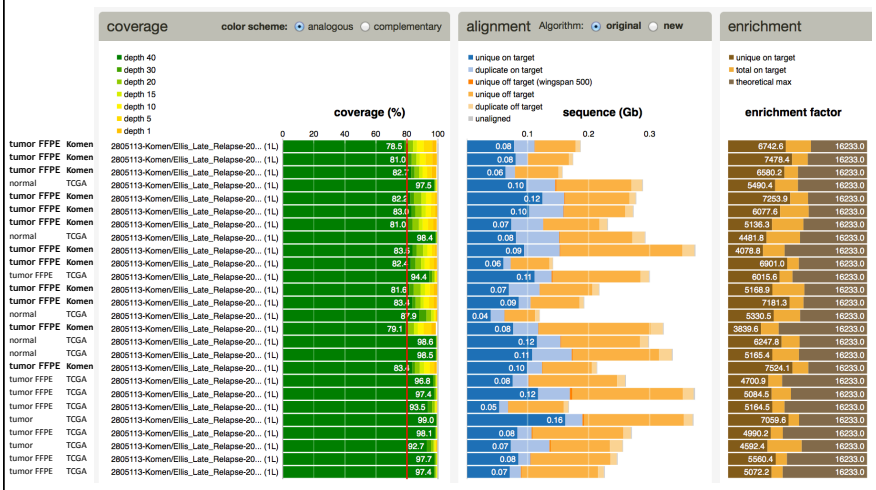
Zoom in to see more detail



http://www.broadinstitute.org/igv

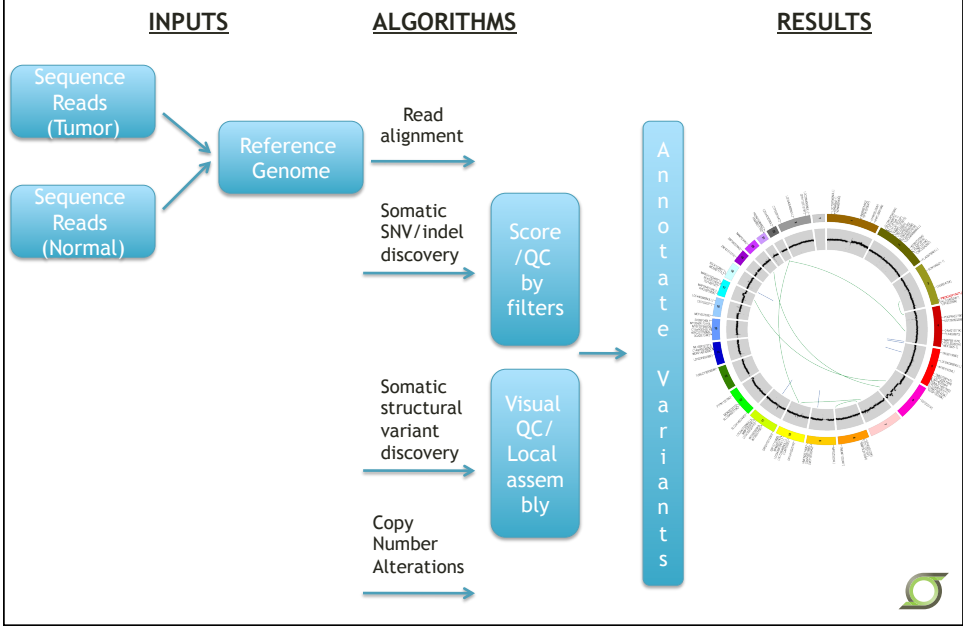
IGV: Somatic Single Nucleotide Variant QC

RefCov: Coverage Depth and Breadth from Hybrid Capture



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

Somatic Variant Discovery Pipeline



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

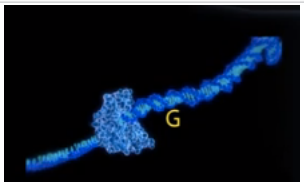
Third Generation Sequencers

Variations on a theme

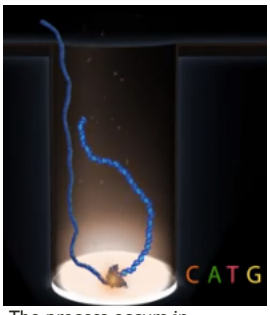
Real Time Sequencing of Single DNA Molecules

DNA:polymerase complex is immobilized at the ZMW bottom

Fluorescent nucleotides are introduced to the ZMW



A light pulse is produced as each fluorescent base is held in the polymerase active site. The phosphate is cleaved during incorporation, releasing the fluorophore.



The process occurs in parallel in all the loaded ZMWs

Pacific Biosciences RS

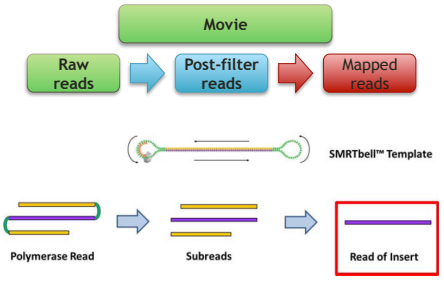
Sample Prep

- Shearing (Covaris/Hydroshear)
- Polish ends
- SMRTbell™ ligation
- Sequencing primer annealing

Library/Polymerase Complex

- DNA polymerase binding
- Load library/polymerase complex onto SMRT cell

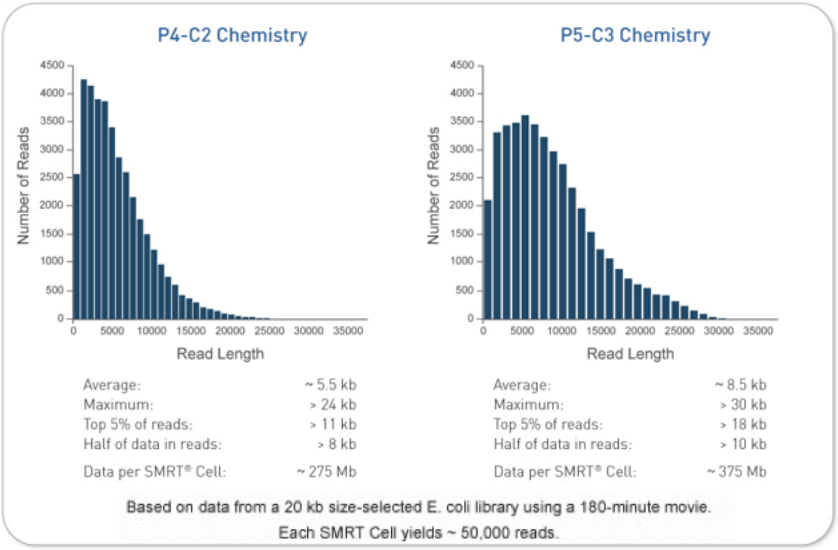
Sequencing



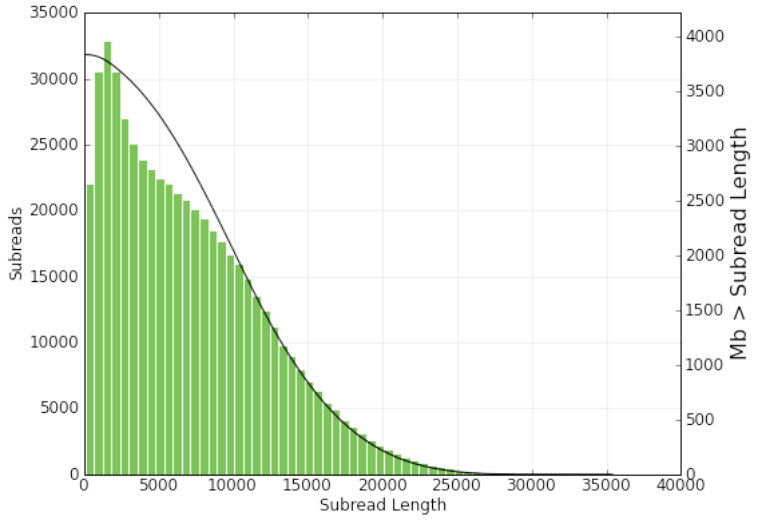
PacBio: 20 kb Library Preparation and Sequencing

- Covaris g-Tube 20 kb shear
- Pacific Biosciences 20 kb library prep
- Sage Science BluePippin size fractionation
  - 8 – 50 kb
  - 15 – 50 kb
- Pacific Biosciences RSII sequencing
  - Polymerase: P5
  - Sequencing chemistry: C3
  - MagBead loading
  - Per SMRT Cell
    - 180 minute collection time
    - "Stage start"

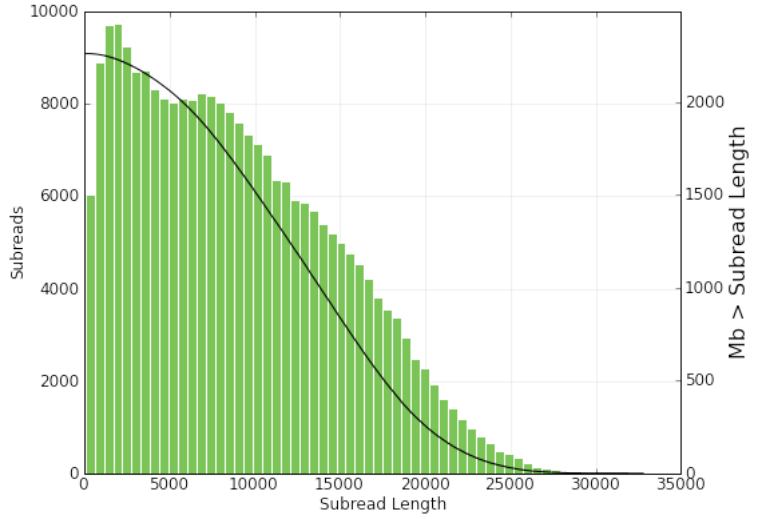
PacBio: Improvements in Polymerase and Chemistry



Chicken 20 kb - BluePippin 8-50 kb



Chicken 20 kb - BluePippin 15-50 kb



Human BAC/fosmid clones sequenced by PacBio platform

clone name	Clone Size (bp)	library size	SMRT cell	Number of mapped Subreads	Error Corrected Coverage Post-Vector/E. Coli Screened	Number of contigs after de novo assembly
ABC11-47241000C4	39755	10 kb	1	50384	121.1X	1
ABC11-47399300K22	38,934	10 kb	1	56599	311X	3
ABC11-49599500A20	41423	10 kb	1	63180	162X	5
ABC12-46674300M3	39380	10 kb	1	57265	157X	2
ABC12-47036800M8	40,000***	10 kb	1	59535	317.2X	1
ABC14-50418300F21	40,000***	10 kb	1	66469	140X	1
ABC7-4028360016	31663	10 kb	1	56042	116X	1
ABC7-42060100J1	36886	10 kb	1	42220	109.3X	1
ABC9-41286700F24	40,000***	10 kb	1	53298	337X	7
ABC9-43817800N19	40,000***	10 kb	1	33745	151.4X	1
ABC9-44010900K17	42398	10 kb	1	47414	117.3X	1
CH17-176P24	207,445	10 kb	1	78003	41X	1*
CH17-194E17	170,000***	10 kb	1	24274	78.6X	1**
CH17-199I12	176,000***	10 kb	1	55588	60X	1
CH17-275L14	223691	10 kb	1	84211	77X	2
CH17-345B22	230,000***	10 kb	1	39245	108.8X	1
CH17-390D12	177,000***	10 kb	1	32540	41X	2
CH17-442P13	150,000***	10 kb	1	56444	32X	8
CH17-90K13	224074	10 kb	1	51909	53X	3
RP11-84A7	189483	10 kb	1	45524	44X	4
WI2-2025H20	37272	10 kb	1	76365	41X	1
WI2-3087P5	39143	10 kb	1	27716	88X	1

*** Estimated clone size based on restriction enzyme digests and/or type of clone (fosmid/BAC)
** This assembly contains 1 human contig plus contaminated bacterial contigs
* This assembly produces 1 short contig due to collapsed repeat within the contig.

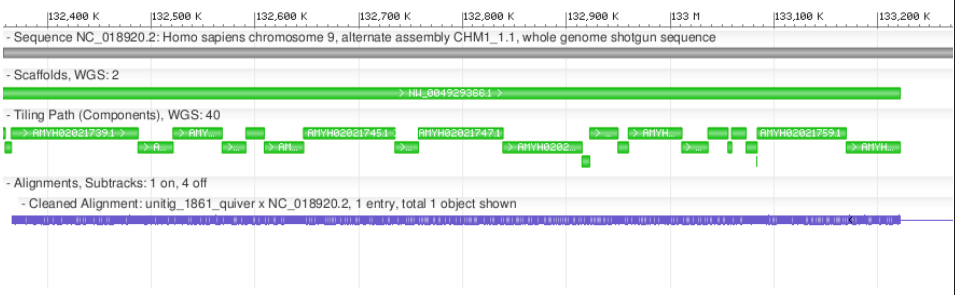
Comparative assemblies with Illumina or PacBio

clone name	Illumina assembly coverage	PacBio PreAssembled read Coverage Post vector/ E.coli Screened	Illumina total contig #	PacBio total contig #	Illumina total contig bases (bp)	PacBio total contig bases (bp)	Illumina N50 contig bases	PacBio N50 contig bases (bp)	% GC
H_GD-281P19	64X	83.0X	93	1	198861	217805	13790	217805	46%
H_GD-280I20	73X	119.7X	20	1	198255	197966	17306	197966	34%
H_GD-358O03	70X	108.3X	66	1	172074	196503	12435	196503	43%
H_GD-433K21	70X	101.8X	90	1	220679	222522	5371	222522	35%
H_GD-196M11	65X	82.1X	33	7	131252	197654	11085	26921	39%
H_GD-219D13	74X	119.5X	25	2	107454	147058	6761	122737	42%
H_GD-389L19	73X	97.9X	20	8	137328	239670	13262	47406	42%
H_GD-266C19	76X	106.3X	19	1	194736	194593	17995	194593	36%

Pac Bio: Long reads improve the Human Reference Genome sequence

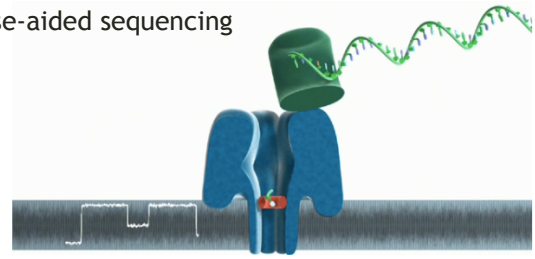
- Since the HRG finished sequence was announced and published in 2004, our group has continued to improve the reference
  - Addition of new content, including novel content from other human genomes
  - Improvement of previously poorly finished regions
  - Finishing of regions between segmental duplications
- Our new approach to HRG improvement will include sequencing haploid human genomes (hydatidiform mole) with Pacific Biosciences long read sequencing
  - One such genome (CHM1) already has 60X coverage from PacBio
  - An assembly of CHM1 is now being compared to the HRG (grCH38)

Alignment of CHM1 PacBio assembly to CHM1_1.1

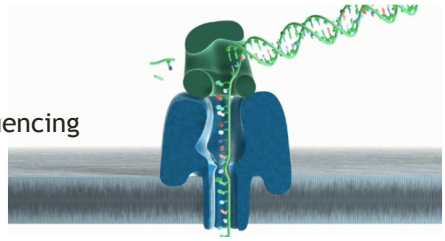


Nanopore Sequencing

Exonuclease-aided sequencing

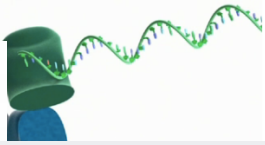


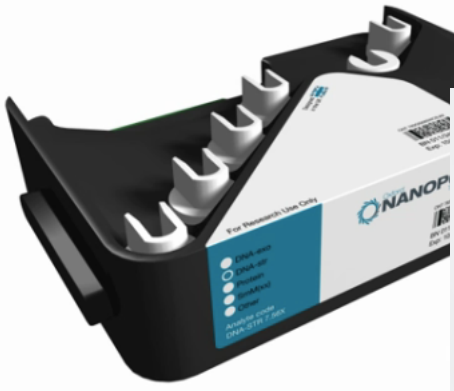
Pore translocation sequencing



Oxford Nanopore Sequencing

Exonuclease-aided sequencing



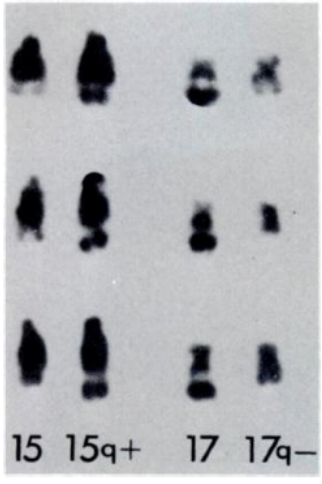


- Variable read lengths
- Electrical current-based detection of triplet nucleotides in pore

Translating the Cancer Genome

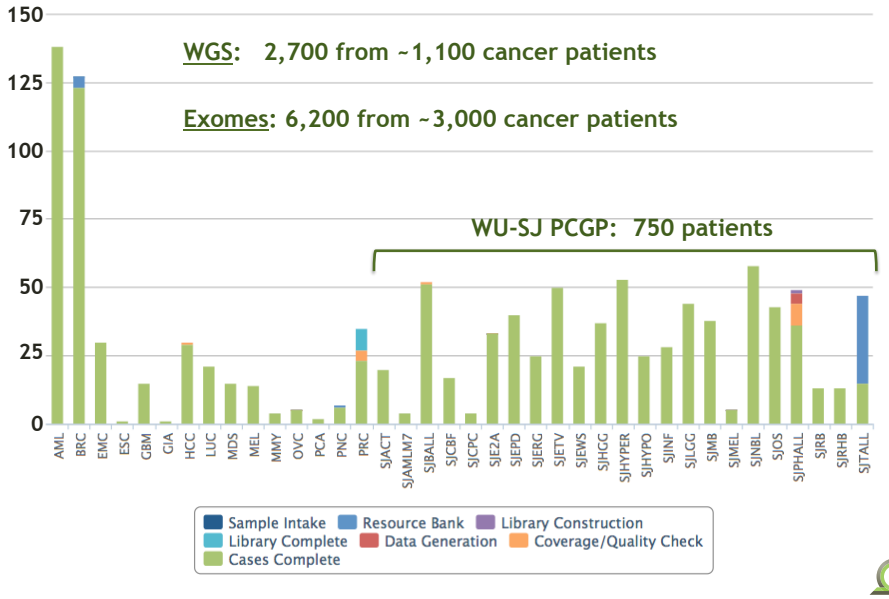
Therapeutic Options via NGS and analysis

Cancer is a Disease of the Genome

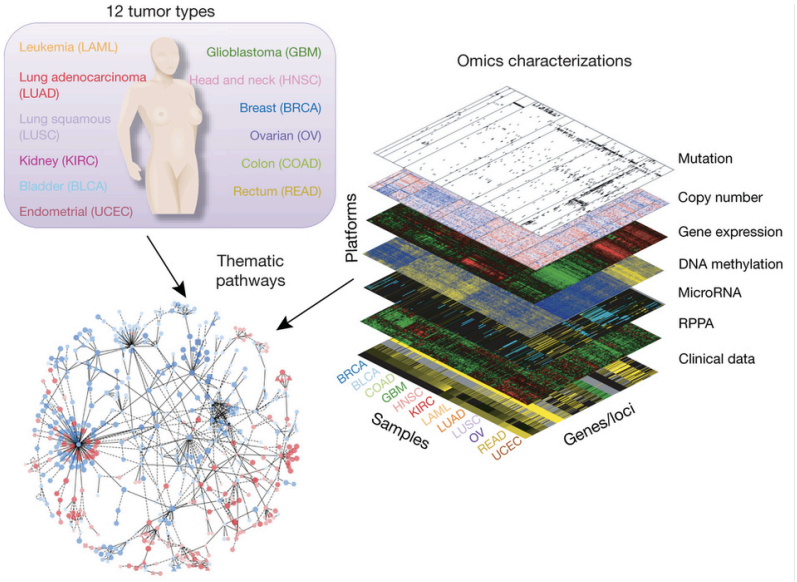


In the early 1970's, Janet Rowley's microscopy studies of leukemia cell chromosomes suggested that specific alterations led to cancer, laying the foundation for cancer genomics.

TGI: Cancer Cases by WGS (March 2014)

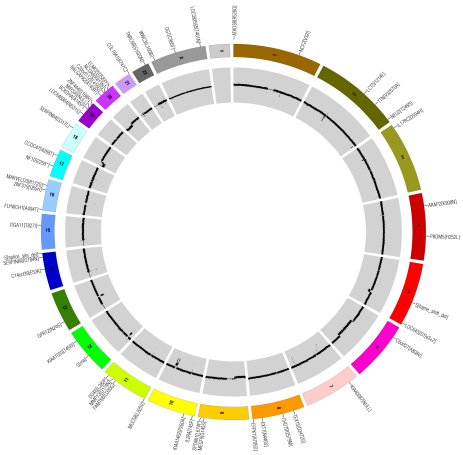


Pan-Cancer Analyses from TCGA



Nature Genetics 45: 1113-1120 (2013)

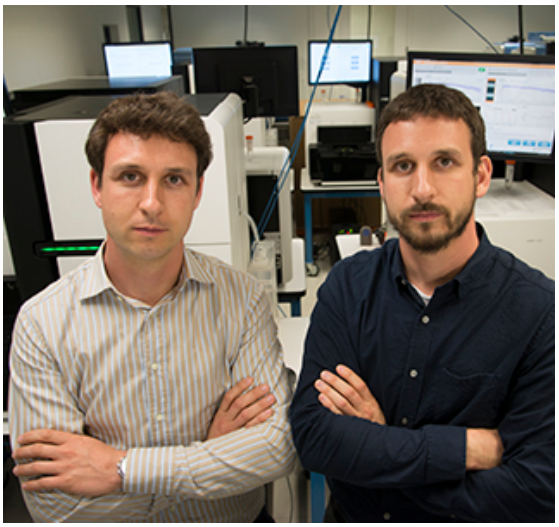
Comprehensive Cancer Genomics



Integrated WGS/Exome/RNA-Seq

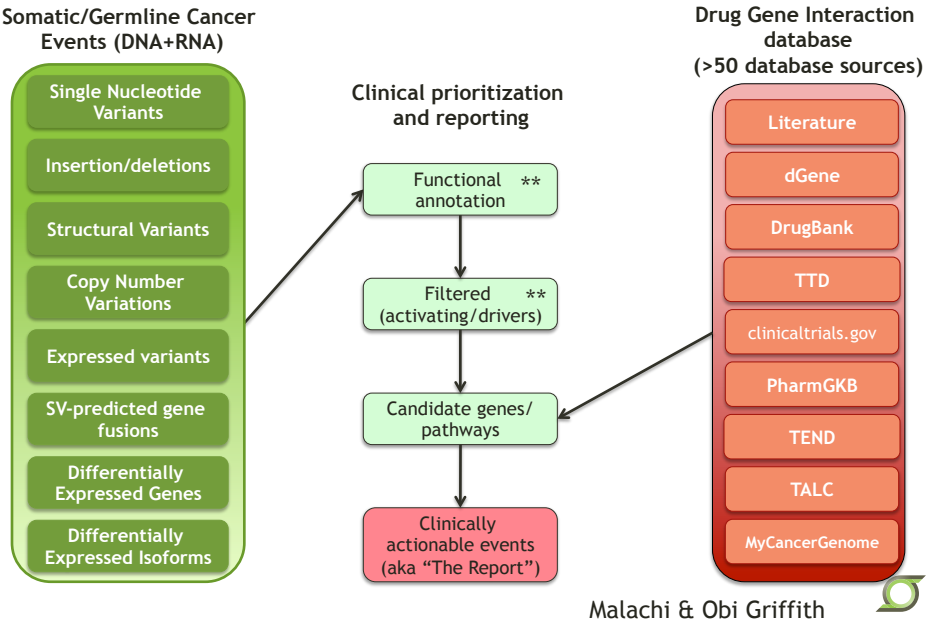
- **WGS** analysis yields:
  - SNVs (single nucleotide variants)
  - CNVs (amplification/deletion)
  - SVs (translocations, inversions)
  - Indels (focused insertions/ deletions)
- **Exome:** validates WGS discoveries, integrated coverage depth allows clonality analysis
- **RNA-Seq:** over-expression metrics, expressed SNVs, gene fusions
- **Clinical Action:** identifying druggable targets

Linking Somatic Variants to Therapies



Obi Griffith, Ph.D. and Malachi Griffith, Ph.D.

Therapeutic Interpretation of Variants

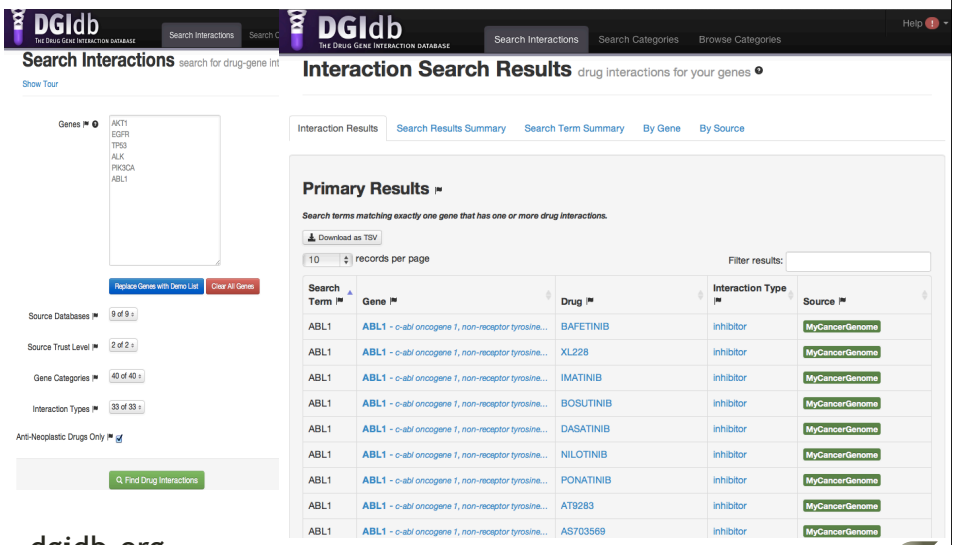


Malachi & Obi Griffith

DoCM: A Database of Canonical Cancer Mutations

- Highly curated database of mutations having a demonstrated association with cancer
- General information about each somatic variant
  - Chromosomal Location
  - Strand
  - Gene
  - Protein impact of variant (annotation)
  - PubMed ID evidence cited, linked
- Easy to access from the web and programmatically through an API

DGIdb: Drug Gene Interaction database

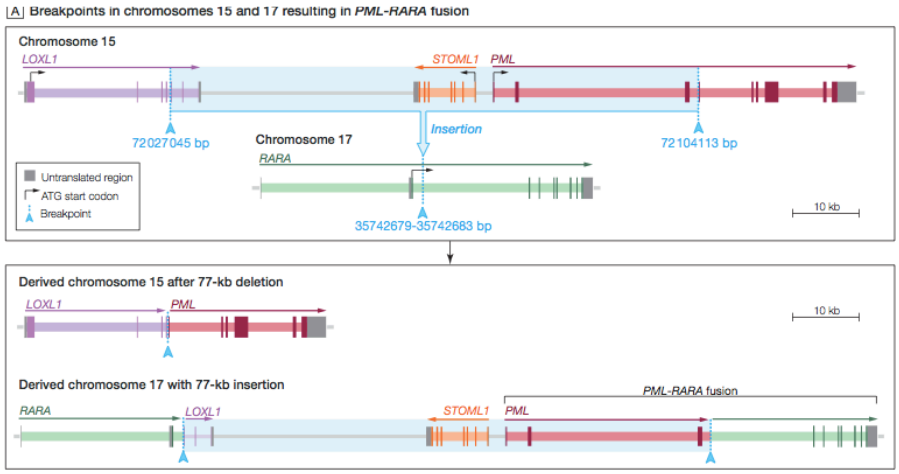


dgidb.org

Griffith, M. et al., Nature Methods 2013

PRELIMINARY COMMUNICATION

Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene



Welch et al., JAMA 2011: 305(15): 1577-1584.

Lukas Wartman, M.D. is Patient "ALL1"

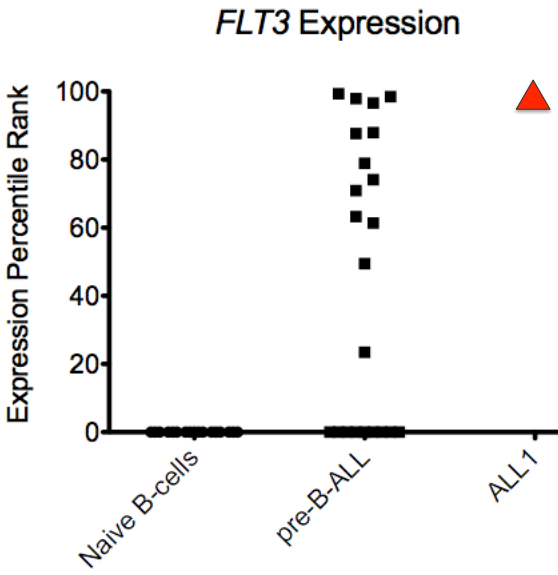
The New York Times

In Treatment for Leukemia, Glimpses of the Future

Second Chance: Lukas Wartman, a leukemia doctor and researcher, developed the disease himself. As he faced death, his colleagues sequenced his cancer genome. The result was a totally unexpected treatment.

By GINA KOLATA
Published: July 7, 2012

FLT3 Over-expression in ALL1

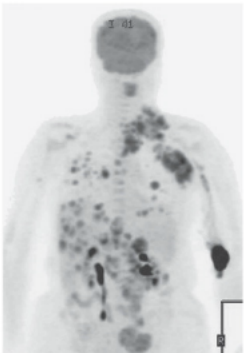


- FLT3 was within the top 1% of all expressed genes.
- Absent a normal comparator, the literature report from Marston identified FLT3 over-expression in pre-B-ALL
- Based on **wt** FLT3 over-expression by the tumor cells, we predicted the cancer would be sensitive to the ***FLT3*** **inhibitor Sunitinib** (Sutent) [DrugBank].

Marston E, et al. Blood 2009 Jan 1;113(1):117-26.

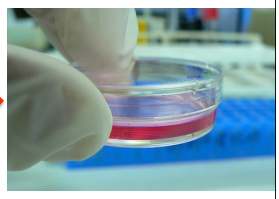
Genome-driven cancer immunotherapy

Sequencing to identify tumor-specific immunoepitopes:



Patient biopsied metastatic melanoma lesions

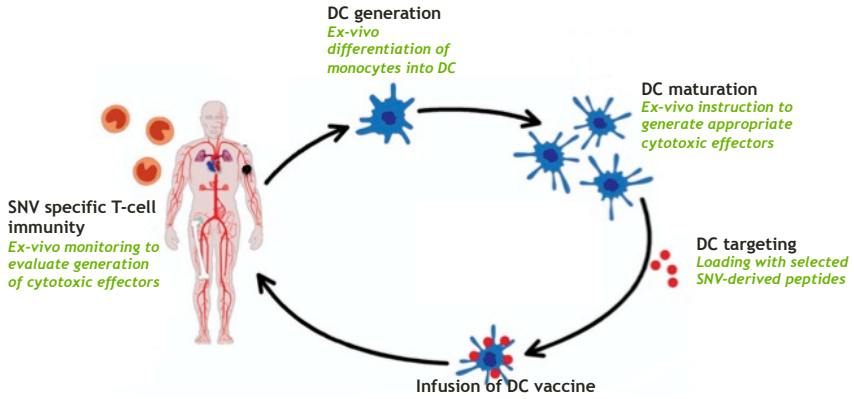
Tumor and germline DNA sequenced, somatic mutations identified; RNA capture verifies expressed mutations and expression level; netMHC algorithm identifies immunoepitopes



Apheresis samples from patient used to verify the algorithmically-identified immunoepitopes that elicit T cell memory

Mardis, Schreiber et al., Nature 2012

Dendritic Cell Vaccine Platform



A dendritic cell-based approach is currently being tested in an FDA approved protocol for **metastatic melanoma** patients:

- Patient 1 has received all three doses of vaccine, and is being monitored
- Patient 2 has received three doses of vaccine, this patient has measurable disease and will be monitored for progression, stability or regression
- Patient 3 has measurable disease, has completed her vaccine infusions early March
- Patients 4 and 5 have genomic analysis completed, in vitro assays completed, GMP peptides underway

Acknowledgements

The Genome Institute
Malachi Griffith, Ph.D.
Obi Griffith, Ph.D.
Ben Ainscough
Zach Skidmore
Avinash Ramu
Allison Regier
Lee Trani
Nick Spies
Vincent Magrini, Ph.D.
Sean McGrath
Ryan Demeter
Jasreet Hundal, M.S.
Jason Walker
David Larson, Ph.D.
Lucinda Fulton
Robert Fulton

Richard K. Wilson, Ph.D.

WUSM/Siteman Cancer Center
Timothy J. Ley, M.D.
Lukas Wartman, M.D.
Peter Westervelt, M.D.
John DiPersio, M.D.
Gerry Linette, M.D.
Beatriz Carreno, M.D., Ph.D.

Thanks also to:
Aaron Quinlan
Gabor Marth
Michael Zody

Our patients and their families