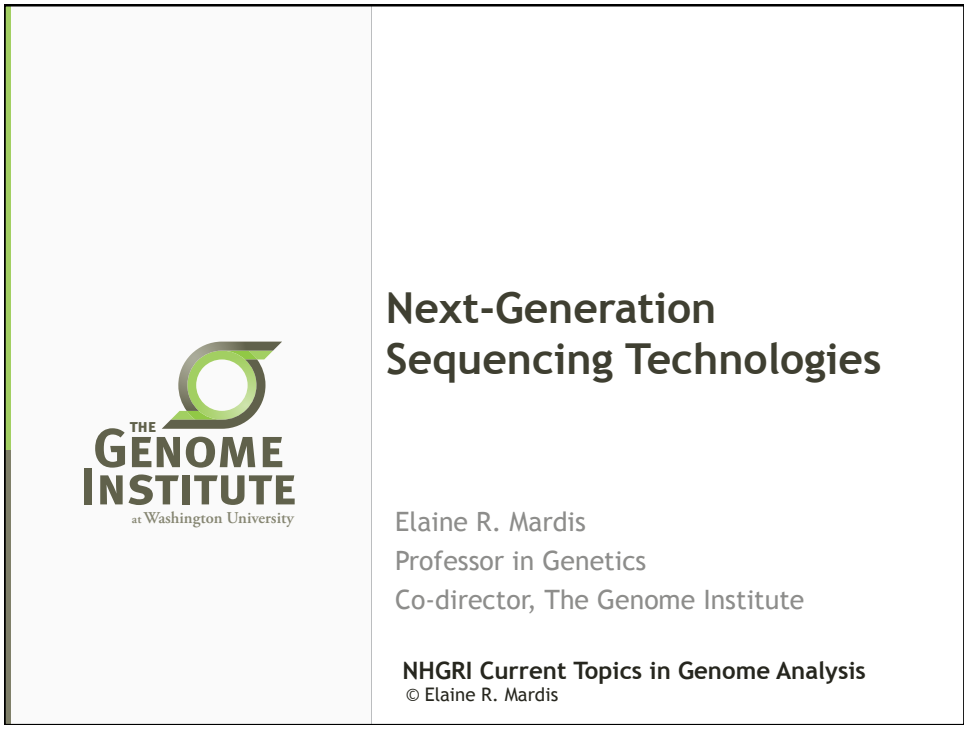
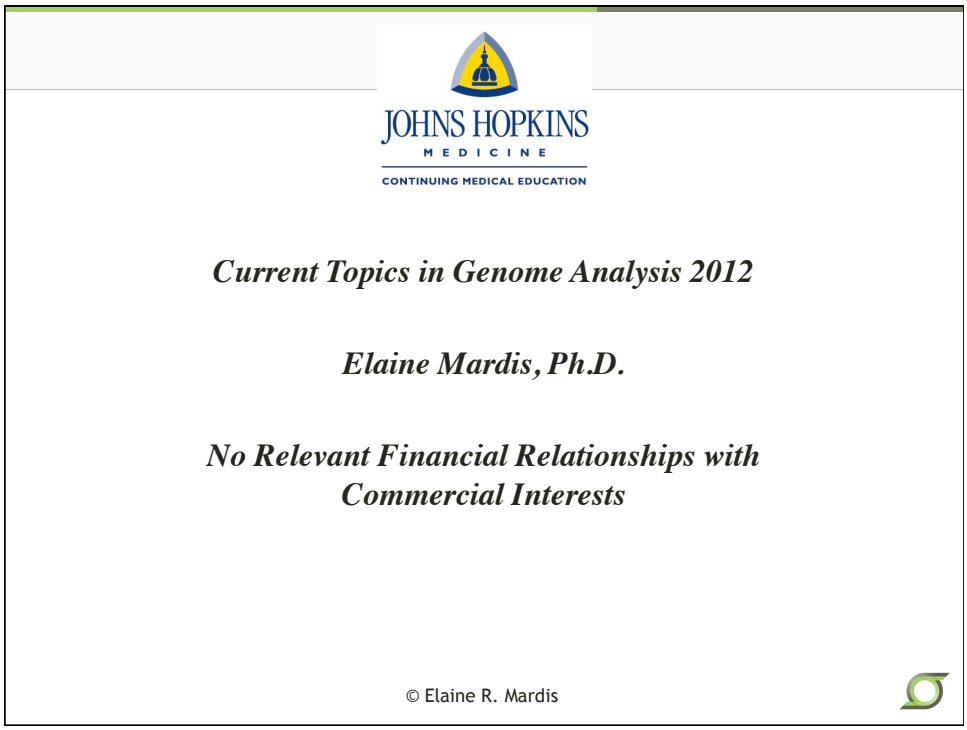

THE GENOME INSTITUTE
at Washington University
Next-Generation Sequencing Technologies
Elaine R. Mardis
Professor in Genetics
Co-director, The Genome Institute
NHGRI Current Topics in Genome Analysis
© Elaine R. Mardis

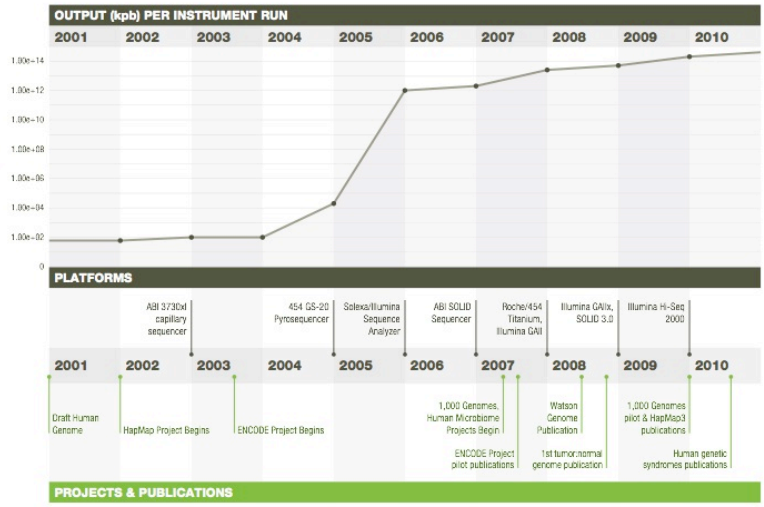

JOHNS HOPKINS
MEDICINE
CONTINUING MEDICAL EDUCATION
Current Topics in Genome Analysis 2012
Elaine Mardis, Ph.D.
No Relevant Financial Relationships with Commercial Interests
© Elaine R. Mardis

Overview

- Next-Generation Sequencing (NGS) Instruments
  - Roche/454
  - Illumina
  - Life Technologies
  - Pacific Biosciences
  - Ion Torrent
  - Oxford Nanopore
- NGS Applications across the spectrum of genomics
  - Examples from our work
  - Future Directions

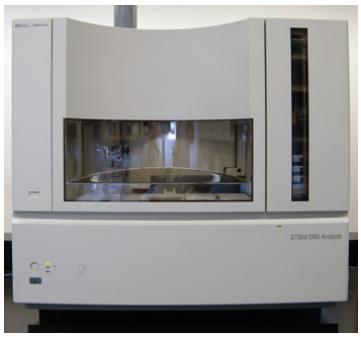
© Elaine R. Mardis

The Trajectory of Throughput: 10 years



E.R. Mardis, Nature (2011) 470: 198-203

Comparative costs: sequencing a human genome



Capillary technology

Applied Biosystems 3730xl (2004)

$15,000,000

Next-gen technology

Illumina HiSeq (2011)

$10,000

© Elaine R. Mardis

Next-generation Sequencer basics

Platforms and their attributes

© Elaine R. Mardis

Next-generation DNA sequencing instruments

- All commercially-available sequencers have the following shared attributes:
  - Random fragmentation of starting DNA, ligation with custom linkers = "a library"
  - Library amplification on a solid surface (either bead or glass)
  - Direct step-by-step detection of each nucleotide base incorporated during the sequencing reaction
  - Hundreds of thousands to hundreds of millions of reactions imaged per instrument run = "massively parallel sequencing"
  - Shorter read lengths than capillary sequencers
  - A "digital" read type that enables direct quantitative comparisons
  - A sequencing mechanism that samples both ends of every fragment sequenced ("paired end" reads)

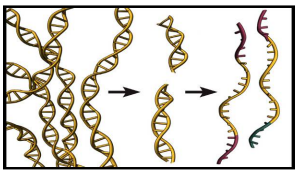
© Elaine R. Mardis

Paired-end reads

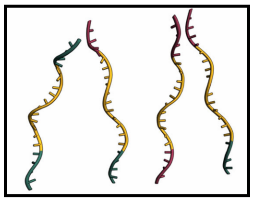
- All next-gen platforms now offer paired end read capability, e.g. sequences can be derived from both ends of the library fragments.
- Differences exist in the _distance_ between read pairs, based on the approach/platform.
  - "paired ends" : linear fragment sequenced at both ends in two separate reactions
  - "mate pairs" : circularized fragment of >1kb, sequenced by a single reaction read or by two separate end reads (platform dependent)
- In general, paired end reads offer advantages for sequencing large and complex genomes because they can be more accurately placed ("mapped") than can single ended short reads.

© Elaine R. Mardis

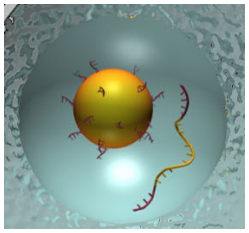
Roche/454 Library Construction and emPCR



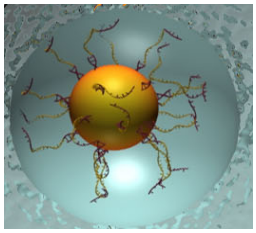
•Random Fragmentation
•Adapter Ligation



Single Stranded Adapter Ligated Library



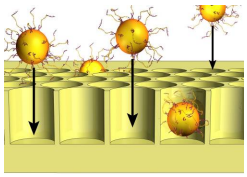
• emPCR for clonal amplification on bead



© Elaine R. Mardis

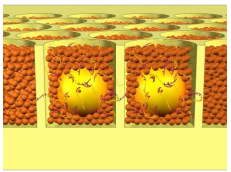
Roche/454 Pyrosequencing

Load beads into PicoTiter™Plate

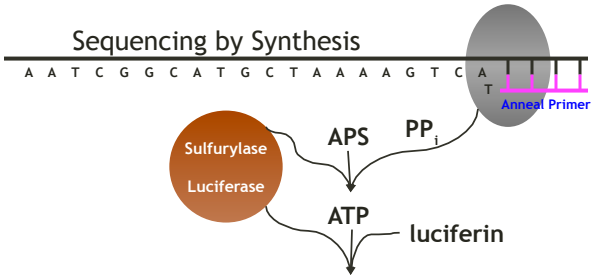


Load Enzyme Beads

Centrifugation



DNA Capture Bead Containing Millions of Copies of a Single Clonal Fragment



© Elaine R. Mardis

454 Instrumentation

Instrument	Run Time (hr)	Read Length (bp)	Yield (Mb/run)	Error Type	Error Rate (%)	Purchase Cost (x1000)
454 FLX+	18-20	700	900	Indel	1	$30 [A]
454 FLX Titanium	10	400	500	Indel	1	$500
454 GS Jr. Titanium	10	400	50	Indel	1	$108

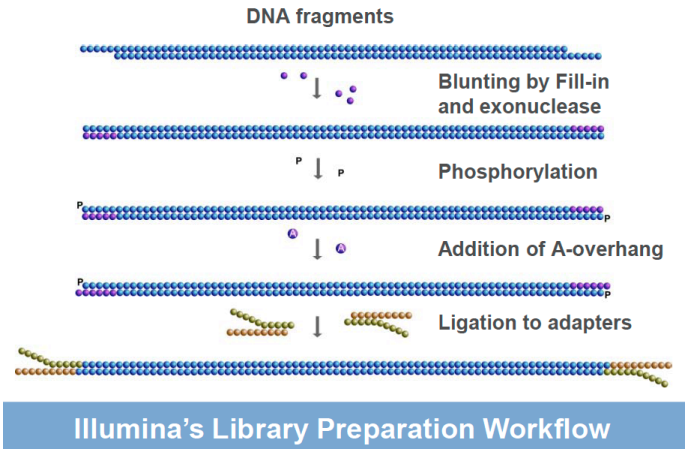
[A] – Requires the 454 FLX Titanium. This is the upgrade cost.

Notable:

- Mate pair paired end reads of 3kb, 8kb and 20 kb separation without an increase in run time.
- Cost per run makes sequencing an entire human genome cost-prohibitive relative to other technologies (~ $20/Mbp)
- Great platform for targeted validation

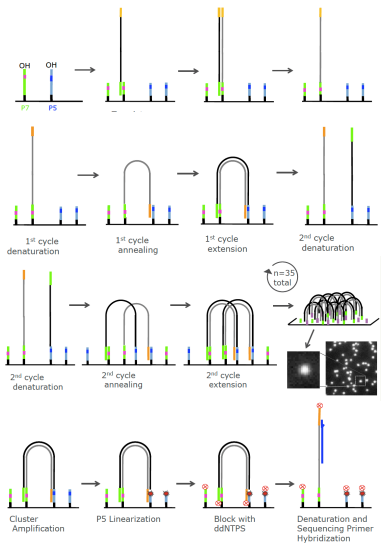
© Elaine R. Mardis

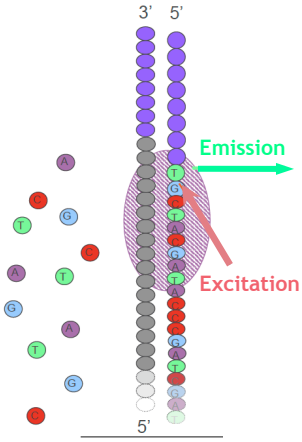
Illumina Sequencing: Library Preparation

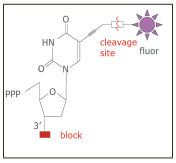


© Elaine R. Mardis

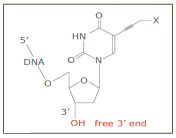
Illumina Sequencing by Synthesis







Incorporate
Detect
De-block
Cleave fluor



© Elaine R. Mardis

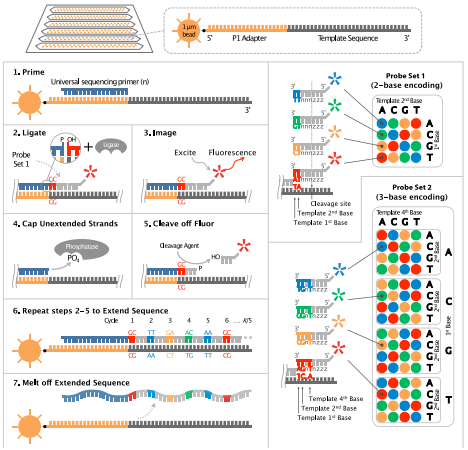
Illumina Instrumentation

- 2010: HiSeq 2000
  - Two flow cells per run
  - 100 Gbp/FC or two genome equivalents per run
  - New scanning mechanics - scans both surfaces of FC lanes
- 2011: HiSeq 2000
  - Improved chemistry (v. 3): increased yield and accuracy
- 2011: MiSeq

Instrument	Run Time (days)	Read Length (bp)	Yield (Gb/run)	Error Type	Error Rate (%)	Purchase Cost (x1000)
GAIIx	14	150 x 150	96	Sub	>0.1	$525
HiSeq 2000	8	100 x 100	200 x 2	Sub	>0.1	$700
HiSeq 2000 v3	10	100 x 100	<600	Sub	>0.1	$700
MiSeq	1	150 x 150	2	Sub	>0.1	$125

© Elaine R. Mardis

Life Technologies: sequencing by ligation



- custom adapter library
- emPCR on magnetic beads
- sequencing by <u>ligation</u> using fluorescent probes from a common primer
- sequential rounds of ligation from a series of primers
- fixed/known nucleotides for each probeset identify two bases each cycle, or "two base encoding"

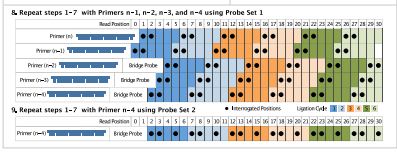


Figure 1. Ligation-based sequencing with 5500 Series SOLiD™ System Exact Call Chemistry.

© Elaine R. Mardis

SOLiD Instrumentation

Instrument	Run Time (days)	Read Length (bp)	Yield (Gb/run)	Error Type	Error Rate (%)	Purchase Cost (x1000)
SOLiD 4	12	50 x 35 PE	71	A-T Bias	>0.06	$475
SOLiD 5500 xl	8	75 x 35 PE 60 x 60 MP	155	A-T Bias	>0.01	$595

<u>5500 xl</u>

- Front-end automation addresses bottlenecks at emPCR, breaking, and enrichment of beads
- 6-lane Flow Chip with independent lanes/2 per run
- Cost per whole genome data set is predicted to be $6K by 2011
- Very high accuracy data due to two-base encoding
- ECC Module – An optional 6th primer that increases accuracy to 99.999%
- Direct conversion of color space to base space
- True paired-end chemistry enabled – Ligation reaction can be used in either direction

© Elaine R. Mardis

Third generation sequencers??

- Recently, new sequencing platforms were introduced.
- The Pacific Biosciences sequencer is a single molecule detection system that marries nanotechnology with molecular biology.
- The Ion Torrent uses pH rather than light to detect nucleotide incorporations.
- The MiSeq is a scaled down version of the HiSeq, with faster chemistry and scanning.
- All offer a faster run time, lower cost per run, reduced amount of data generated relative to 2nd Gen platforms, and the potential to address genetic questions in the clinical setting.

© Elaine R. Mardis

Comparisons to Third-Generation Sequencers

Company	Platform Name	Sequencing	Amplification	Run Time
Roche	454 Ti	DNA Polymerase "Pyrosequencing"	emPCR	10 hours
Illumina	Hi-Seq/ MiSeq	DNA Polymerase	Bridge amplification	10 days/ 24 hours
Life	SOLiD/ 5500	DNA Ligase	emPCR	12 days
Ion Torrent	PGM	Synthesis H^+ detection	emPCR	2 hours
Pacific Biosciences	RS	Synthesis	NONE	45 min

© Elaine R. Mardis

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

Movie 1 (v1.1 & v1.2): Raw reads → Post-filter reads → Mapped reads

Movie 2 (only v1.2): Raw reads → Post-filter reads → Mapped reads

Zero Mode Waveguides (ZMWs)
Ver. 1.1 = 1 x 45,000 per SMRT cell
Ver. 1.2 = 2 x 75,000 per SMRT cell

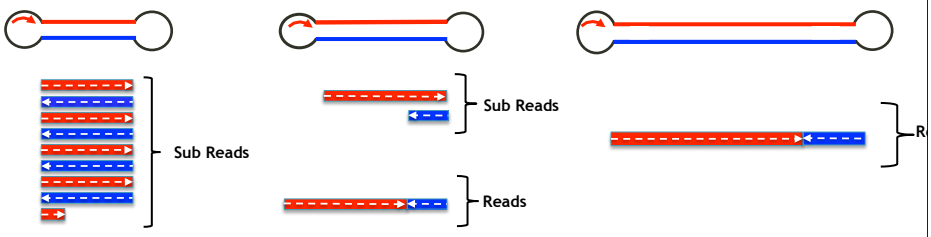
© Elaine R. Mardis

The SMRTbell™ Library Types

Circular Consensus
Small ~250bp
Single Movie
Multiple passes
Many sub-reads

Standard
Large ~2kbp
Single Movie
Few sub-reads

VLR ("very long read")
Larger ~6kbp
Single 45 minute Movie
Single long read provides linking information



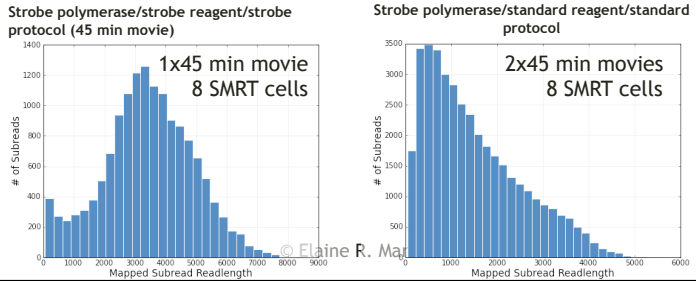
Consensus Read

© Elaine R. Mardis

PacBio RS Instrumentation

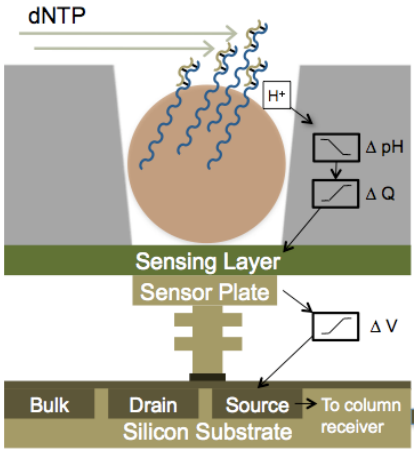
Instrument	Run Time (Hours)	Read Length (bp)	Yield (Mb)	Error Type	Error Rate (%)	Purchase Cost (x1000)
RS	14 (~8 SMRTCells)	2500	45 per SMRTCell	Insertions	15	$695

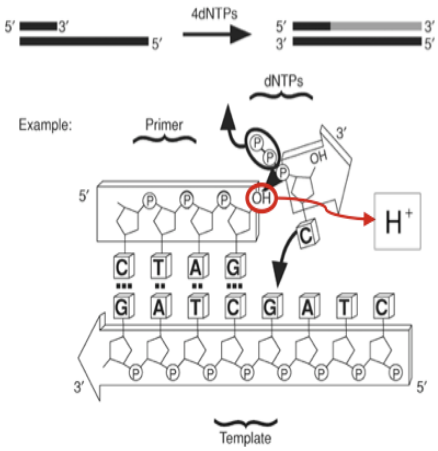
mean mapped sub-read accuracy: 86.2%
mean mapped sub-read length: 3,416 bp
maximum mapped read length: 8,580 bp
maximum 95th percentile mapped read length: 5,807 bp

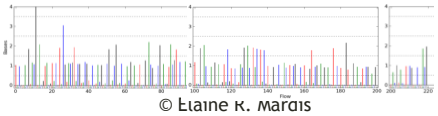


© Elaine R. Mardis

ION Torrent Personal Genome Machine (PGM)

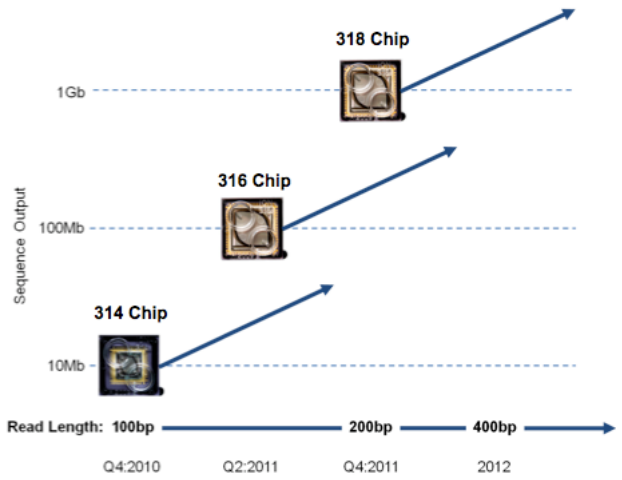






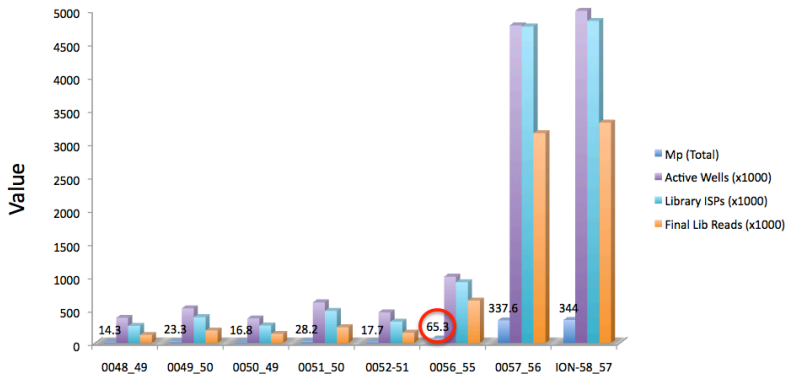
© Elaine R. Mardis

Ion Torrent Yield Trajectory



© Elaine R. Mardis

Ion Torrent Yield with Automation Modules

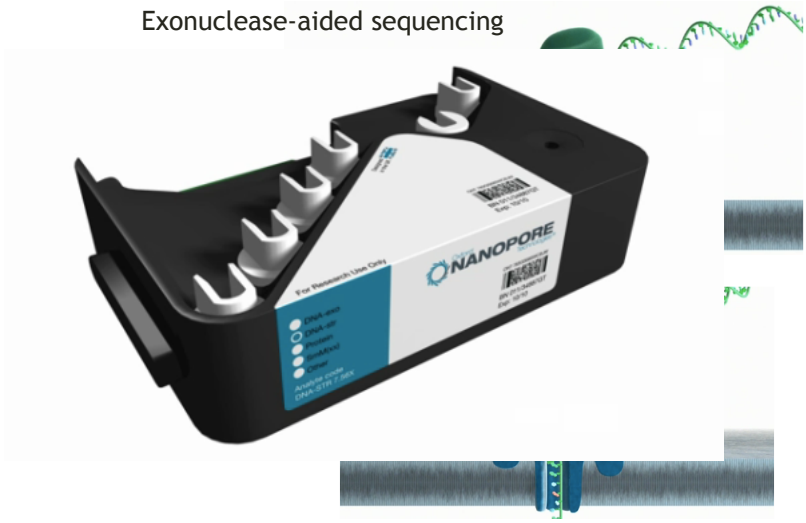


	0048_49	0049_50	0050_49	0051_50	0052-51	0056_55	0057_56	ION-58_57
Chip	**314**	**314**	**314**	**314**	**314**	**314**	**316**	**316**
Sample	***Ef***	***Ef***	***Ef***	***Ef***	***Ef***	***Ec***	***Ec***	***Ec***
Emulsion	**Man**	**Man**	**Man**	**Man**	**Man**	**1-T**	**1-T**	**1-T**
Enrichment	**Man**	**Man**	**Man**	**Man**	**Man**	**ES**	**ES**	**ES**

© Elaine R. Mardis

Oxford Nanopore Sequencing

Exonuclease-aided sequencing



© Elaine R. Mardis

Applying Next Generation Sequencing

- Genomes: re-sequencing or *de novo*
- point mutation/indel/structural variation discovery
- Protein:DNA binding
  - Chromatin IP/histone binding
  - Nucleosome/transcription factor binding, etc.
- ncRNA discovery/sequencing/variants
- Transcriptome sequencing (RNA-seq)
- Genome-wide methylation of DNA (Methyl-seq)
- Clinical sequencing for therapeutic decisions

E.R. Mardis, Annual Reviews in Genetics & Genomics (2008)
E.R. Mardis, Nature (2011) 470: 198-203

© Elaine R. Mardis

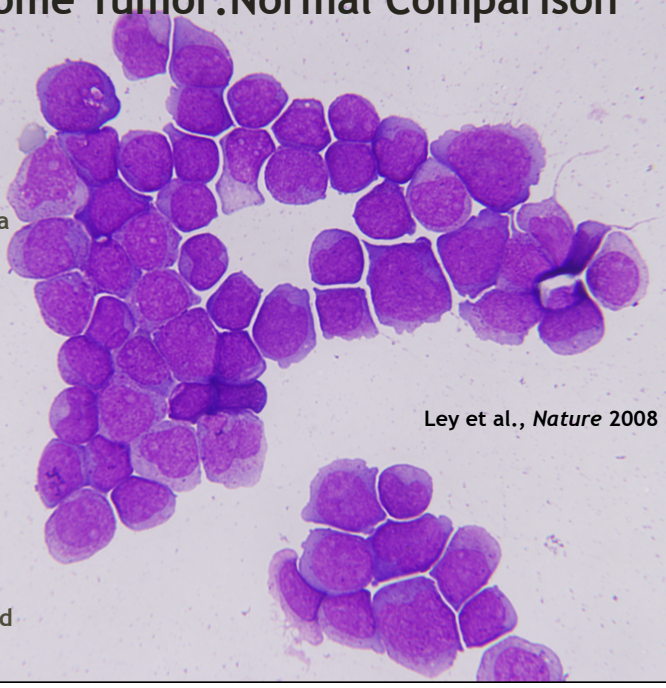
Whole Genome Sequencing: Data Production and Alignment

- Prepare paired end libraries as whole genome fragment/shotgun by random shearing of genomic DNA, adapter ligation, size selection.
- Produce paired end data from each end of billions of library fragments, over-sampling about 30-fold to cover at a depth sufficient to find all types of genome alterations.
- Computer programs align the read pair sequences onto the reference genome and several algorithms are used to discover variants genome-wide.

© Elaine R.

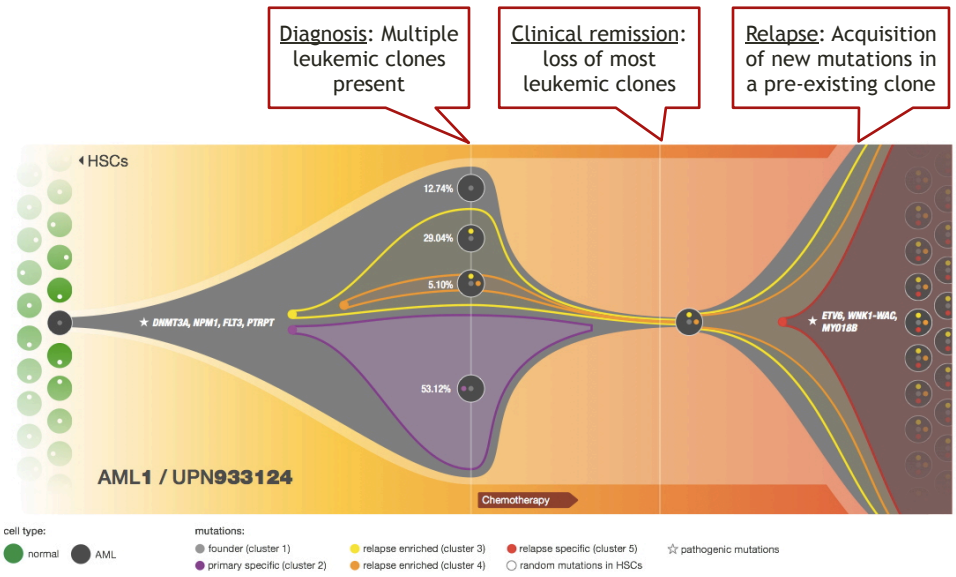
Whole Genome Tumor:Normal Comparison

- Caucasian female, mid-50s at diagnosis
- *De novo* M1 AML
- Family history of AML and lymphoma
- 100% blasts in initial BM sample
- Relapsed and died at 23 months
- Normal cytogenetics
- Informed consent for whole genome sequencing
- Solexa sequencer, 32 bp unpaired reads
- 10 somatic mutations detected



Ley et al., *Nature* 2008

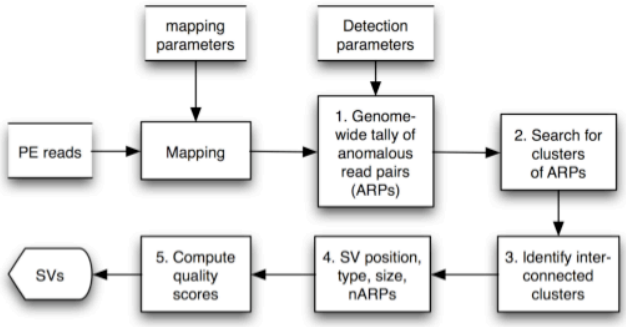
Disease progression model for Patient AML1

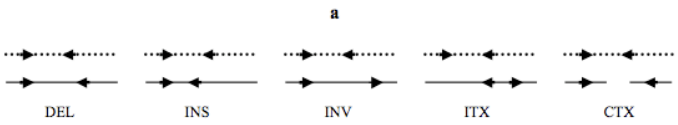


© Elaine R. Mardis

BreakDancer: detecting somatic structural variation

- Read pair analysis with BreakDancer identifies putative SVs based on sets of read pairs that don't map as expected, with respect to distance apart and/or orientation.
- The identified reads are used to produce an assembly, then validated.

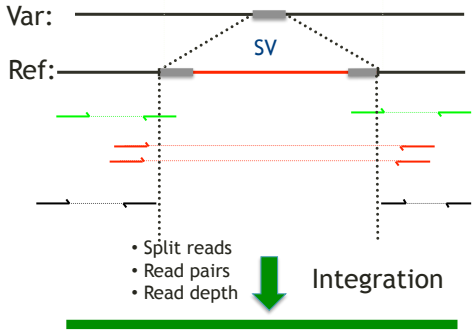




K. Chen et al., Nature Methods 6: 677-81 (2009)

© Elaine R. Mardis

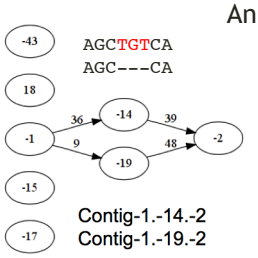
TIGRA_SV: assembling SVs to nucleotide resolution



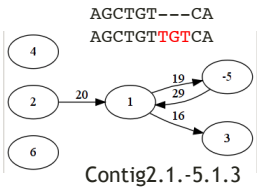
TIGRA_SV targeted local Assembly

Features:
- *de Bruijin* graphic approach
- Multiple Kmer, read threading
- Decode all alleles

An indel

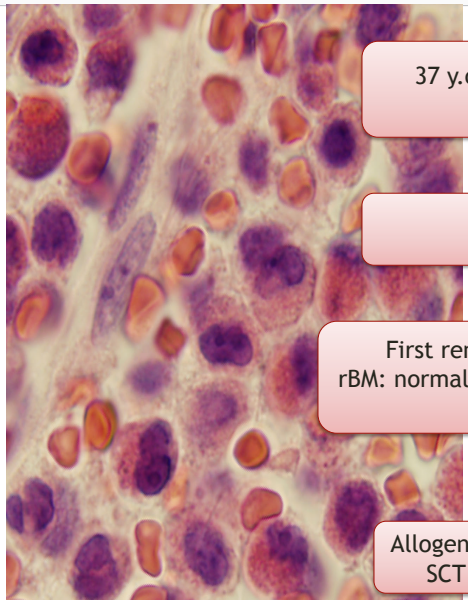


A tandem duplication



© Elaine R. Mardis

Clinical case: "AML52"



37 y.o. female with *de novo* AML; M3 morphology

↓ Chemo + ATRA

Complex cytogenetics, persistent leukemia

↓ Chemo only

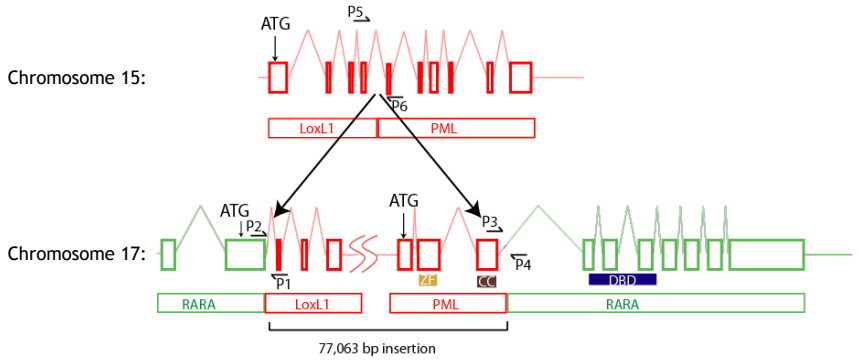
First remission, referred to WU for SCT. rBM: normal morphology, cytogenetics; negative for PML/RARA.

??? → Allogeneic SCT | Consolidation + ATRA

© Elaine R. Mardis

An insertion-derived fusion of PML and RARA

Tumor



PML-RARA: PML exons 1-3 fused to RARA exons 3-9 (bcr3 variant in frame)

RARA-LoxL1: fusion out of frame

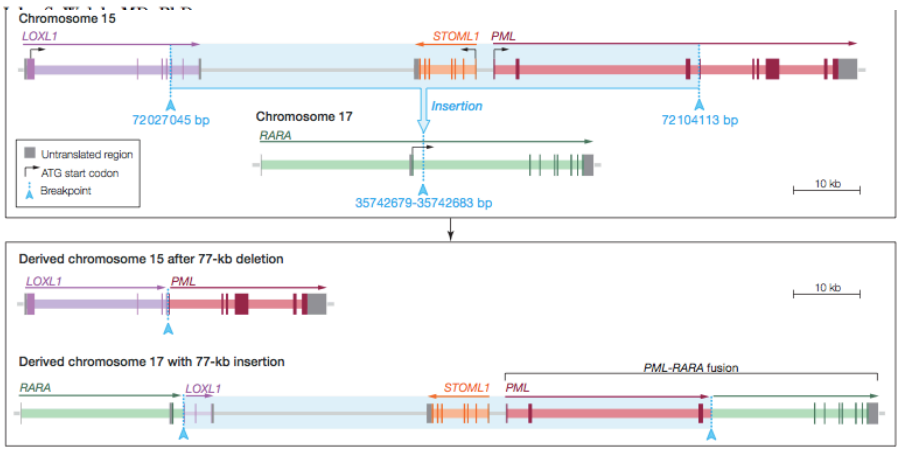
LoxL1-PML: truncated protein - premature stop in novel LoxL1 exon 5a

© Elaine R. Mardis

THE Genome CENTER AT WASHINGTON UNIVERSITY

PRELIMINARY COMMUNICATION

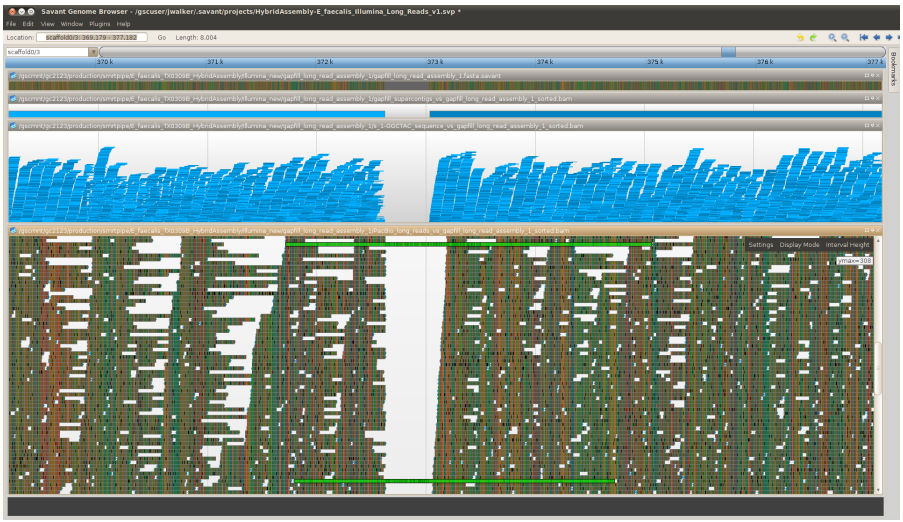
Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene



Welch et al., JAMA April 20, 2011

© Elaine R. Mardis

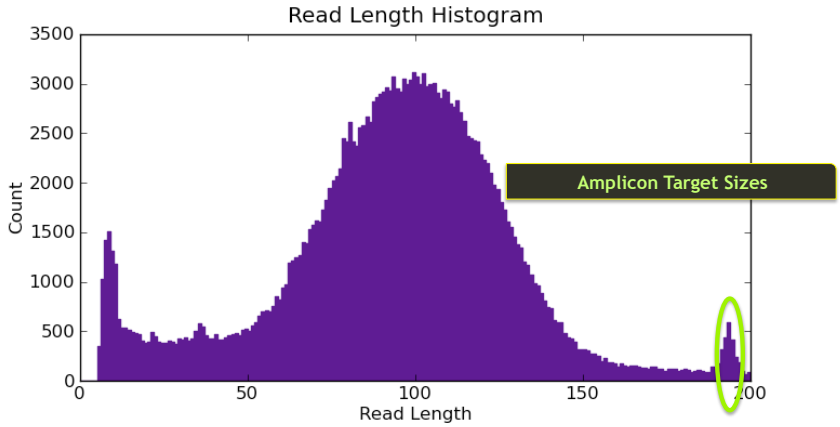
Combining Platforms: *de novo* Assembly



Two VLR PacBio reads contiguate an Illumina assembly gap

© Elaine R. Mardis

Rapid Genotyping by Ion Torrent

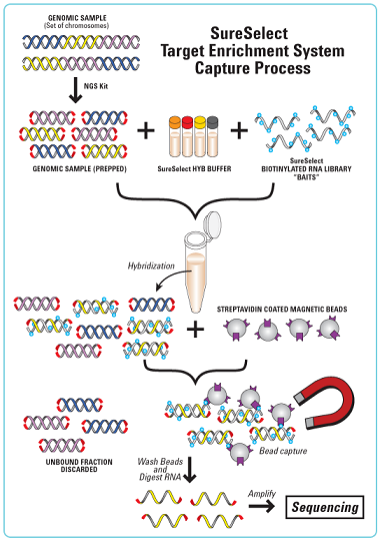


Total Number of Bases [Mbp]	20.35
‣ Number of Q17 Bases [Mbp]	12.49
‣ Number of Q20 Bases [Mbp]	9.66
Total Number of Reads	218,782
Mean Length [bp]	93
Longest Read [bp]	202

© Elaine R. Mardis

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 species' reference genome.
- **Custom capture reagents** can be synthesized to target specific loci that may be of interest in a clinical context.



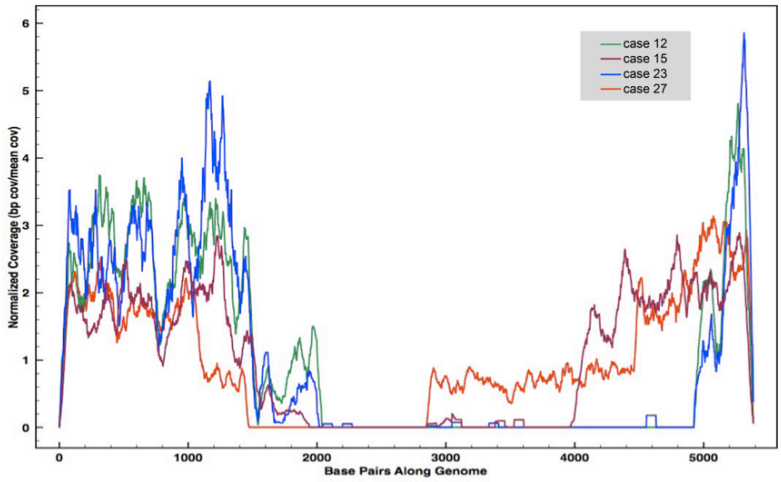
© Elaine R. Mardis

Merkel Cell Polyoma Virus Capture

- Merkel Cell Polyoma virus
  - MCPyV shows frequent genomic deletions and sequence mutations that make it difficult to amplify the virus from cases of MCC by PCR
  - The circular genome does not contain a defined linearization sequence
  - Only FFPE material available for majority of cases
- For proof-of-principle experiments:
  - Biotinylated PCR amplicons designed to target entire 5Kb viral genome
  - Hybrid capture and sequencing
  - Analysis to identify insertion points in human genomes

© Elaine R. Mardis

Viral Coverage Plots (4 FFPE Samples)



Duncavage et al., JMD 2011

© Elaine R. Mardis

Finding the Junction (Integration Site) using SLOPE

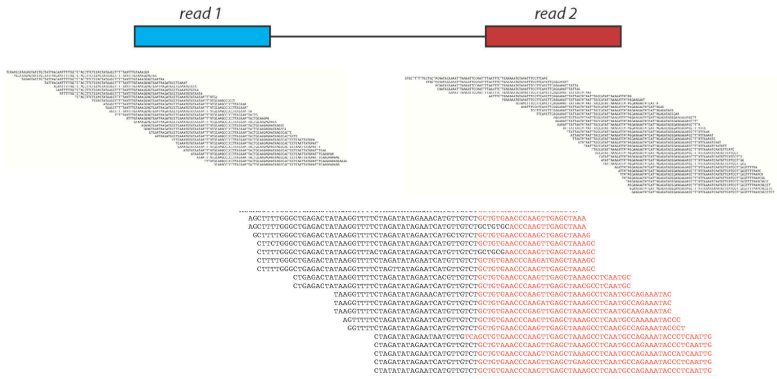


Table 3. MCC Case Clinical Characteristics, PCR-Verified Viral Insertion Sites, and Viral Deletions

Sample	Type	Age of block (years)	Sex	Site	3′ insertion site	5′ insertion site	Viral genomic deletion size
12	Primary	6	M	Buttocks	ch8: 65568962	ch8: 65566806	3.0 kb
23	Metastasis	5	M	Back	ch8: 65568962	ch8: 65566806	3.0 kb
27	Metastasis	2	F	Right arm	ch9: 121417276	undetermined	2.4 kb
15	Metastasis	6	M	Small bowel	ch6: 19684666	ch6: 19684859	1.3 kb

© Elaine R. Mardis

RNA Sequencing

RNA isolate

polyA priming, RT, ds DNA

Fragment, RT w/randoms

SAGE tags

Size selection for nc RNA classes

Adapter-ligated fragments for next-gen sequencing

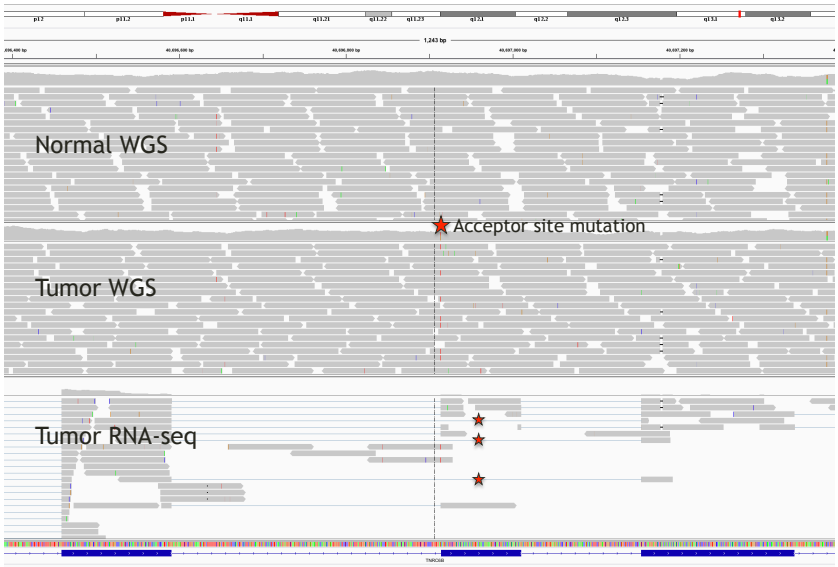
↓

Alignment to reference database & discovery

-Expression levels
-Novel splice isoforms
-Allelic bias in transcription
-Fusion transcripts

© Elaine R. Mardis

TNRC6B splice site mutation -> alt. splicing



© Elaine R. Mardis

Metastatic breast cancer (to brain)

Human Microbiome Project

Goals:
-Defining Core Microbiomes in "healthy" subjects
-Catalog changes in Microbiomes with disease

Subjects → Samples → Microbial Communities

Microbial Communities → Community Phenotypes

Microbial Communities → Sources of strains → Reference Sequences → Microbial Communities

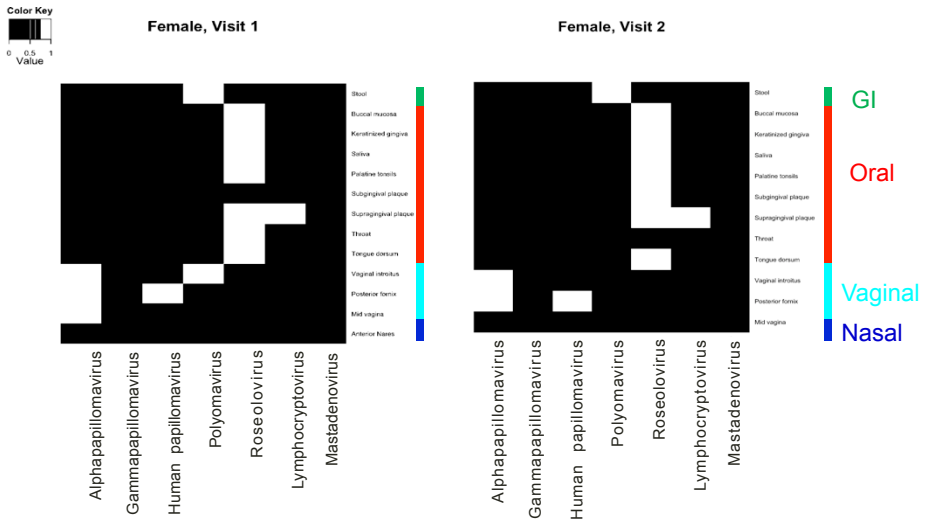
Microbial Communities → Metagenomics → WGS, 16S rRNA

WGS → Reference Sequences

Data Coordination, Analysis and Concordance to Clinical Data

© Elaine R. Mardis

Stability of virome over time



© Elaine R. Mardis

Kristine Wylie, WUGI

400 MRSA Genomes

© Elaine R. Mardis

Conclusions

- 2nd and 3rd generation sequencing instruments are revolutionizing biological research.
- Earliest impacts have been on cancer genomics and metagenomics.
- The extreme need for bioinformatics-based analytical approaches to interpret these large data sets has revitalized the field and introduced statistical and mathematical rigor.
- Integration across data sets from DNA, RNA, methylation, proteomics, etc. presents the next challenge but provides comprehensive analytical power to inform biology.
- With newer instruments, clinical applications have potential for implementation, with appropriate interpretive algorithms.

© Elaine R. Mardis

Acknowledgements

- **The Genome Institute**
  Vince Magrini
  Todd Wylie
  Sean McGrath
  Amy Ly
  Jason Walker
  Jasreet Hundal
  Lisa Cook
  Li Ding
  George Weinstock
  Richard K. Wilson
- **Clinical Collaborators (WUSM)**
  Tim Ley
  Phil Tarr
  John Pfeifer

© Elaine R. Mardis