



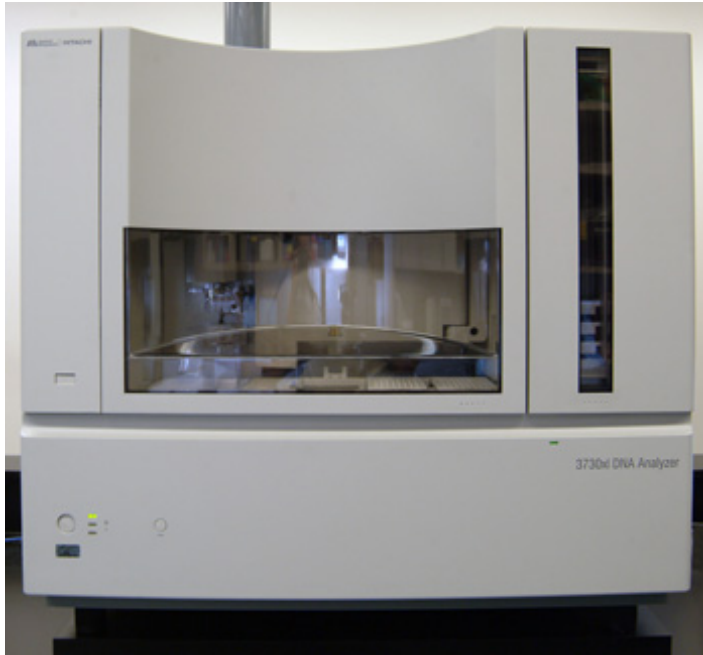
DNA Sequencing 2012

Richard K. Wilson, Ph.D.

Professor of Genetics

Director, The Genome Institute

Sequencing a human genome...



“Old technology”

Applied Biosystems 3730xl
(2004)

\$15,000,000

2-3 years



“Next-gen technology”

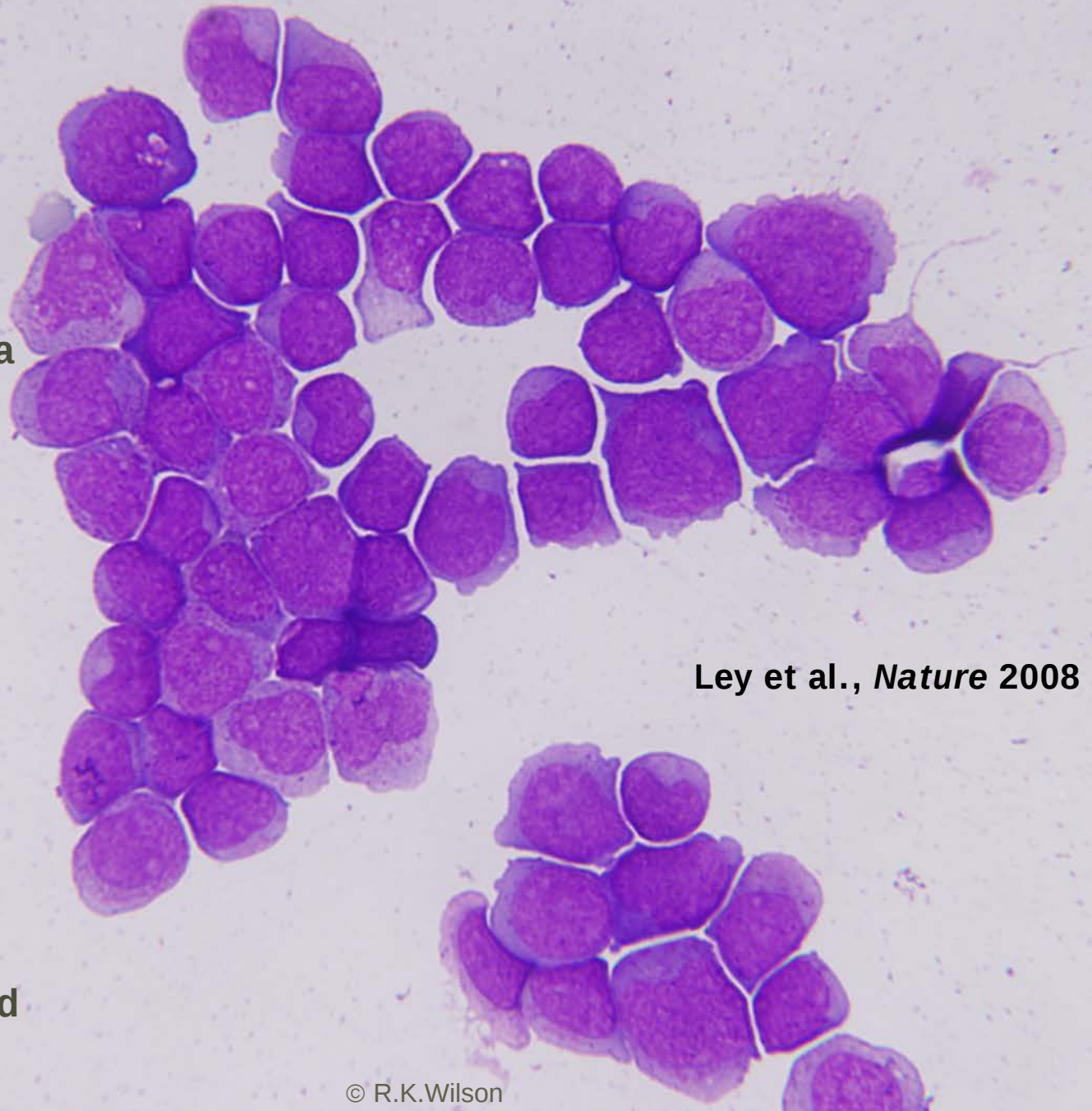
Illumina HiSeq 2000
(2012)

\$5,000

2-3 weeks

“AML1”

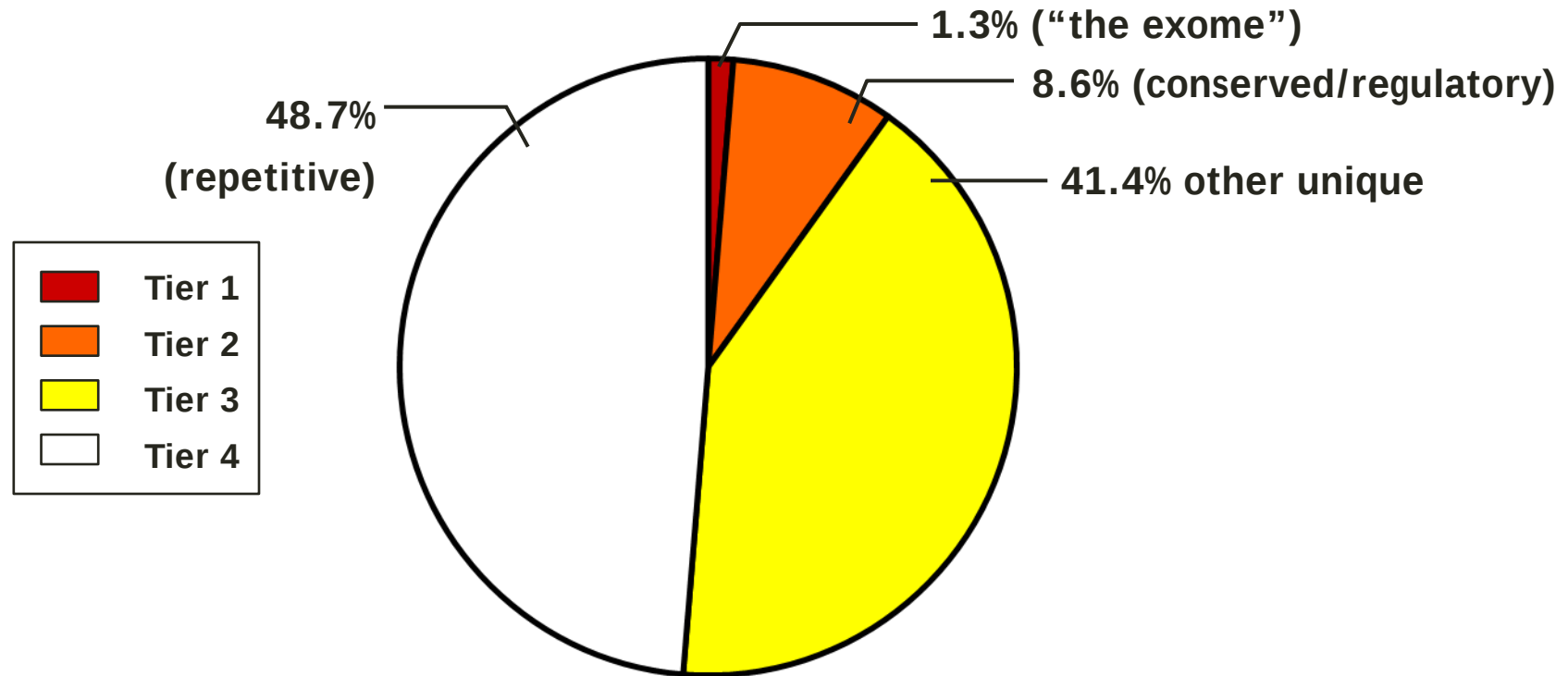
- Caucasian female, mid-50s at diagnosis
- *De novo* M1 AML
- Family history of AML and lymphoma
- 100% blasts in initial BM sample
- Relapsed and died at 23 months
- Normal cytogenetics
- Informed consent for whole genome sequencing
- Solexa sequencer, 32 bp unpaired reads
- 10 Tier 1 somatic mutations detected



Ley et al., *Nature* 2008

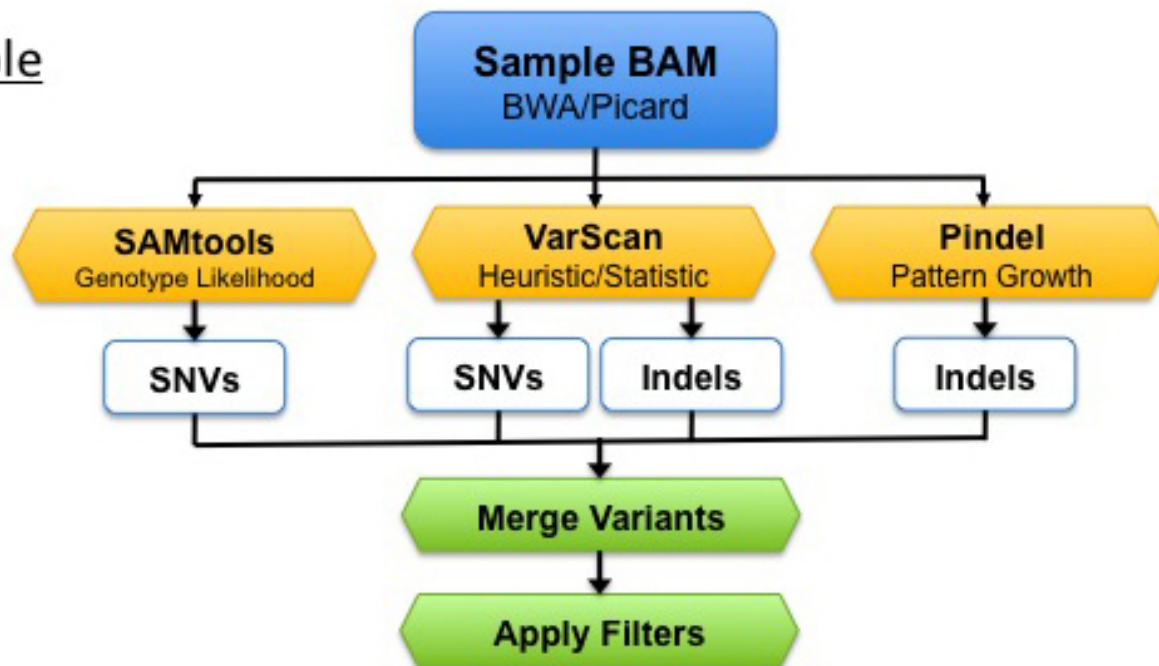
Sequencing *and analyzing* a human genome...

% of the Human Genome in each annotation tier

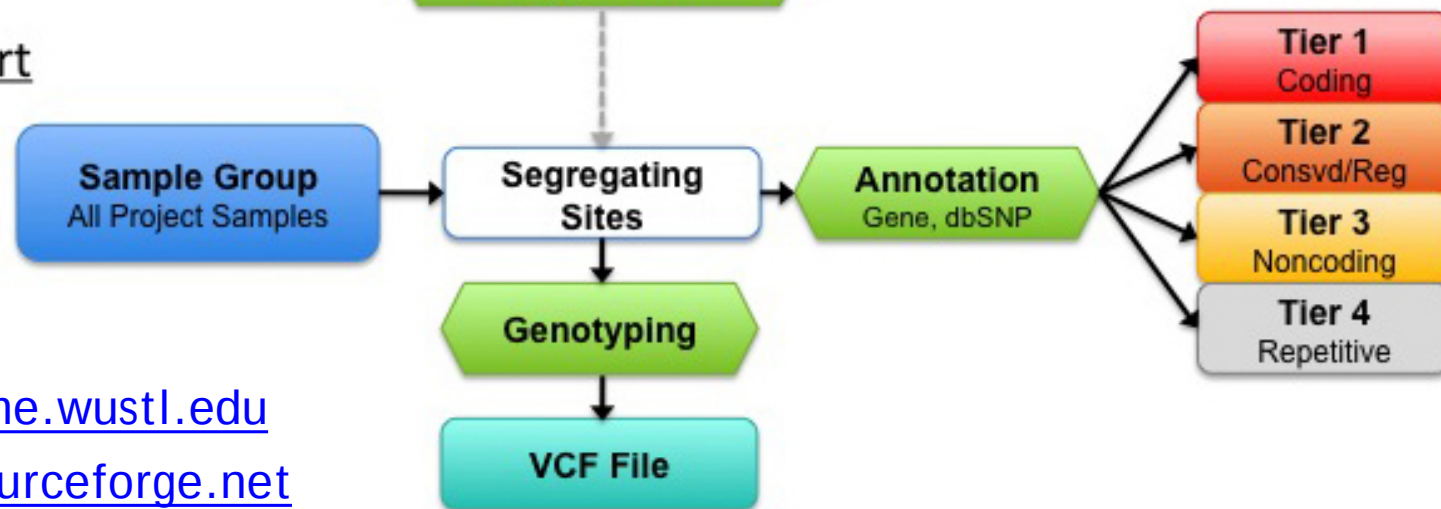


Variant detection in individuals and cohorts

Per Sample



Per Cohort



gmt.genome.wustl.edu
varscan.sourceforge.net

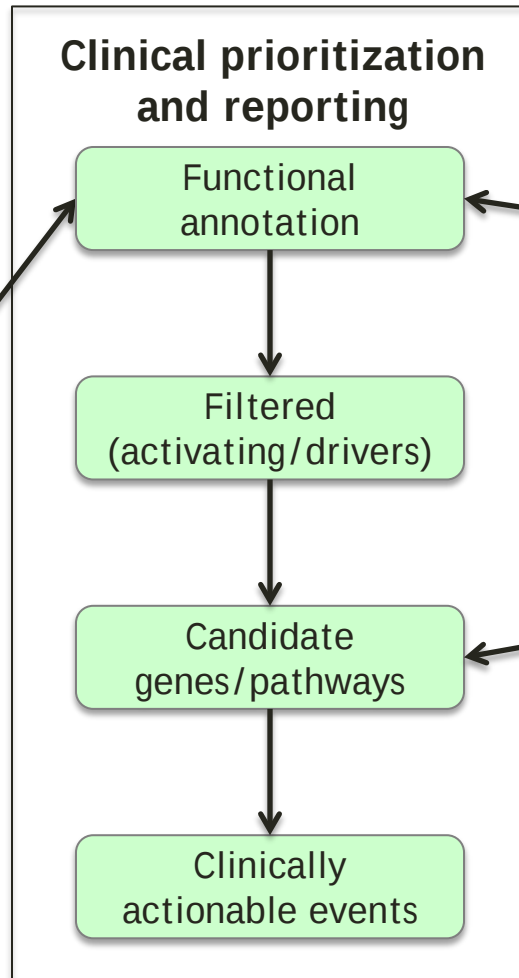


A comprehensive genome analysis pipeline

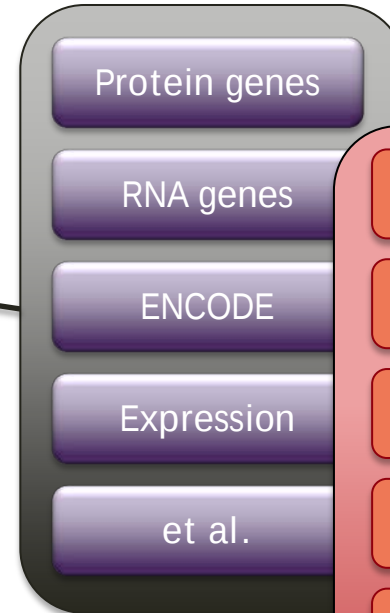
Somatic/Germline Features



WU ClinSeq Pipeline



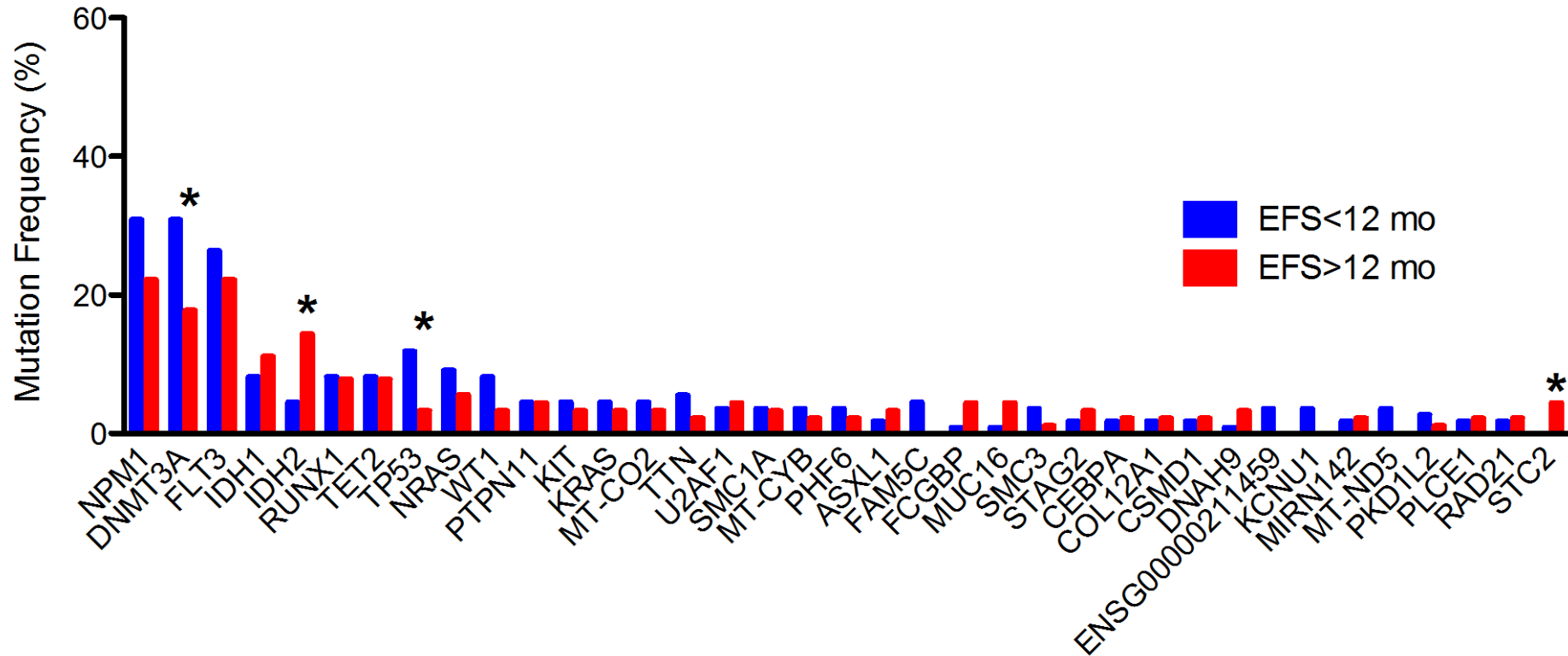
NCBI/EBI



TGI Drug-Gene database (24 DBs)



Somatic mutations in 200 AML patients

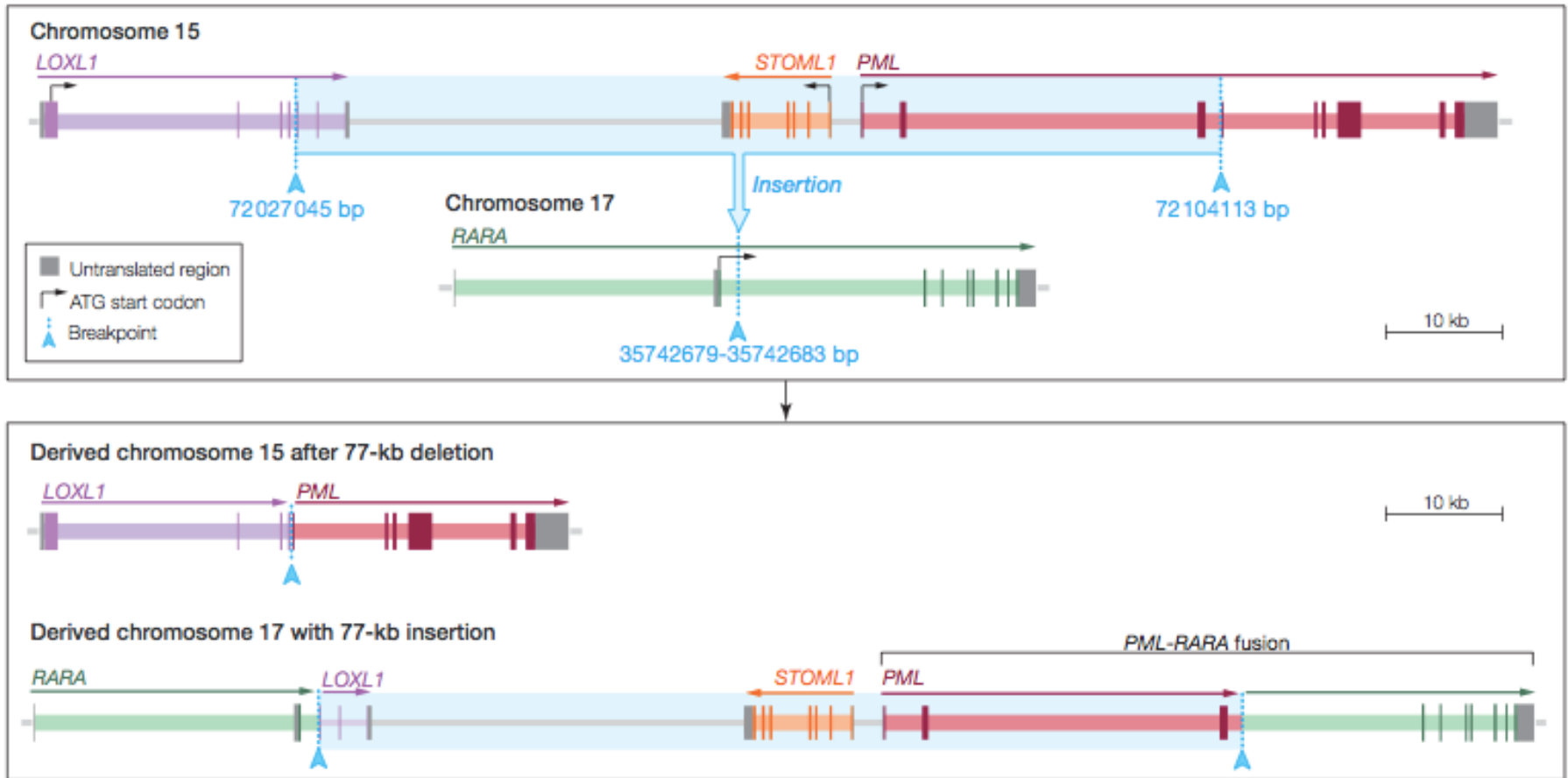


Welch et al., in preparation



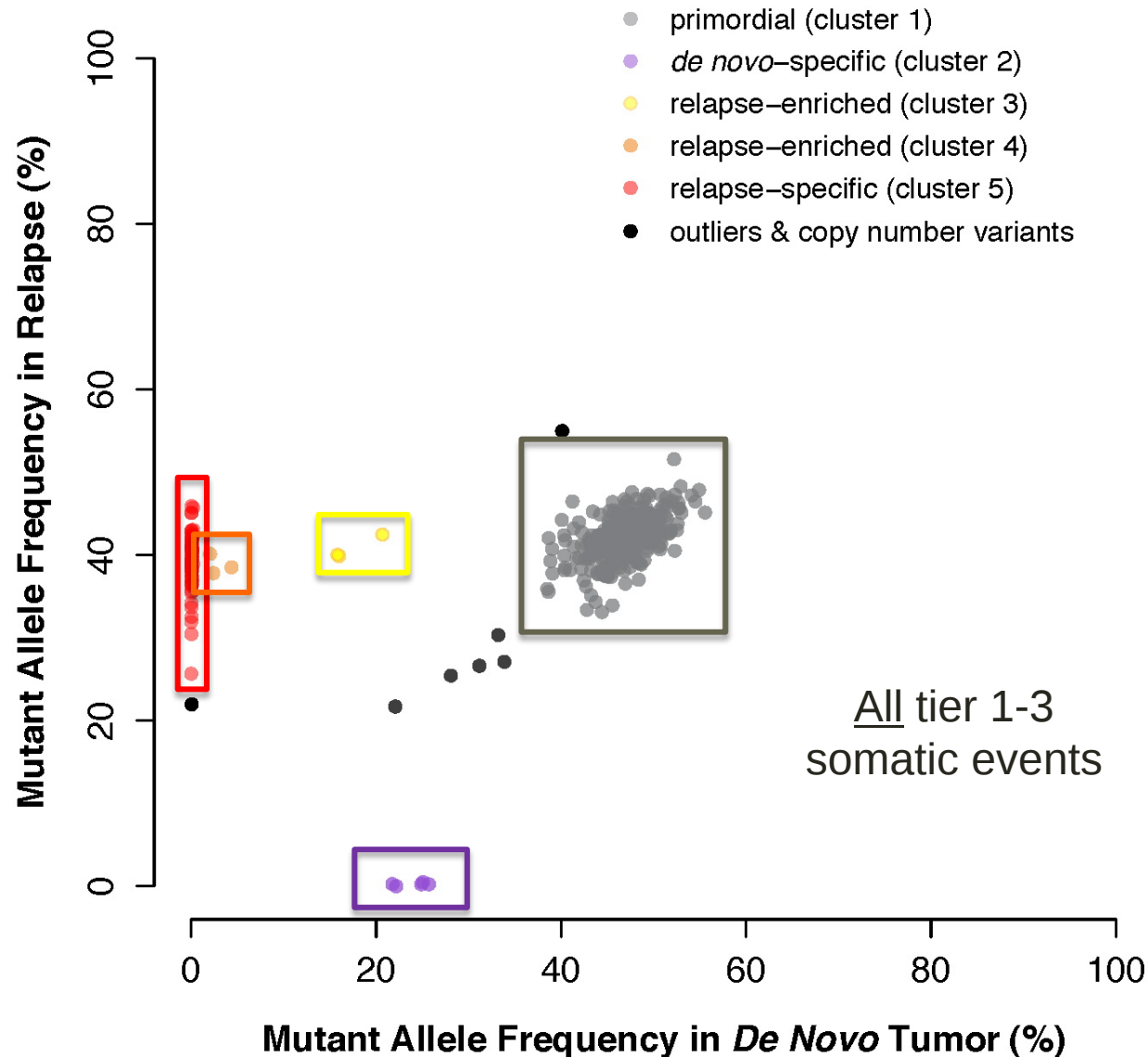
Use of Whole-Genome Sequencing to Diagnose a Cryptic Fusion Oncogene

A Breakpoints in chromosomes 15 and 17 resulting in *PML-RARA* fusion



Welch et al., JAMA 2011

Deep digital sequencing in patient AML1 (relapse)

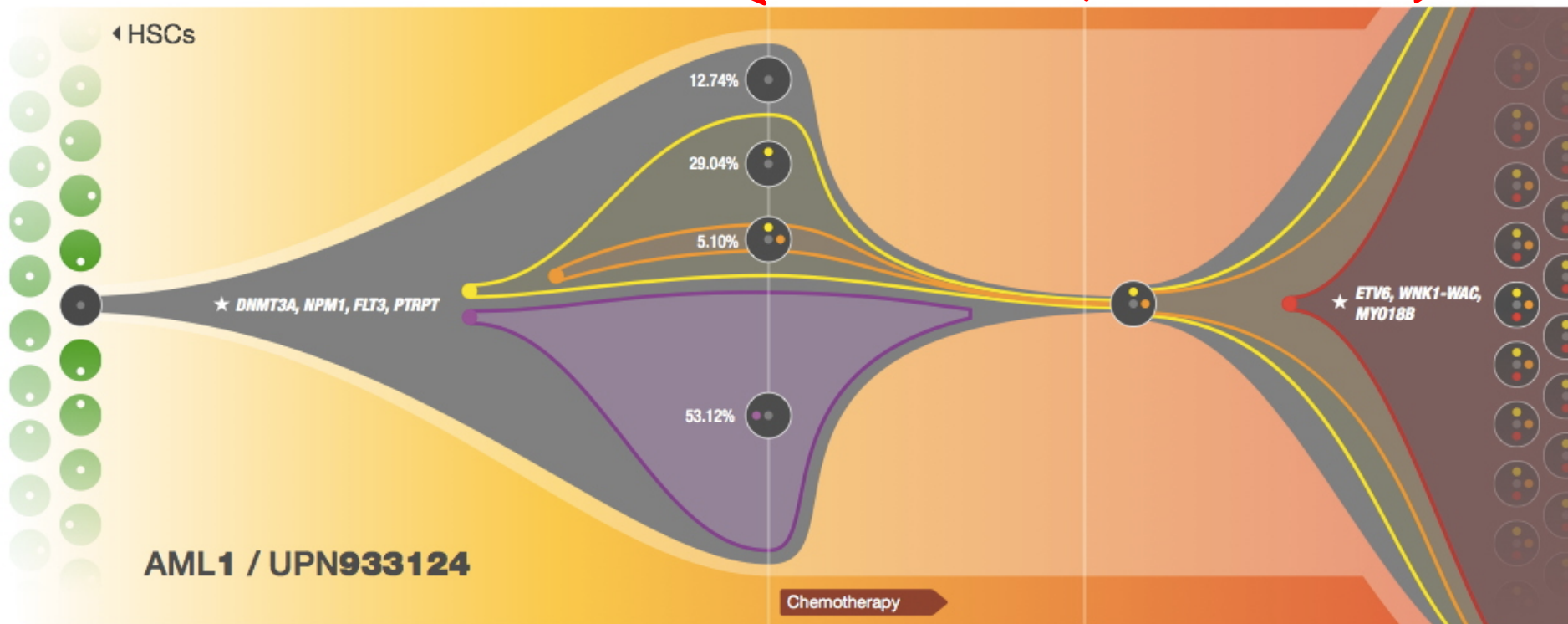


Disease progression model for Patient AML1

Diagnosis: Multiple
leukemic clones
present

Clinical remission:
loss of some
leukemic clones

Relapse: Acquisition
of new mutations in
a pre-existing clone



cell type:

normal AML

mutations:

founder (cluster 1)

primary specific (cluster 2)

relapse enriched (cluster 3)

relapse enriched (cluster 4)

relapse specific (cluster 5)

random mutations in HSCs

☆ pathogenic mutations

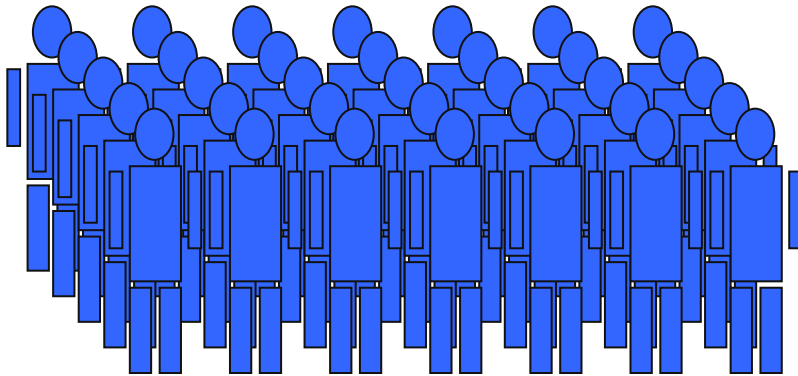
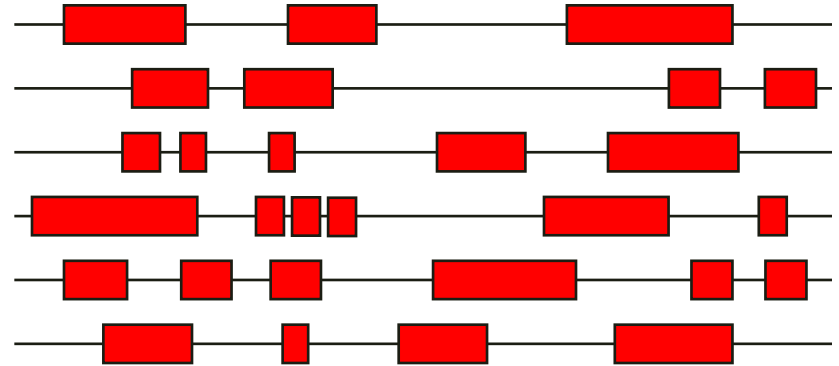
Genomic opportunities in large cohorts

- Sequencing options...



Targeted sequencing (hybrid capture)

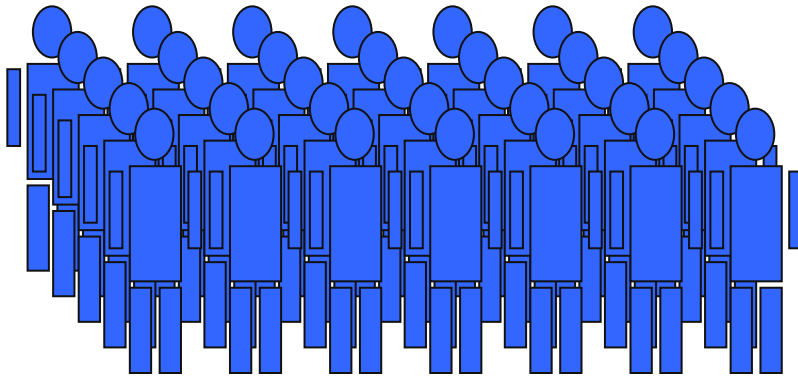
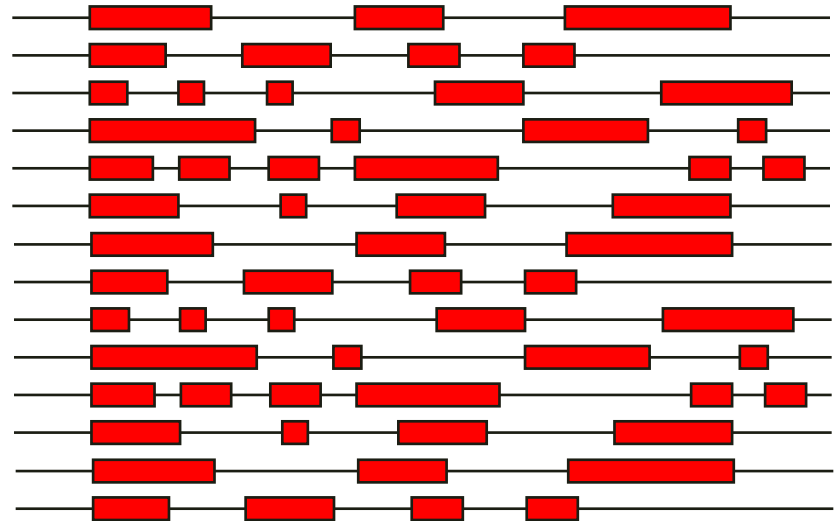
list of candidate
genes and/or
regions of interest
(e.g. GWAS peaks)



large collection of patient
samples

Exome sequencing (hybrid capture)

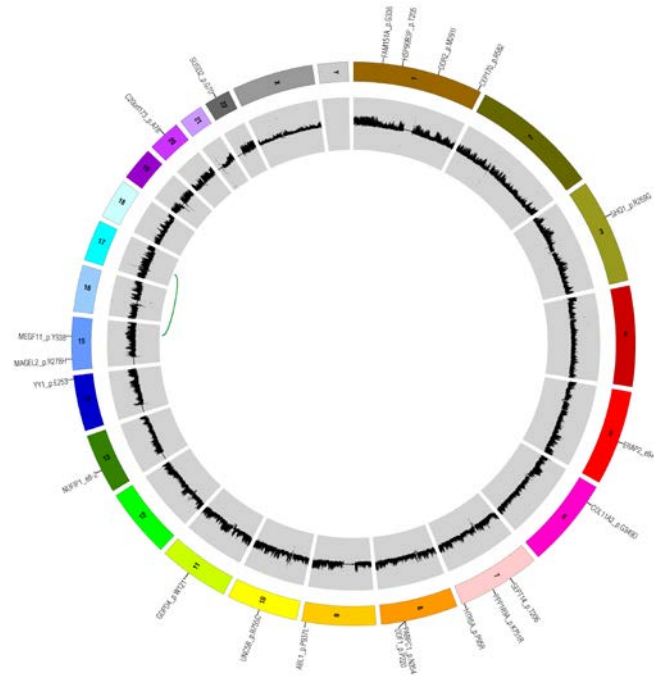
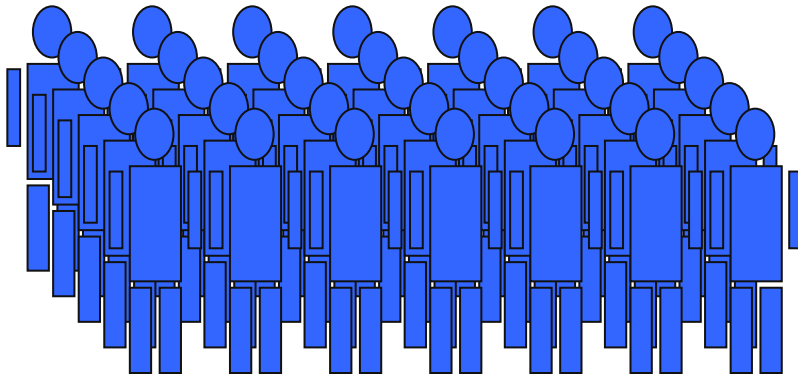
Ideally all CCDS
exons & selected
RNA genes



large collection of patient
samples

Whole genome sequencing

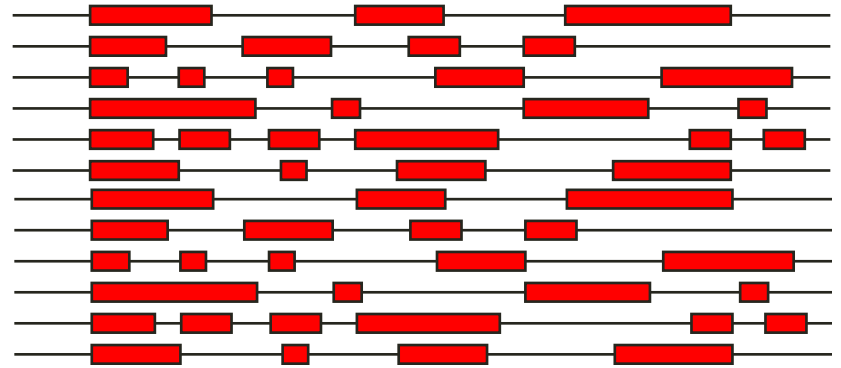
complete genome
sequences aligned
to reference HGS



large collection of patient
samples

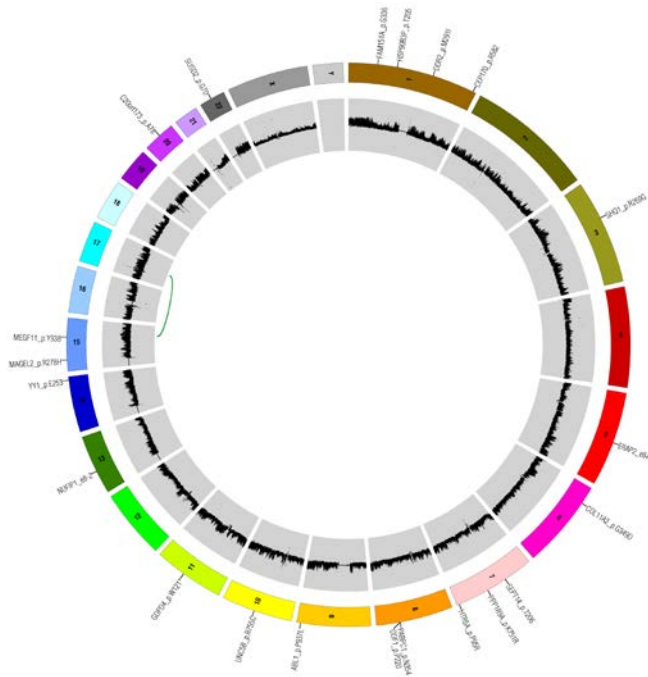
Whole Genome or Exome sequencing?

- Exome sequencing costs less (~1/6 WGS)
- Simplified analysis (60 Mbp)
- Sequence more samples
- “Low-hanging fruit”



VS.

- Non-exonic variants (“tier 2/3”) may play a role in human disease
- WGS resolves fine structure around deleted genes/exons
- WGS covers exons not/poorly covered by exome reagents
- WGS resolves SV, CNV, indels not detected by SNP arrays



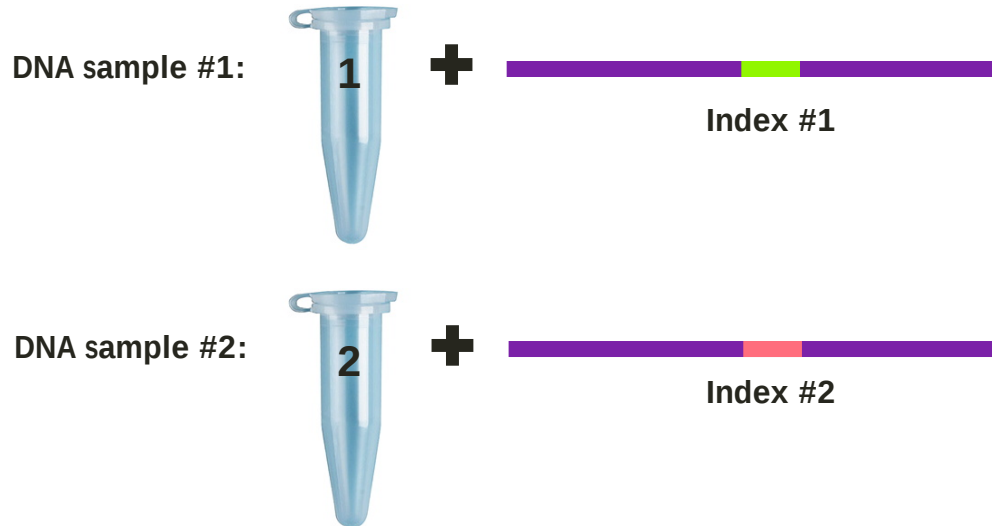
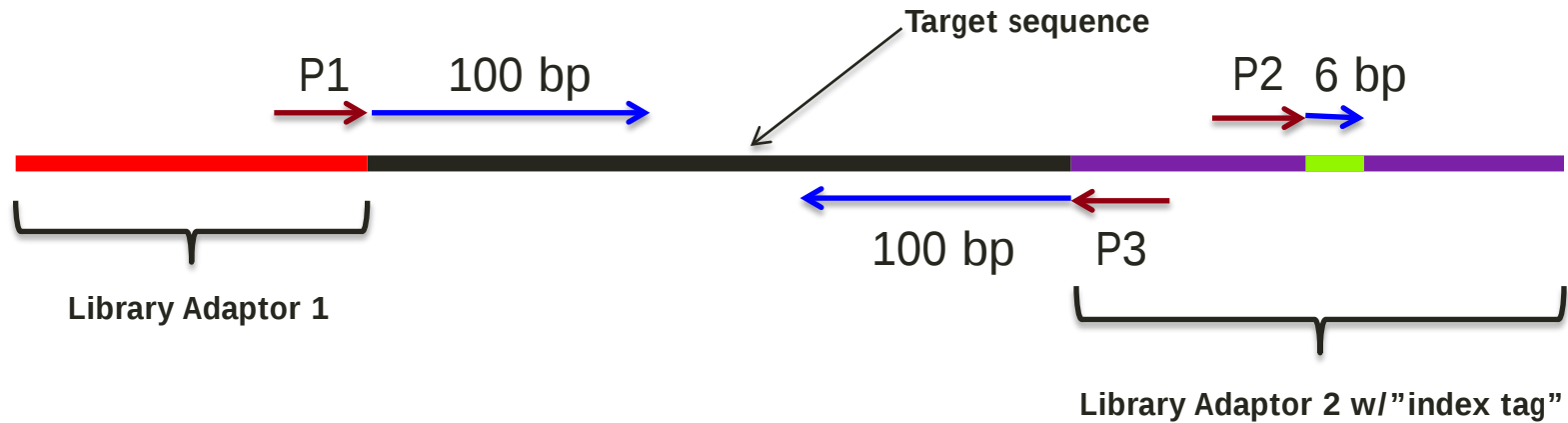
Exome sequencing reagents (relative to “WuSpace”)

	% Product Unique	% Product Shared	% CDS Not Targeted	% CDS Targeted
NimbleGen v2 (35.9 Mb)	8.3%	91.7%	30.1%	69.9%
NimbleGen v3 (63.6 Mb)	42.2%	57.8%	22.2%	77.8%
Agilent SS 50Mb (51.5 Mb)	32.1%	67.9%	25.9%	74.1%
Illumina TruSeq v1 (62.1 Mb)	42.5%	57.5%	24.4%	75.6%

- WuSpace (47 Mbp) consists of all CDS exons and RNA annotations from NCBI GenBank 37c and Ensembl v58. Includes: 38,551 gene names, 120,141 transcript names, 27,062 RNAs, 941,210 CDS exons. A/K/A “tier 1” for WGS analysis.



Multiplexed DNA sequencing (“indexed”)



Library construction

Pool samples

Hybrid capture

DNA sequencing

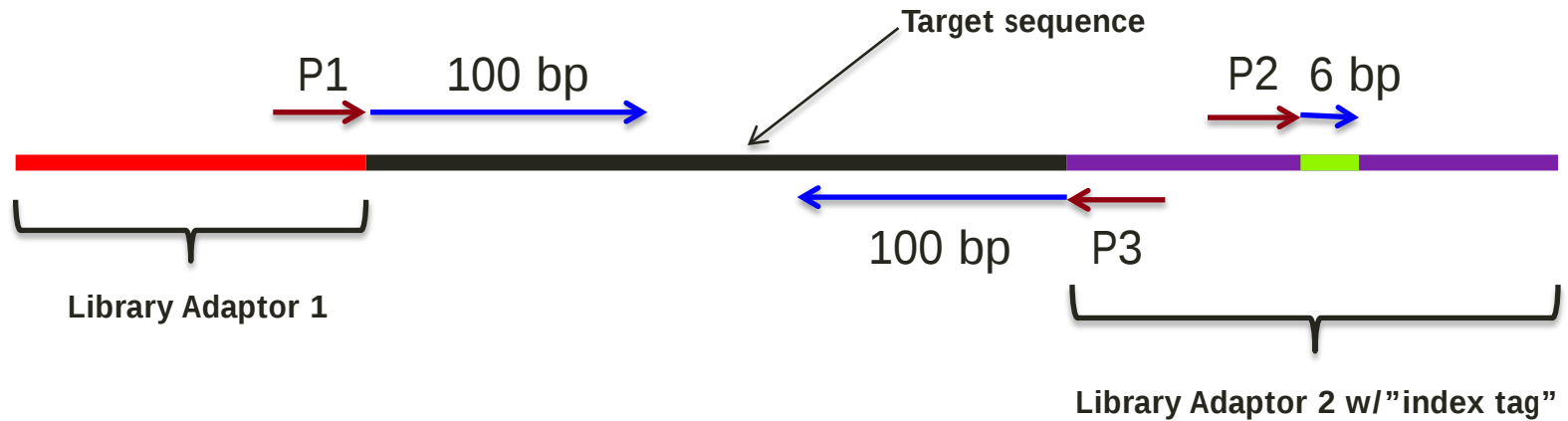
De-multiplex

Align to reference

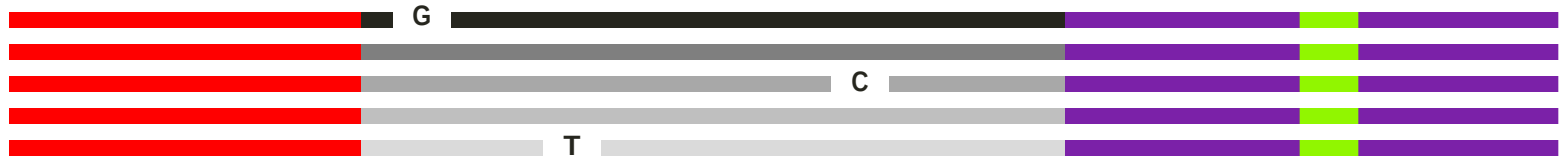
Variant detection



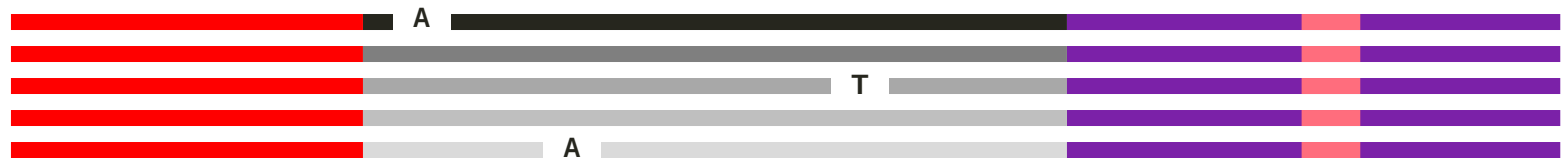
Multiplexed DNA sequencing (“indexed”)



DNA sample #1:



DNA sample #2:



Multiplexed DNA sequencing (“indexed”)

96 Indexed
DNA samples



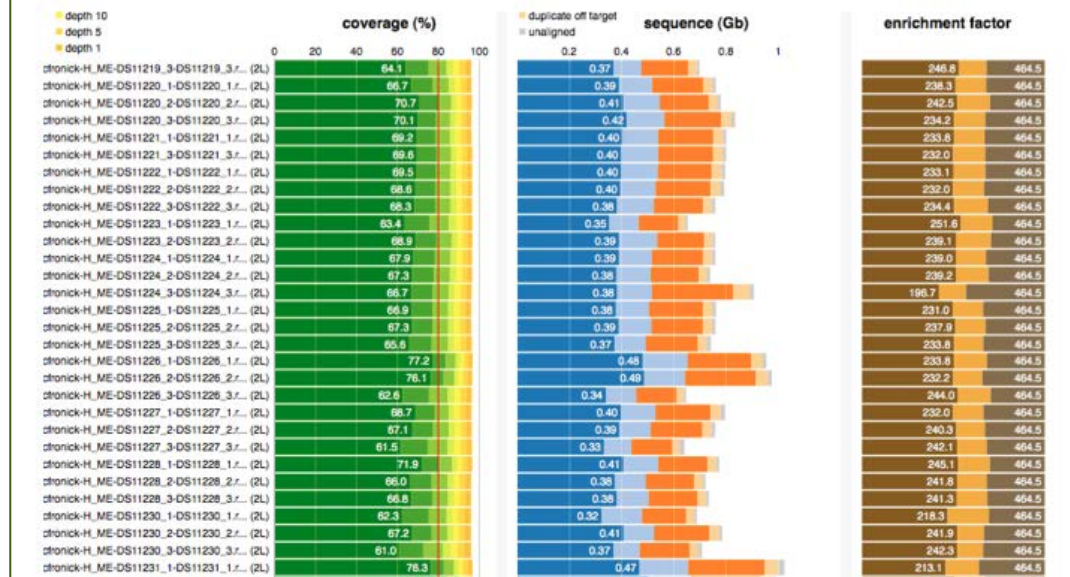
Capture
probes



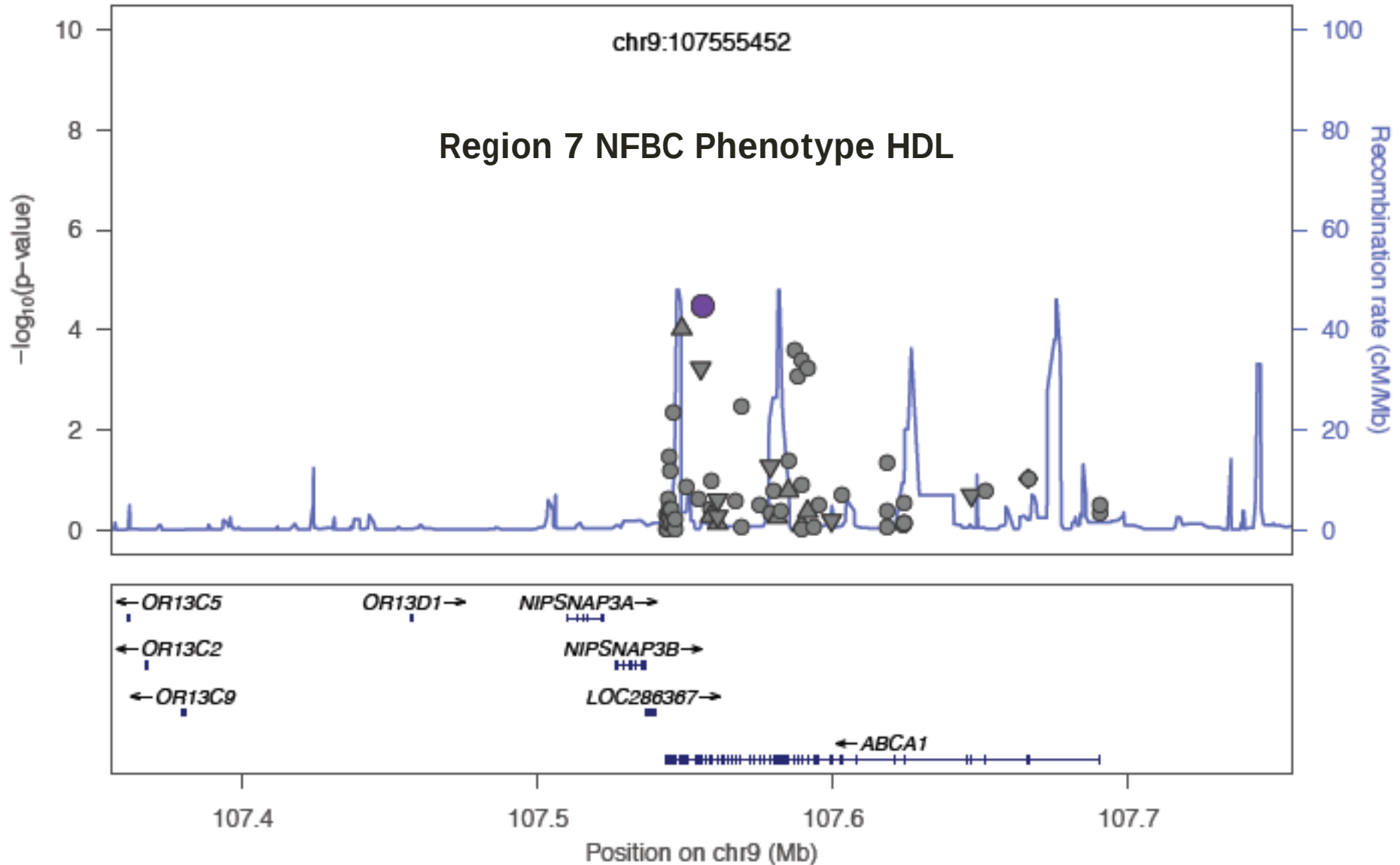
WU Indexed Capture Projects

- ASMS: 0.25 Mb, 7,000 DNAs
- AMD: 1 Mb, 3,400 DNAs
- Arthritis 1: 0.4 Mb, 2,800 DNAs
- Arthritis 2: 1.5 Mb, 2,800 DNAs
- Cleft Lip: 6.6 Mb, 5,600 DNAs

96 Samples 5ug input 2 lanes of data



Targeted sequencing for Metabolic Syndrome



What can be done for \$10M? (Data production)

- Targeted sequencing (indexed hybrid capture)
 - 0.5-4 Mb/100-1,500 genes: ~50,000 samples (~\$200/DNA)
 - 4-8 Mb/1,500-3,000 genes: ~33,000 samples (~\$300/DNA)
- Exome sequencing (commercial reagents, 60 Mb)
 - 10,000 samples (<\$1,000/exome; indexed, 5 DNAs/lane)
- Whole genome sequencing (~30x coverage)
 - 2,000 genomes (~\$5,000/genome)
- Costs include library production, capture & reagents, sequence production, data processing & storage, initial variant detection.
- Costs do not include higher level analyses or validation.





How many samples must be sequenced?

- Definitions:
 - Discovery: detecting at least one occurrence of the variant
 - Recurrency: detecting occurrence in two or more samples
- Given a study size of 1,000:
 - At 1% frequency, a variant is detected essentially with 100% power (discovery and recurrency), as are discovery events at 0.5%
 - At 0.5% frequency, recurrency is detected with ~96% power
 - Very rare events at 0.1% can still be discovered with ~63% power
- Actual power for disease will be somewhat lower, assuming the underlying disease mechanisms act through combinations of events, e.g. in pathways



How many samples do we need to sequence?

