THE GENOME INSTITUTE
at Washington University

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"

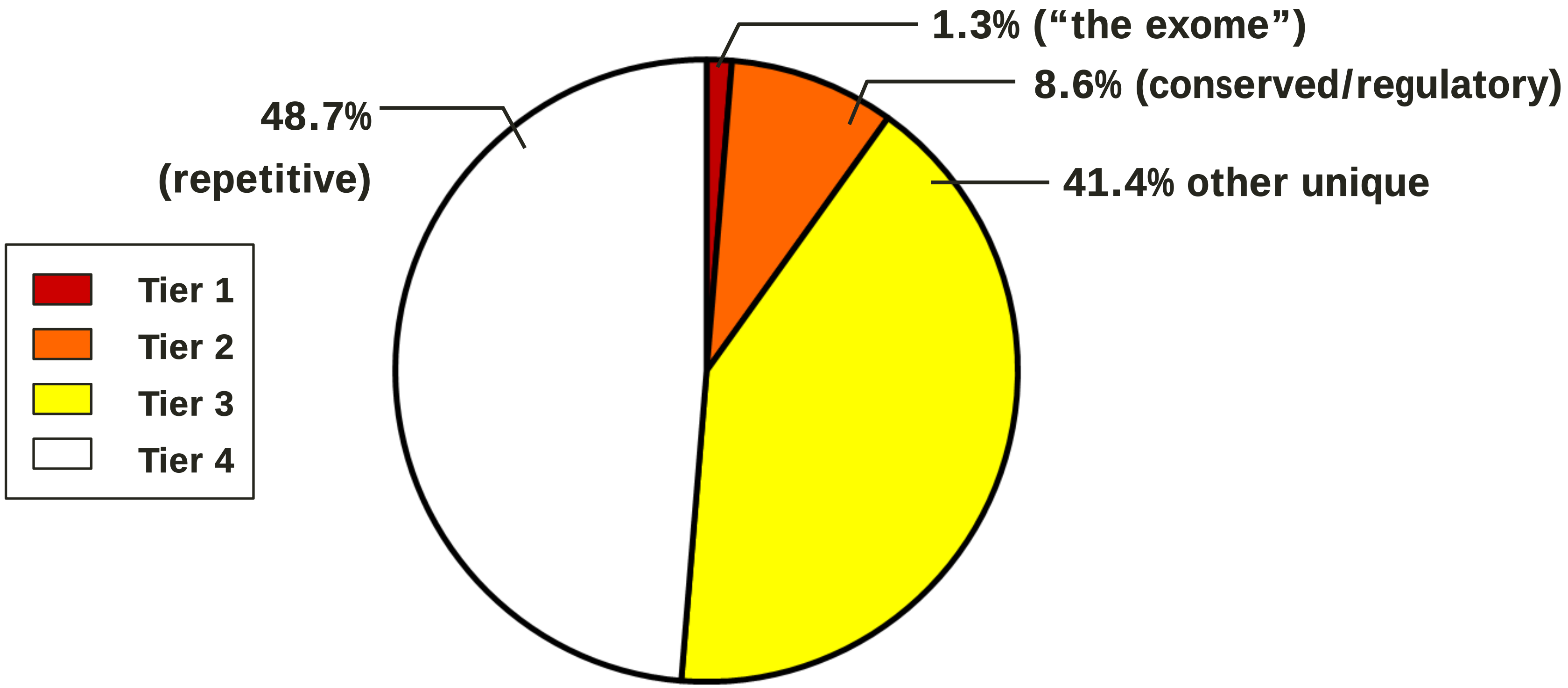
- Caucasian female, mid-50s at diagnosis
- *De novo* M1 AML
- Family history of AML and lymphoma
- 100% blasts in initial BM sample
- Relapsed and died at 23 months
- Normal cytogenetics
- Informed consent for whole genome sequencing
- Solexa sequencer, 32 bp unpaired reads
- 10 Tier 1 somatic mutations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

Ley et al., *Nature* 2008

© R.K.Wilson

Sequencing *and analyzing* a human genome...

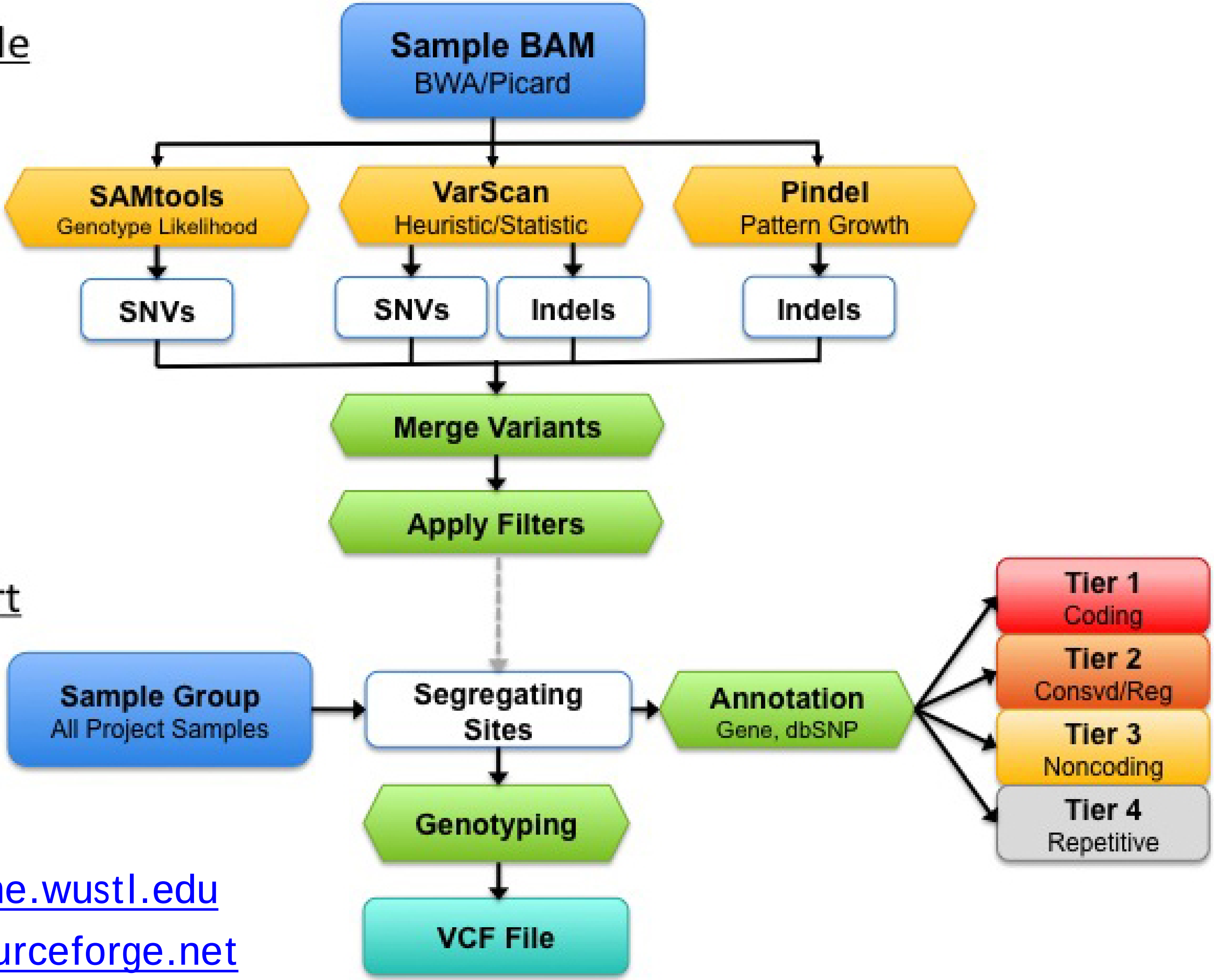
% of the Human Genome in each annotation tier



© R.K.Wilson

Variant detection in individuals and cohorts

Per Sample



Sample BAM
BWA/Picard

SAMtools
Genotype Likelihood

VarScan
Heuristic/Statistic

Pindel
Pattern Growth

SNVs

SNVs

Indels

Indels

Merge Variants

Apply Filters

Per Cohort

Sample Group
All Project Samples

Segregating Sites

Annotation
Gene, dbSNP

Tier 1
Coding

Tier 2
Consvd/Reg

Tier 3
Noncoding

Tier 4
Repetitive

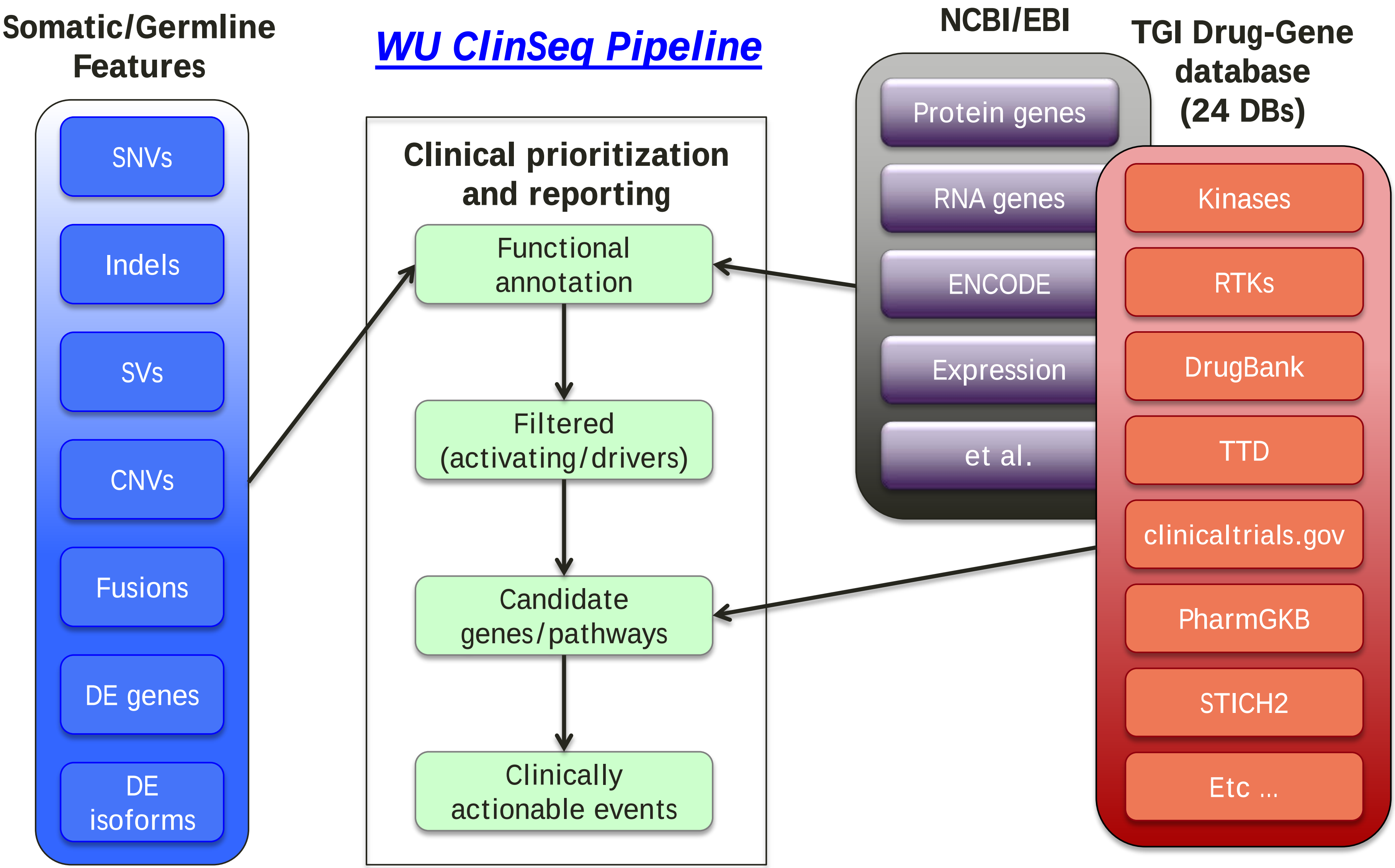
Genotyping

VCF File

gmt.genome.wustl.edu

varscan.sourceforge.net

A comprehensive genome analysis pipeline



Somatic/Germline Features

- SNVs
- Indels
- SVs
- CNVs
- Fusions
- DE genes
- DE isoforms

WU ClinSeq Pipeline

Clinical prioritization and reporting

Functional annotation → Filtered (activating/drivers) → Candidate genes/pathways → Clinically actionable events

NCBI/EBI

- Protein genes
- RNA genes
- ENCODE
- Expression
- et al.

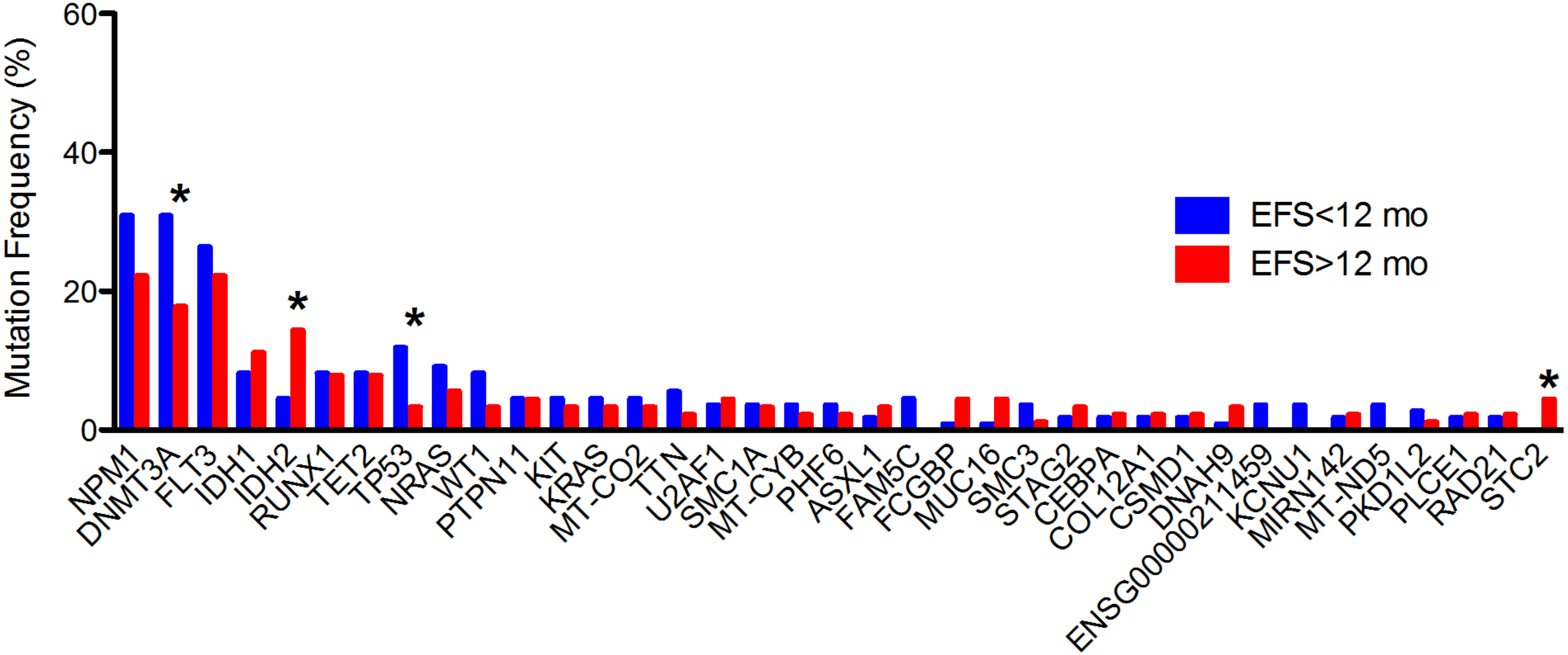
TGI Drug-Gene database (24 DBs)

- Kinases
- RTKs
- DrugBank
- TTD
- clinicaltrials.gov
- PharmGKB
- STICH2
- Etc ...

© R.K.Wilson

M. Griffith

Somatic mutations in 200 AML patients

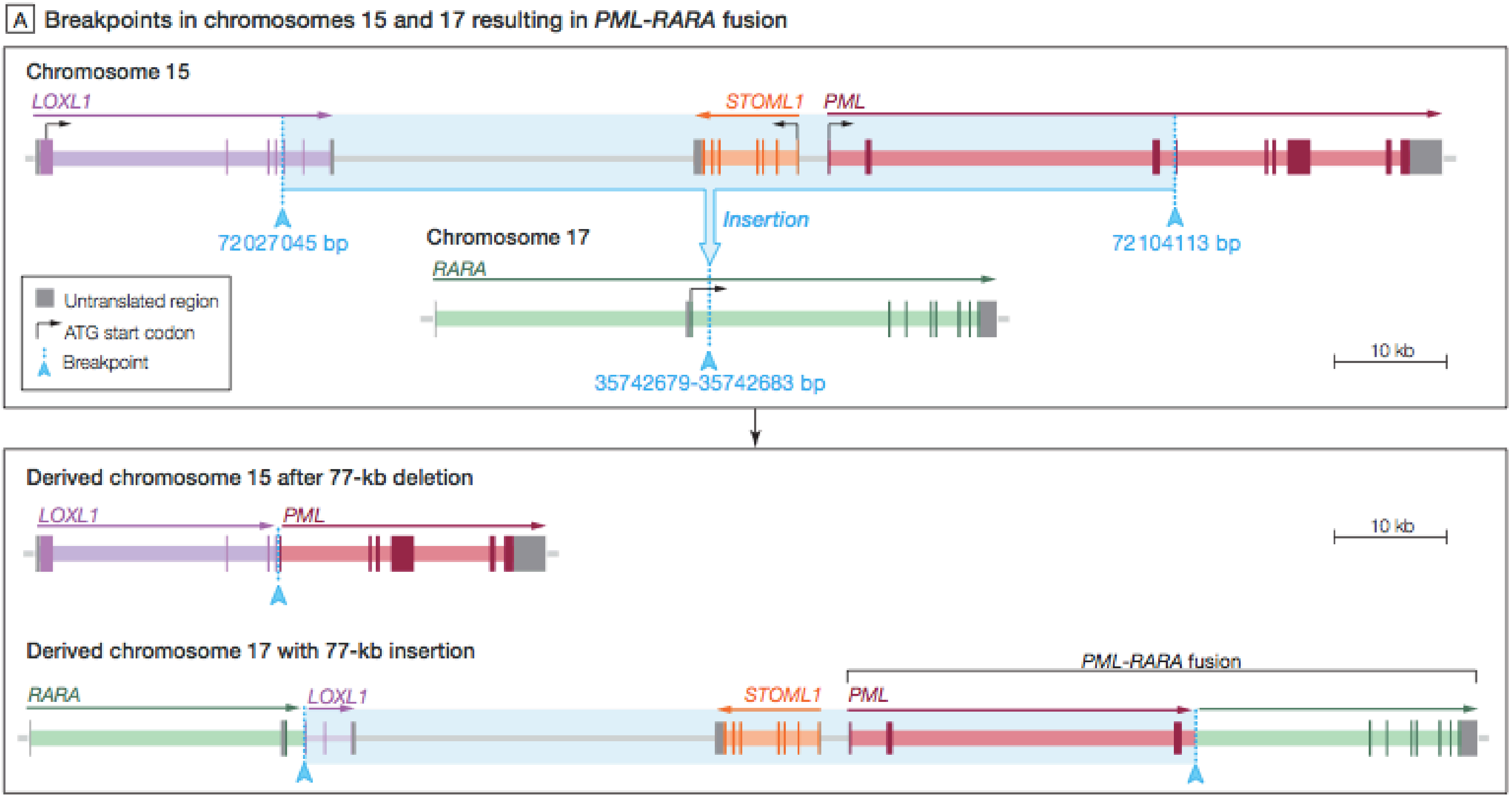


Welch et al., in preparation

JAMA®
The Journal of the American Medical Association

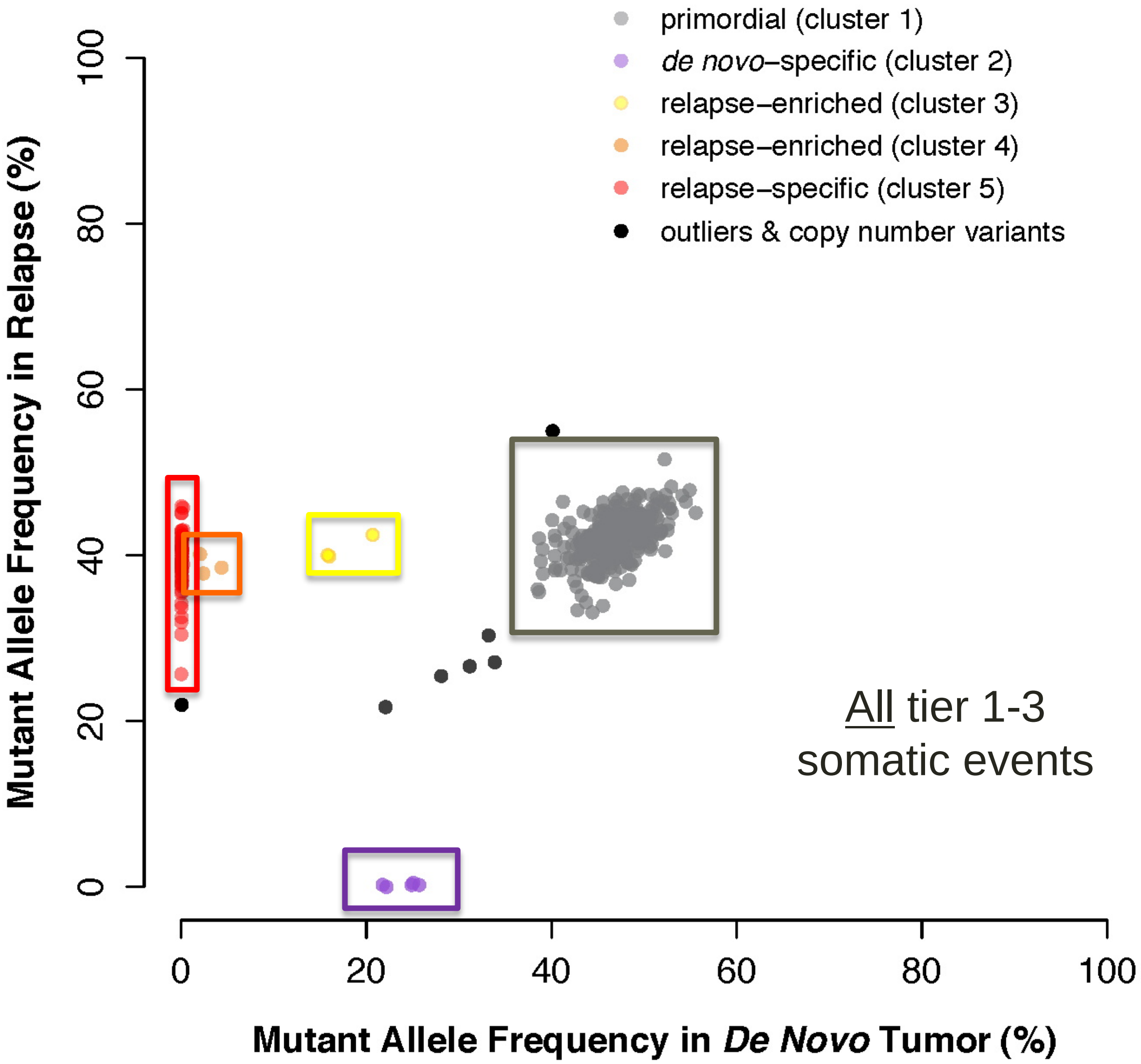
JAMA. 2011;305(15):1577-1584. doi: 10.1001/jama.2011.497

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



Welch et al., *JAMA* 2011

Deep digital sequencing in patient AML1 (relapse)



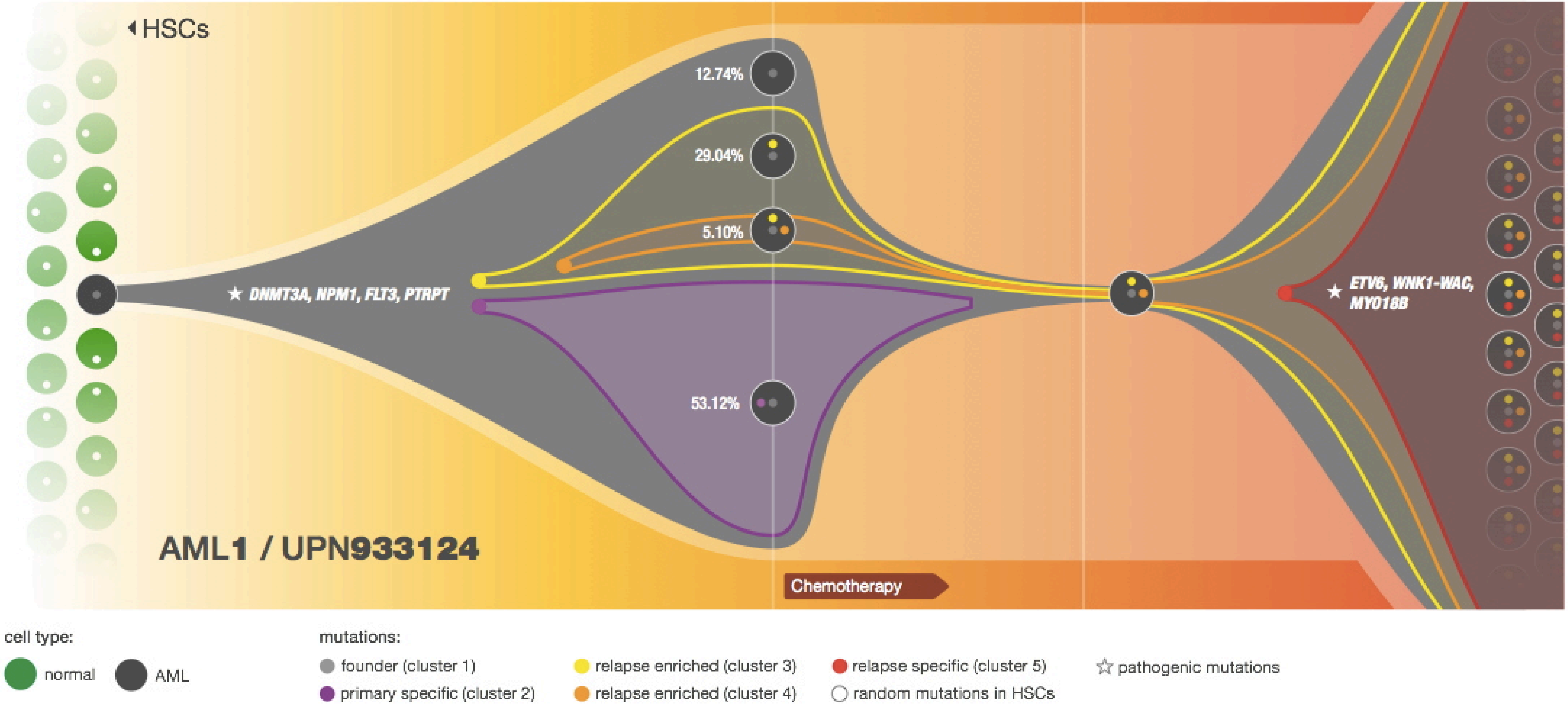
All tier 1-3 somatic events

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


HSCs
12.74%
29.04%
5.10%
53.12%
★ DNMT3A, NPM1, FLT3, PTRPT
★ ETV6, WNK1-WAC, MYO18B
AML1 / UPN933124
Chemotherapy
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

© R.K.Wilson

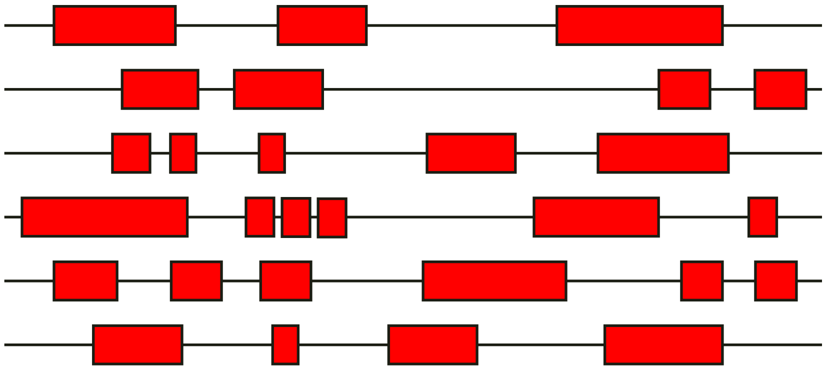
Ding et al., *Nature* 2012

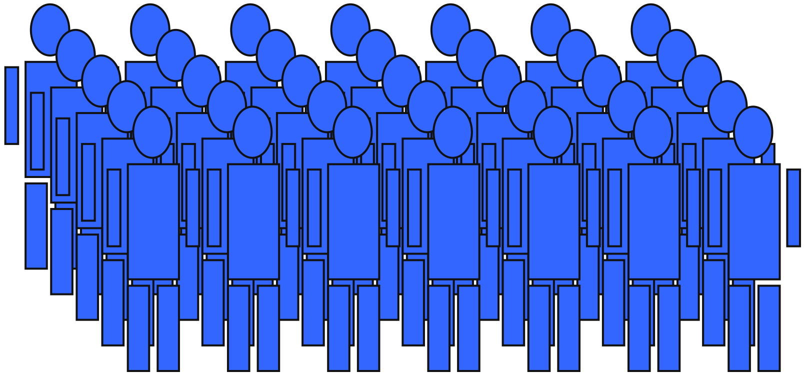
Genomic opportunities in large cohorts

- Sequencing options…

Targeted sequencing (hybrid capture)

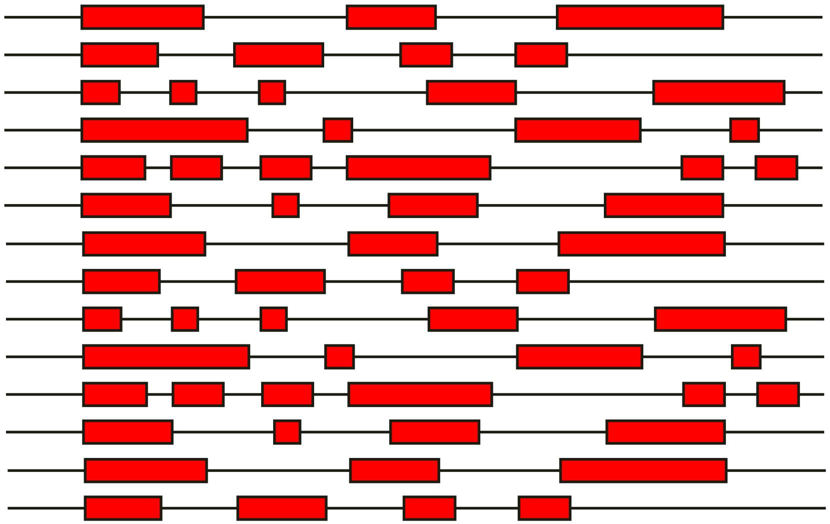
list of candidate
genes and/or
regions of interest
(e.g. GWAS peaks)



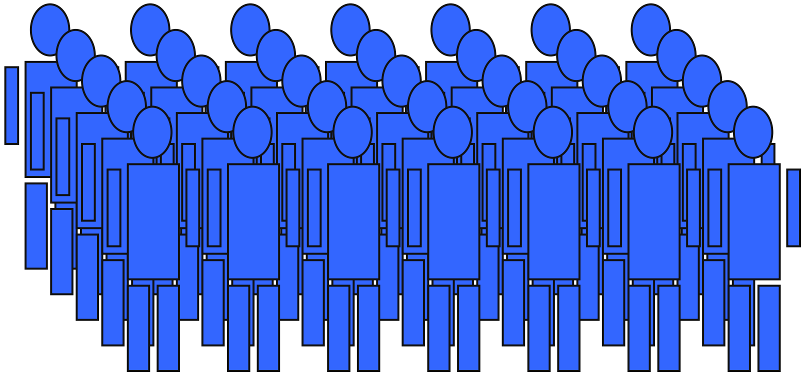


large collection of patient
samples

Exome sequencing (hybrid capture)

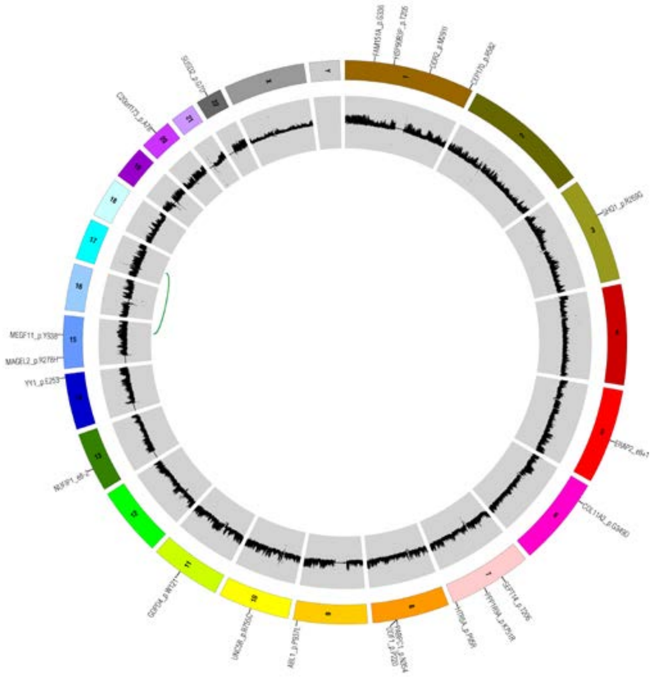


Ideally all CCDS exons & selected RNA genes

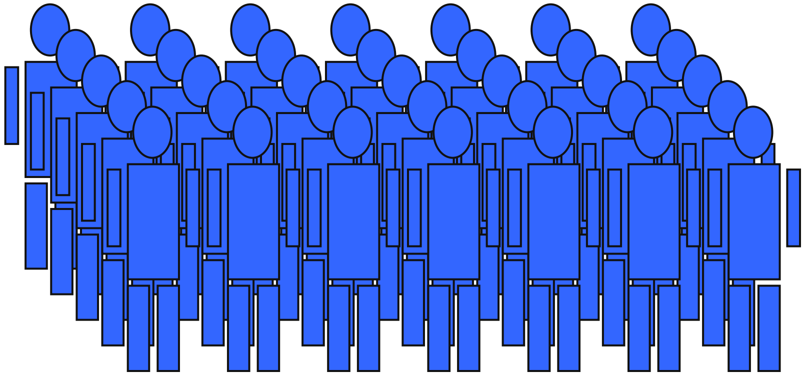


large collection of patient samples

Whole genome sequencing



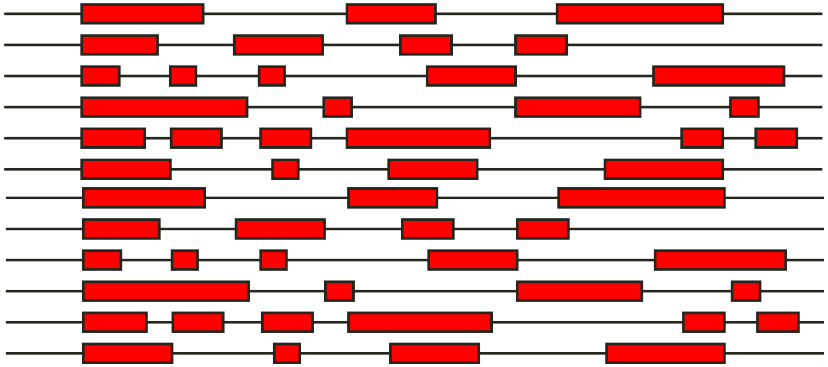
complete genome sequences aligned to reference HGS



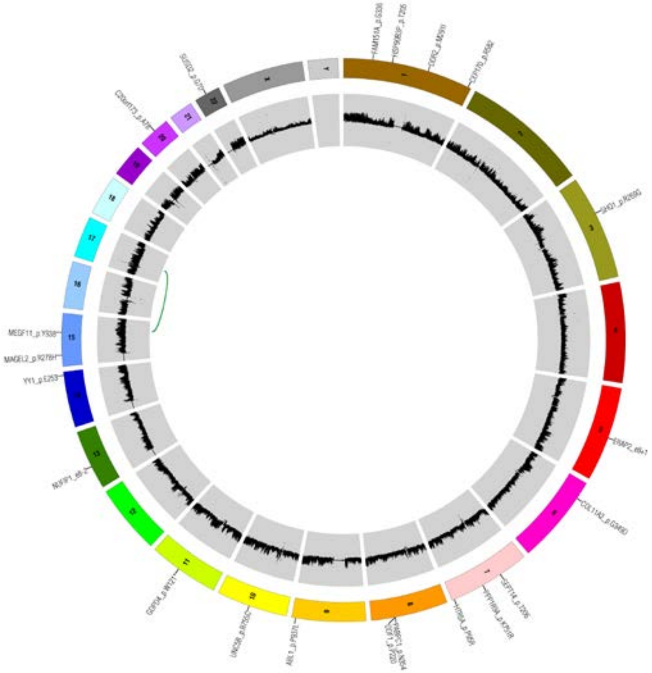
large collection of patient samples

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

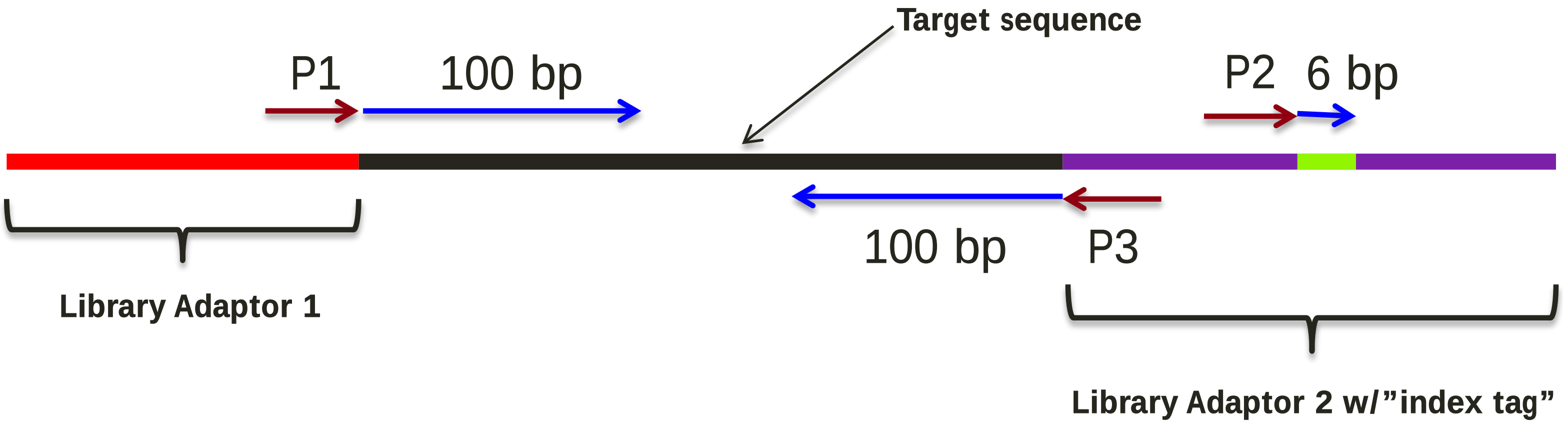
Exome sequencing reagents (relative to “WuSpace”)

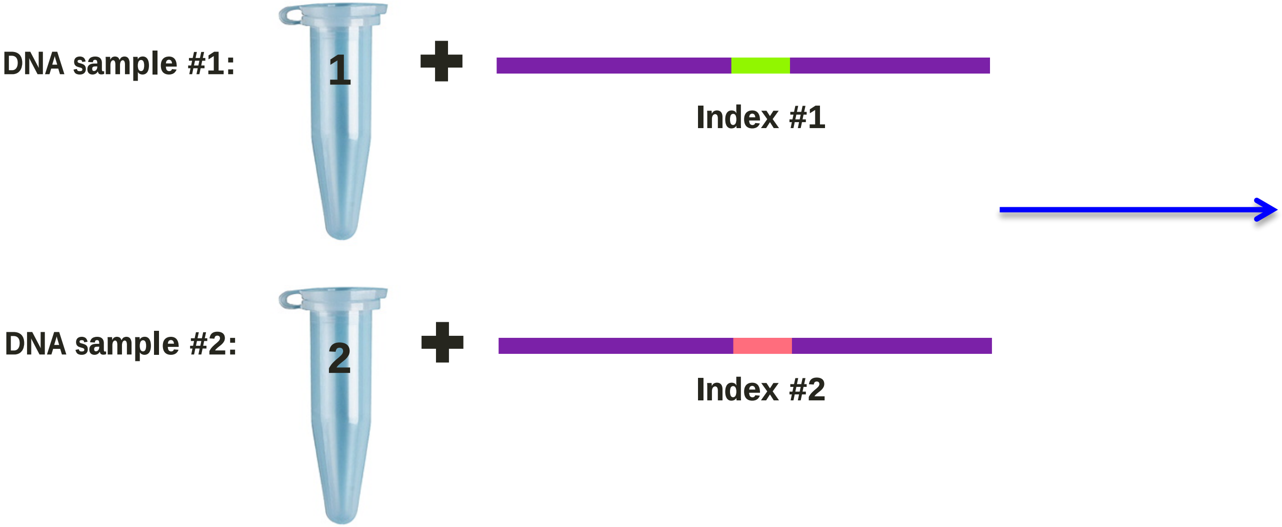
	% Product Unique	% Product Shared	% CDS Not Targeted	% CDS Targeted
NimbleGen v2 (35.9 Mb)	8.3%	91.7%	30.1%	69.9%
NimbleGen v3 (63.6 Mb)	42.2%	57.8%	22.2%	77.8%
Agilent SS 50Mb (51.5 Mb)	32.1%	67.9%	25.9%	74.1%
Illumina TruSeq v1 (62.1 Mb)	42.5%	57.5%	24.4%	75.6%

- 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.

T. Wylie, J. Walker

Multiplexed DNA sequencing (“indexed”)





Library construction

Pool samples

Hybrid capture

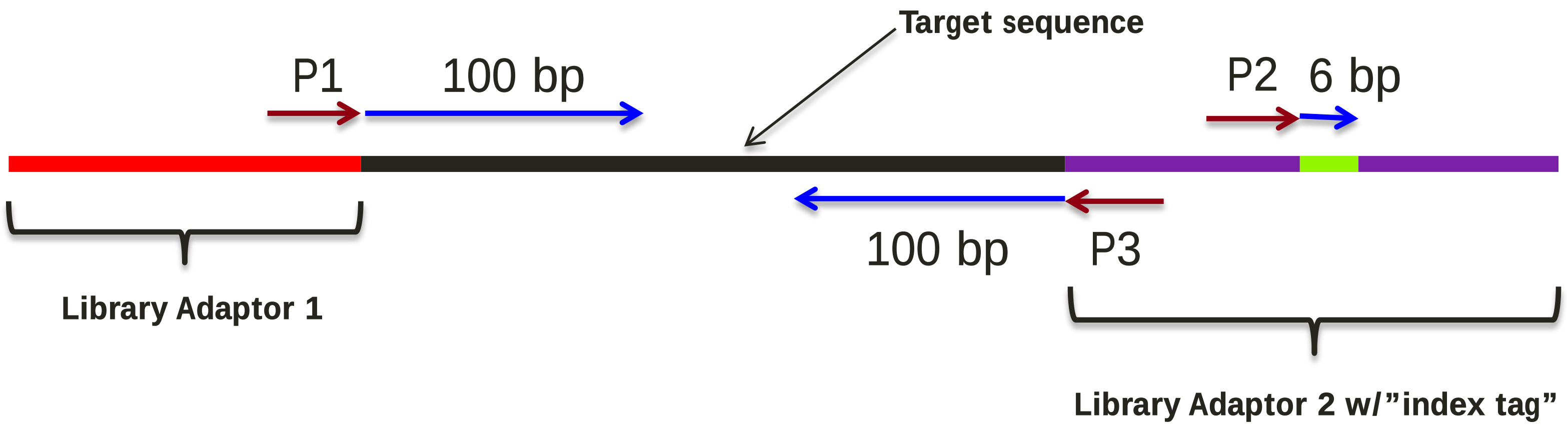
DNA sequencing

De-multiplex

Align to reference

Variant detection

Multiplexed DNA sequencing ("indexed")



Target sequence

P1 100 bp

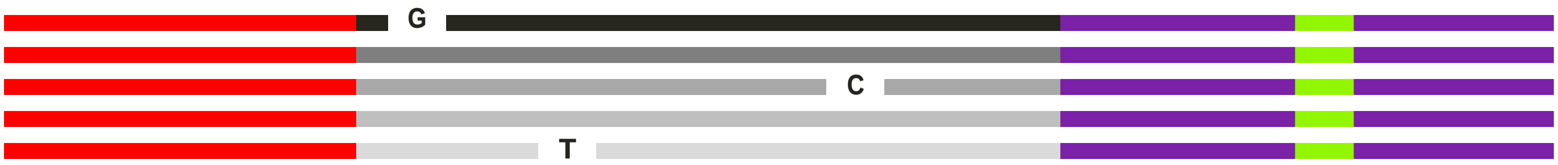
P2 6 bp

100 bp P3

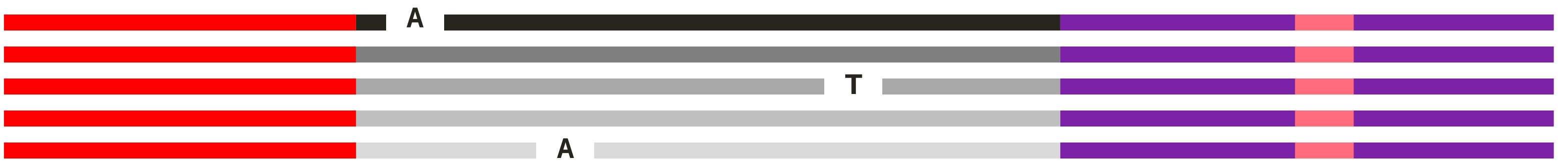
Library Adaptor 1

Library Adaptor 2 w/"index tag"

DNA sample #1:



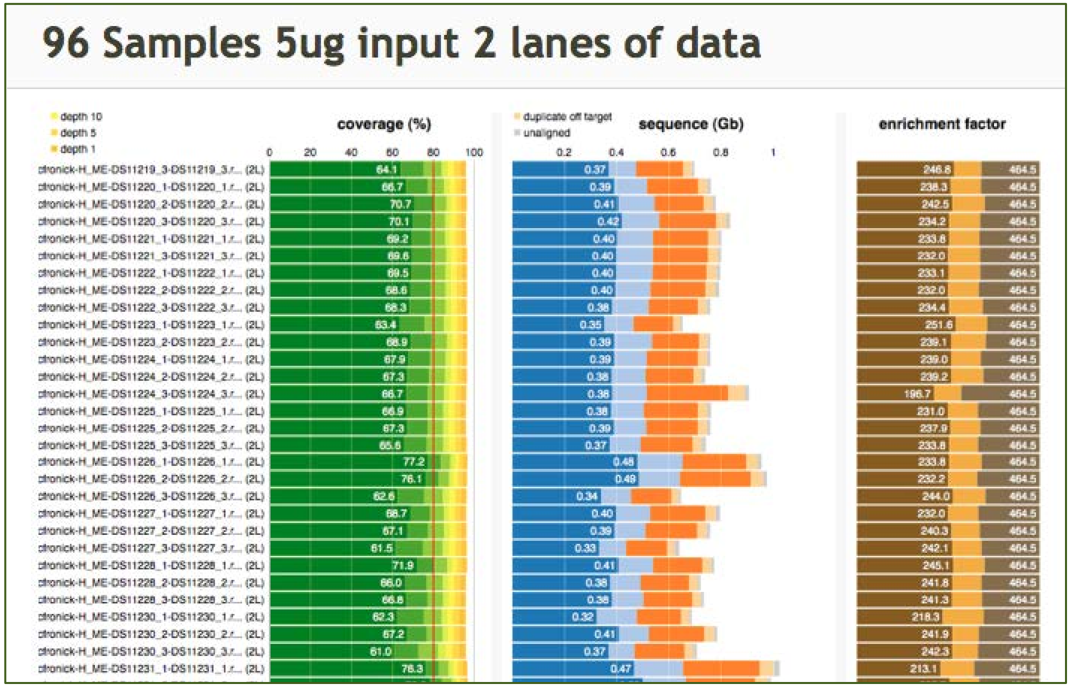
DNA sample #2:



Multiplexed DNA sequencing (“indexed”)

96 Indexed
DNA samples

Capture
probes



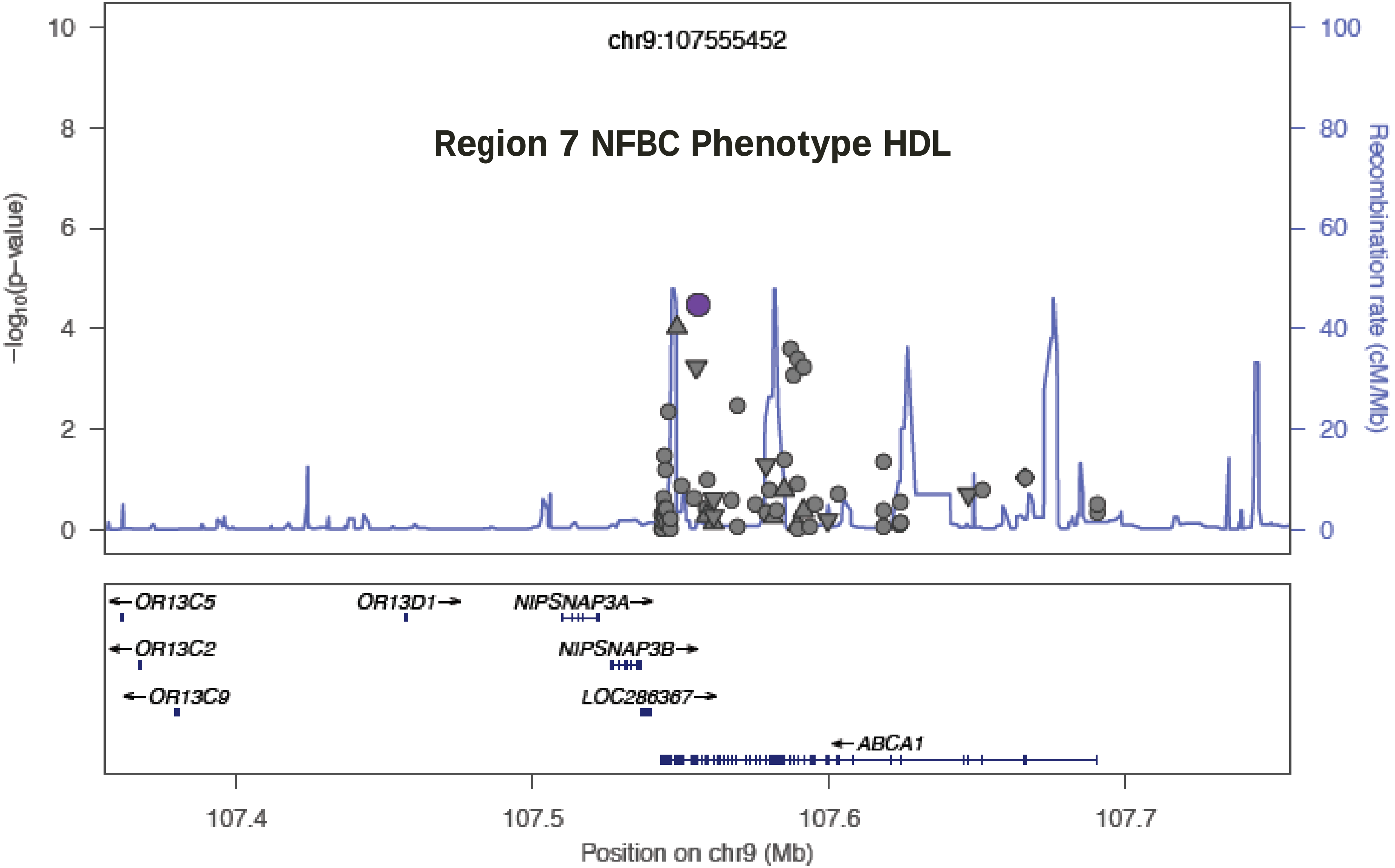
96 Samples 5ug input 2 lanes of data

	coverage (%)	sequence (Gb)	enrichment factor	
ctronick-H_ME-DS11219_3-DS11219_3.r... (2L)	64.1	0.37	246.8	464.5
ctronick-H_ME-DS11220_1-DS11220_1.r... (2L)	66.7	0.39	238.3	464.5
ctronick-H_ME-DS11220_2-DS11220_2.r... (2L)	70.7	0.41	242.5	464.5
ctronick-H_ME-DS11220_3-DS11220_3.r... (2L)	70.1	0.42	234.2	464.5
ctronick-H_ME-DS11221_1-DS11221_1.r... (2L)	69.2	0.40	233.8	464.5
ctronick-H_ME-DS11221_3-DS11221_3.r... (2L)	69.6	0.40	232.0	464.5
ctronick-H_ME-DS11222_1-DS11222_1.r... (2L)	69.5	0.40	233.1	464.5
ctronick-H_ME-DS11222_2-DS11222_2.r... (2L)	68.6	0.40	232.0	464.5
ctronick-H_ME-DS11222_3-DS11222_3.r... (2L)	68.3	0.38	234.4	464.5
ctronick-H_ME-DS11223_1-DS11223_1.r... (2L)	63.4	0.35	251.6	464.5
ctronick-H_ME-DS11223_2-DS11223_2.r... (2L)	68.9	0.39	239.1	464.5
ctronick-H_ME-DS11224_1-DS11224_1.r... (2L)	67.9	0.39	239.0	464.5
ctronick-H_ME-DS11224_2-DS11224_2.r... (2L)	67.3	0.38	239.2	464.5
ctronick-H_ME-DS11224_3-DS11224_3.r... (2L)	66.7	0.38	196.7	464.5
ctronick-H_ME-DS11225_1-DS11225_1.r... (2L)	66.9	0.38	231.0	464.5
ctronick-H_ME-DS11225_2-DS11225_2.r... (2L)	67.3	0.39	237.9	464.5
ctronick-H_ME-DS11225_3-DS11225_3.r... (2L)	65.6	0.37	233.8	464.5
ctronick-H_ME-DS11226_1-DS11226_1.r... (2L)	77.2	0.48	233.8	464.5
ctronick-H_ME-DS11226_2-DS11226_2.r... (2L)	76.1	0.49	232.2	464.5
ctronick-H_ME-DS11226_3-DS11226_3.r... (2L)	62.6	0.34	244.0	464.5
ctronick-H_ME-DS11227_1-DS11227_1.r... (2L)	68.7	0.40	232.0	464.5
ctronick-H_ME-DS11227_2-DS11227_2.r... (2L)	67.1	0.39	240.3	464.5
ctronick-H_ME-DS11227_3-DS11227_3.r... (2L)	61.5	0.33	242.1	464.5
ctronick-H_ME-DS11228_1-DS11228_1.r... (2L)	71.9	0.41	245.1	464.5
ctronick-H_ME-DS11228_2-DS11228_2.r... (2L)	66.0	0.38	241.8	464.5
ctronick-H_ME-DS11228_3-DS11228_3.r... (2L)	66.8	0.38	241.3	464.5
ctronick-H_ME-DS11230_1-DS11230_1.r... (2L)	62.3	0.32	218.3	464.5
ctronick-H_ME-DS11230_2-DS11230_2.r... (2L)	67.2	0.41	241.9	464.5
ctronick-H_ME-DS11230_3-DS11230_3.r... (2L)	61.0	0.37	242.3	464.5
ctronick-H_ME-DS11231_1-DS11231_1.r... (2L)	78.3	0.47	213.1	464.5

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

Targeted sequencing for Metabolic Syndrome



S. Service

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

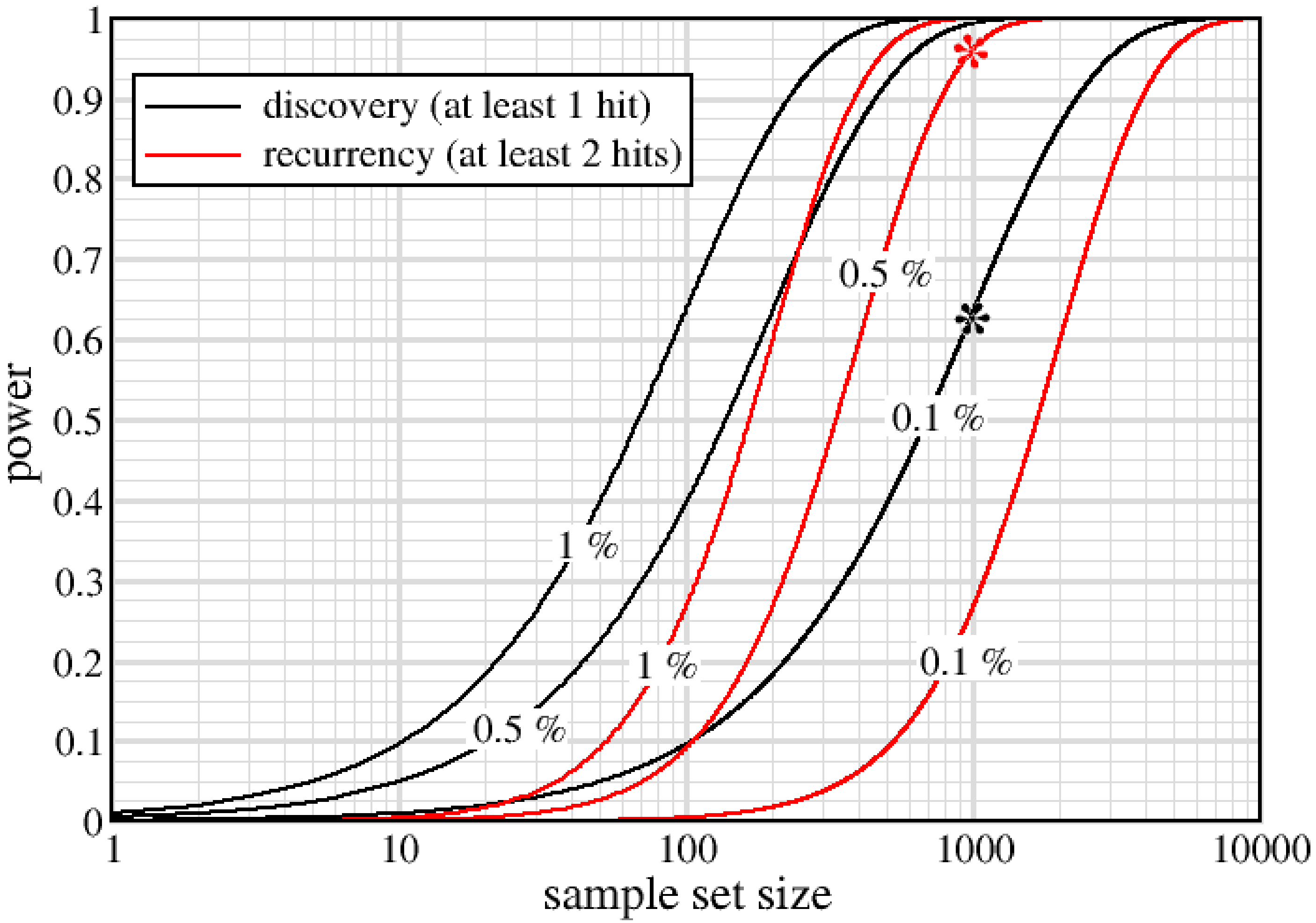
- Targeted sequencing (indexed hybrid capture)
  - 0.5-4 Mb/100-1,500 genes: ~50,000 samples (~$200/DNA)
  - 4-8 Mb/1,500-3,000 genes: ~33,000 samples (~$300/DNA)
- Exome sequencing (commercial reagents, 60 Mb)
  - 10,000 samples (<$1,000/exome; indexed, 5 DNAs/lane)
- Whole genome sequencing (~30x coverage)
  - 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

M. Wendl

How many samples do we need to sequence?



© R.K.Wilson