SCD: Sequencing Options

Richard K. Wilson, Ph.D.
Professor of Genetics
Director, The Genome Institute

SCD: WU proposals (2010-11)

- Ley & Townes (NHLBI RFA: triaged)
  - Clinical phenotyping, banking (skin) of ~25 patients. Initial focus on families with sibs having dramatically different outcomes.
  - Perform WGS on all patients, correlate phenome:genome findings to establish genes/markers associated with severe disease.
  - Generate iPS lines for all patients, perform WGS before & after gene correction with homologous recombination (i.e. determine effectiveness, safety of an iPS approach to SCD).
- Ley, Townes & Wilson (NHGRI grant: CIP pending)
  - Clinical phenotyping, banking (skin) of ~1000 patients.
  - Phase 1: WGS of families with sibs of dramatically different outcomes, correlate phenome:genome findings to establish genes/markers associated with severe disease.
  - Phase 2: WGS of up to 1000 SCD patients.

© R.K.Wilson

Sequencing a human genome…

"Old technology"

Applied Biosystems 3730xl (2004)

$15,000,000

2-3 years

"Next-gen technology"

Illumina HiSeq (Dec 2011)

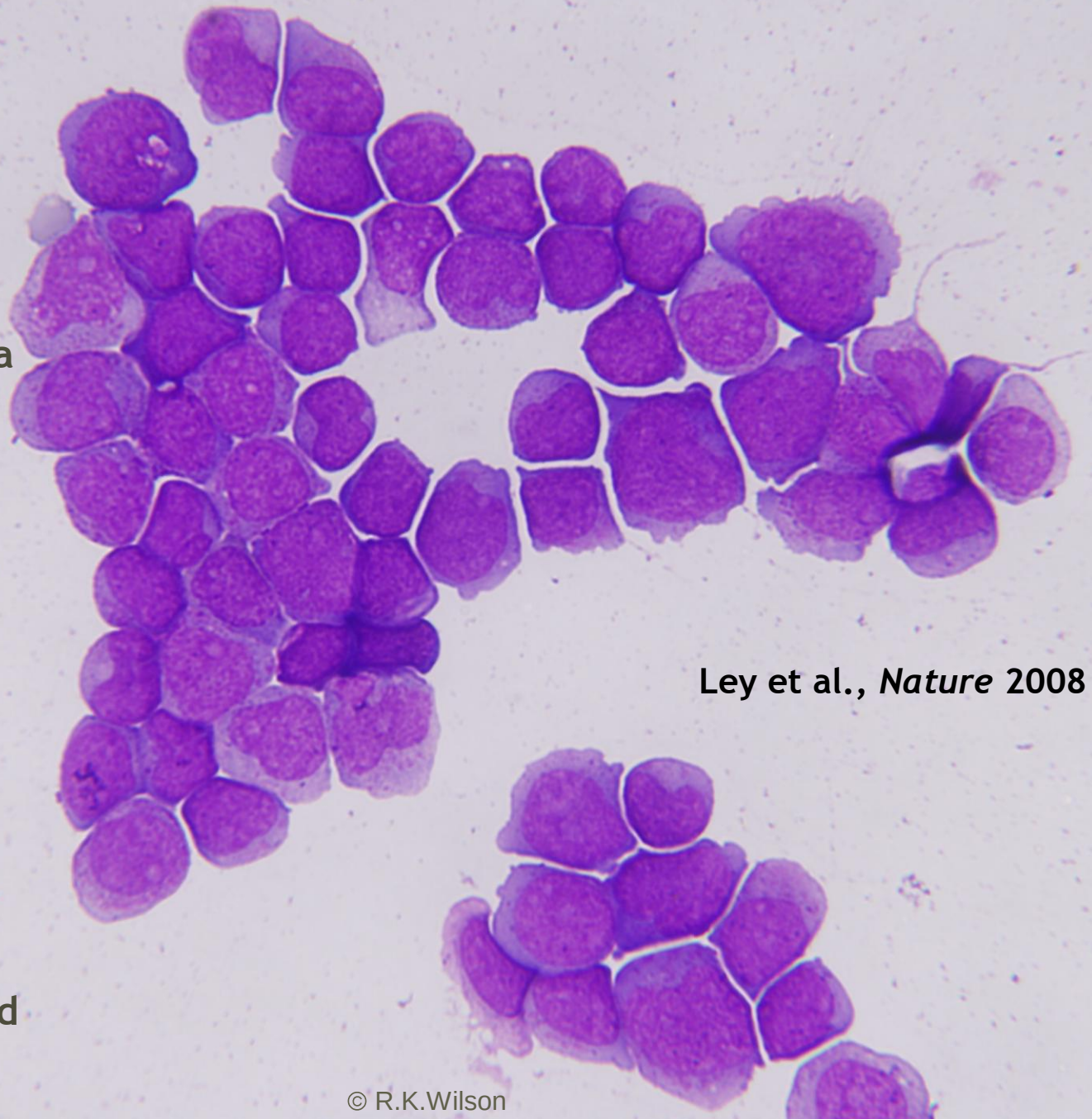
$10,000

2-3 weeks

© R.K.Wilson

"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 somatic mutations detected

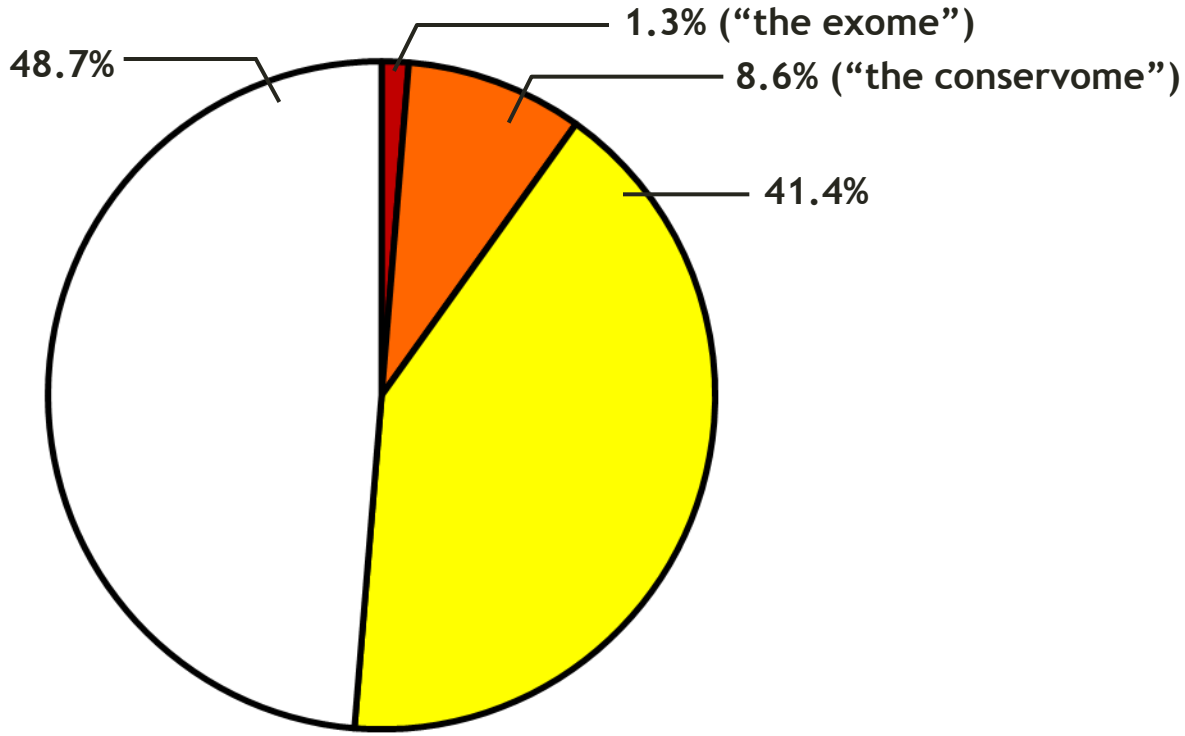


Ley et al., *Nature* 2008

© R.K.Wilson

Sequencing *and analyzing* a human genome…

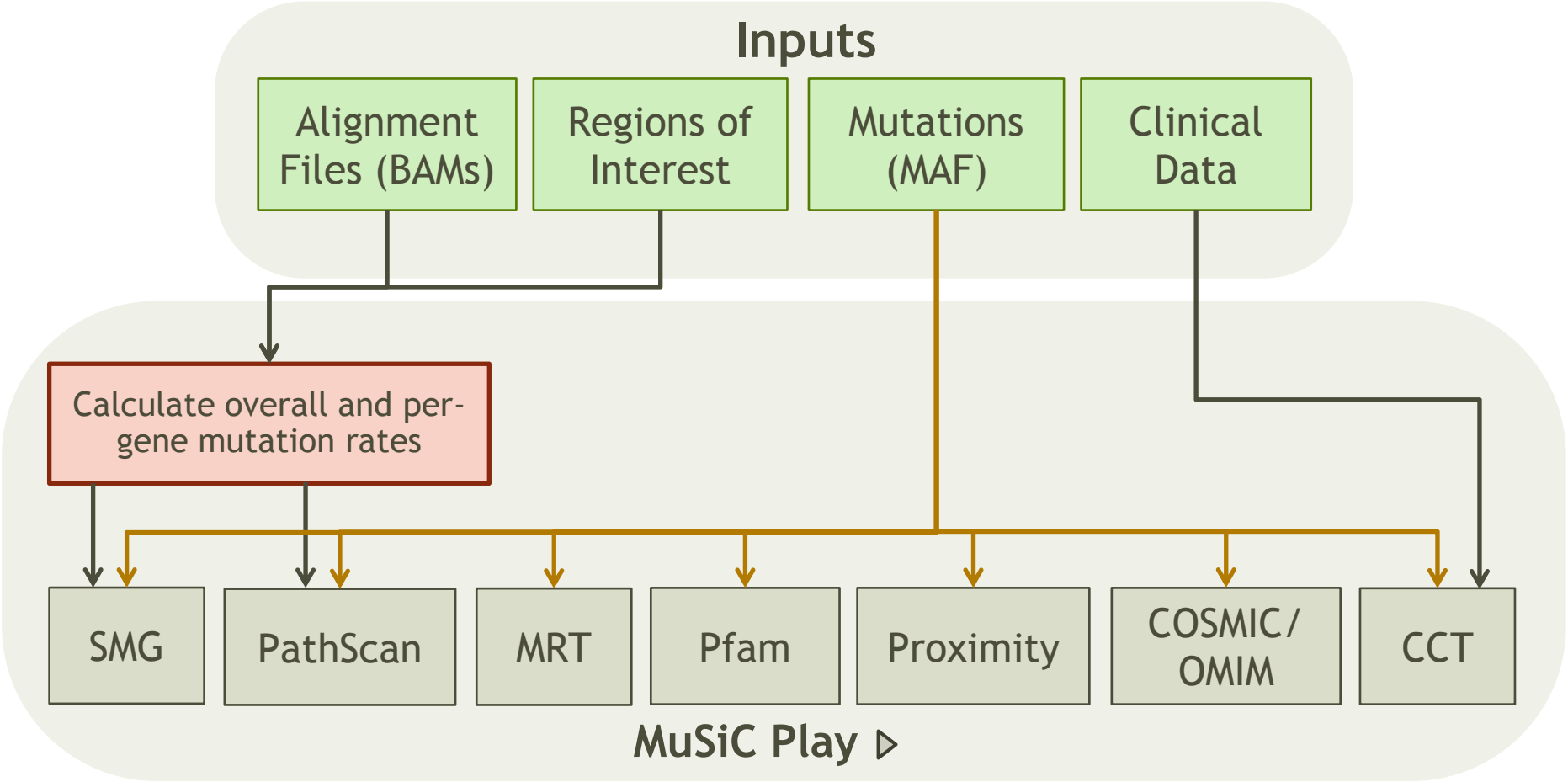
% of the Human Genome in each annotation tier



- Tier 1: WuSpace
- Tier 2: mammalian conserved
- Tier 3: non-repetitive
- Tier 4: repetitive

© R.K.Wilson

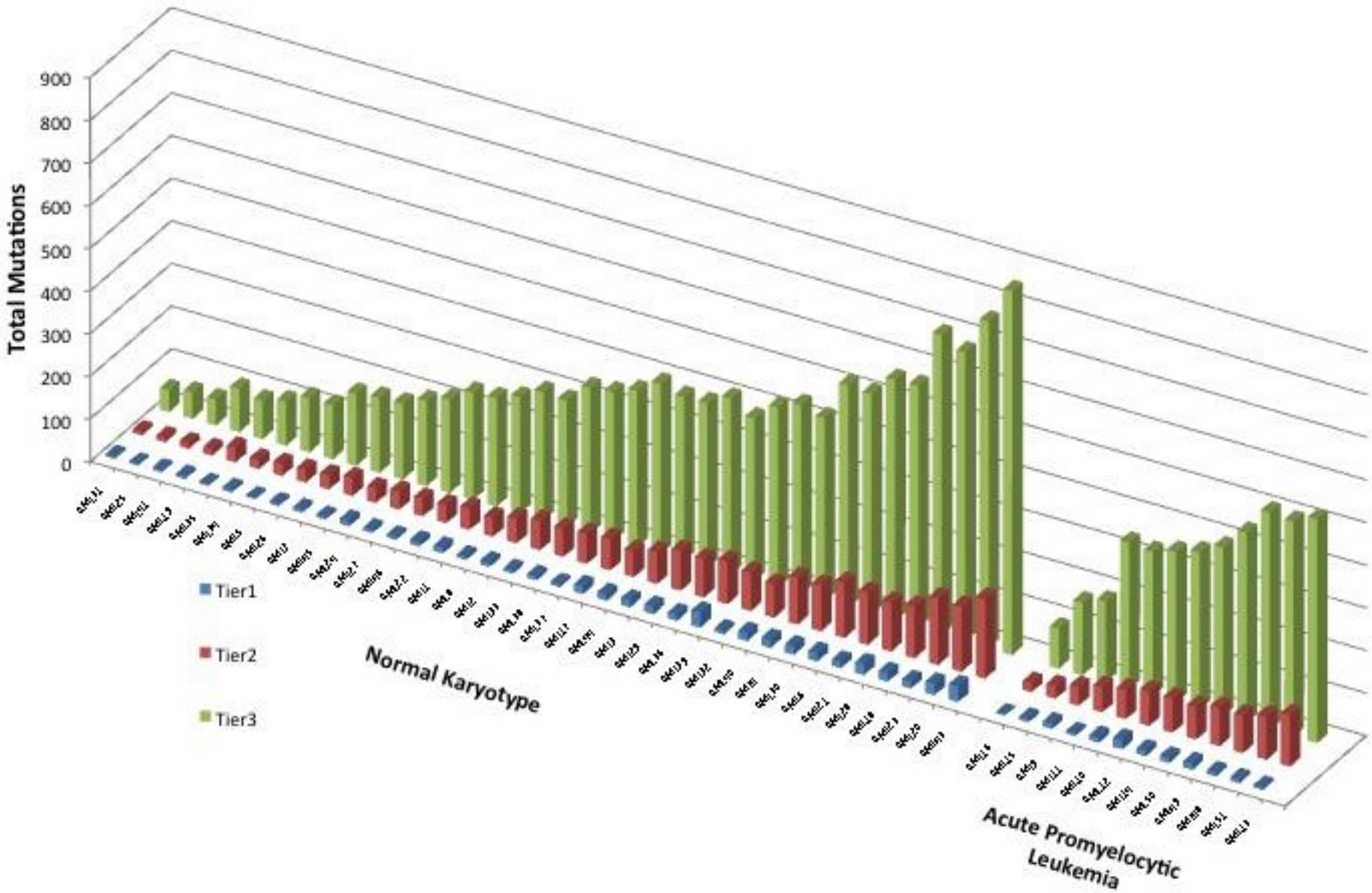
Analysis and Discovery - The MuSiC Suite



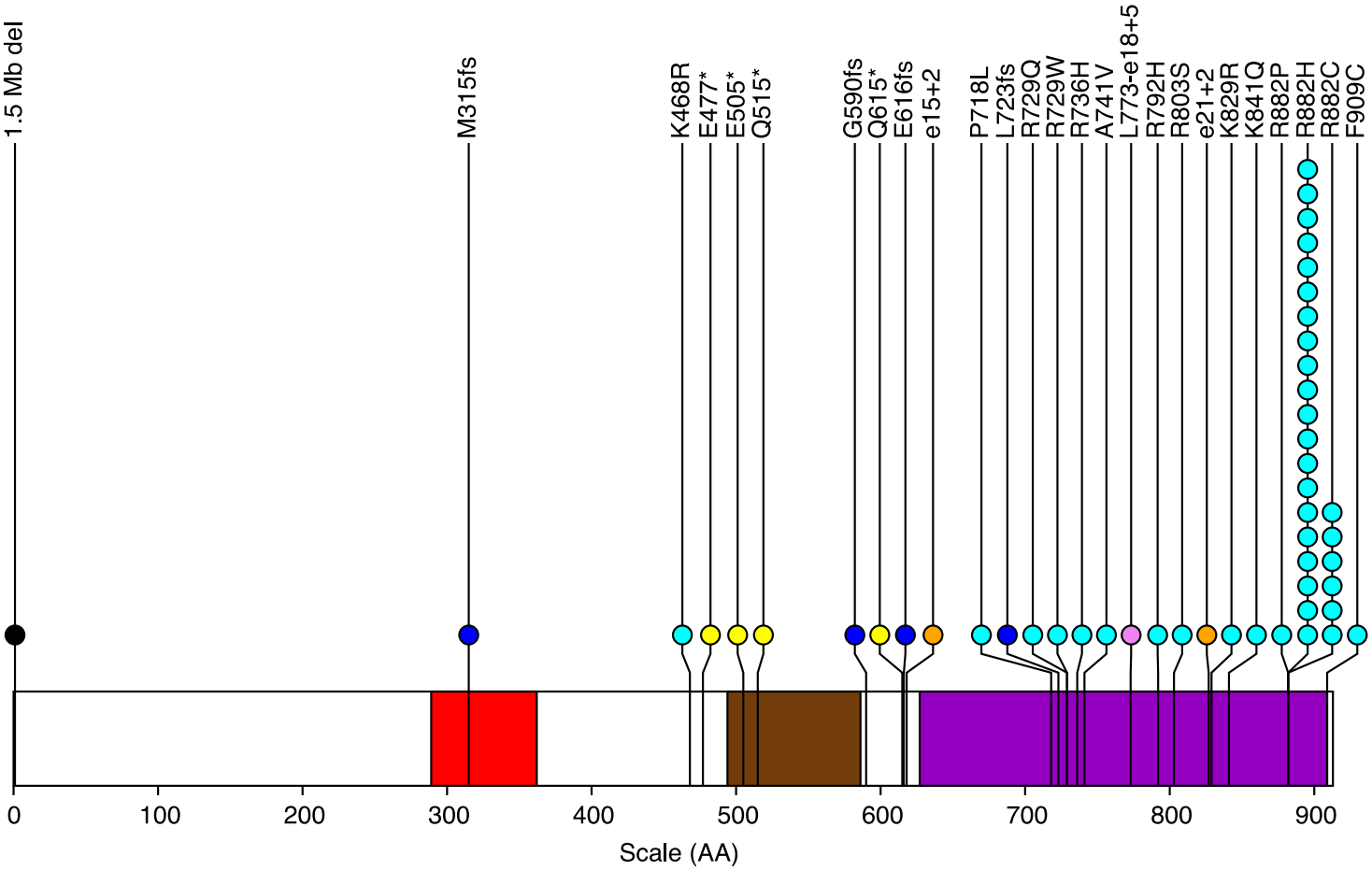
- MuSiC (Mutational Significance in Cancer) is a suite of statistical tools that can also run as a fully automated downstream analysis pipeline.
- Available at: **http://gmt.genome.wustl.edu/genome-music/**

© R.K.Wilson

Genome-wide somatic mutations in 50 AML patients



© R.K.Wilson



DNMT3A

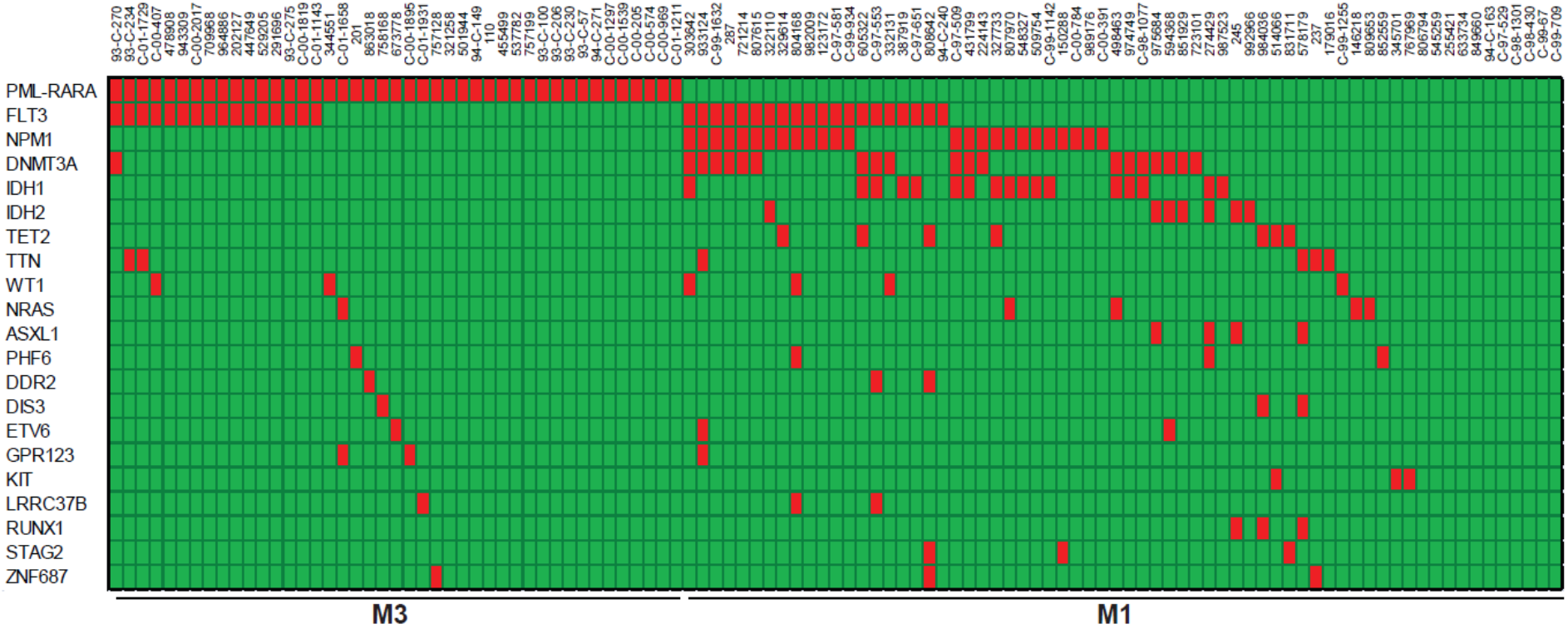
- frame shift deletion
- missense
- nonsense
- splice site substitution
- splice site deletion
- whole gene deletion

- MTase
- PHD
- PWWP

- ***DNMT3A*** **mutations are present in 22% of *de novo* AMLs, and 34% of cytogenetically normal patients.**
- ***DNMT3A*** **mutations are strongly associated high-risk AML.**

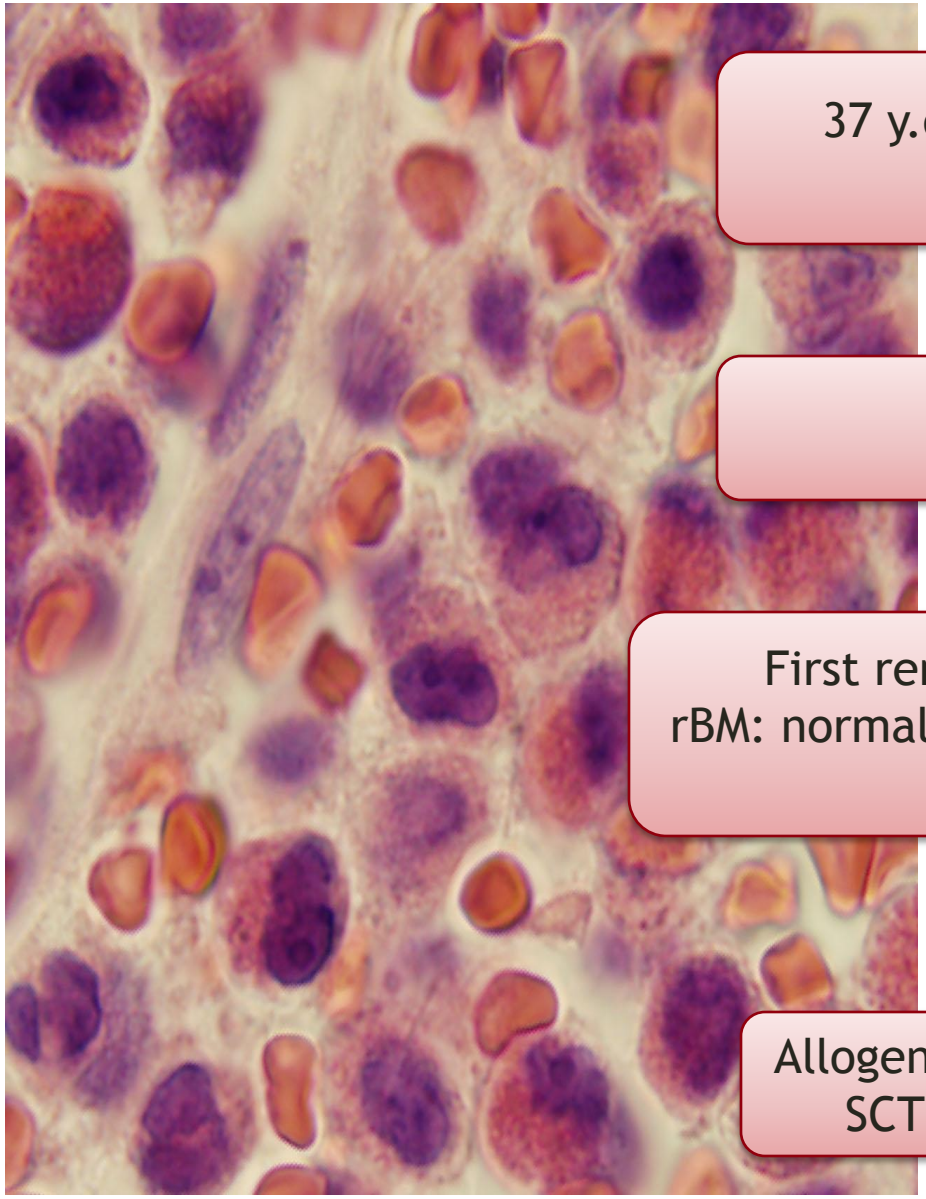
© R.K.Wilson

The emerging genomic landscape of AML



© R.K.Wilson

Structural Variation: WGS in a 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

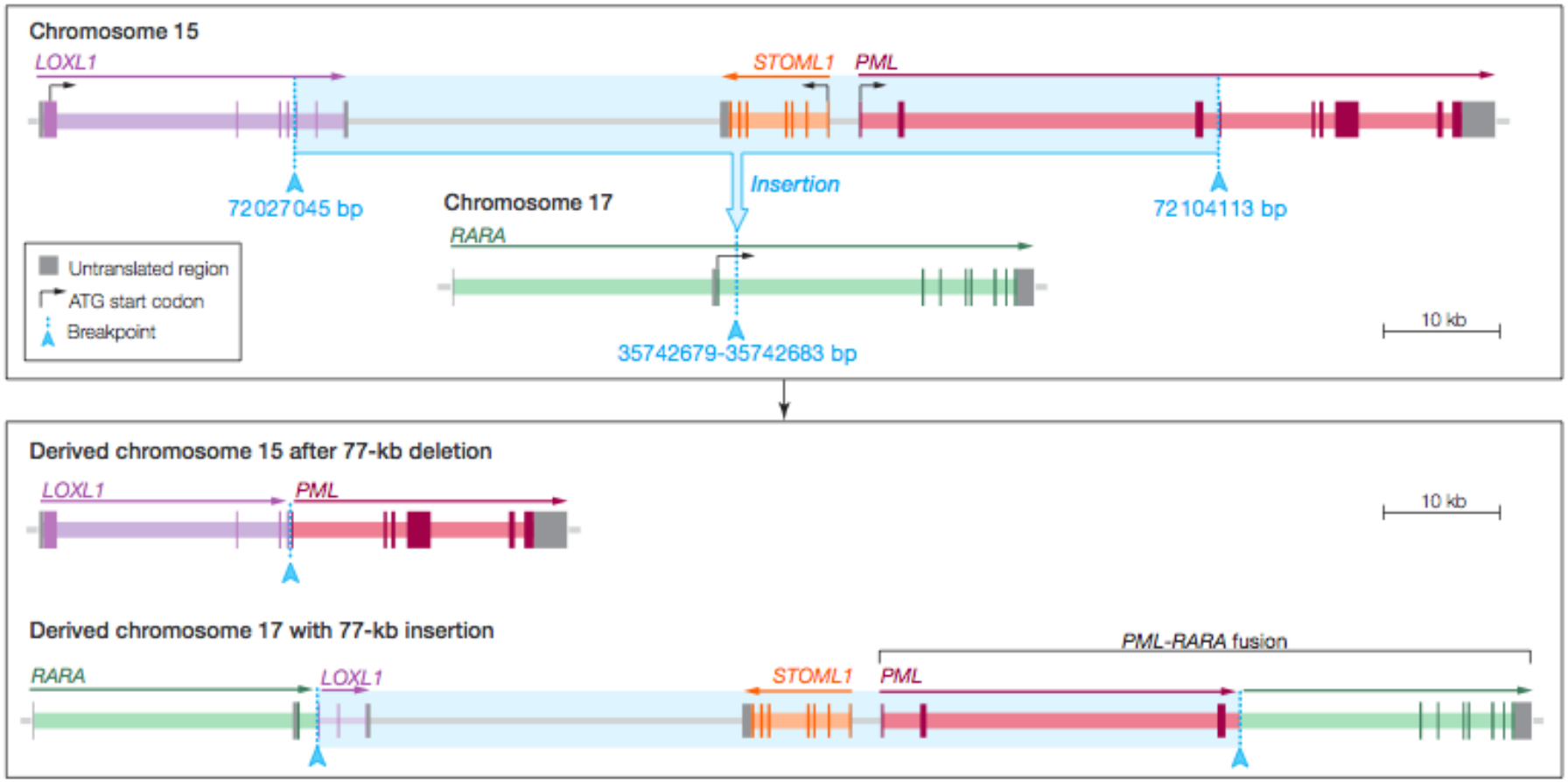
??? ↘ Consolidation + ATRA

© R.K.Wilson

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

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



Welch et al., JAMA April 20, 2011

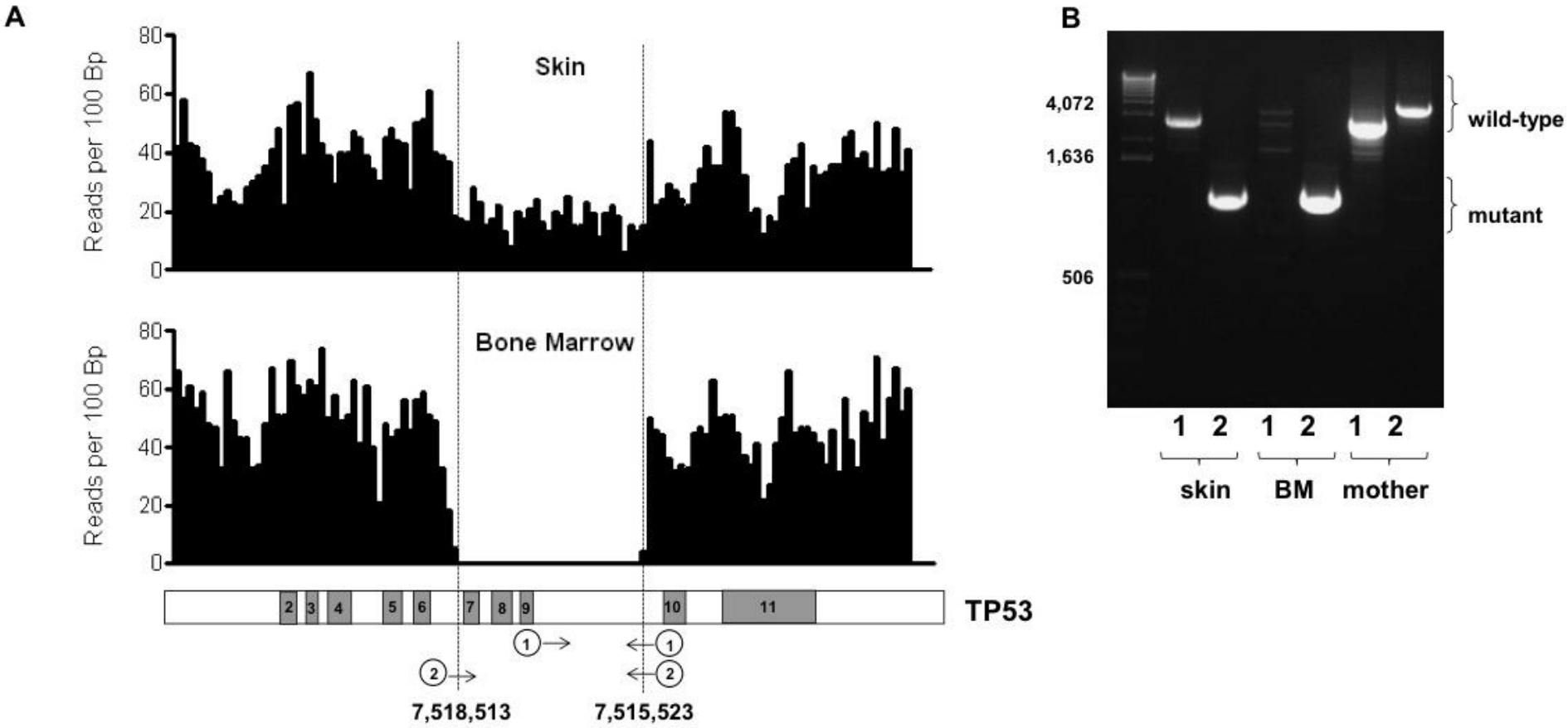
© R.K.Wilson

Focal deletion: WGS in a clinical case (tAML1)

- 37 y.o. female presented with T2N1 breast cancer ER/PR/Her2+. Rx with MRM, ACE chemotherapy and local radiotherapy. BRCA1 and BRCA2 status normal.
- At age 39: Stage III-C ovarian cancer diagnosed. Rx with TAHBSO, carboplatinum and Taxol.
- At age 43: locally recurrent ovarian CA. Rx with 5 cycles of carboplatinum and Taxol.
- 2 months after completing chemotherapy, presented with t-AML/respiratory failure. Expired 9 days after presentation.
- Detailed family history did not suggest inherited cancer susceptibility. Patient has three minor children.

JAMA
The Journal of the American Medical Association
JAMA. 2011;305(15):1568–1576. doi: 10.1001/jama.2011.473

Identification of a Novel *TP53* Cancer Susceptibility Mutation Through Whole-Genome Sequencing of a Patient With Therapy-Related AML



Link et al., JAMA April 20, 2011

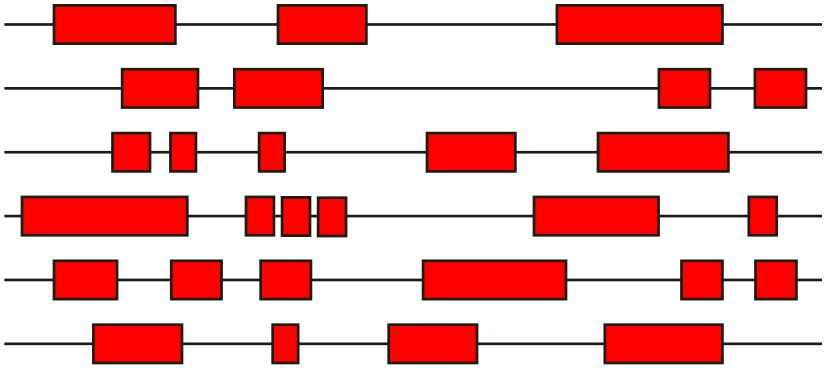
© R.K.Wilson

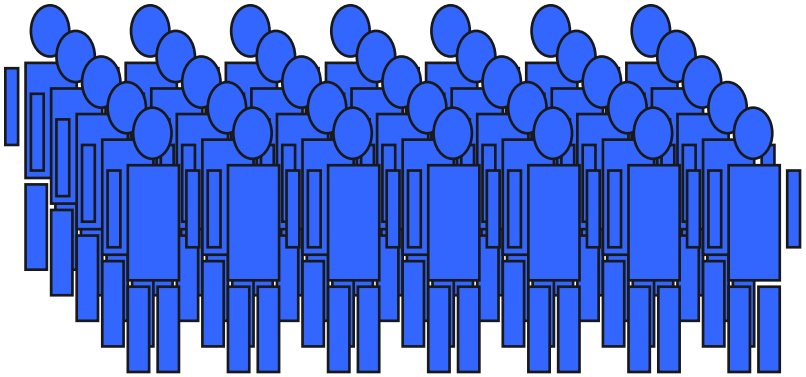
Genomic opportunities in Sickle Cell Disease

- Sequencing options…

Targeted sequencing (hybrid capture)

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



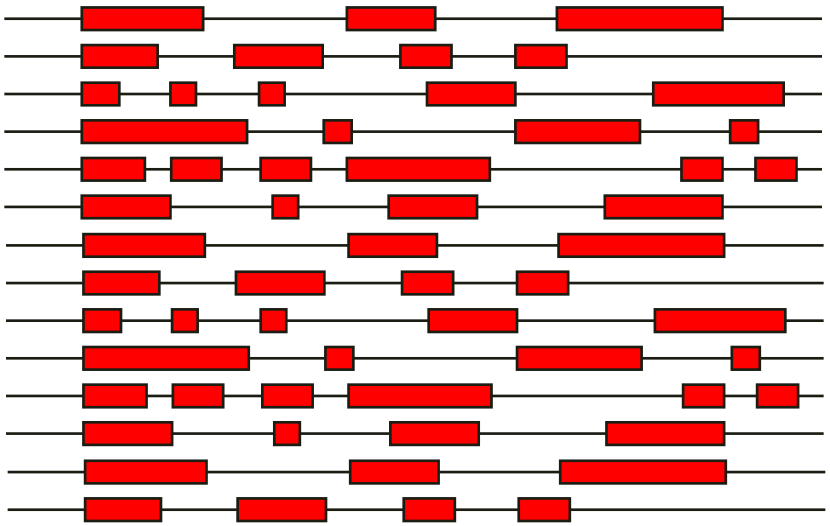


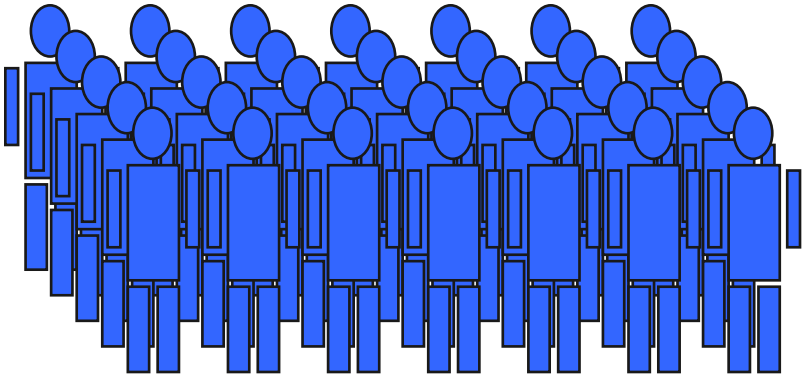
large collection of patient samples

© R.K.Wilson

Exome sequencing (hybrid capture)

Ideally all CCDS exons & selected RNA genes



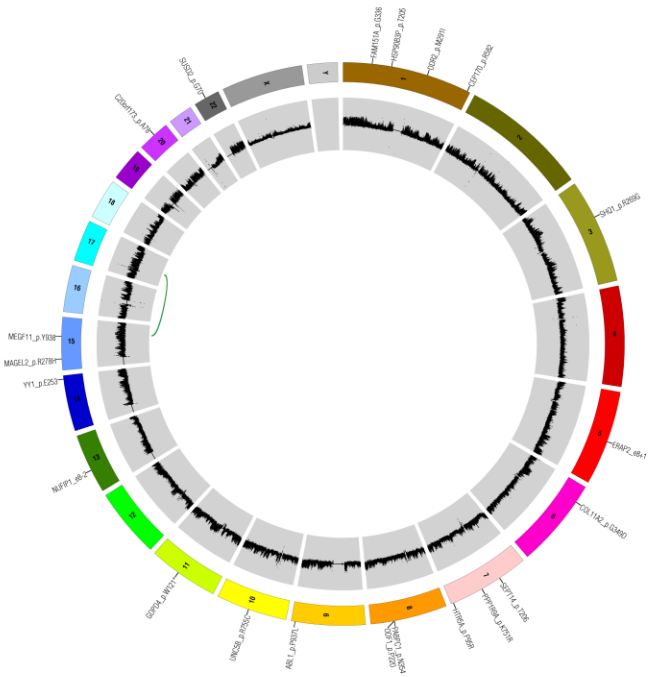


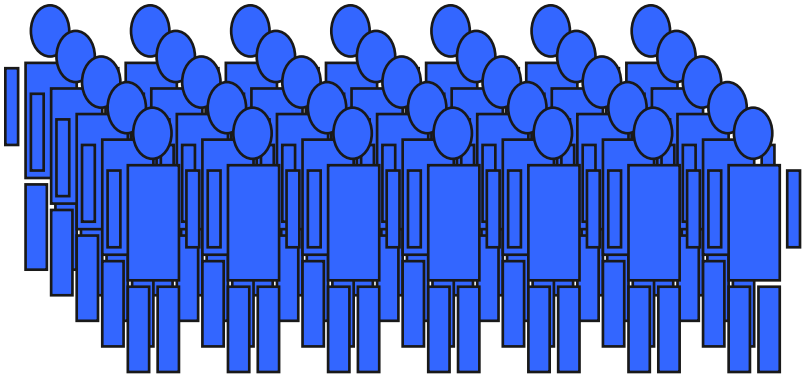
large collection of patient samples

© R.K.Wilson

Whole genome sequencing

complete genome sequences aligned to reference HGS



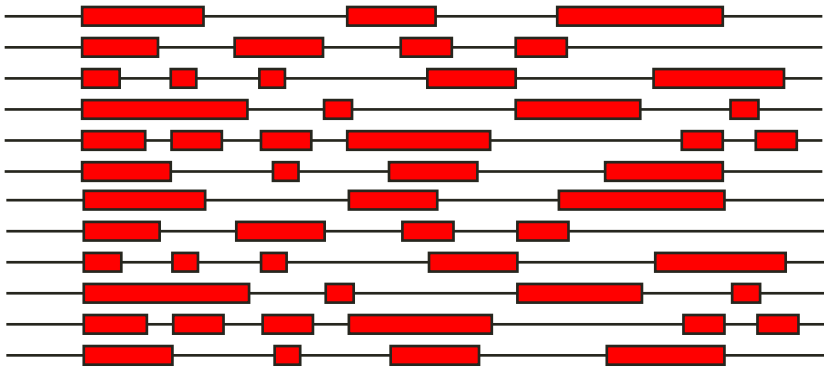


large collection of patient samples

© R.K.Wilson

Whole Genome or Exome sequencing?

- Exome sequencing costs less (~1/6 WGS)
- Simplified analysis (50 Mbp)
- Sequence more samples
- “Low-hanging fruit”



vs.

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

© R.K.Wilson

Exome sequencing reagents (relative to TCGA CCDS)

	% Product Unique	% Product Shared	% CDS Not Targeted	% CDS Targeted
NimbleGen v2 (35.9 Mb)	11.0%	89.0%	2.9%	97.1%
Mystery reagent (63.6 Mb)	48.2%	51.8%	0.1%	99.9%
Agilent SS 50Mb (51.5 Mb)	37.0%	63.0%	1.4%	98.6%
Illumina TruSeq v1 (62.1 Mb)	49.8%	50.2%	5.4%	94.6%

- TCGA CCDS (34 Mbp) is an intersection of the Agilent SSv2 target space and CDS exons. Currently the agreed-upon comparator for exome data produced by the TCGA GSCs.

© R.K.Wilson

T. Wylie, J. Walker

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%
Mystery reagent (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.

© R.K.Wilson

T. Wylie, J. Walker

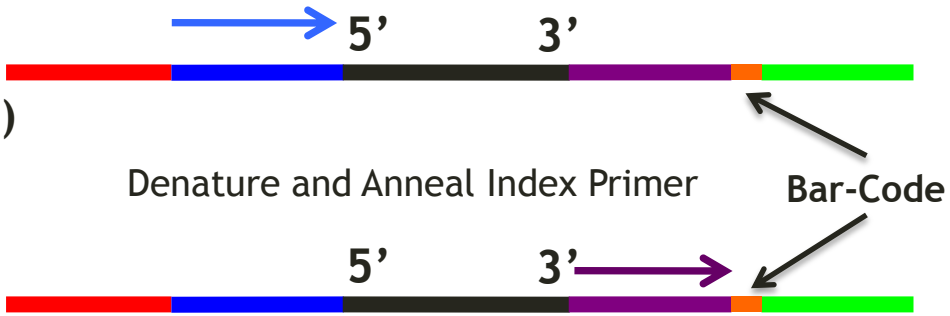
FLT3 coverage in an AML tumor sample

© R.K.Wilson

T. Wylie, J. Walker

Multiplexed libraries for targeted sequencing

Forward Primer:
Sequences 5' end of insert (100 bp)

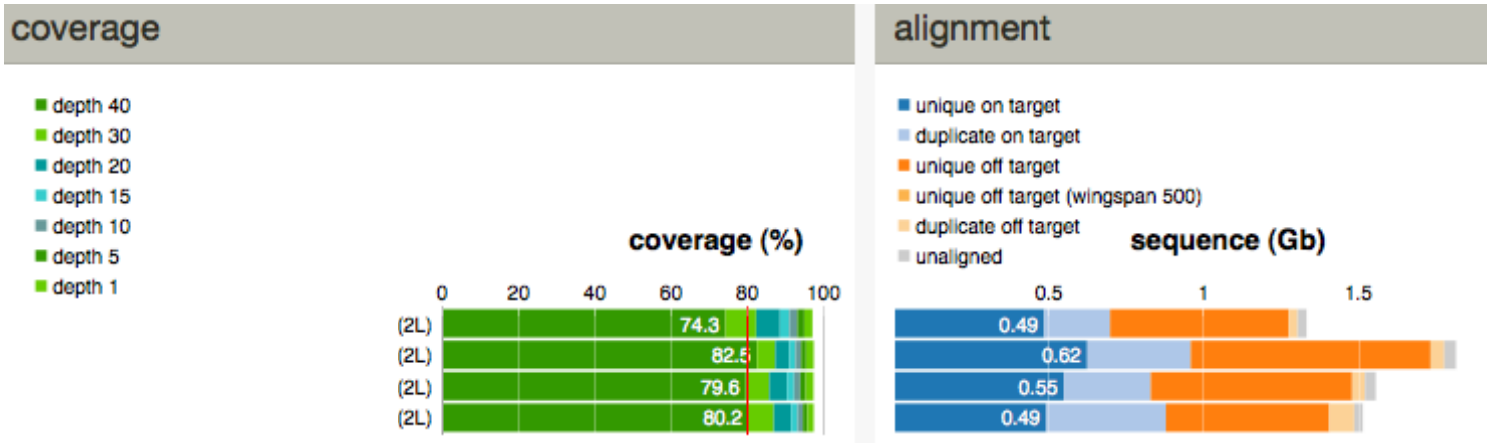


Index Primer:
Sequences bar-code (6bp)

Denature, Bridge Amplify, Anneal Reverse Primer

Reverse Primer:
Sequences 3' end of Insert (100 bp)

De-Multiplexing Indexed Reads in Hybrid Selection and Coverage Evaluation



V. Magrini

Targeted sequencing for Metabolic Syndrome

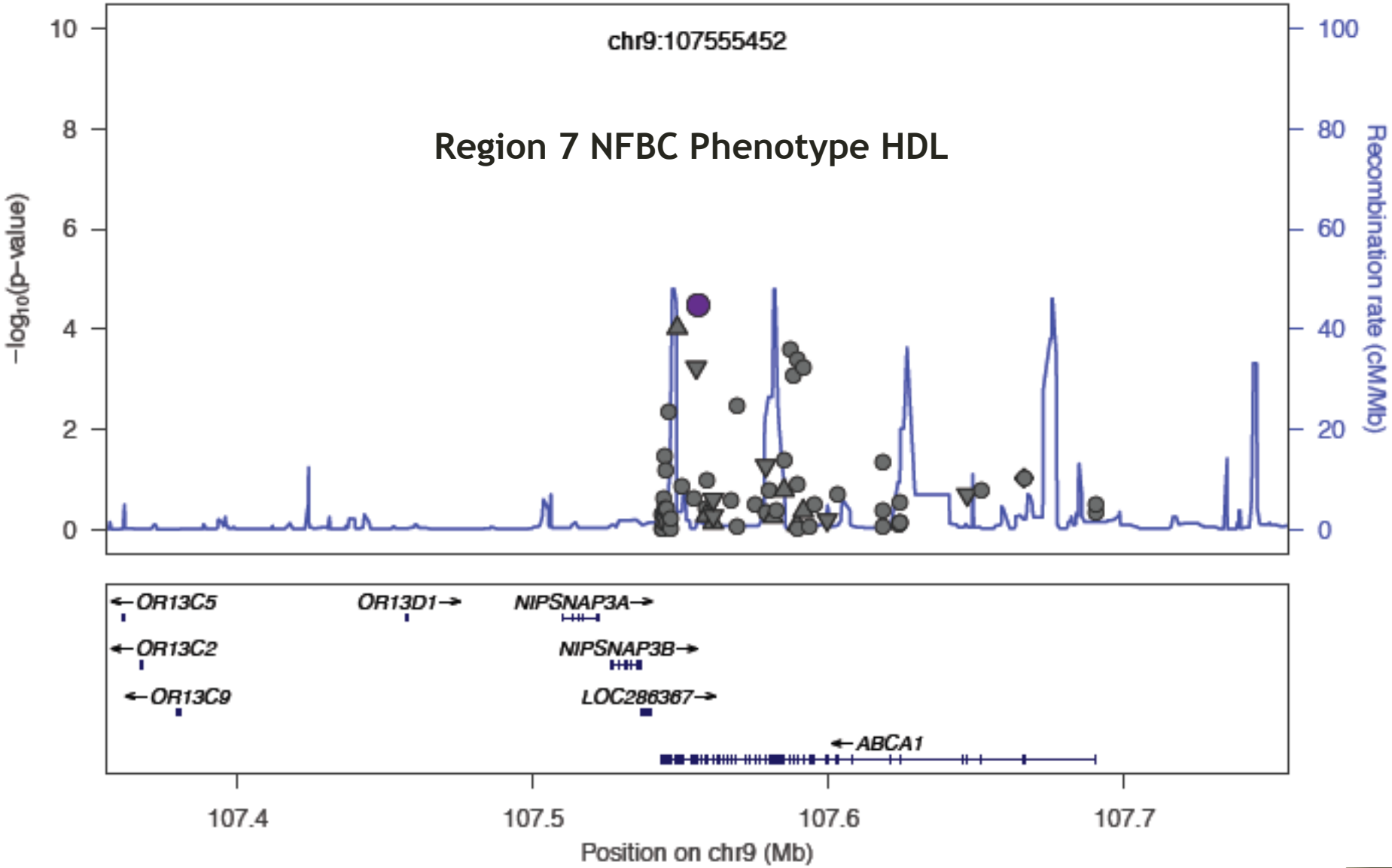
- Targeted capture of 79 ROIs* in 6,965 samples
  - 5,127 North Finland Birth Cohort (NFBC)
  - 1,838 Finnish-U.S. Investigation of NIDDM Genetics (FUSION)
- First-pass data complete for 6,188 samples (89%)
  - >95% genotype goncordance vs. SNP array data
  - >80% ROI coverage >20x
- Full exome sequencing also completed
  - 600 NFBC, 400 FUSION sequenced, analysis in progress…

Collaborators: N. Freimer & M. Boehnke

*** Total of ~0.5 Mbp**

© R.K.Wilson

Targeted sequencing for Metabolic Syndrome



© R.K.Wilson

S. Service

How many samples do we need to sequence?

- 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

© R.K.Wilson

What can we do for $10M? (Data production/analysis)

- Targeted sequencing (custom hybrid capture)
  - 0.5 Mbp/100 genes: 33,000 samples
  - 3.0 Mbp/600 genes: 32,000 samples
  - 6.0 Mbp/1200 genes: 29,000 samples
- Exome sequencing (commercial reagents, 60 Mbp)
  - 6,300 samples (~$1,500/sample)
- Whole genome sequencing (~30x coverage)
  - 1,000 samples (~$9,600/sample)
- Costs include library production, capture & reagents, sequence production, data processing & variant detection.
- Sequencing costs will continue to decrease…

© R.K.Wilson