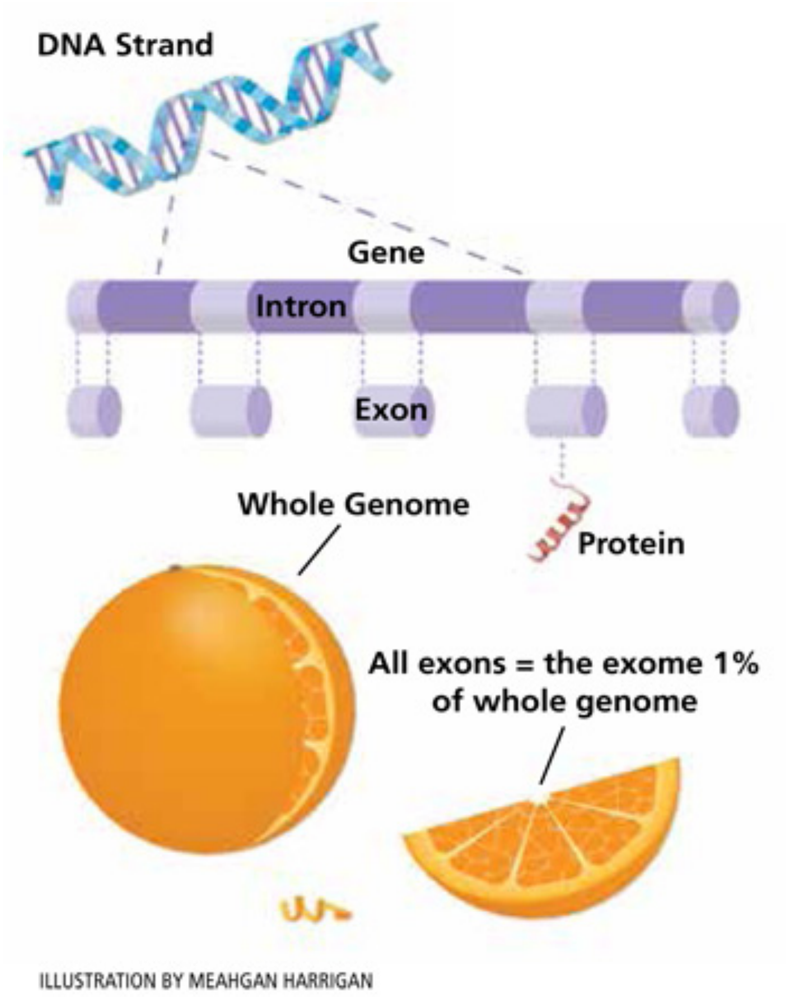
Exome Sequencing Data Analysis

Chunhua Yan, Ph.D.

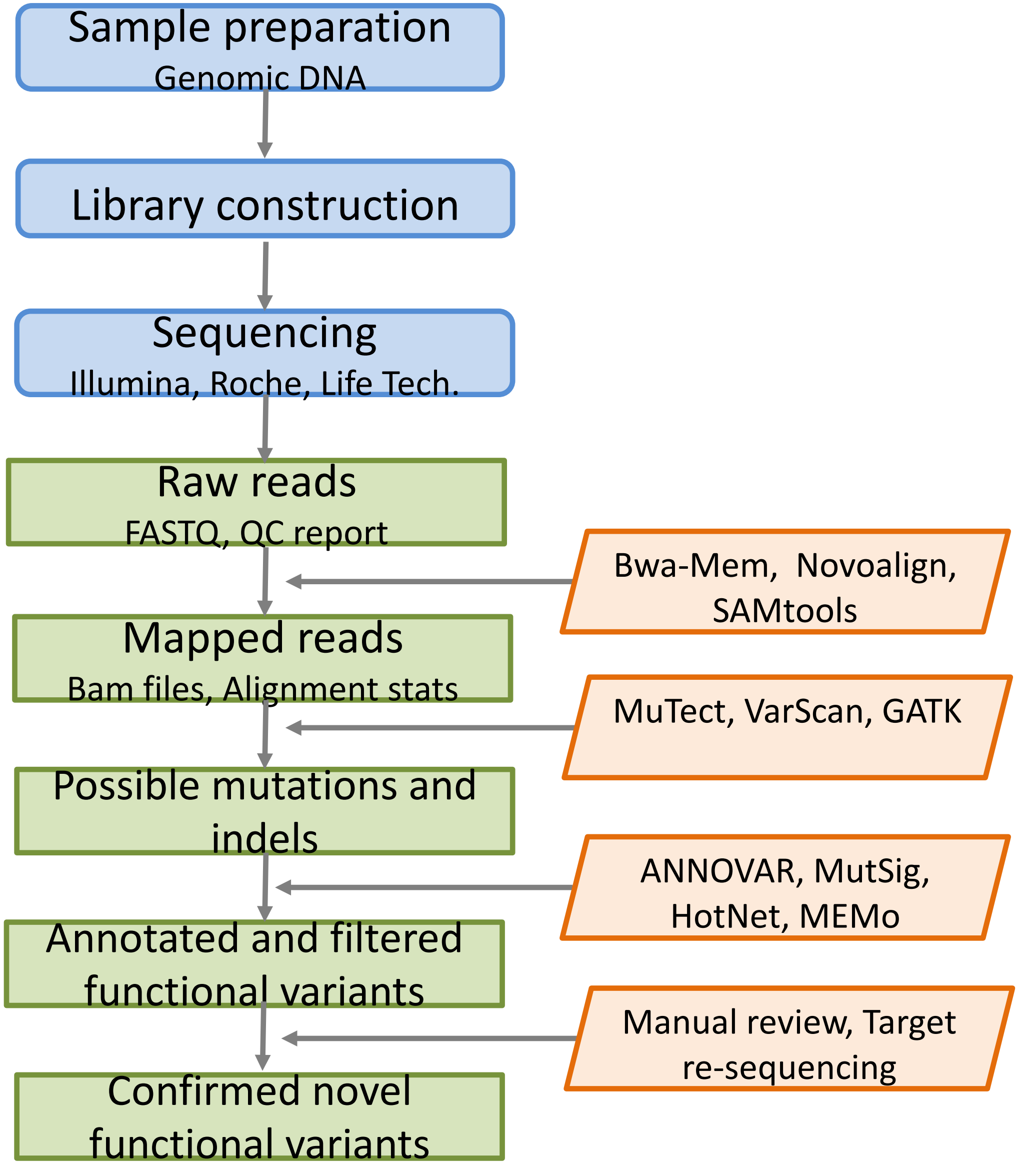
NCI CBIIT

2/21/2017

Exome sequencing workflow


DNA Strand
Gene
Intron
Exon
Protein
Whole Genome
All exons = the exome 1% of whole genome
ILLUSTRATION BY MEAHGAN HARRIGAN

http://www.dana-farber.org


Sample preparation
Genomic DNA

Library construction

Sequencing
Illumina, Roche, Life Tech.

Raw reads
FASTQ, QC report

Bwa-Mem, Novoalign, SAMtools

Mapped reads
Bam files, Alignment stats

MuTect, VarScan, GATK

Possible mutations and indels

ANNOVAR, MutSig, HotNet, MEMo

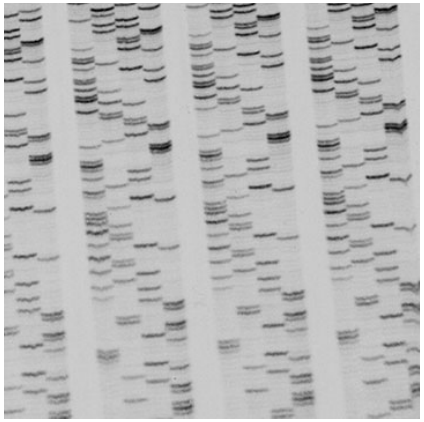
Annotated and filtered functional variants

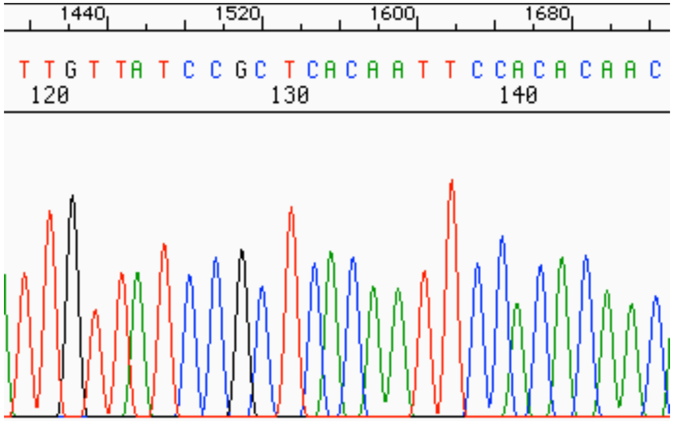
Manual review, Target re-sequencing

Confirmed novel functional variants

1) Next generation sequencing (NGS)
    - ❖ Platform
    - ❖ Productivity
2) Exome sequencing
3) Experimental design
4) Mutation study resources

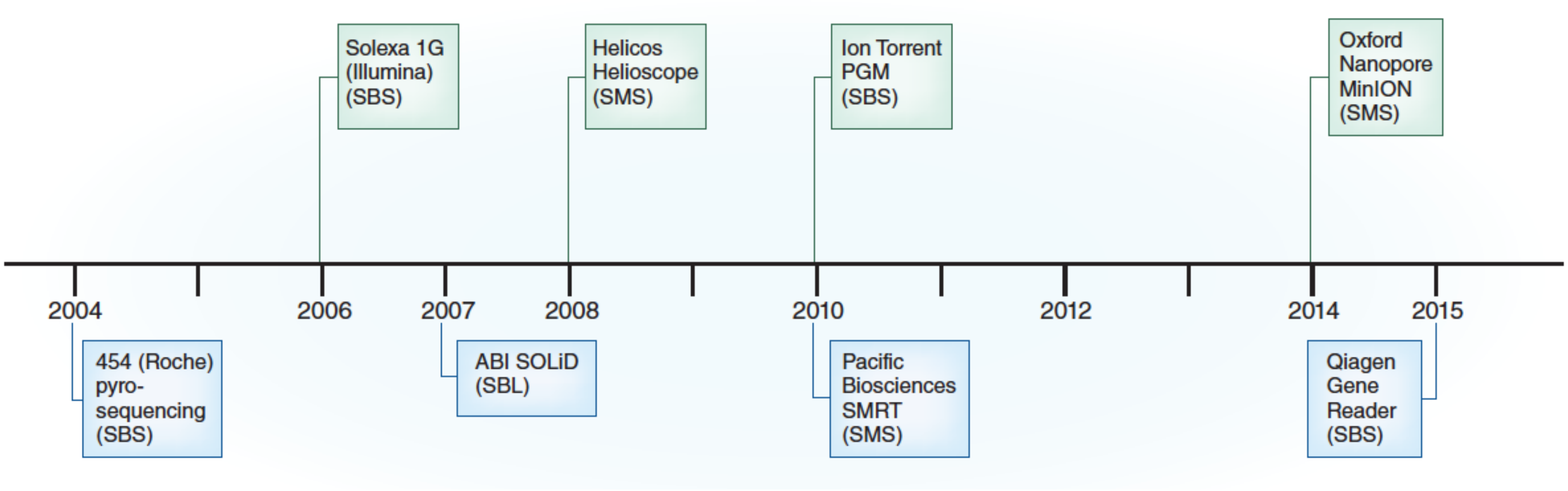
Sanger sequencing: dye-terminator sequencing, 1977-present





Next generation sequencing technology

2004-present



2004 – 454 (Roche) pyro-sequencing (SBS)

2006 – Solexa 1G (Illumina) (SBS)

2007 – ABI SOLiD (SBL)

2008 – Helicos Heliscope (SMS)

2010 – Ion Torrent PGM (SBS)

2010 – Pacific Biosciences SMRT (SMS)

2014 – Oxford Nanopore MinION (SMS)

2015 – Qiagen Gene Reader (SBS)

illumina®

ThermoFisher SCIENTIFIC

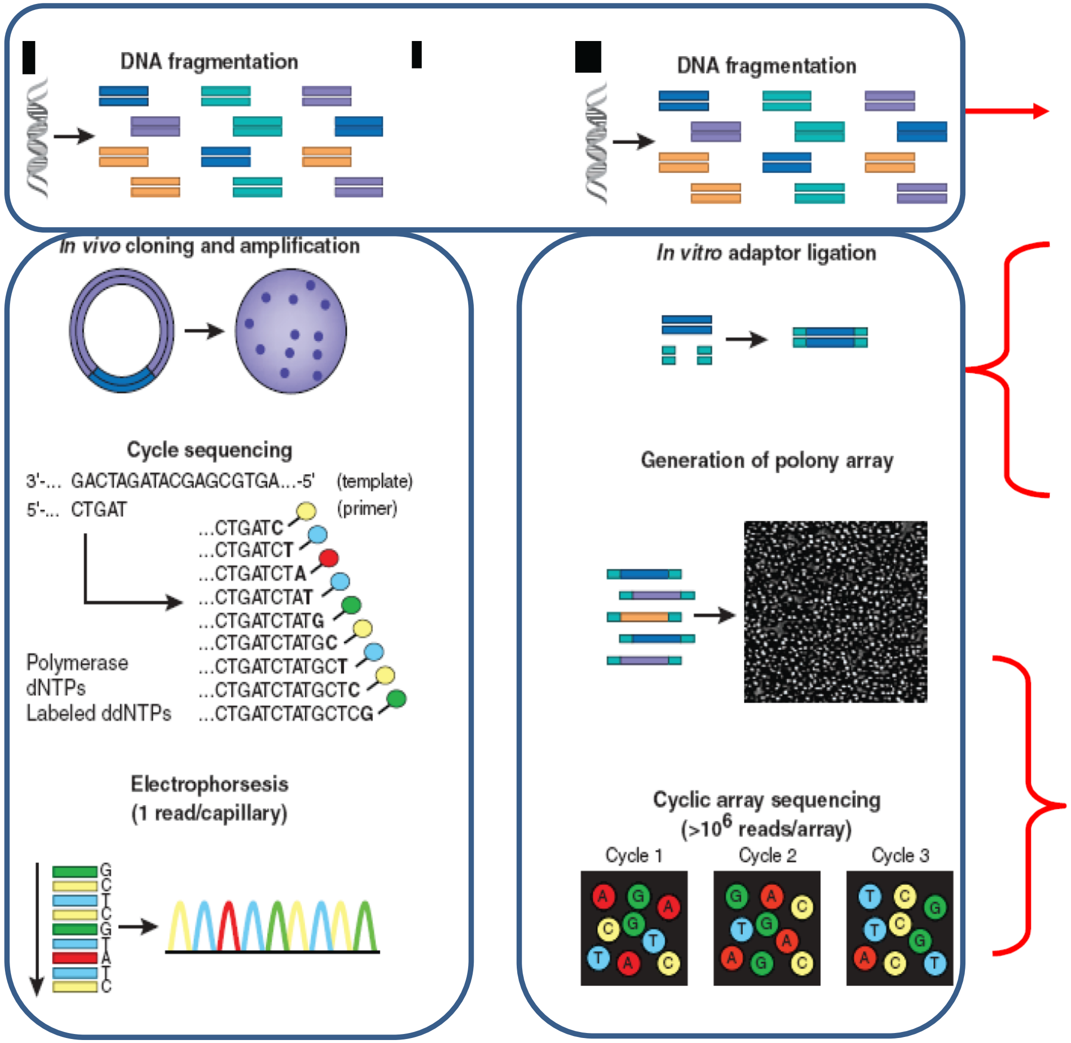
PACBIO®

Mardis ER. (2017) Nat Protoc. 12(2):213-218.

Comparison of sequencing methods

Sanger sequencing

Next generation sequencing



DNA is fragmented

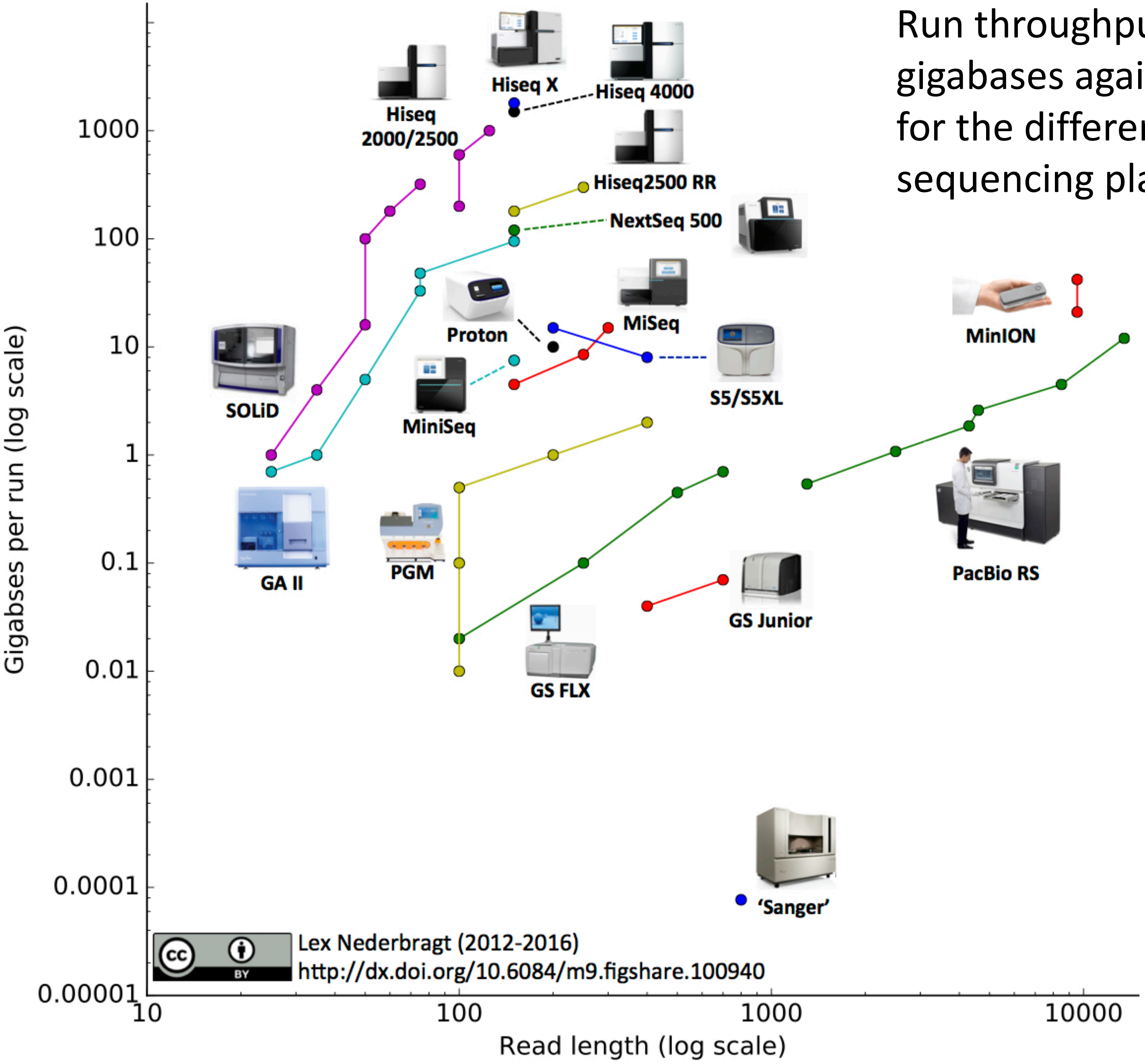
Adaptors ligated to fragments (**Library construction**)

Clonal amplification of fragments on a solid surface (**Bridge PCR** or **Emulsion PCR**)

Direct step-by-step detection of each nucleotide base incorporated
during the sequencing reaction

http://ueb.vhir.org/NGS; https://www.qiagen.com

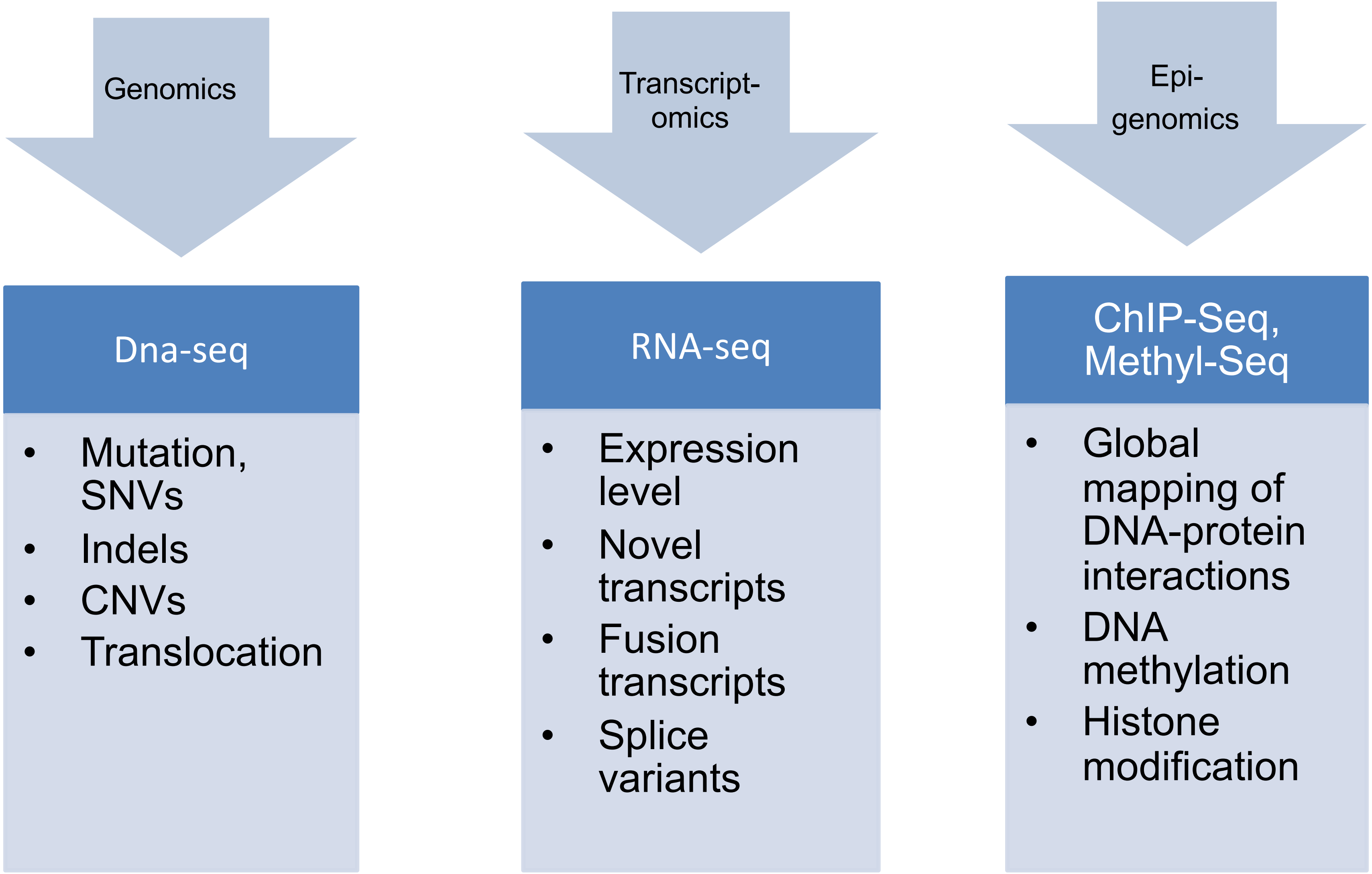
Run throughput in gigabases against length for the different sequencing platforms



Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

https://flxlexblog.wordpress.com

Next generation sequencing applications



Genomics

Dna-seq

- Mutation, SNVs
- Indels
- CNVs
- Translocation

Transcriptomics

RNA-seq

- Expression level
- Novel transcripts
- Fusion transcripts
- Splice variants

Epigenomics

ChIP-Seq, Methyl-Seq

- Global mapping of DNA-protein interactions
- DNA methylation
- Histone modification

1) Next generation sequencing (NGS)
2) Exome sequencing
    - Benefit
    - Capture technology
3) Experimental design
4) Mutation study resource

Whole exome sequencing: Why?

- Focuses on the part of the genome we understand best, the exons of genes
- About 85% of known mutations in Mendelian diseases affect the exome
  - Nonsense, missense, splice, indel mutations
- Depending on the annotation and coverage of flanking sequencing: ~35-60Mb => 1-2% of human genomes
- There are ~200,000 coding exons in ~20,000 genes
- A whole exome is 1/6 the cost of whole genome and 1/15 the amount of data

Biesecker, L. et al. (2011) Genome Biology 12:128

Exome sequencing balances the coverage and cost

Sanger	Targeted	Exome	Whole Genome
• Accurate • Cheap per exon • High turn-around	• Optimization possible • Low chance of incidental findings • Easy analysis	• No bias for genes • Standardized workflow • Re-use of performed exomes to interpret new ones	• No sequencing bias • Detect SVs and SNVs
•Low diagnostic yield for genetically heterogeneous diseases	•Design and re-design required •Different designs for different disorders	• No non-coding regions • Sequencing bias • Incidental findings	• Data analysis bottleneck • Interpretation of non-coding regions • Expensive, time-consuming

http://webinar.sciencemag.org/webinar/archive/exome-sequencing-today%E2%80%99s-lab

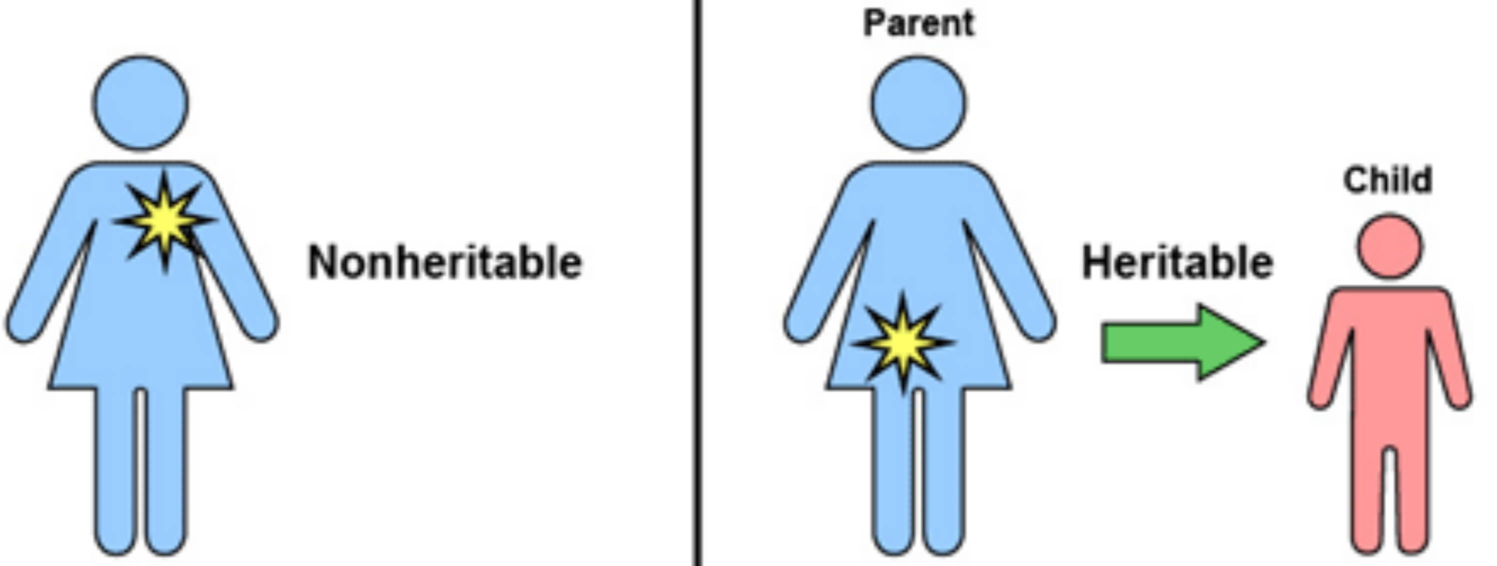
Exome sequencing detects mutations

Somatic mutations

- Occur in ***nongermline*** tissues
- Cannot be inherited

Germline mutations

- Present in egg or sperm
- Can be inherited
- Cause cancer family syndrome



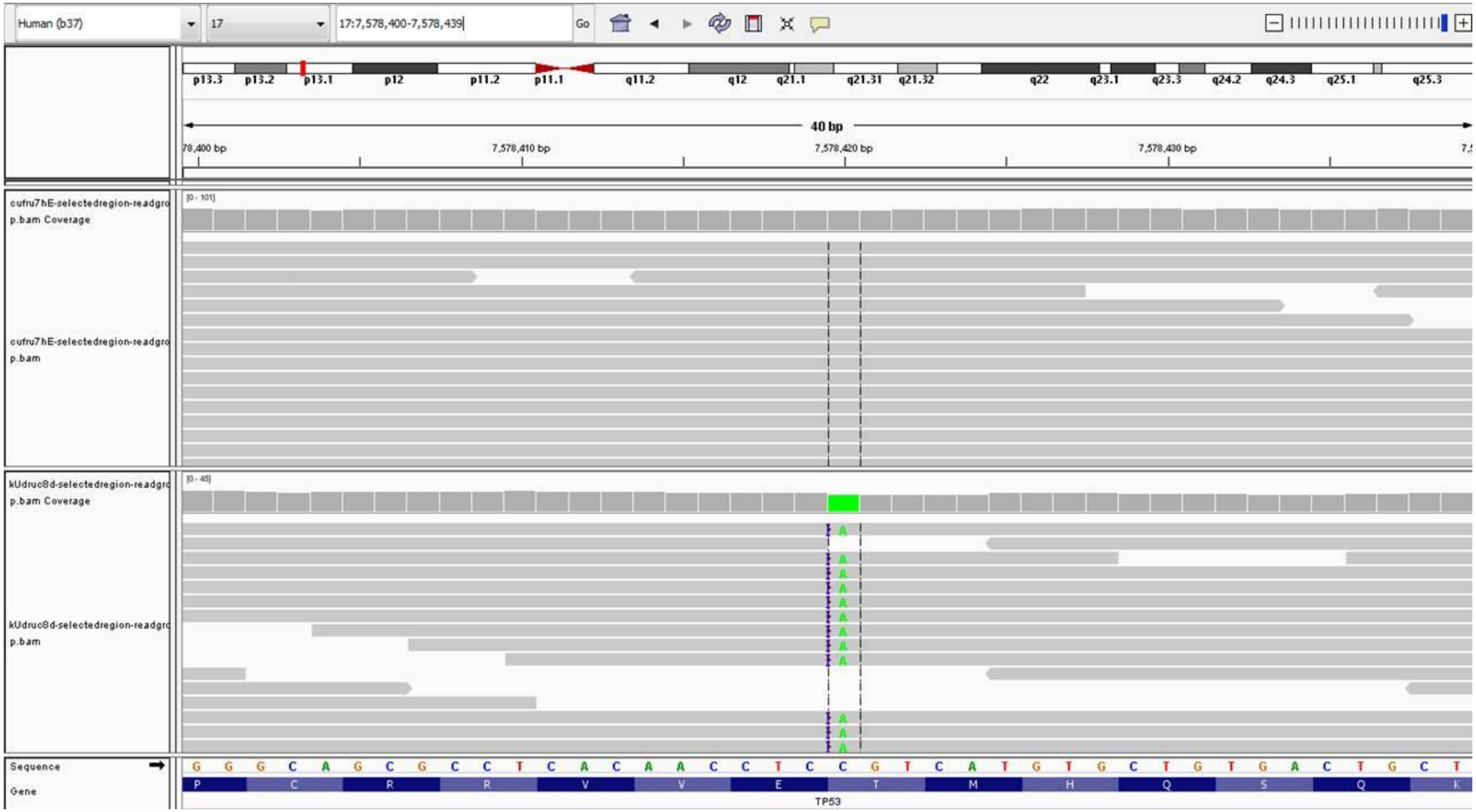
Mutation in tumor only
(for example, breast)

Mutation in
egg or sperm

All cells
affected in
offspring

Adapted from the National Cancer Institute and the American Society of Clinical Oncology

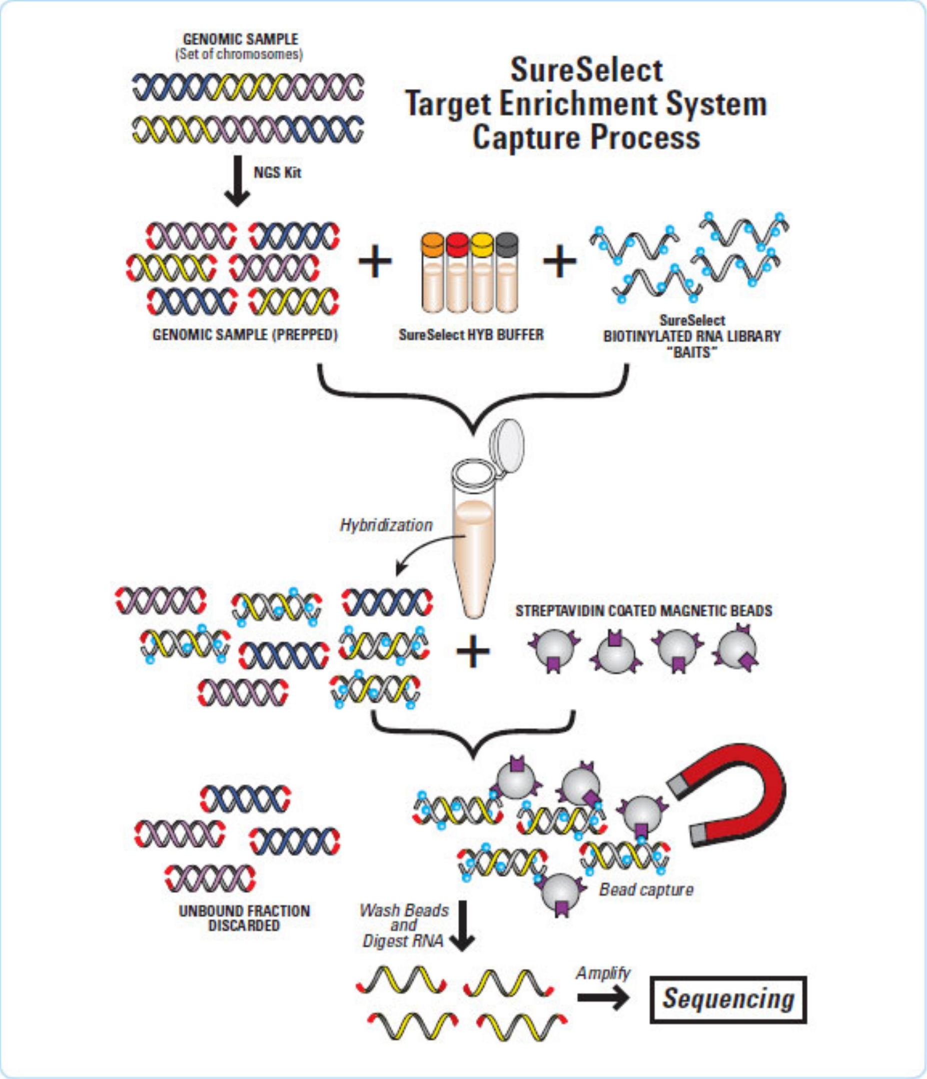
Somatic mutation calls require tumor-normal paired samples


Human (b37)
17
17:7,578,400-7,578,439
Go
p13.3 p13.2 p13.1 p12 p11.2 p11.1 q11.2 q12 q21.1 q21.31 q21.32 q22 q23.1 q23.3 q24.2 q24.3 q25.1 q25.3
40 bp
78,400 bp
7,578,410 bp
7,578,420 bp
7,578,430 bp
cufru7hE-selectedregion-readgro p.bam Coverage
[0 - 101]
cufru7hE-selectedregion-readgro p.bam
kUdruc8d-selectedregion-readgro p.bam Coverage
[0 - 45]
kUdruc8d-selectedregion-readgro p.bam
Sequence
Gene
G G G C A G C G C C T C A C A A C C T C C G T C A T G T G C T G T G A C T G C T
P C R R V V E T M H Q S Q K
TP53

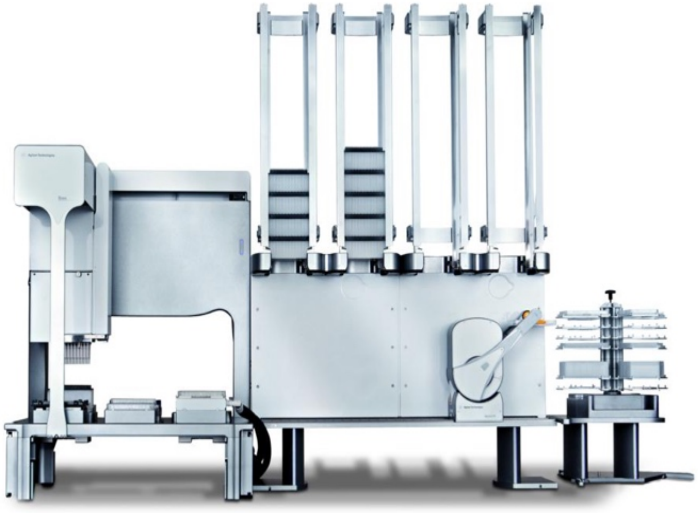
normal

tumor

SureSelect Target Enrichment workflow



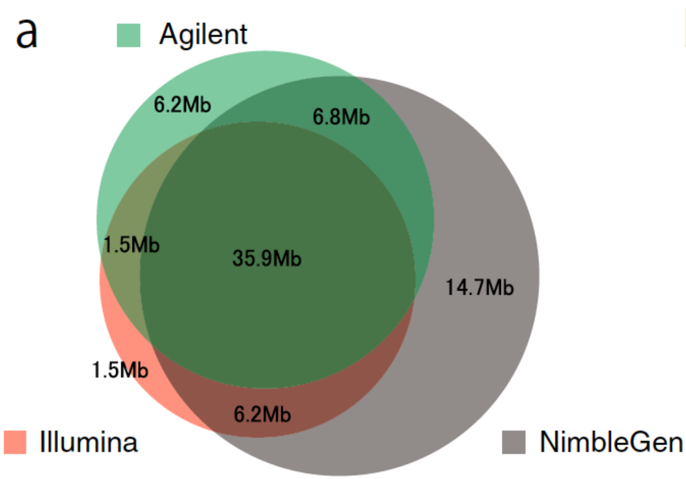
http://www.genomics.agilent.com

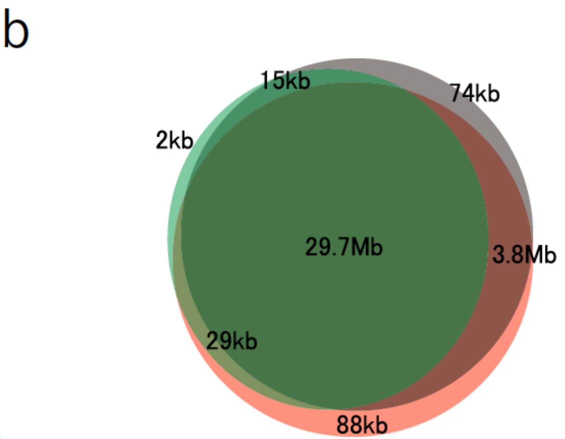


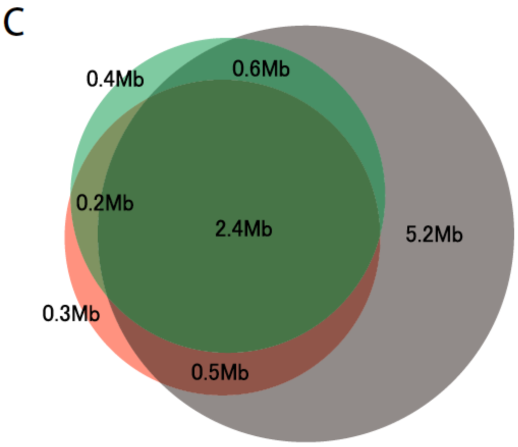
Application	Number of samples/week*	
	Manual	Automated
Whole Genome Sequencing (library preparation only)	100	960
Target Enrichment	20-40	192

Automated NGS Sample Preparation

Comparison of commercial human whole-exome capture platforms







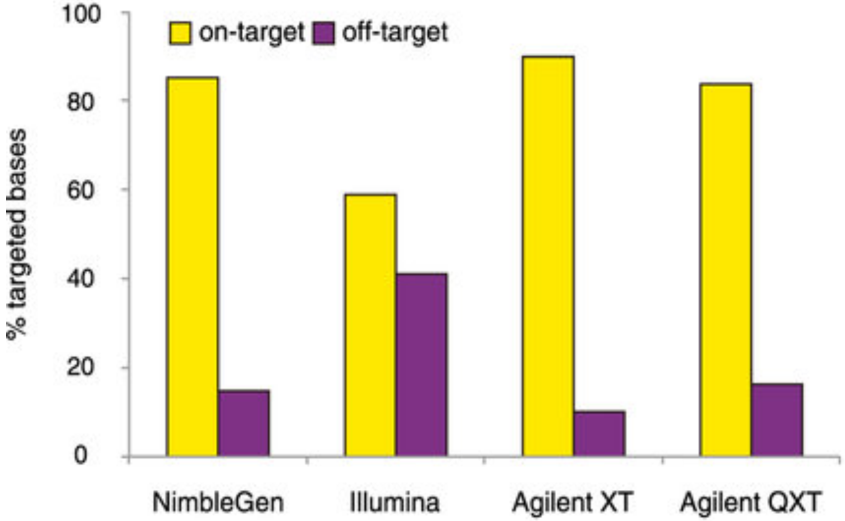
(a) Targeted genomic regions;
(b) Targeted coding regions;
(c) Targeted untranslated regions.

NimbleGen:	63,564,965 bases
Agilent:	50,390,601 bases
Illumina:	45,112,692 bases

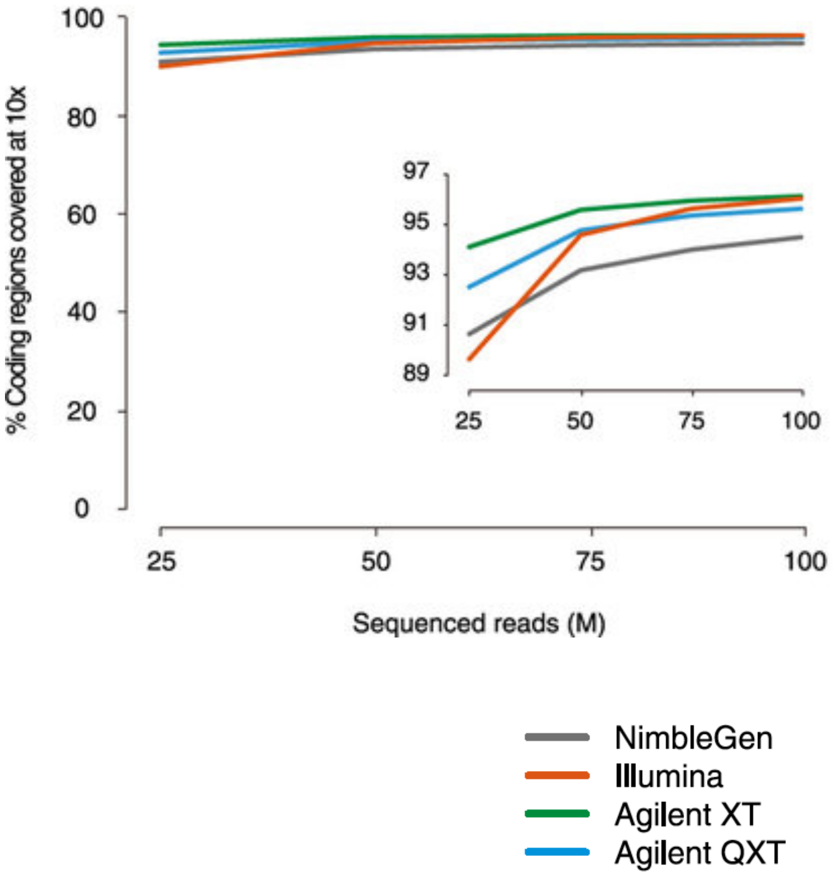
Shigemizu D et al. (2015) Sci Rep. 5:12742.

Coverage of target regions

On-target enrichment



%Coding regions covered at 10x at different read depth



Shigemizu D et al. (2015) Sci Rep. 5:12742.

1) Next generation sequencing (NGS)
2) Exome sequencing
3) Experimental design
   - ❖Sample size
   - ❖Sequencing coverage
4) Mutation study resource

Whole exome DNA sources

Tumor DNA

- Fresh frozen (FF)
- Paraffin embedded tissue (FFPE)
- Cell line
- Single cell
- cfDNA

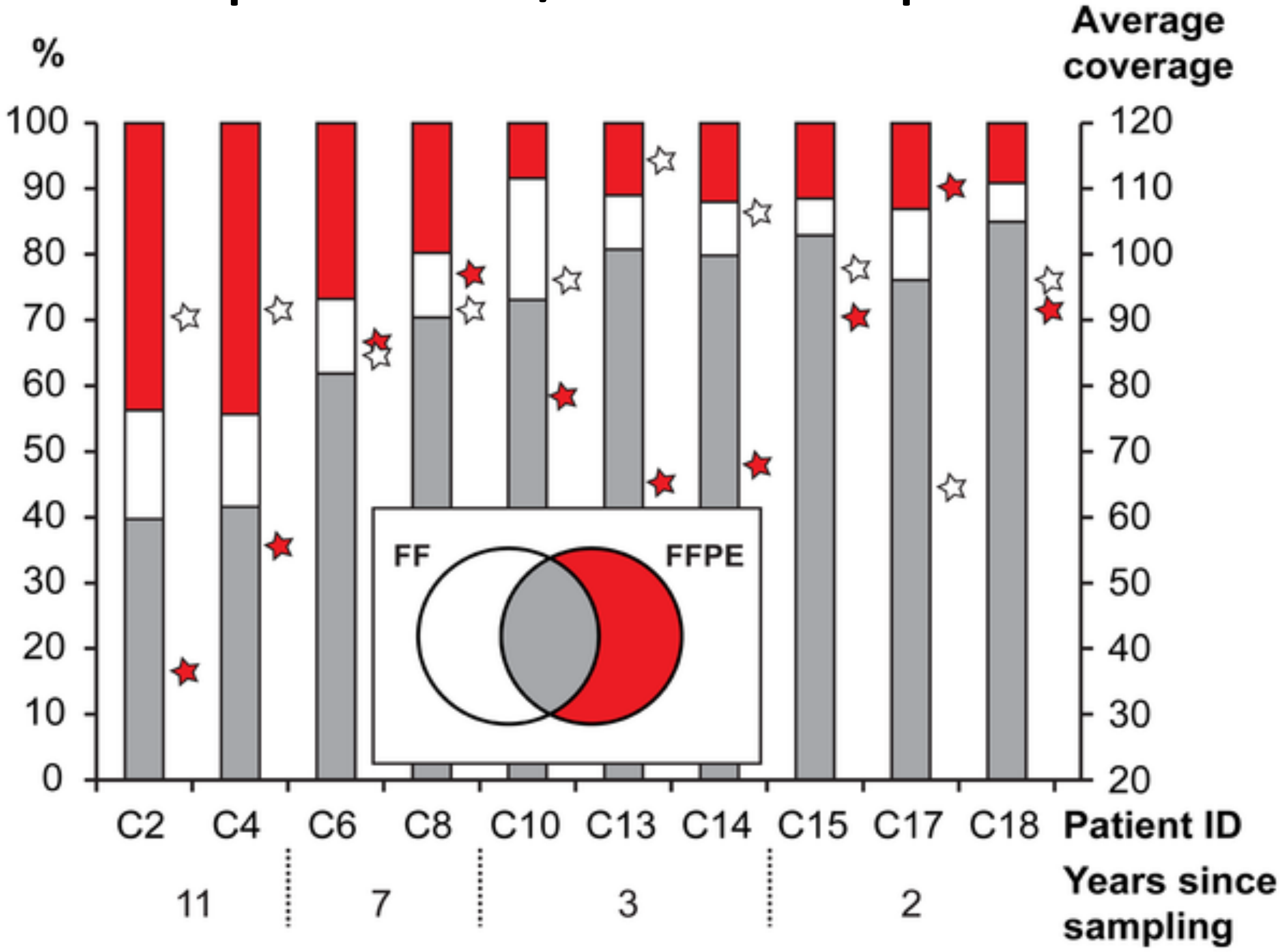
Normal tissue

- Blood
- Neighboring tissue

Clinical information

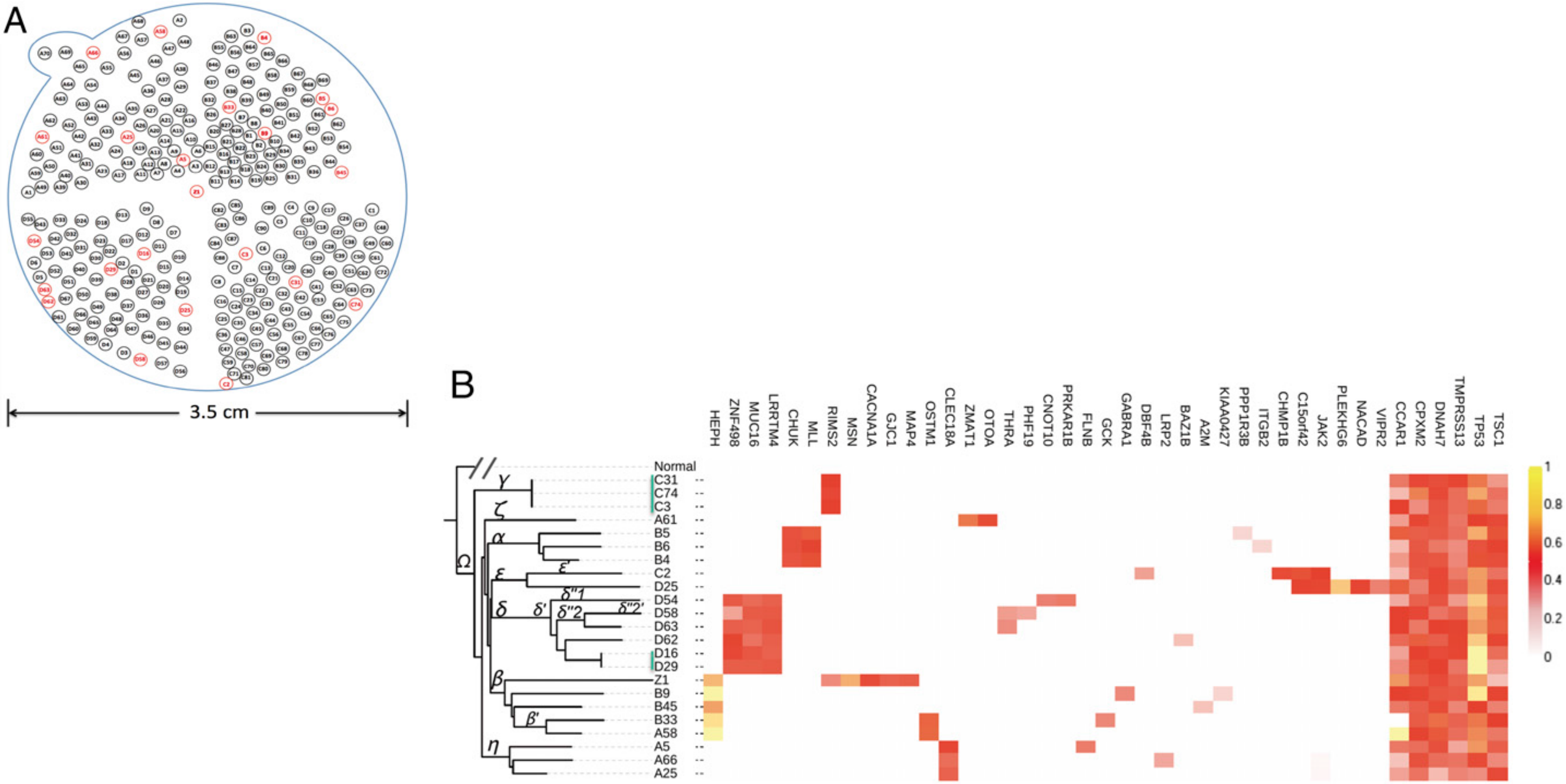
- Date of diagnosis
- Malignancy stage
- Location of primary tumor
- Location of metastatic tumor
- Therapies

Variants detected in exome sequencing data from the paired FF/FFPE samples



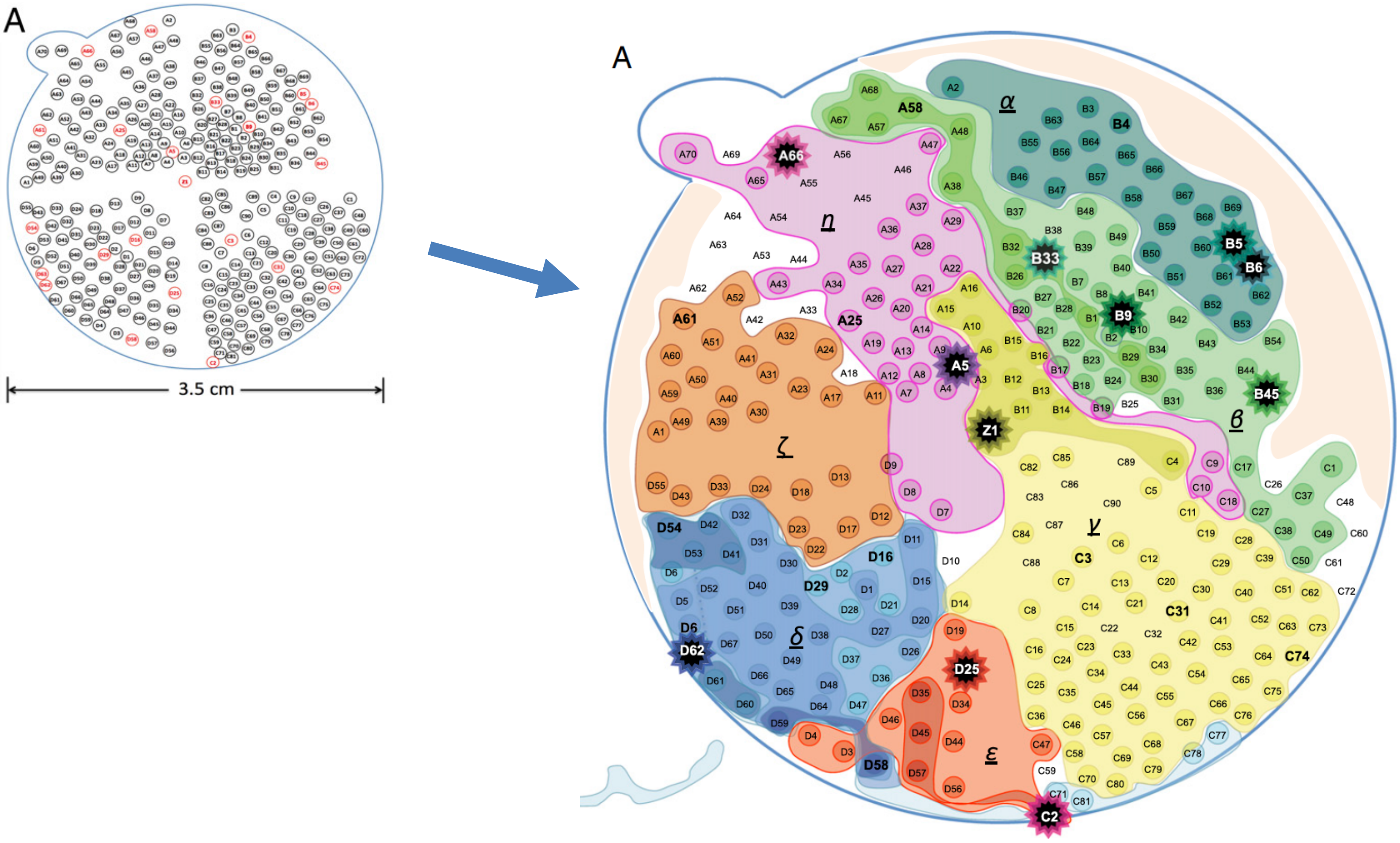
Hedegaard, J. et al. (2014) PLoS ONE 9(5): e98187.

High genetic diversity in a single tumor (HCC)



Ling, S. (2015) Proc Natl Acad Sci U S A. 112(47):E6496-505

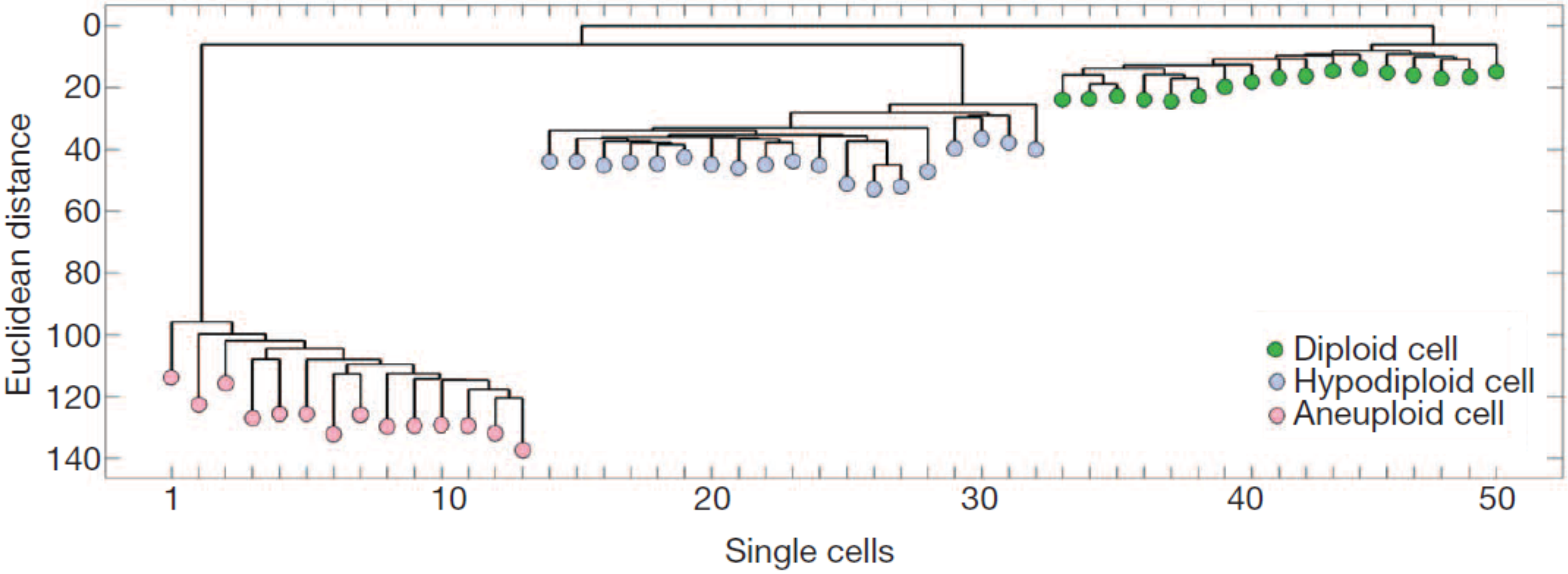
Map of the mutation clones



Ling, S. (2015) Proc Natl Acad Sci U S A. 112(47):E6496-505

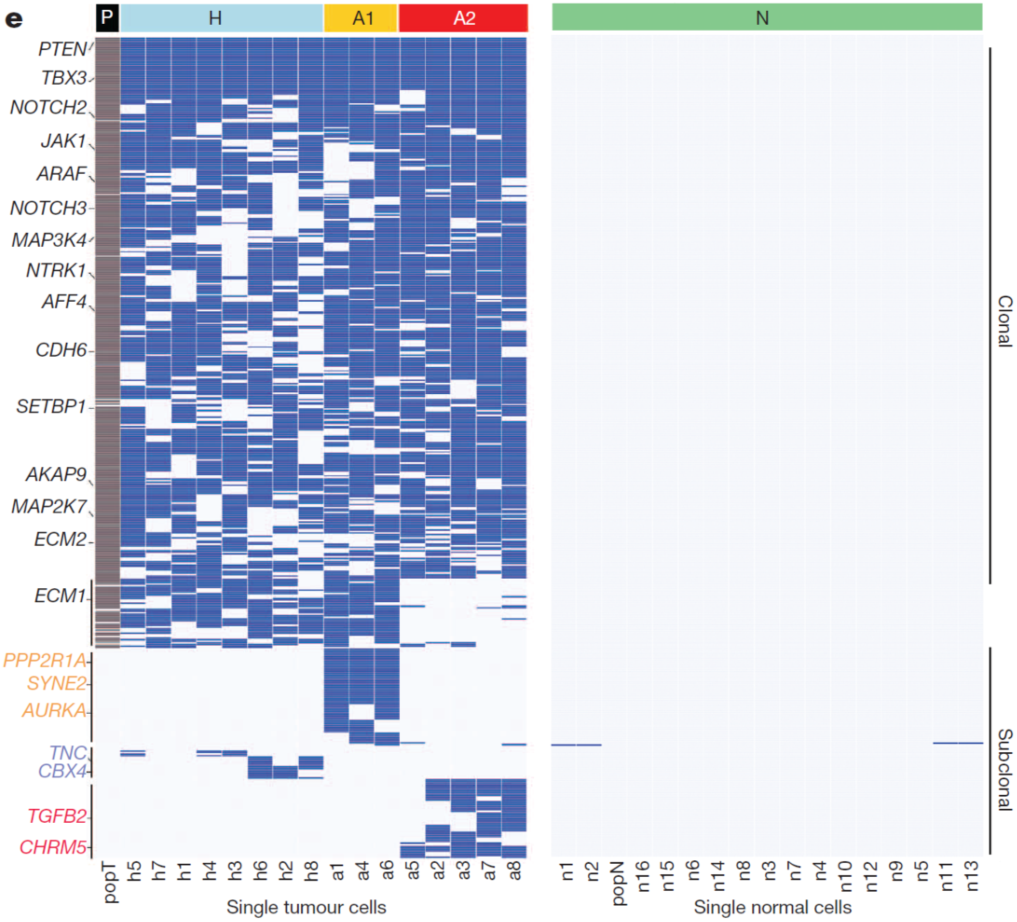
Single cell exome sequencing demonstrates the sample heterogeneity

Triple-negative breast cancer (TNBC)



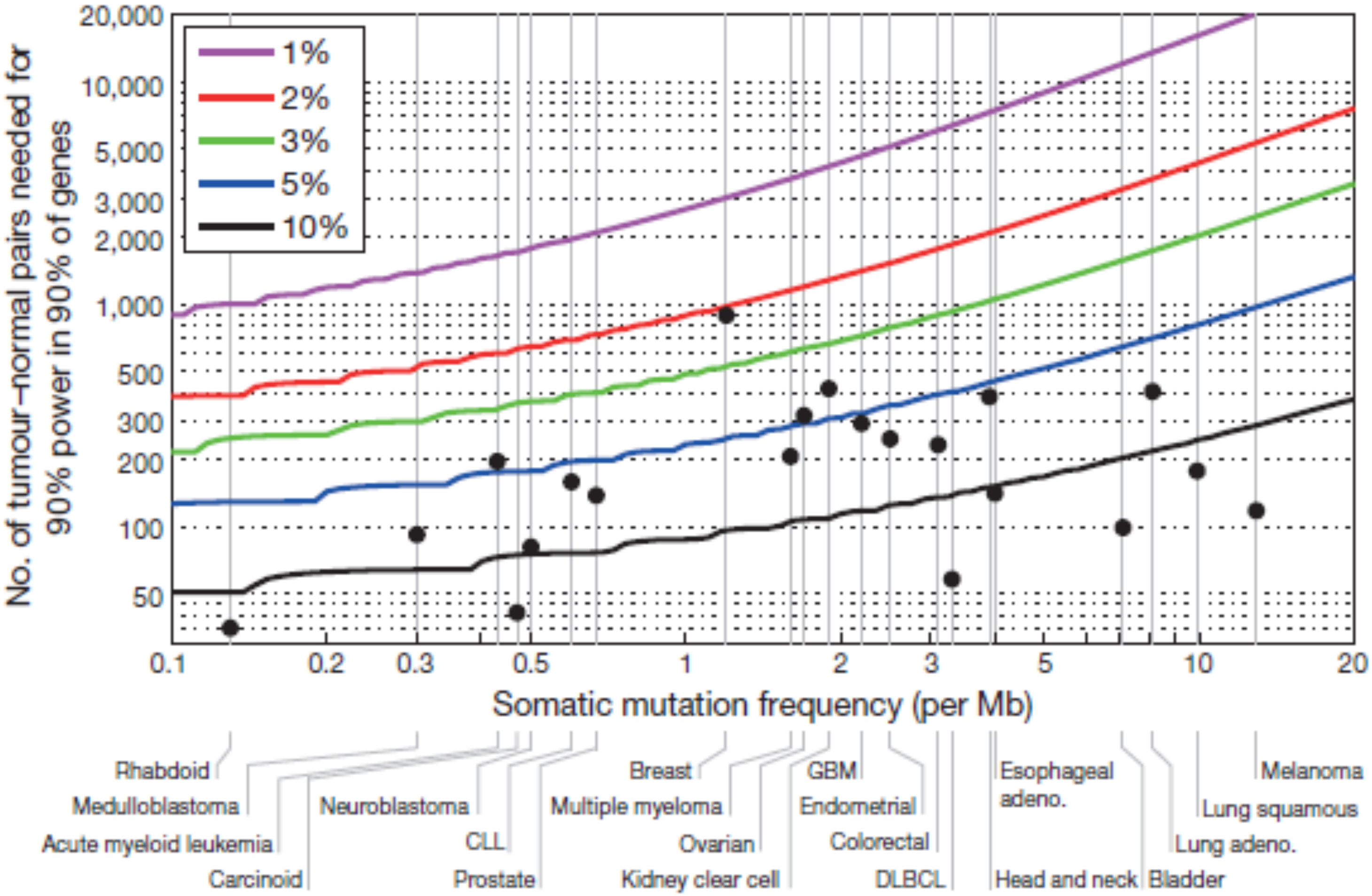
Wang, Y. et al. (2014) Nature. 512(7513):155-60.

Certain mutations only occur in a subset of TNBC cell populations



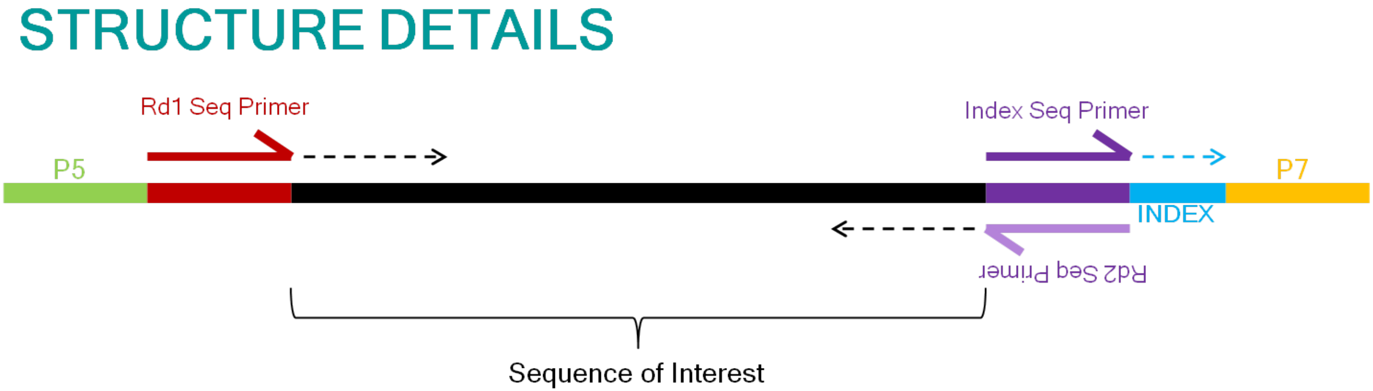
Wang, Y. et al. (2014) Nature. 512(7513):155-60.

The number of samples needed to detect significantly mutated genes

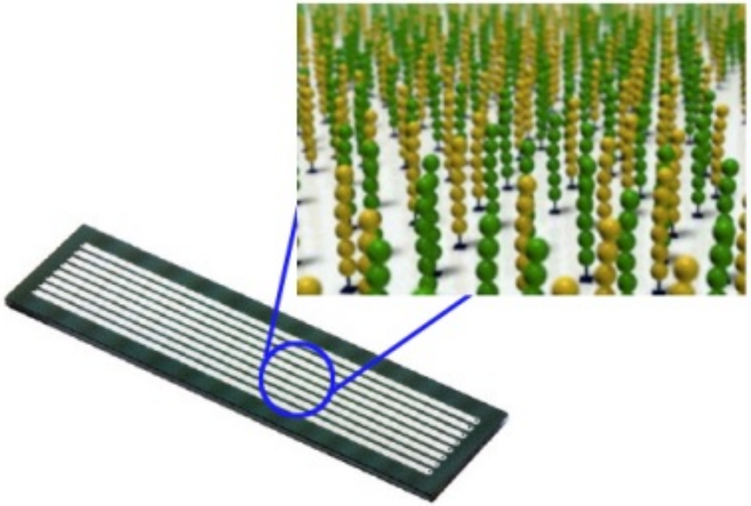


Lawrence, M.S. et al. (2014) Nature. 505(7484):495-501

Sequencing terminology



1. Insert
2. Read
3. Single Read (SR)
4. Paired End (PE)
5. Multiplexing
6. Flowcell
7. Lane

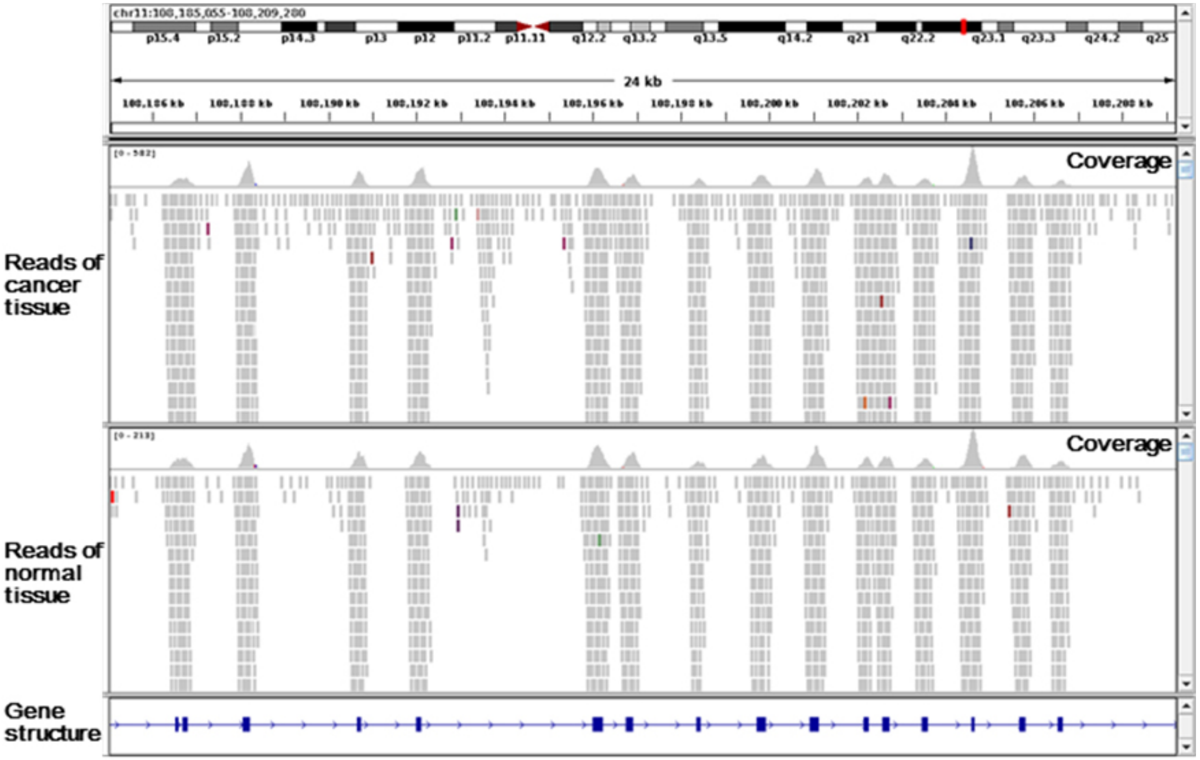


Normand, R. et al. (2013) Methods Mol Biol. 1038:1-26.

Sequencing coverage

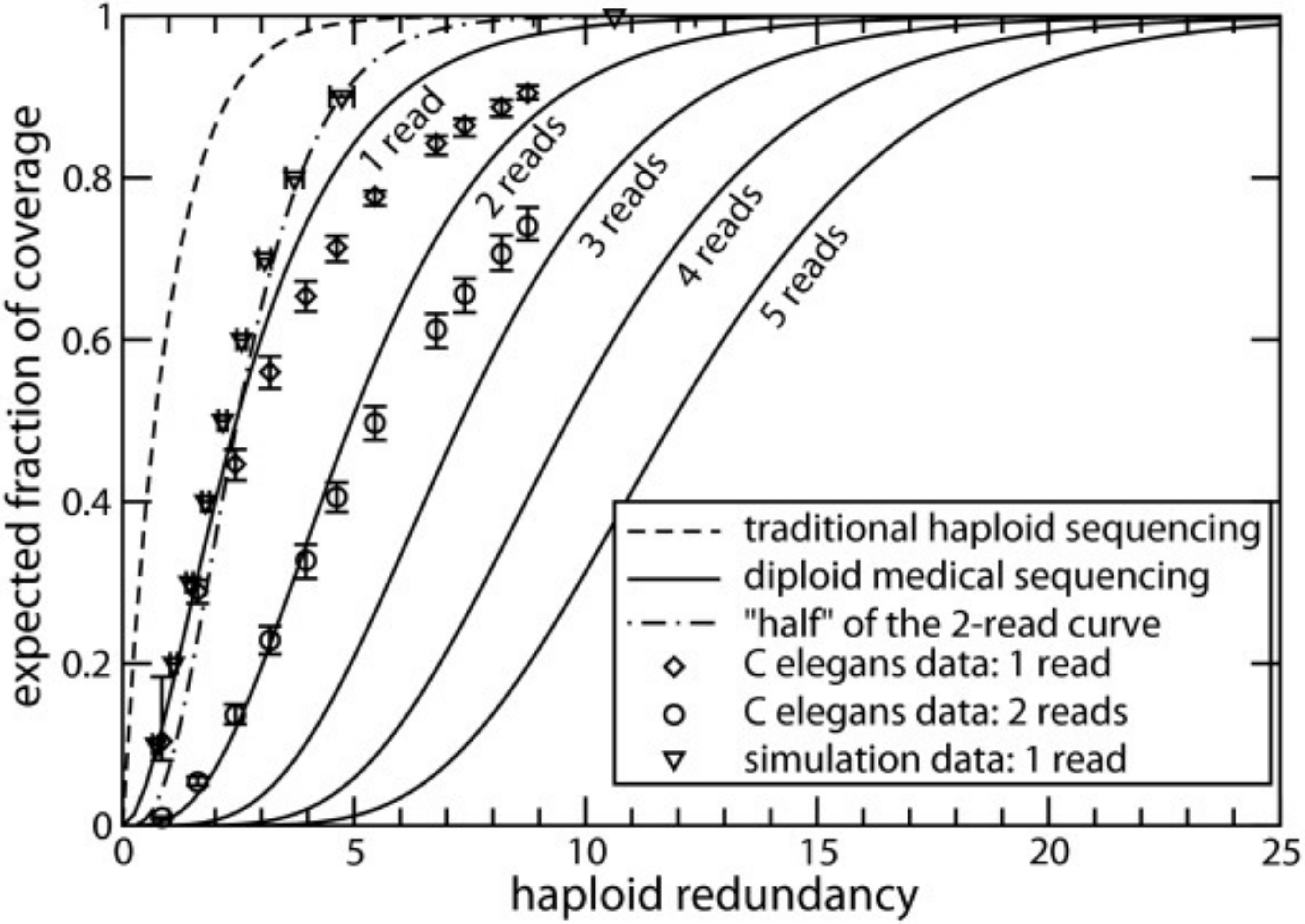
Average coverage =

read length × number of mapped reads/ genome size



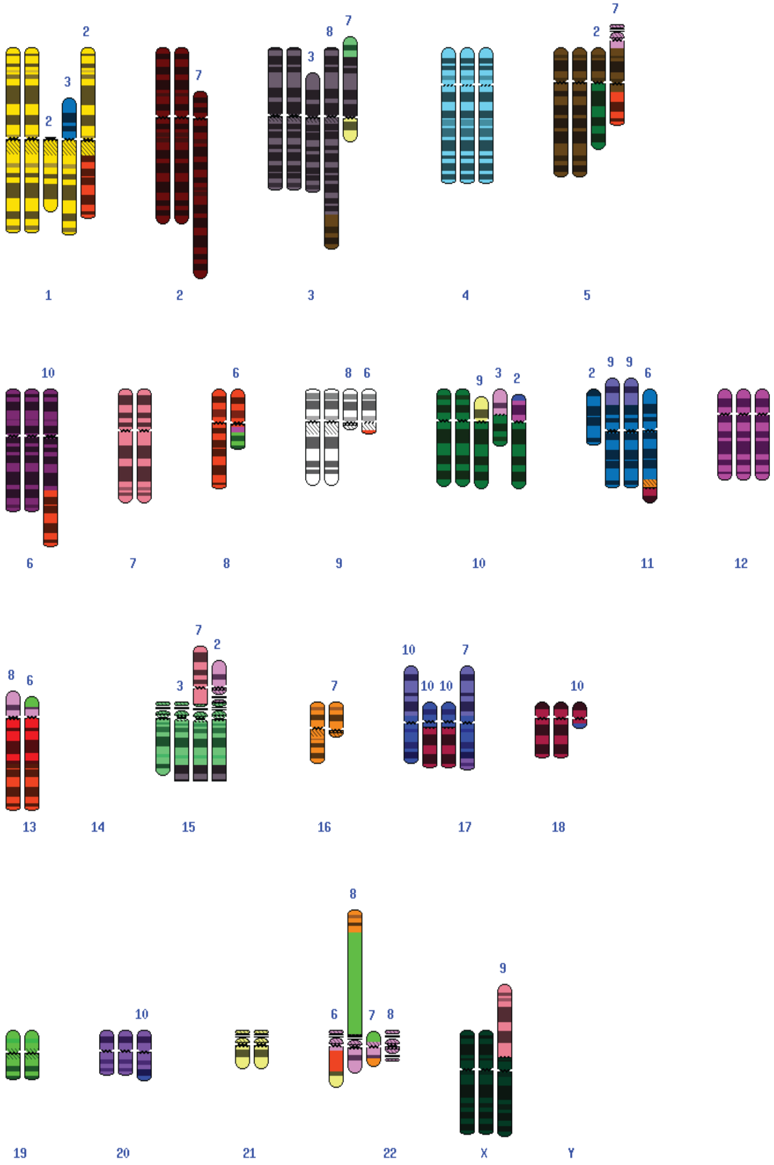
Normand, R. et al. (2013) Methods Mol Biol. 1038:1-26.

Diploid genome and coverage



Wendl, M.C. et al. (2008) BMC Bioinformatics. 9:239.

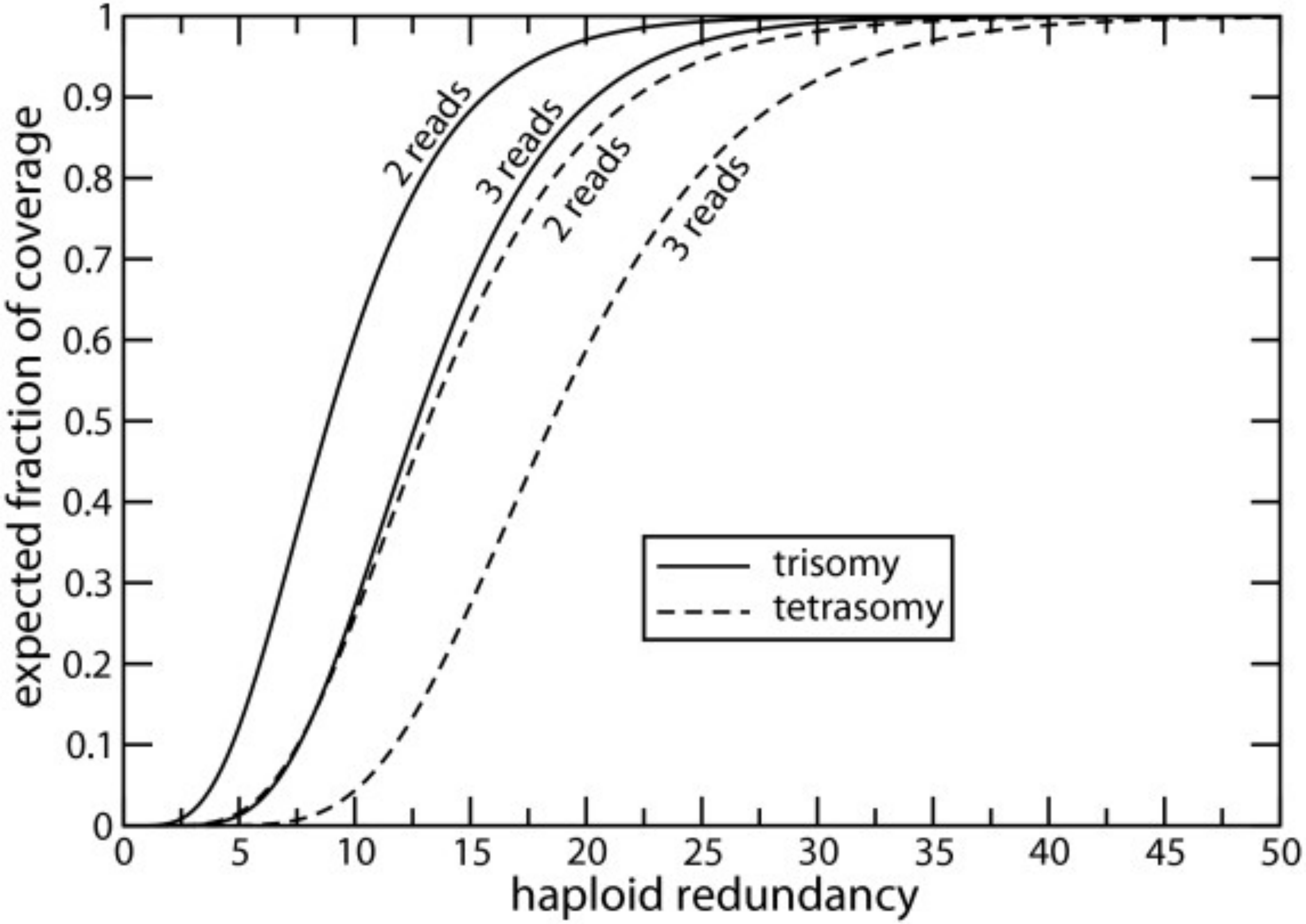
The complexity of cancer genome



OVCAR-3, NCI60 cell line, Ovarian cancer

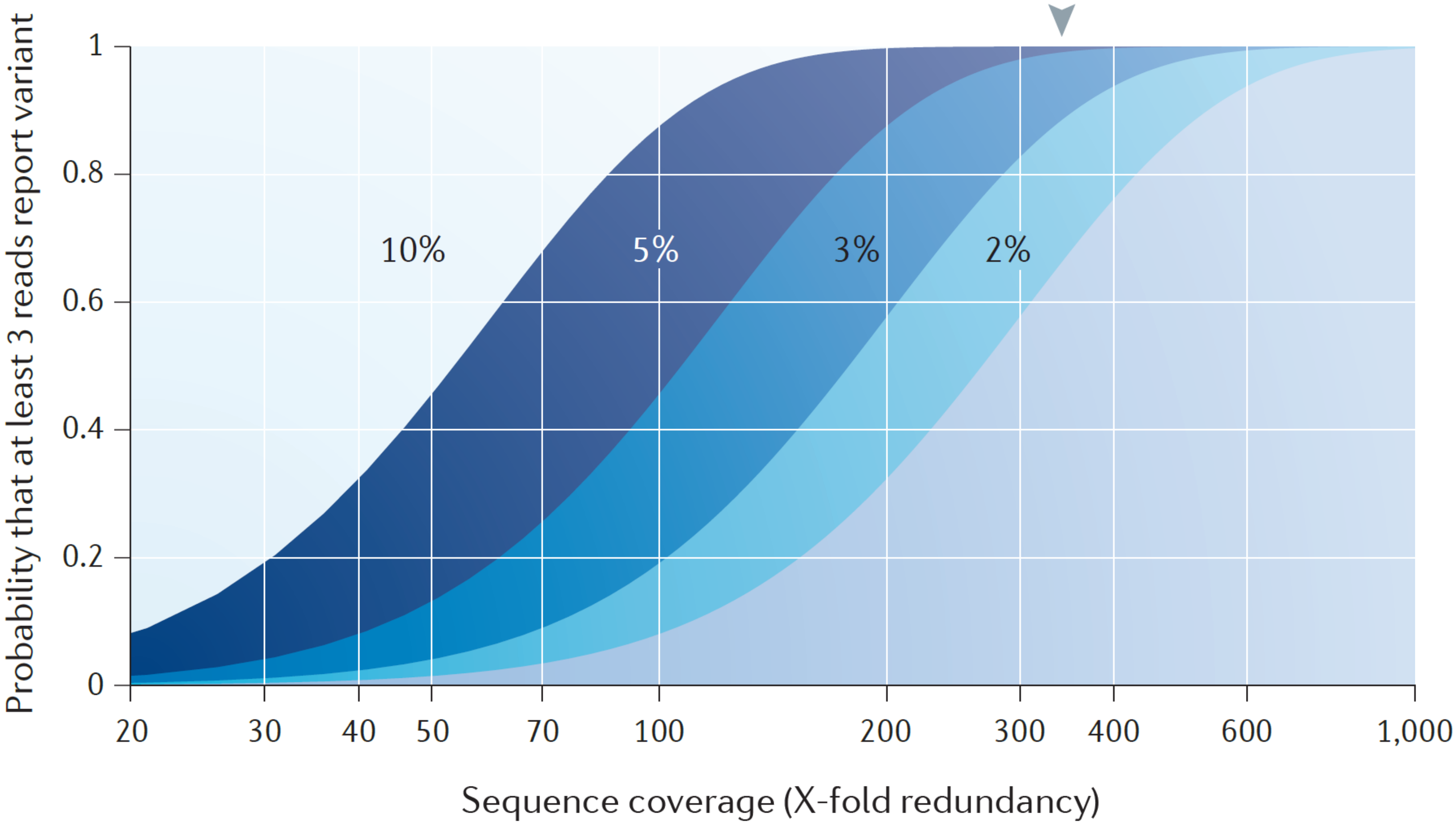
http://www.ncbi.nlm.nih.gov/projects/sky/skyquery.cgi

Polyploid genome and coverage



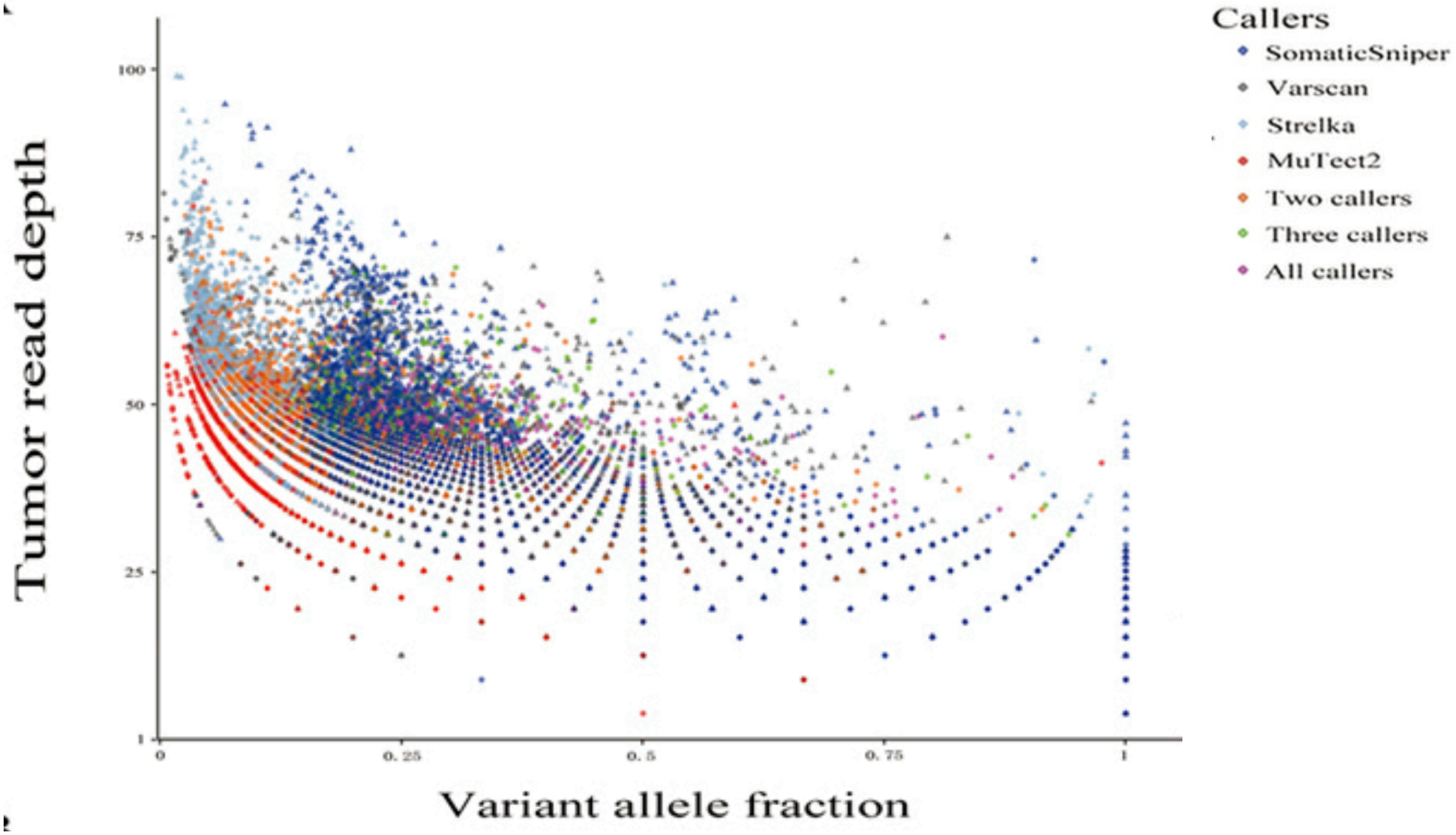
Wendl, M.C. et al. (2008) BMC Bioinformatics. 9:239.

High coverage is needed for low tumor fraction samples



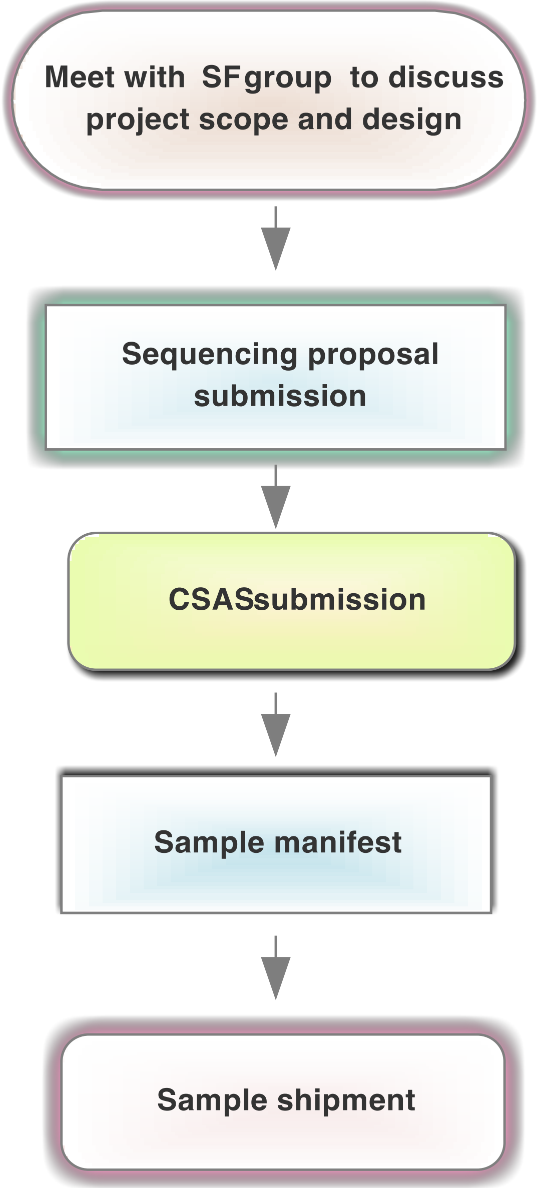
Ding, L. et al. (2014) Nat Rev Genet. 15(8):556-70

The depth-VAF scatter plot of SNV candidates in WES



Cai L, et al. (2016)Sci Rep. 6:36540.

Steps to bring in projects to CCR–SF



Meet with SFgroup to discuss project scope and design

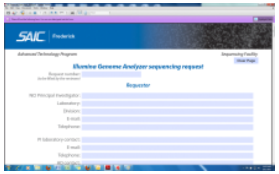
Sequencing proposal submission

CSASsubmission

Sample manifest

Sample shipment

Make sure to consult your bioinformatics for experiment design



SAIC | Frederick

Advanced Technology Program

Sequencing Facility

Illumina Genome Analyzer sequencing request

Request number:
(to be filled by the reviewer)

Requestor

NCI Principal Investigator:

Laboratory:

Division:

E-mail:

Telephone:

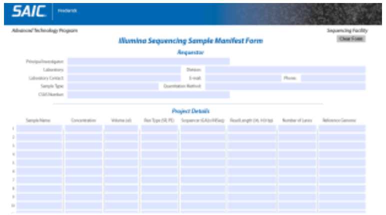
PI laboratory contact:

E-mail:

Telephone:

NCI contact:

https://ostr.cancer.gov/resources/fnl-cores/sequencing-facility



SAIC | Frederick

Advanced Technology Program

Sequencing Facility

Illumina Sequencing Sample Manifest Form

Requestor

Principal Investigator:

Laboratory: Division:

Laboratory Contact: E-mail: Phone:

Sample Type: Quantitation Method:

CSAS Number:

Project Details

	Sample Name	Concentration	Volume (ul)	Run Type (SR, PE)	Sequencer (GAIIx/HiSeq)	Read Length (36, 101 bp)	Number of Lanes	Reference Genome
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								

Courtesy of Yongmei Zhao, CCR-SF

leidos
Leidos Biomedical Research, Inc.

Cancer Research Technology Program

Sequencing Facility

Clear Form

Illumina Sequencing Sample Manifest Form

Requestor

Principal Investigator:

Laboratory: Division:

Laboratory Contact: E-mail: Phone:

Sample Type: Quantitation Method:

CSAS Number: Application: □ Chip Seq □ mRNA Seq* □ Total RNA Seq** □ micro RNA □ gDNA □ other

*Poly A tail selection

**Capture mRNA and long non-coding RNA species i.e. lincRNA, snRNA, snoRNA, however higher sequencing depth is required

Project Details

	Sample Name	Concentration	Volume (ul)	Run Type (SR, PE)	Sequencer (GA2x/HiSeq)	Read Length (36, 103 bp)	Number of Lanes	Reference Genome
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								

Comments

Please complete the Sample Manifest form and e-mail it to the attention of **Jyoti Shetty** at ***shettyju@mail.nih.gov*** prior to shipping your samples.
Please include any Quality Control documentation available such as gel images or electropherograms.

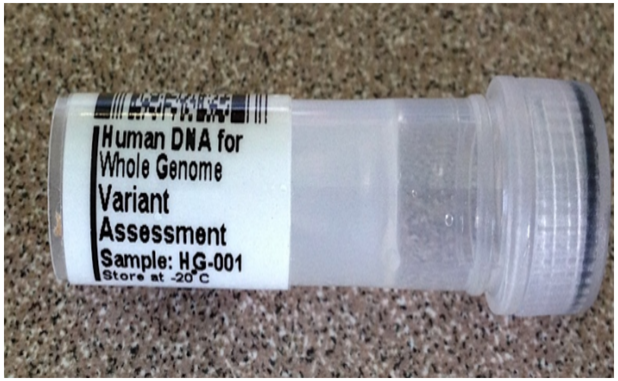
NATIONAL CANCER INSTITUTE

1) Next generation sequencing (NGS)
2) Exome sequencing
3) Experimental design
4) Mutation study resources
    - ❖Genome in a Bottle
    - ❖DREAM mutation challenge

Genome in a Bottle Consortium

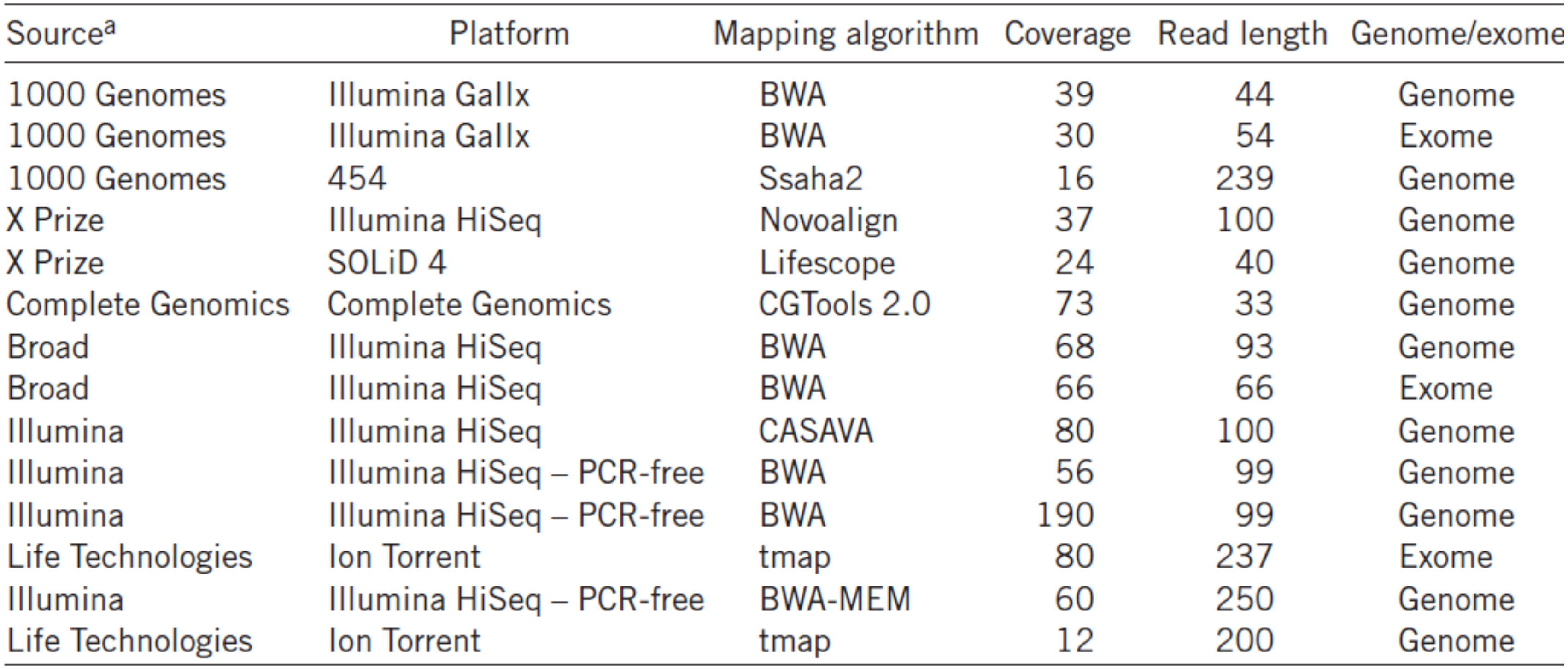
- No widely accepted set of metrics to characterize the fidelity of variant calls from NGS…
- *Genome in a Bottle Consortium* is developing standards to address this…
  - well-characterized human genomes as *Reference Materials* (RMs)
    - characterized and disseminated by NIST
  - tools and methods to use these RMs
    - Global Alliance for Genomics and Health Benchmarking Team

Genome in a Bottle Consortium



http://genomeinabottle.org

The data sets for NA12878 are available at the Genome in a Bottle ftp site at NCBI

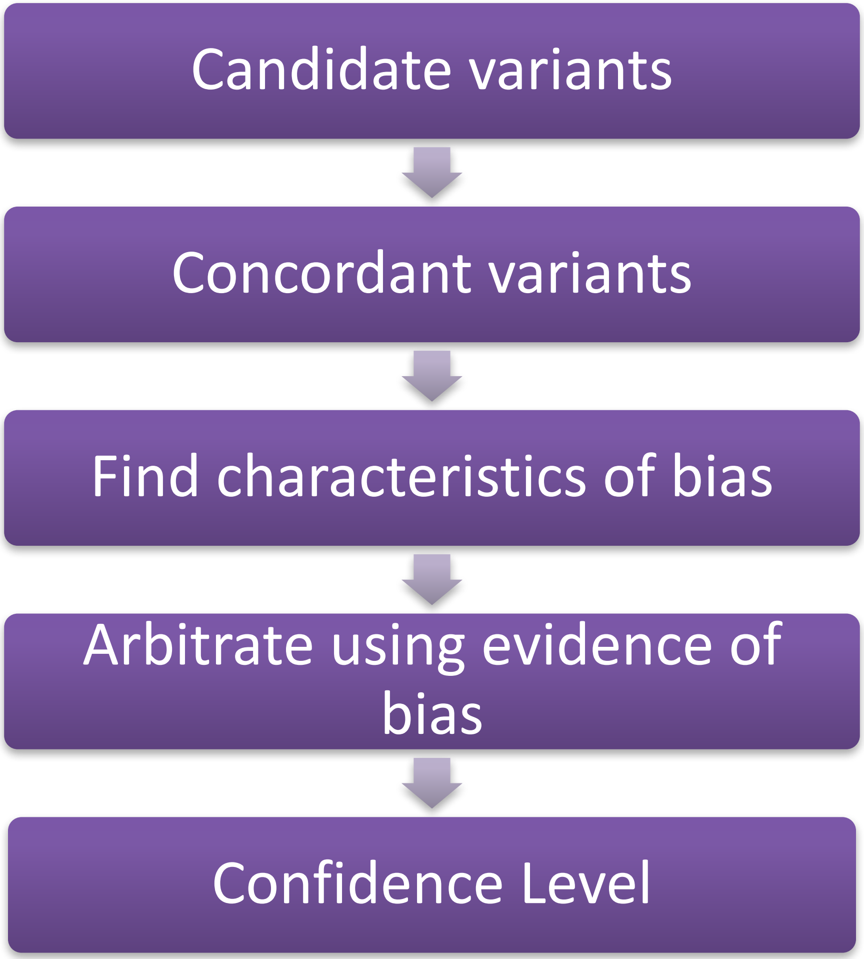


Source[a]	Platform	Mapping algorithm	Coverage	Read length	Genome/exome
1000 Genomes	Illumina GaIIx	BWA	39	44	Genome
1000 Genomes	Illumina GaIIx	BWA	30	54	Exome
1000 Genomes	454	Ssaha2	16	239	Genome
X Prize	Illumina HiSeq	Novoalign	37	100	Genome
X Prize	SOLiD 4	Lifescope	24	40	Genome
Complete Genomics	Complete Genomics	CGTools 2.0	73	33	Genome
Broad	Illumina HiSeq	BWA	68	93	Genome
Broad	Illumina HiSeq	BWA	66	66	Exome
Illumina	Illumina HiSeq	CASAVA	80	100	Genome
Illumina	Illumina HiSeq – PCR-free	BWA	56	99	Genome
Illumina	Illumina HiSeq – PCR-free	BWA	190	99	Genome
Life Technologies	Ion Torrent	tmap	80	237	Exome
Illumina	Illumina HiSeq – PCR-free	BWA-MEM	60	250	Genome
Life Technologies	Ion Torrent	tmap	12	200	Genome

ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878

Zook, J.M. et al. (2014) Nat Biotechnol. 32(3):246-51.

Integration methods to establish benchmark variant calls



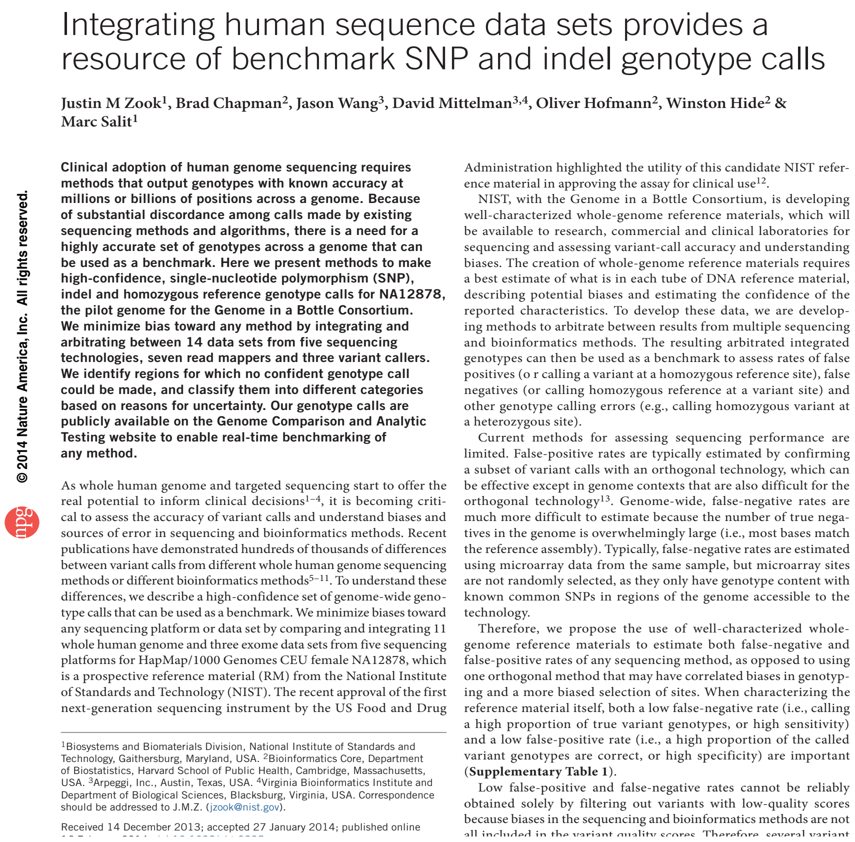
Candidate variants

Concordant variants

Find characteristics of bias

Arbitrate using evidence of bias

Confidence Level



Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin M Zook[1], Brad Chapman[2], Jason Wang[3], David Mittelman[3,4], Oliver Hofmann[2], Winston Hide[2] & Marc Salit[1]

Clinical adoption of human genome sequencing requires methods that output genotypes with known accuracy at millions or billions of positions across a genome. Because of substantial discordance among calls made by existing sequencing methods and algorithms, there is a need for a highly accurate set of genotypes across a genome that can be used as a benchmark. Here we present methods to make high-confidence, single-nucleotide polymorphism (SNP), indel and homozygous reference genotype calls for NA12878, the pilot genome for the Genome in a Bottle Consortium. We minimize bias toward any method by integrating and arbitrating between 14 data sets from five sequencing technologies, seven read mappers and three variant callers. We identify regions for which no confident genotype call could be made, and classify them into different categories based on reasons for uncertainty. Our genotype calls are publicly available on the Genome Comparison and Analytic Testing website to enable real-time benchmarking of any method.

As whole human genome and targeted sequencing start to offer the real potential to inform clinical decisions[1–4], it is becoming critical to assess the accuracy of variant calls and understand biases and sources of error in sequencing and bioinformatics methods. Recent publications have demonstrated hundreds of thousands of differences between variant calls from different whole human genome sequencing methods or different bioinformatics methods[5–11]. To understand these differences, we describe a high-confidence set of genome-wide genotype calls that can be used as a benchmark. We minimize biases toward any sequencing platform or data set by comparing and integrating 11 whole human genome and three exome data sets from five sequencing platforms for HapMap/1000 Genomes CEU female NA12878, which is a prospective reference material (RM) from the National Institute of Standards and Technology (NIST). The recent approval of the first next-generation sequencing instrument by the US Food and Drug Administration highlighted the utility of this candidate NIST reference material in approving the assay for clinical use[12].

NIST, with the Genome in a Bottle Consortium, is developing well-characterized whole-genome reference materials, which will be available to research, commercial and clinical laboratories for sequencing and assessing variant-call accuracy and understanding biases. The creation of whole-genome reference materials requires a best estimate of what is in each tube of DNA reference material, describing potential biases and estimating the confidence of the reported characteristics. To develop these data, we are developing methods to arbitrate between results from multiple sequencing and bioinformatics methods. The resulting arbitrated integrated genotypes can then be used as a benchmark to assess rates of false positives (o r calling a variant at a homozygous reference site), false negatives (or calling homozygous reference at a variant site) and other genotype calling errors (e.g., calling homozygous variant at a heterozygous site).

Current methods for assessing sequencing performance are limited. False-positive rates are typically estimated by confirming a subset of variant calls with an orthogonal technology, which can be effective except in genome contexts that are also difficult for the orthogonal technology[13]. Genome-wide, false-negative rates are much more difficult to estimate because the number of true negatives in the genome is overwhelmingly large (i.e., most bases match the reference assembly). Typically, false-negative rates are estimated using microarray data from the same sample, but microarray sites are not randomly selected, as they only have genotype content with known common SNPs in regions of the genome accessible to the technology.

Therefore, we propose the use of well-characterized whole-genome reference materials to estimate both false-negative and false-positive rates of any sequencing method, as opposed to using one orthogonal method that may have correlated biases in genotyping and a more biased selection of sites. When characterizing the reference material itself, both a low false-negative rate (i.e., calling a high proportion of true variant genotypes, or high sensitivity) and a low false-positive rate (i.e., a high proportion of the called variant genotypes are correct, or high specificity) are important (**Supplementary Table 1**).

Low false-positive and false-negative rates cannot be reliably obtained solely by filtering out variants with low-quality scores because biases in the sequencing and bioinformatics methods are not all included in the variant quality scores. Therefore, several variant

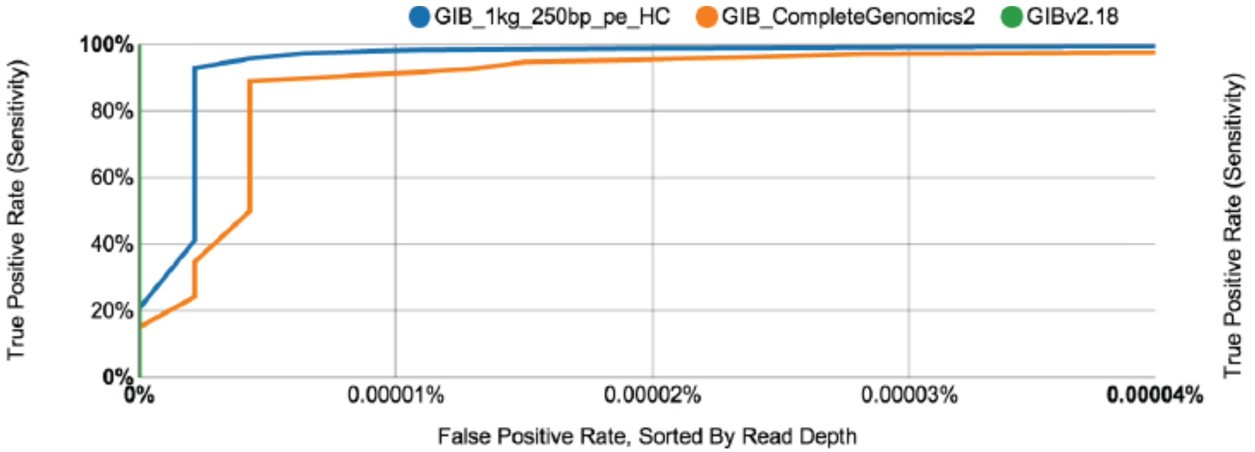
[1] Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA. [2] Bioinformatics Core, Department of Biostatistics, Harvard School of Public Health, Cambridge, Massachusetts, USA. [3] Arpeggi, Inc., Austin, Texas, USA. [4] Virginia Bioinformatics Institute and Department of Biological Sciences, Blacksburg, Virginia, USA. Correspondence should be addressed to J.M.Z. (jzook@nist.gov).

Received 14 December 2013; accepted 27 January 2014; published online

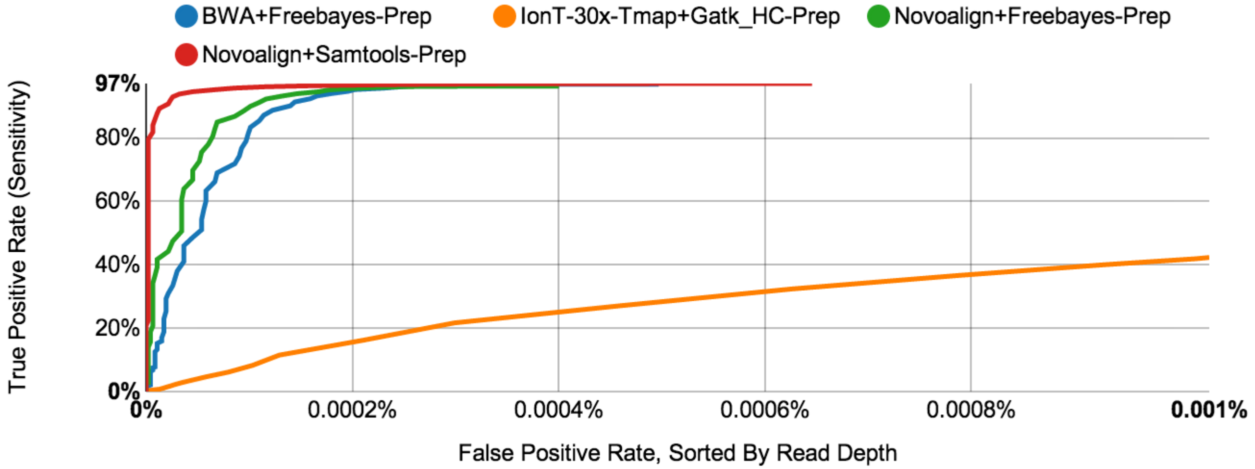
Zook, JM et al (2014) Nat Biotechnol. 32(3):246-51.

http://www.slideshare.net/GenomeInABottle/presentations

~2.7M high confident snps are detected by multiple algorithms



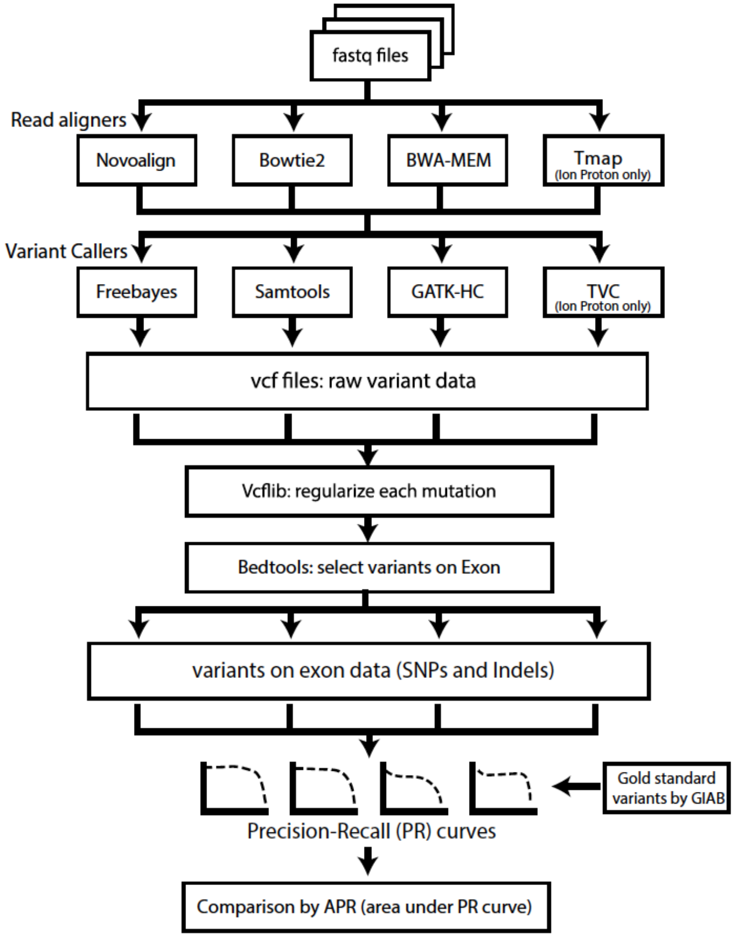
Exome SNPs



Exome SNPs

Zook, J.M. et al. (2014) Nat Biotechnol. 32(3):246-51.

Systematic comparison of variant calling pipelines using GIAB



fastq files

Read aligners

Novoalign

Bowtie2

BWA-MEM

Tmap (Ion Proton only)

Variant Callers

Freebayes

Samtools

GATK-HC

TVC (Ion Proton only)

vcf files: raw variant data

Vcflib: regularize each mutation

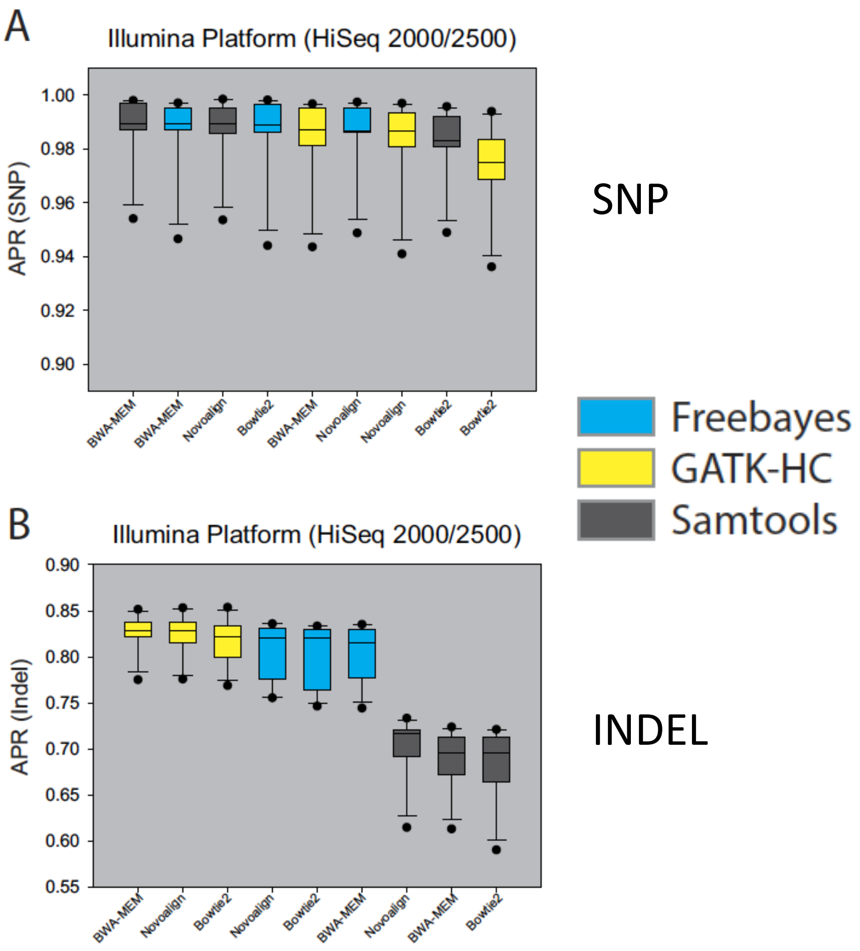
Bedtools: select variants on Exon

variants on exon data (SNPs and Indels)

Precision-Recall (PR) curves

Gold standard variants by GIAB

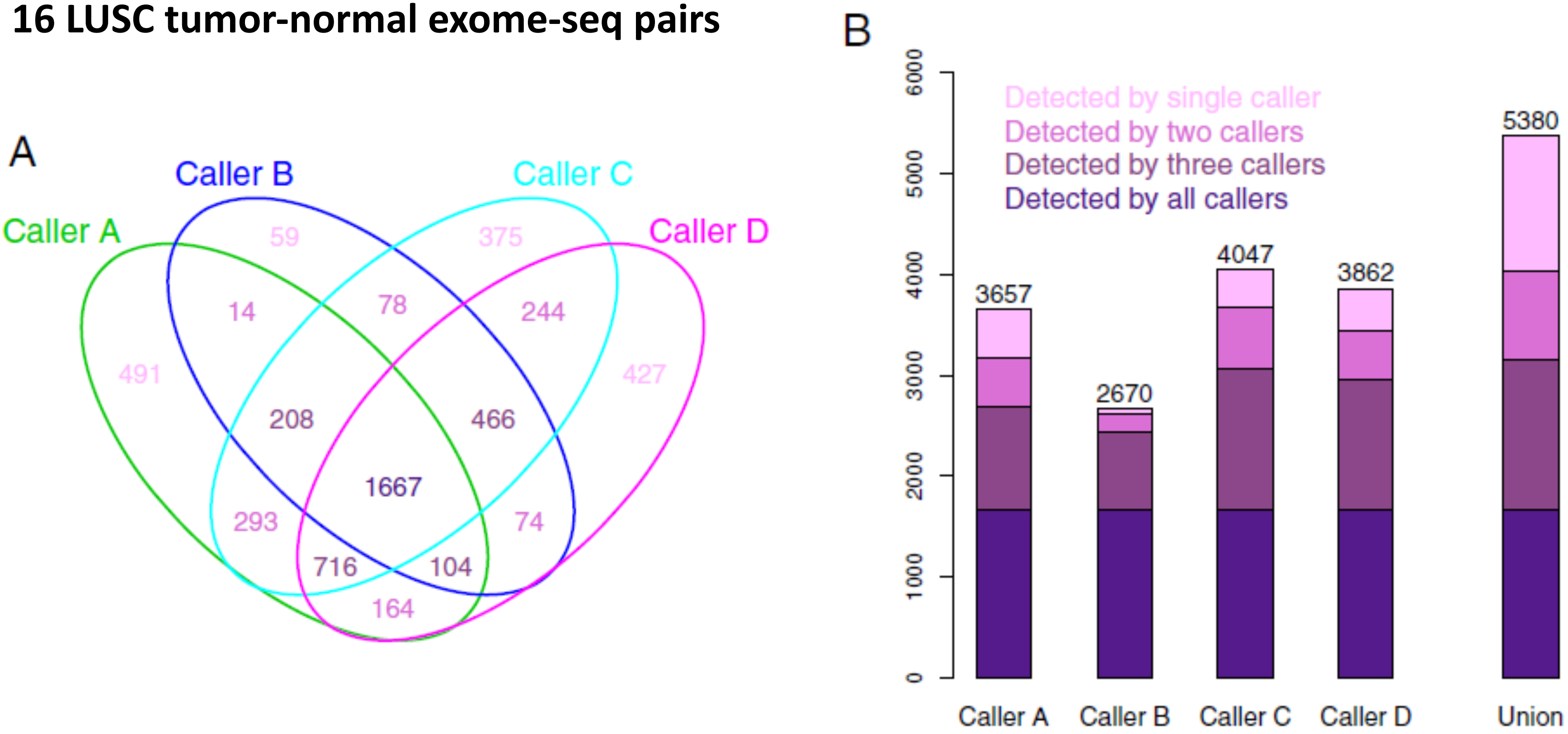
Comparison by APR (area under PR curve)



Hwang S, Kim E, Lee I, Marcotte EM. (2015) Sci Rep. 5:17875.

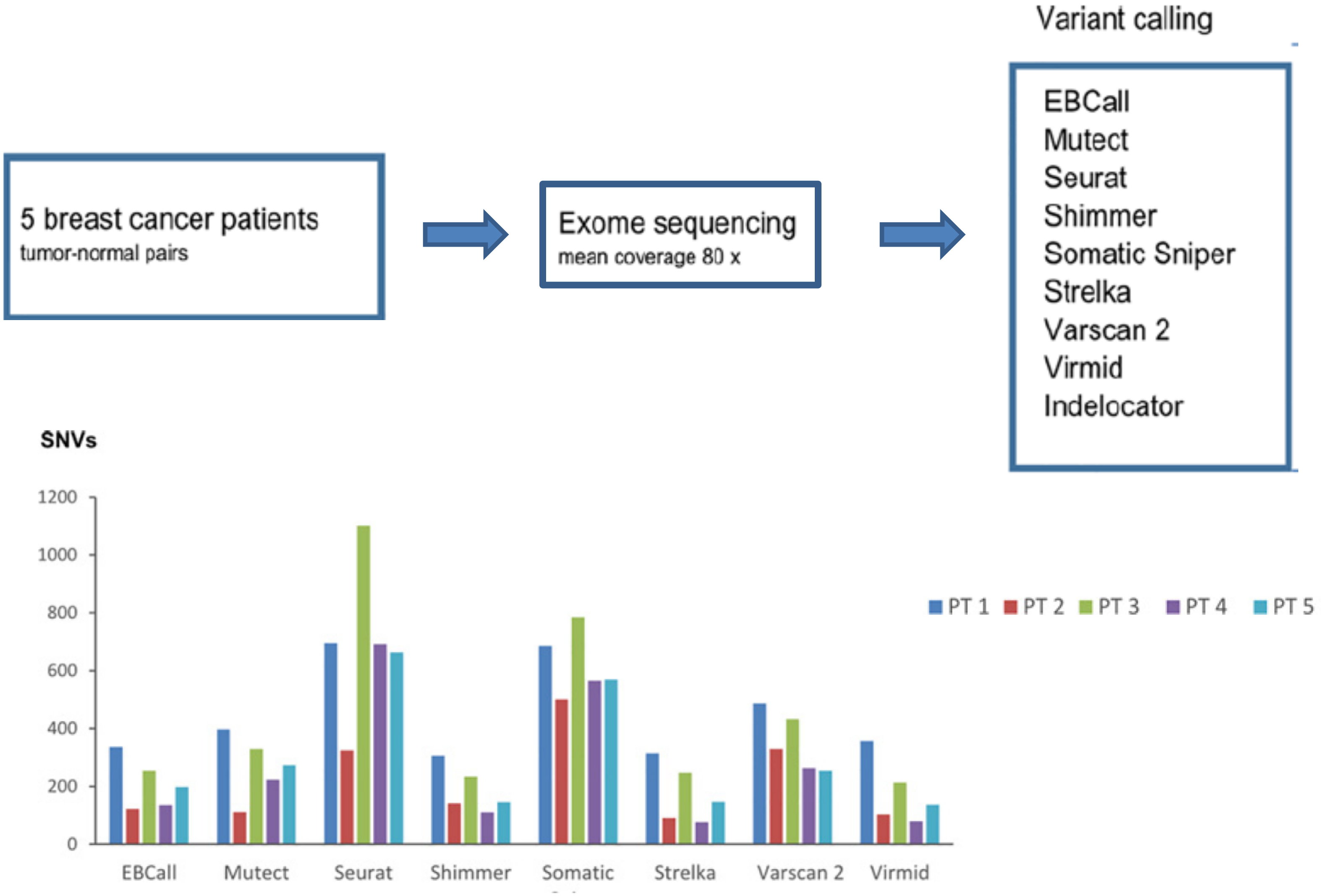
The mutation caller performance varies drastically, 2013

16 LUSC tumor-normal exome-seq pairs



Kim, S.Y., Speed, T.P. (2013) BMC Bioinformatics. 10;14:189.

Evaluation of somatic mutation callers 2016



5 breast cancer patients
tumor-normal pairs

Exome sequencing
mean coverage 80 x

Variant calling

- EBCall
- Mutect
- Seurat
- Shimmer
- Somatic Sniper
- Strelka
- Varscan 2
- Virmid
- Indelocator

Krøigård AB et al (2016) PLoS One. 11(3):e0151664.

DREAM
CHALLENGES
powered by Sage Bionetworks
CONTACT US | NEWS
CHALLENGES | ABOUT DREAM | OUR COMMUNITY | PUBLICATIONS
DREAM Challenges

ICGC-TCGA-DREAM Somatic Mutation Calling Challenge

ICGC-TCGA-DREAM Somatic Mutation Calling Challenge

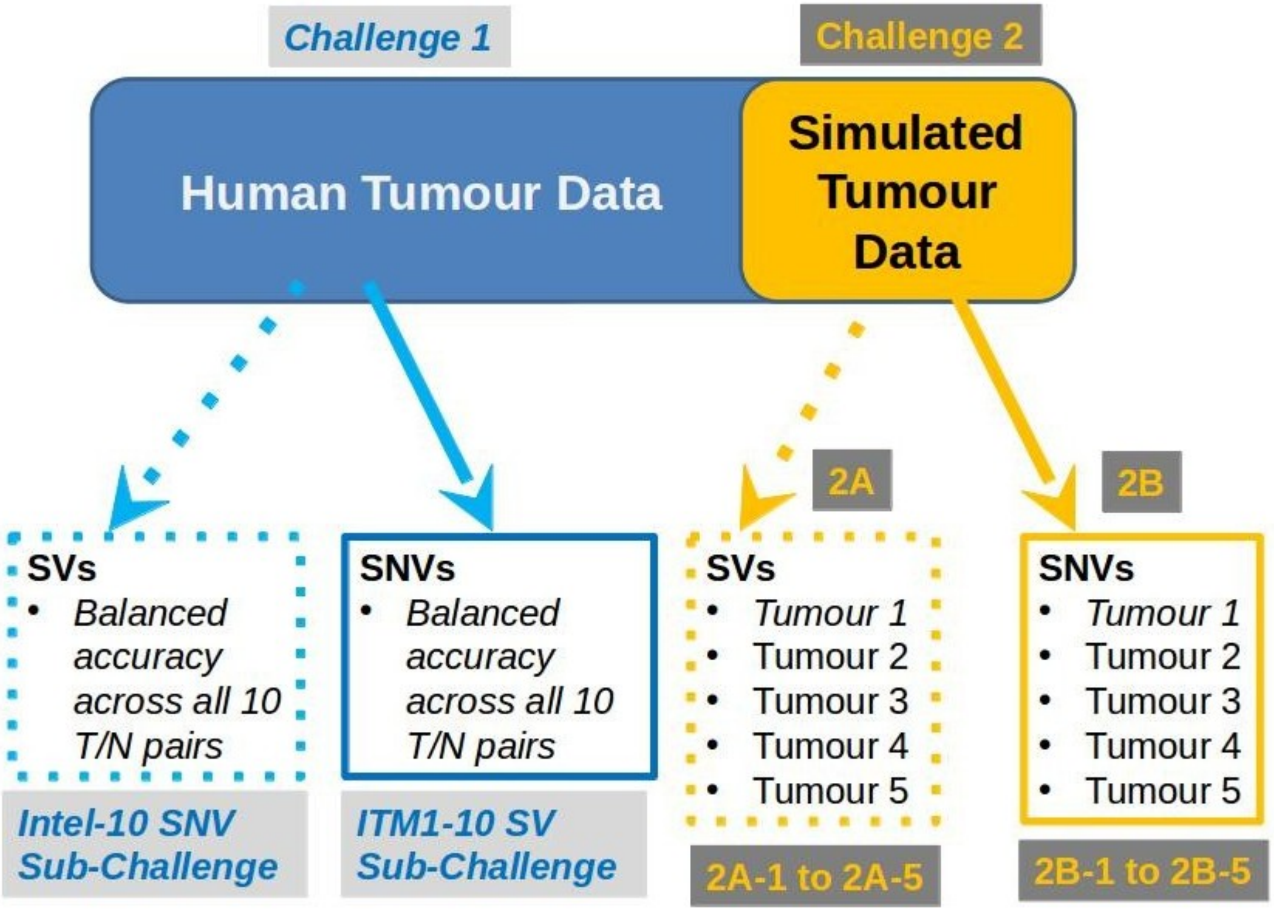
DEC 12, 2013 – NOV 1, 2014 **Closed**

ICGC INOVA Join the future of health. OICR Ontario Institute for Cancer Research

The Cancer Genome Atlas Understanding genomics to improve cancer care UNIVERSITY OF CALIFORNIA SANTA CRUZ

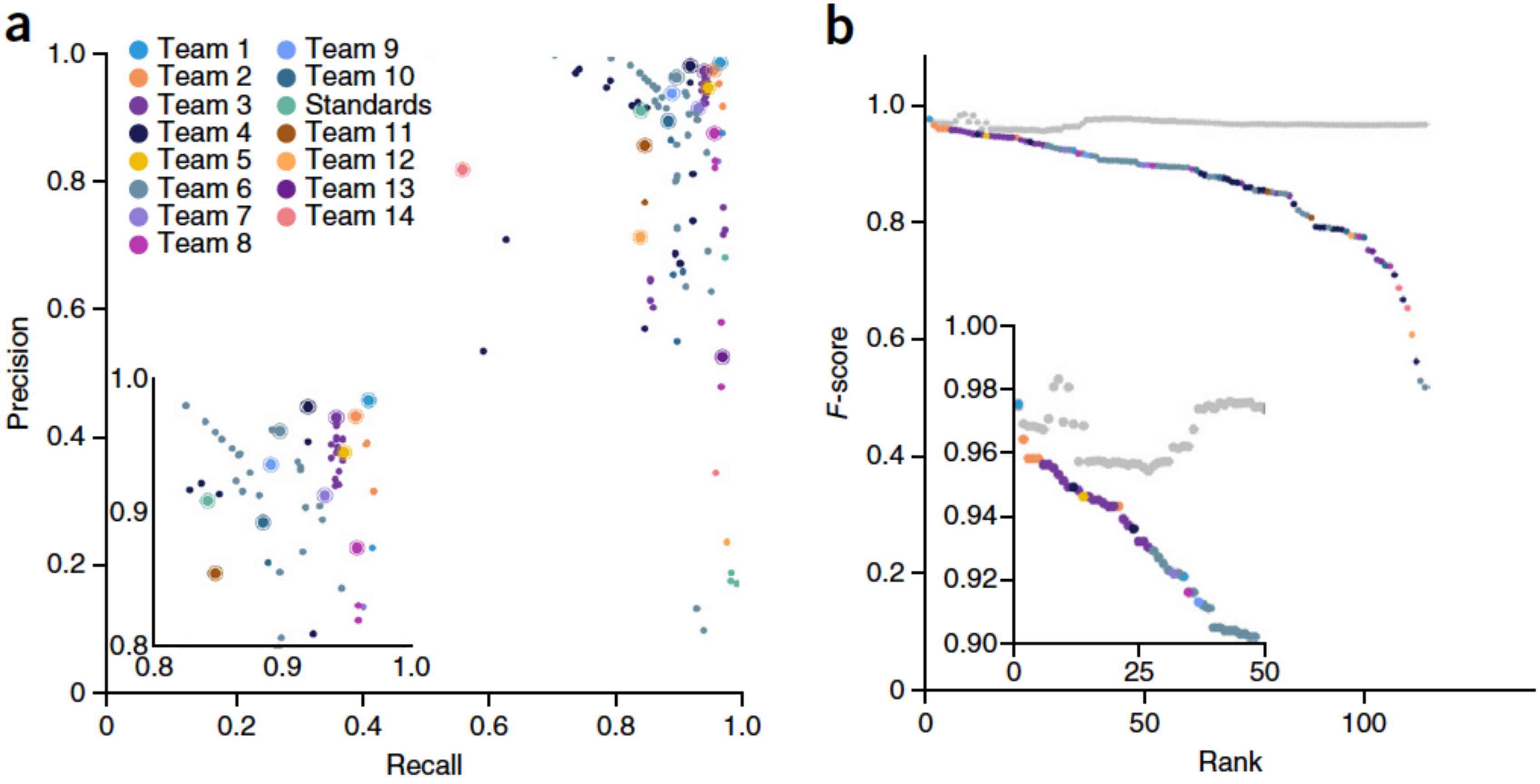
http://dreamchallenges.org

Challenge data and assessment



http://dreamchallenges.org

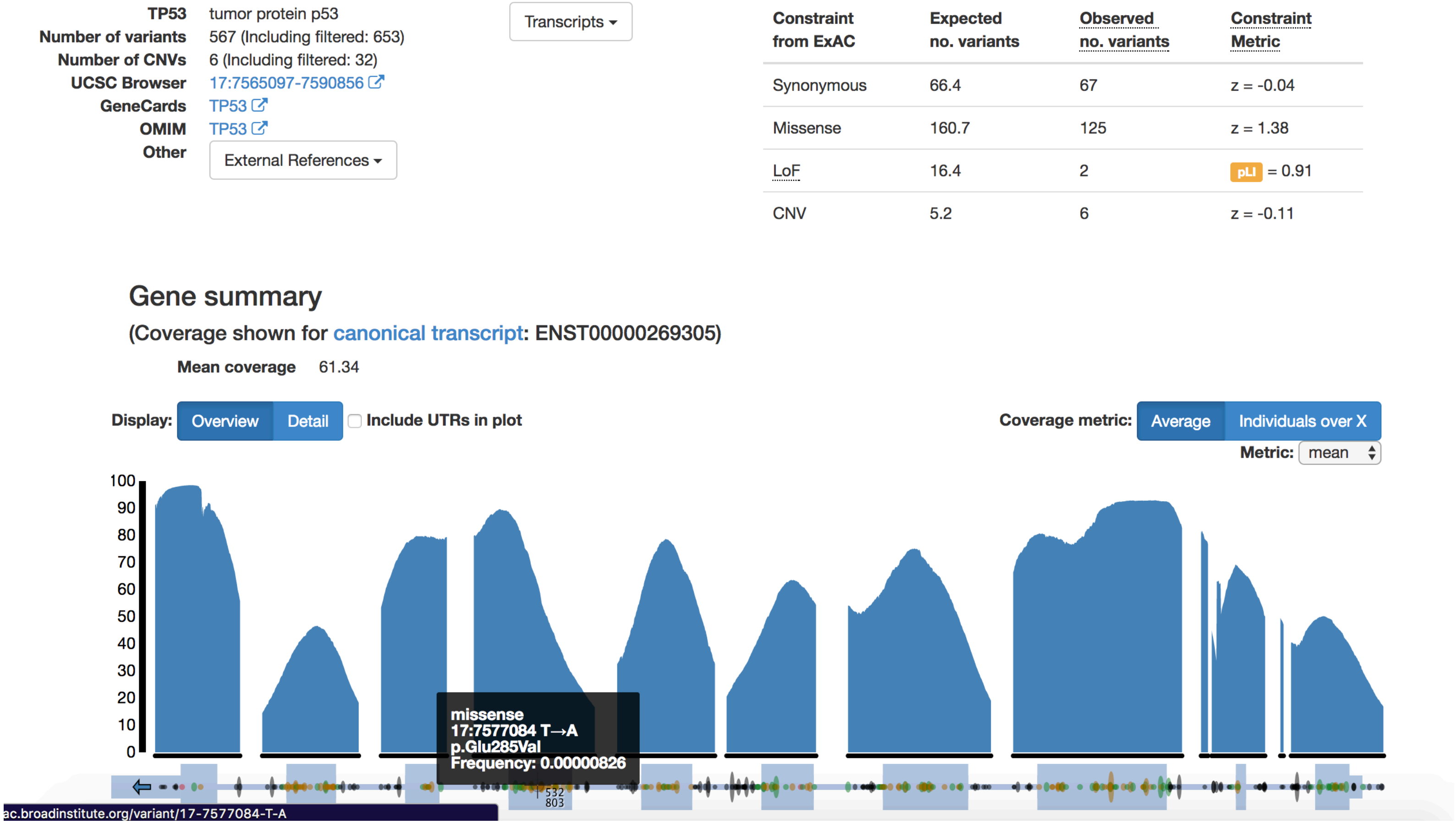
Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection



Ewing, A.D. et al. (2015) Nat Methods. 12(7):623-30.

nature methods

ExAc: Exome Aggregation Consortium



TP53	tumor protein p53
Number of variants	567 (Including filtered: 653)
Number of CNVs	6 (Including filtered: 32)
UCSC Browser	17:7565097-7590856
GeneCards	TP53
OMIM	TP53
Other	External References ▾

Transcripts ▾

Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	66.4	67	z = -0.04
Missense	160.7	125	z = 1.38
LoF	16.4	2	pLI = 0.91
CNV	5.2	6	z = -0.11

Gene summary

(Coverage shown for canonical transcript: ENST00000269305)

Mean coverage 61.34

Display: Overview Detail □ **Include UTRs in plot**

Coverage metric: Average Individuals over X

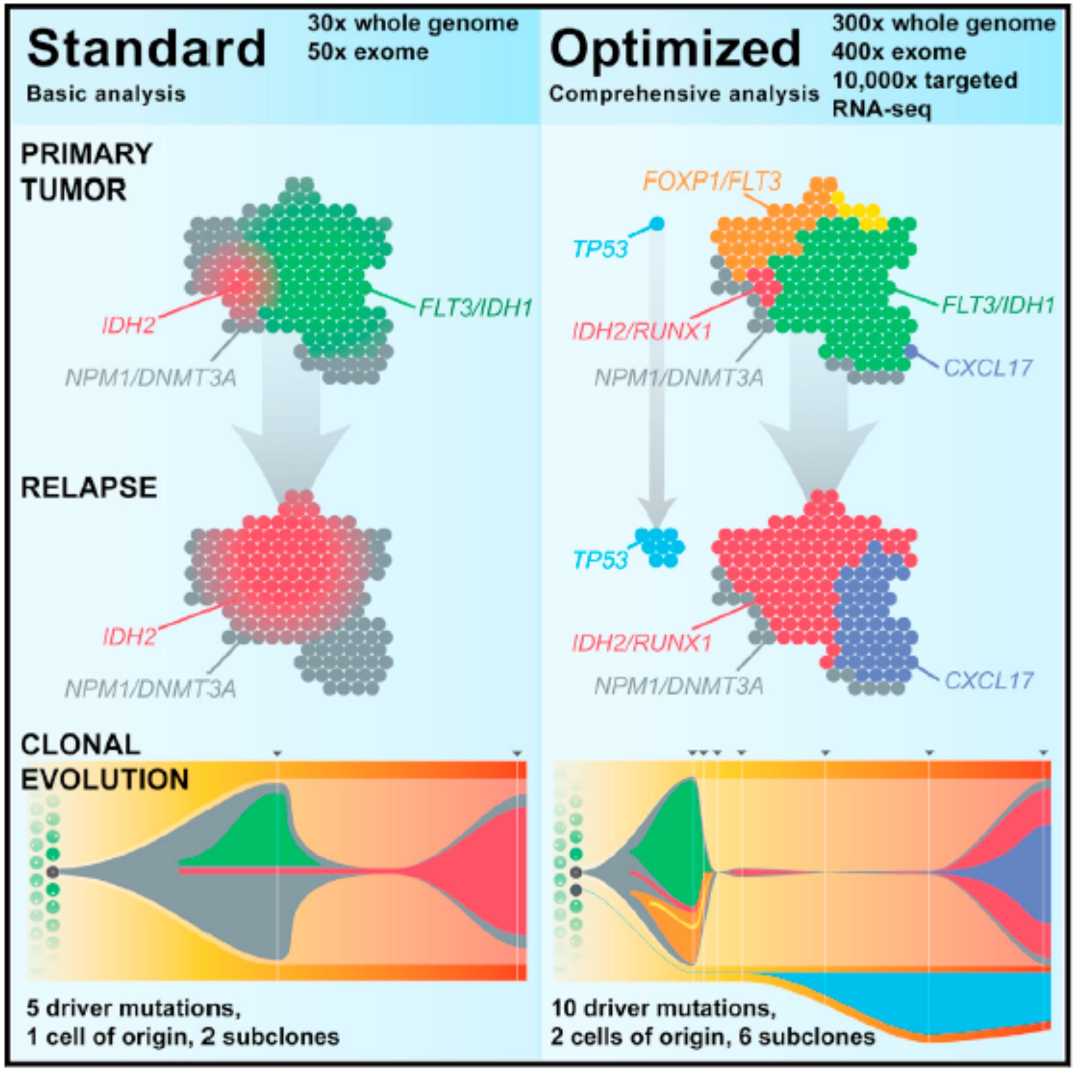
Metric: mean

missense
17:7577084 T→A
p.Glu285Val
Frequency: 0.00000826

http://exac.broadinstitute.org

Optimizing Cancer Genome Sequencing and Analysis

Graphical Abstract



Authors

Malachi Griffith, Christopher A. Miller,
Obi L. Griffith, ..., Elaine R. Mardis,
Timothy J. Ley, Richard K. Wilson

1) Sample and case selection
2) Matched normal samples
3) Library construction
4) Sequencing platform
5) Sequencing depth
6) Exome-seq
7) Whole genome sequecing
8) Targeted sequencing
9) Sequence alignment
10) Variant calling
11) Subclonal inference
12) RNA-seq

Griffith M. et al (2015) Cell Syst. 1(3):210-223.