Computational Approaches for Cancer Genome Analysis with Next-Generation Sequencing

Matthew Meyerson, M.D., Ph.D.

Dana-Farber Cancer Institute
Broad Institute of Harvard and MIT
Harvard Medical School

Conflicts of interest

Receive research support from Genentech

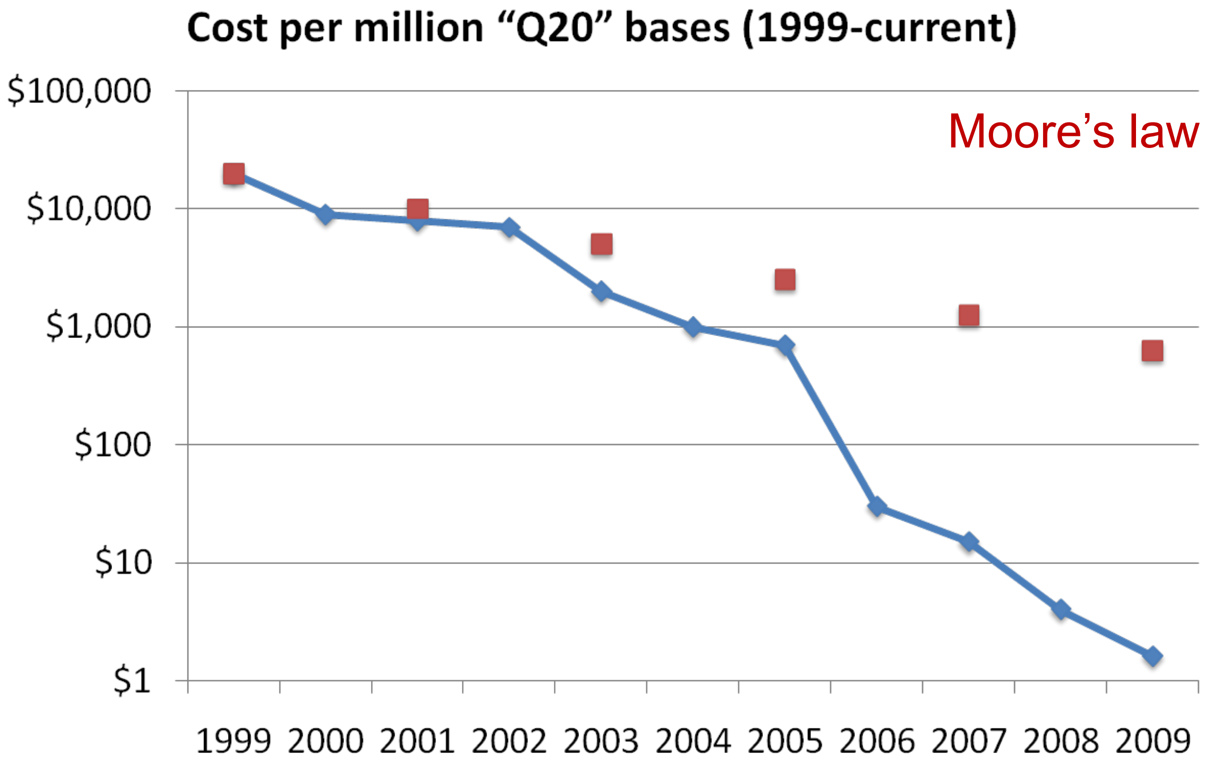
Receive research support from, and consult for, Novartis

Founding advisor, consultant for, and equity holder in, Foundation Medicine

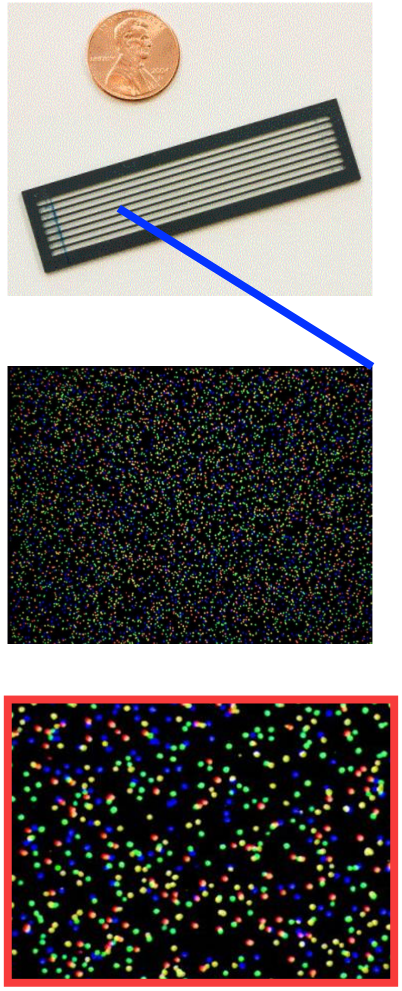
Inventor on patent for use of EGFR mutations as method of diagnosis for lung cancer, licensed to Genzyme Genetics

Why use next-generation sequencing to analyze cancer genomes?

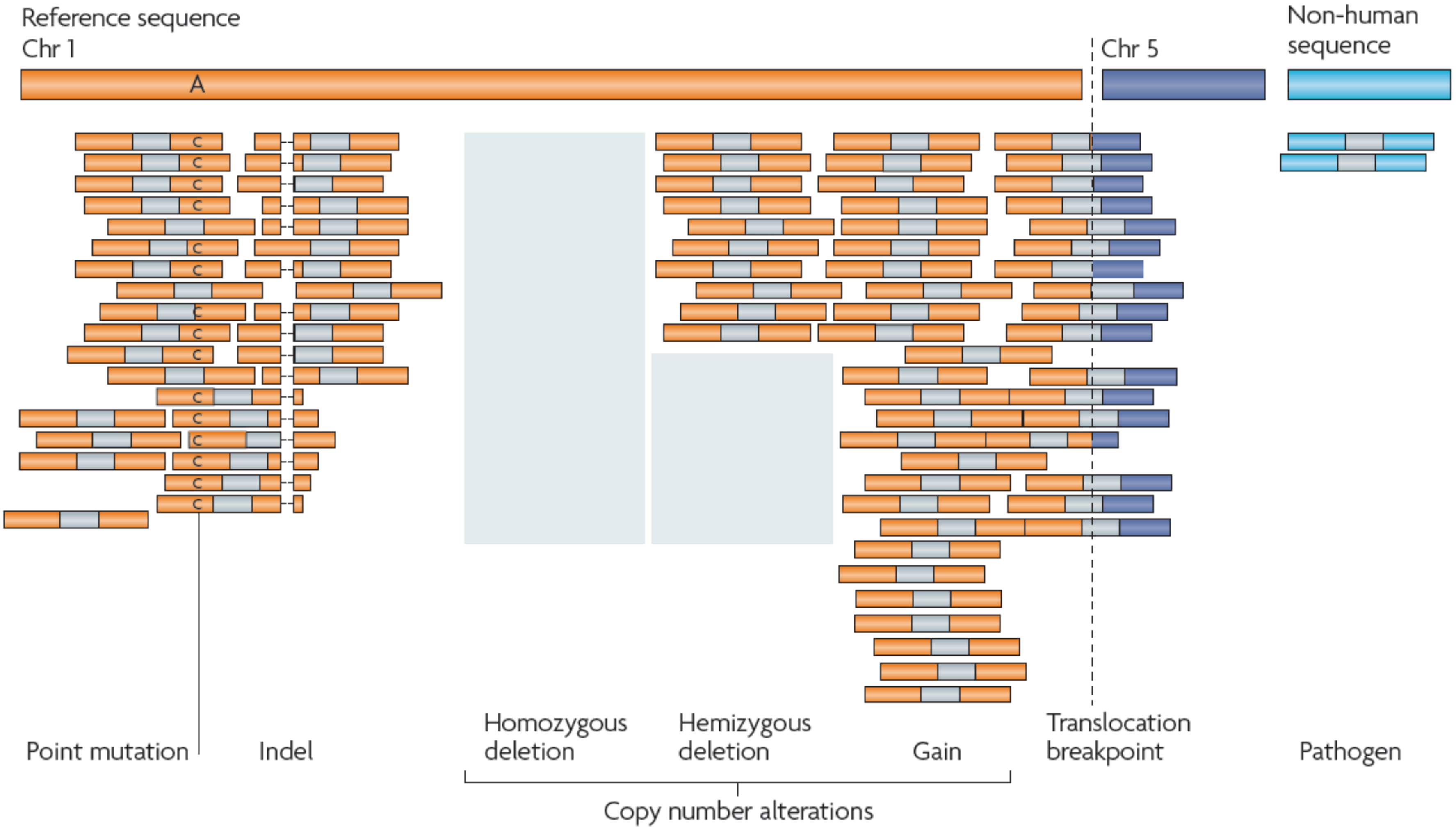
Why sequence? Technology gets better and cheaper…



Cost per million “Q20” bases (1999-current)



Why sequence? Next-generation sequencing allows us to detect all classes of genome alterations



muTect
Kristian Cibulskis

Indelocator
Andrey Sivachenko

SegSeq
Derek Chiang, Gordon Saksena

dRanger
Mike Lawrence
Yotam Drier

PathSeq
Alex Kostic

Gad Getz

Unique features of cancer genomes

Normal and cancer genomes

"Happy families are all alike; every unhappy family is unhappy in its own way".

Leo Tolstoy, *Anna Karenina*

Normal genomes are all (mostly) alike; every cancer genome is abnormal in its own way.

Somatic genome alterations in cancer

Somatic alterations are the major cause of cancer

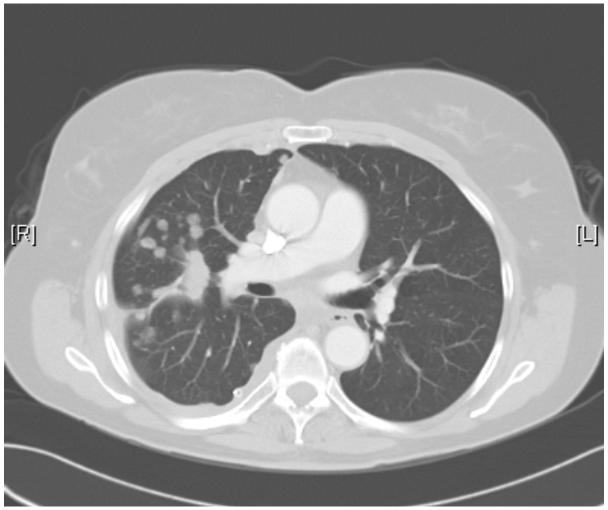
Definition: genome alterations present in the cancer but not in the germ-line

Somatic alterations provide target for therapy

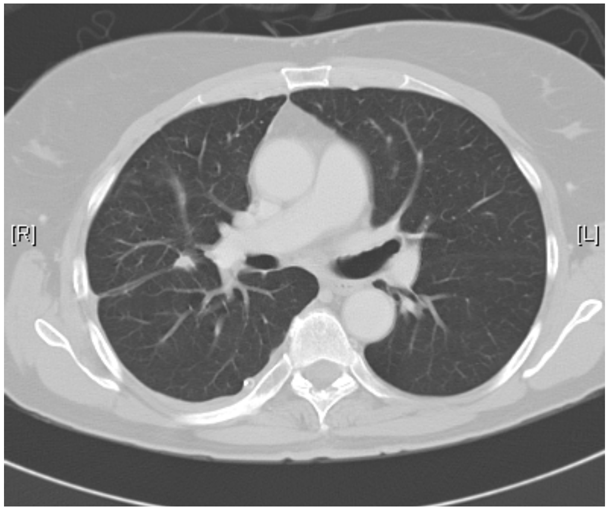
Because these alterations are present only in the tumor, there can be a large "therapeutic window" where toxicity to cancer vastly exceeds toxicity to normal cells

Example: a patient with lung adenocarcinoma, with a somatic *EGFR* deletion mutant in exon 19 (thanks to Bruce Johnson, M.D., DFCI)

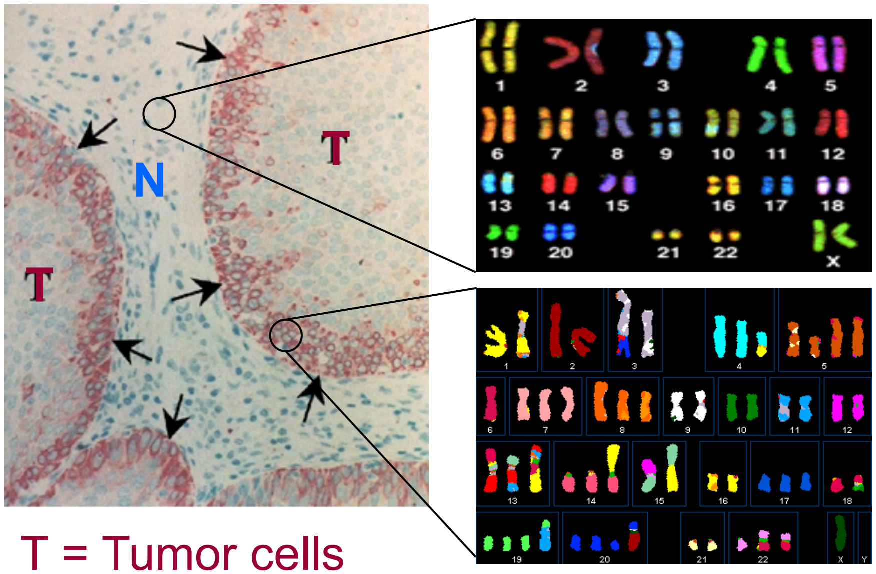
Before treatment



After 2 months erlotinib treatment



Cancer samples represent complex mixtures of cells with distinct genomes

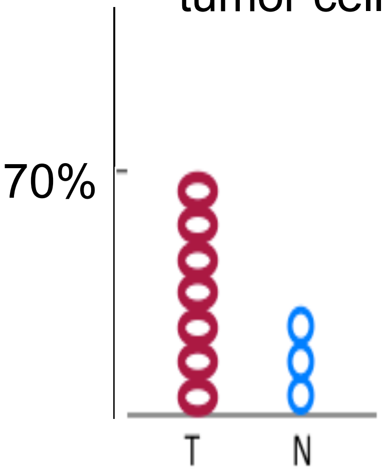


T = Tumor cells

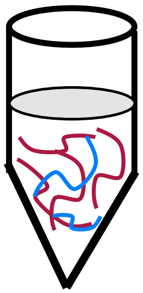
N = Normal cells

Ploidy = mass of DNA in units of normal haploid genome mass. Here ~2.7.

Purity = fraction of tumor cells



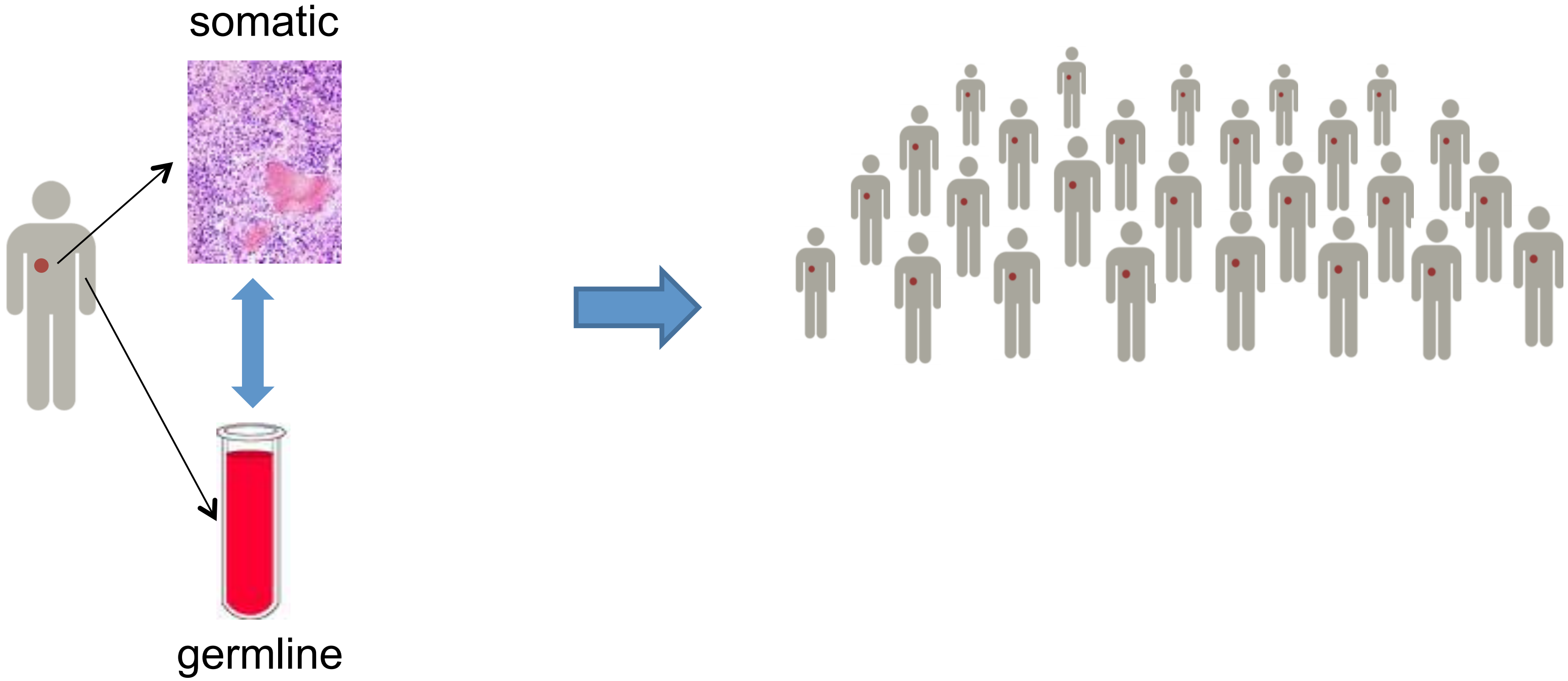
Aliquot of mixed tumor and normal DNA



Because next-generation sequencing is digital and not analog, it is possible to dissect the cancer specific signal from the normal signal by computational analysis of sequence counts at every base position

Goals of cancer genome computational analysis: discovery of cancer genes

Individual **Population**



somatic

germline

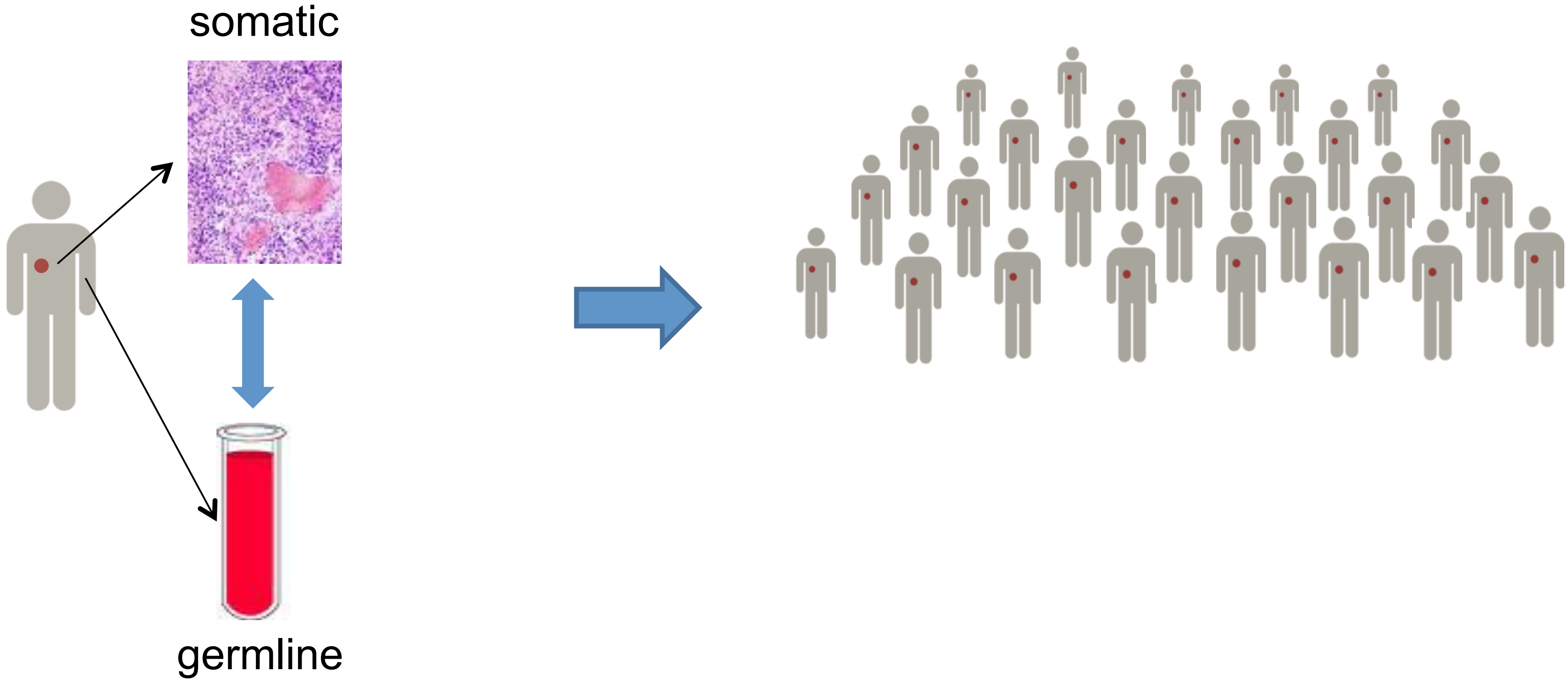
What is the full set of genome alterations within the cancer (and germ-line)—mutations, copy number, translocations, etc?

(1) Which genome alterations are **statistically significant** in the population?
(2) In which **genes** and **pathways** do these alterations occur?

Goals of cancer genome computational analysis: diagnosis

Individual

Population



somatic

germline

What **actionable genome alterations** are carried in the germ-line or somatically altered in the tumor of a particular patient?

(1) Do these alterations predict the natural history of the cancer, inc. **prognosis**?
(2) Do these alterations predict the **response to specific therapies** in clinical trials?

Suppose you have a collection of next-generation sequencing data: what do you do?

Steps of cancer genome analysis with next-generation sequencing

Getting started

Data quality control

Alignment

Variant calling

Visualization

Artifact removal

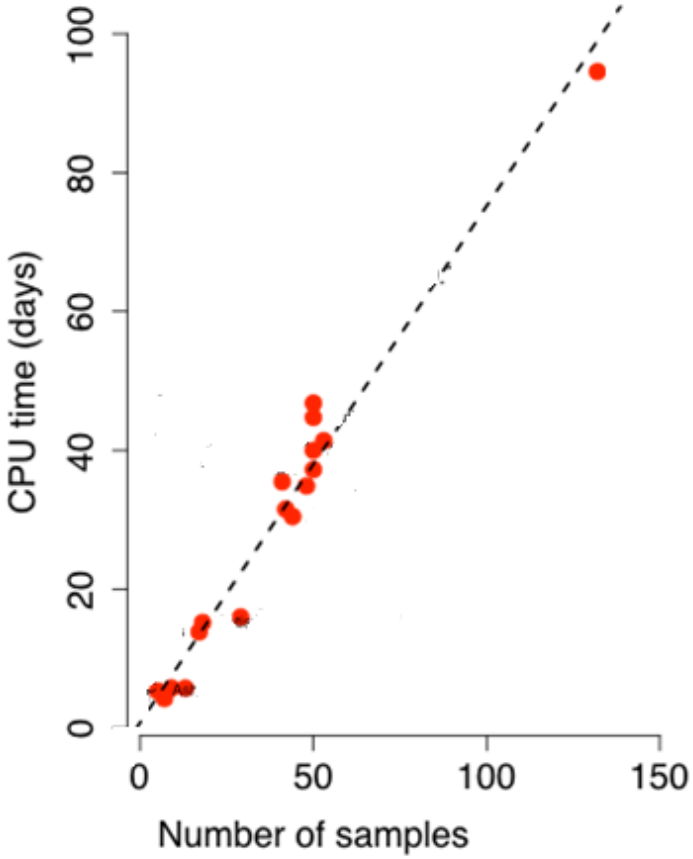
Significance analysis

Analysis of public data sets

Getting started with next-generation sequencing analysis of cancer: some choices

- Hardware
  - Build a cluster
  - Use the cloud
  - Contract it out
- Software
  - Publically available tools
  - Commercial tools
- People
  - Collaborate
  - Build a team
  - Contract it out

Getting started: CPU and storage costs for next-generation sequencing



Storage requirements

Data type	Target	Storage
Per-sample		
Exome	32 Mb	30-50 Gb
Genome	2.85 Gb	250 Gb
Complete Project		
200 exome pairs	32 Mb	20 Tb
50 genome pairs	2.85 Gb	25 Tb

In general, need access to a cluster or a cloud to obtain enough CPU power

Kiran Garimella and Mark DePristo

Getting started: Publically available software tools for next-gen sequence analysis of cancer

Category	Method	URL
Alignment		
	MAQ	http://maq.sourceforge.net
	BWA	http://bio-bwa.sourceforge.net
	ELAND	http://www.illumina.com
	SSAHA2	http://www.sanger.ac.uk/resources/software/ssaha2
	Bowtie	http://bowtie-bio.sourceforge.net/index.shtml
	SOAP2	http://soap.genomics.org.cn
	SHRiMP	http://compbio.cs.toronto.edu/shrimp
	Corona Lite	http://solidsoftwaretools.com/gf/project/corona
	BFAST	http://bfast.sourceforge.net
Mutation calling		
	GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
	SNVMix	http://www.bcgsc.ca/platform/bioinfo/software/SNVMix
	CASAVA	http://www.illumina.com/software/genome_analyzer_slftware.ilmn
	Samtools	http://samtools.sourceforge.net
	Unified genotyper	http://www.broadinstitute.org/gsa/wiki/index.php/unified_genotyper
	VarScan	http://varscan.sourceforge.net
Indel calling		
	Pindel	http://www.ebi.ac.uk/~kye/pindel
Copy number analysis		
	CBS	http://www.bioconductor.org
	SegSeq	http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=182
Pathogen detection		
		http://www.broadinstitute.org/software/pathseq/
Visualization		
	CIRCOS	http://mkweb.bcgsc.ca/circos
	IGV	http://www.broadinstitute.org/igv

Meyerson, Gabriel, Getz, *Nat Rev Genetics,* 2010

Getting started: people's qualities needed to analyze next-gen cancer genome sequence data

Necessary knowledge and attitudes may be achieved by one person or by communication within a team

Understanding the features of the cancer genome

- Heterogeneity, purity, altered ploidy, somatic nature of mutations

Understanding and applying statistical principles

- Significance analysis, outliers, error models

Enjoying diving into the data

- Visualizing, browsing, annotating, exploring…

Ability to store, retrieve and manipulate data

- Databases, file systems, input/output, nomenclature

Ability to automate analytical processes

- Even when using off-the-shelf software, ability to write simple scripts is needed

Steps of cancer genome analysis with next-generation sequencing

Getting started

Data quality control

Alignment

Variant calling

Visualization

Artifact removal

Significance analysis

Analysis of public data sets

Data quality control: how do you know if your sequence data is worth analyzing?

Is it the right sample?

- Species matching?
- Tumor/normal genotype matching?
- Gender and other fingerprint matching?
- Similarity to other known tumor genomes?

Is the raw sequence quality sufficient?

- Quality scores from instrument run
- Internal positive controls (e.g. PhiX174 control for Illumina)

Does the sequence align to the proper reference?

- Degree of alignment to genome, transcriptome, or exome reference

Is coverage of the desired targets sufficient?

- On-target percentage for hybrid capture
- Library complexity (# of unique input DNA molecules)

Steps of cancer genome analysis with next-generation sequencing

Getting started

Data quality control

Alignment

Variant calling

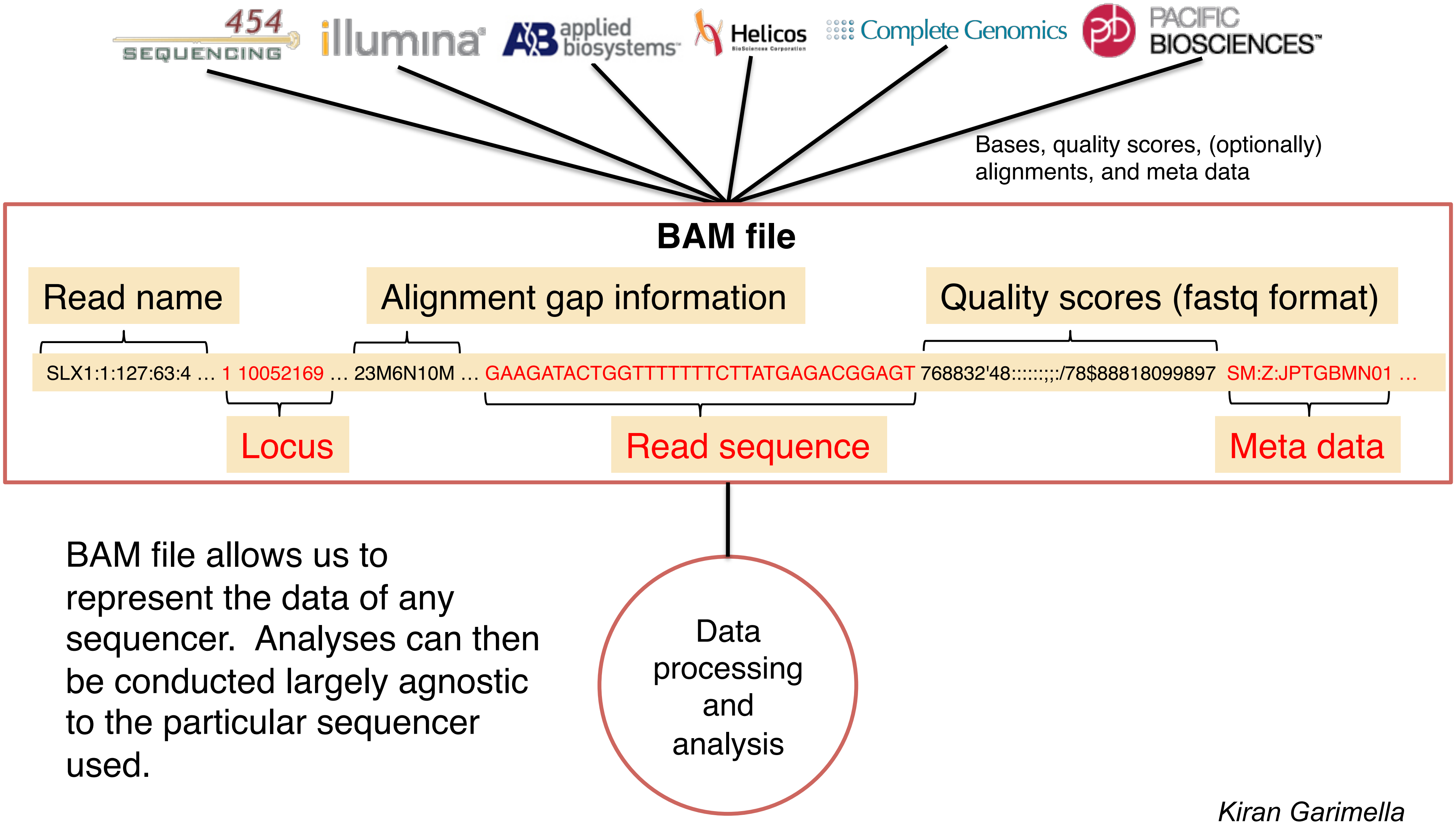
Visualization

Artifact removal

Significance analysis

Analysis of public data sets

BAM files are a standard format for sequencer-agnostic analyses



Bases, quality scores, (optionally) alignments, and meta data

BAM file

Read name

Alignment gap information

Quality scores (fastq format)

SLX1:1:127:63:4 … 1 10052169 … 23M6N10M … GAAGATACTGGTTTTTTTCTTATGAGACGGAGT 768832'48::::::;;:/78$88818099897 SM:Z:JPTGBMN01 …

Locus

Read sequence

Meta data

Data processing and analysis

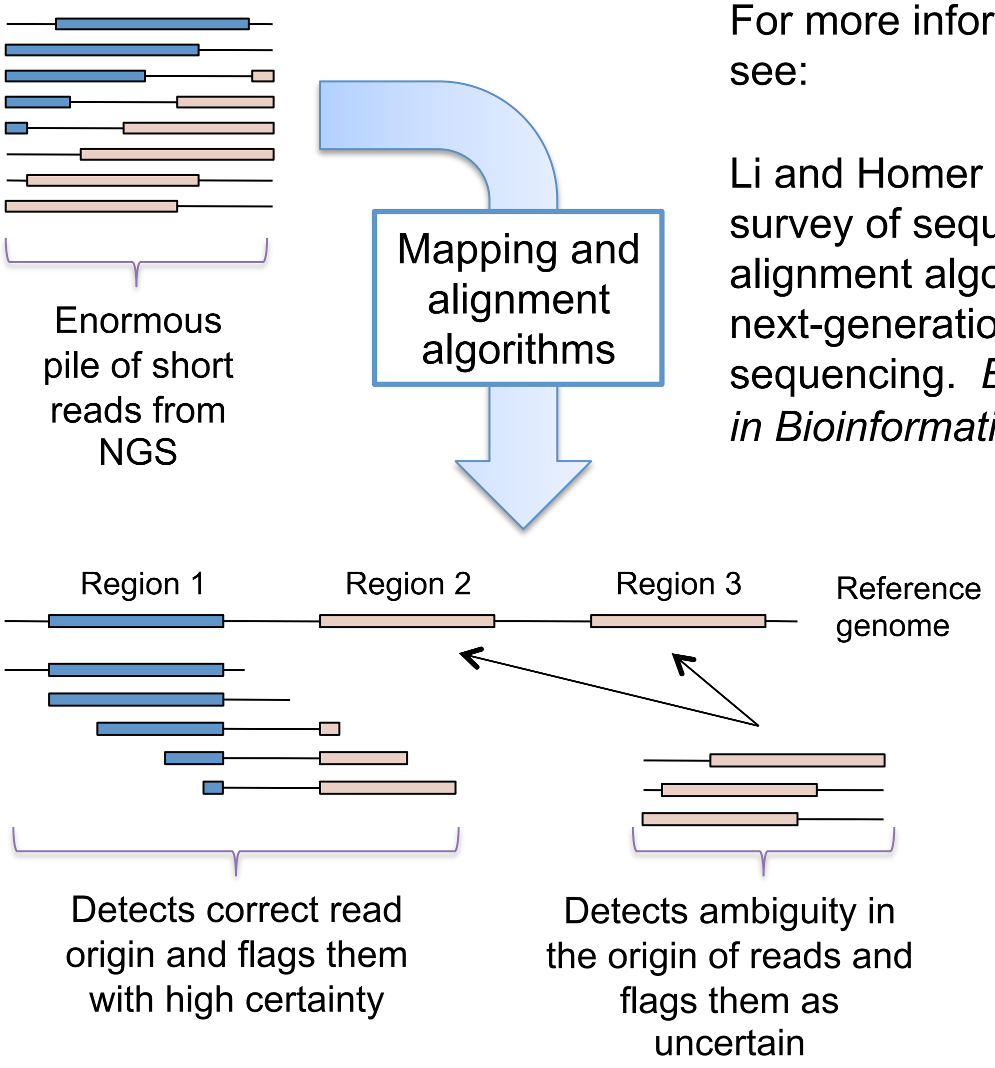
BAM file allows us to represent the data of any sequencer. Analyses can then be conducted largely agnostic to the particular sequencer used.

Kiran Garimella

Accurate alignment and mapping is key

For more information see:

Li and Homer (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*.



Enormous pile of short reads from NGS

Mapping and alignment algorithms

Region 1

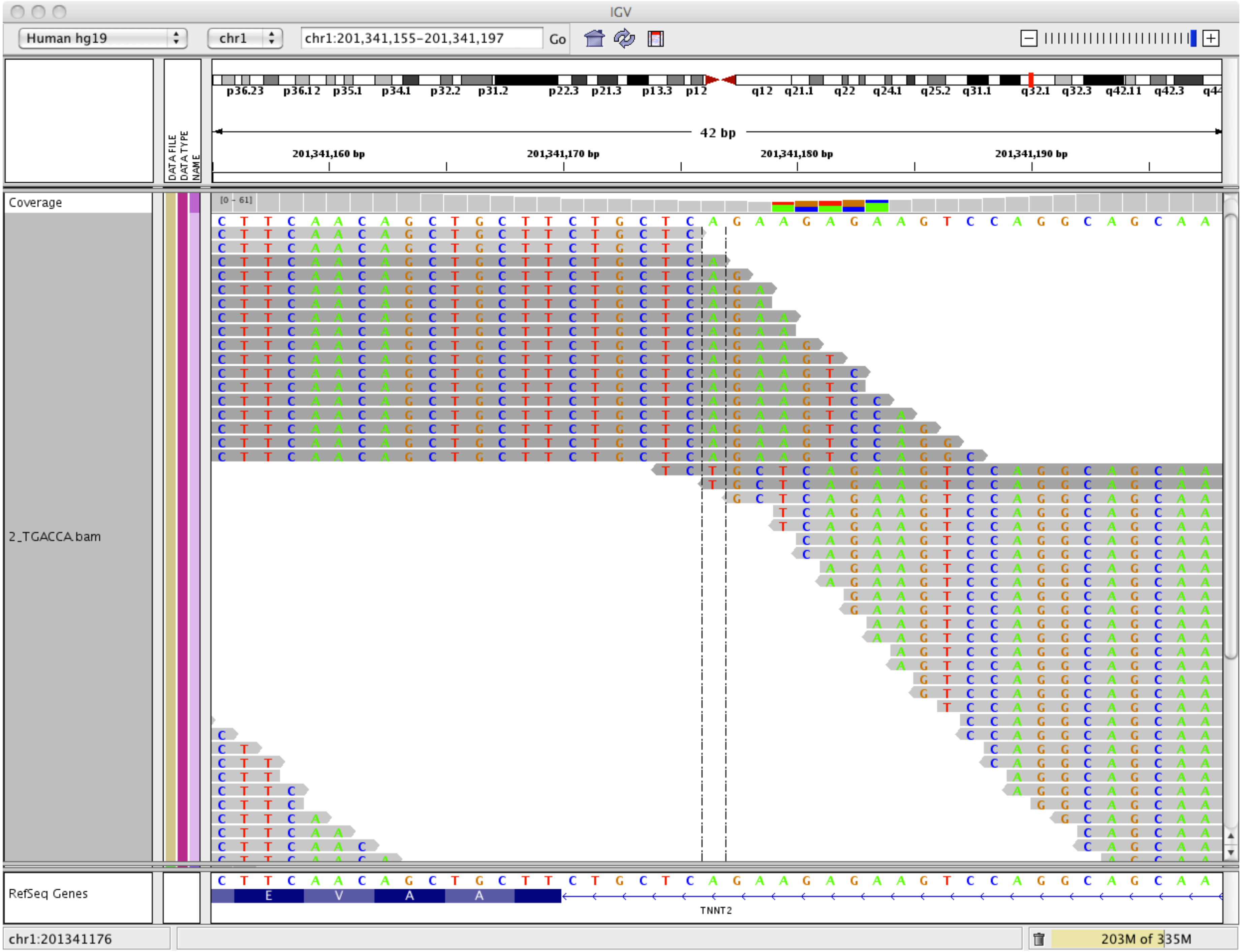
Region 2

Region 3

Reference genome

Detects correct read origin and flags them with high certainty

Detects ambiguity in the origin of reads and flags them as uncertain

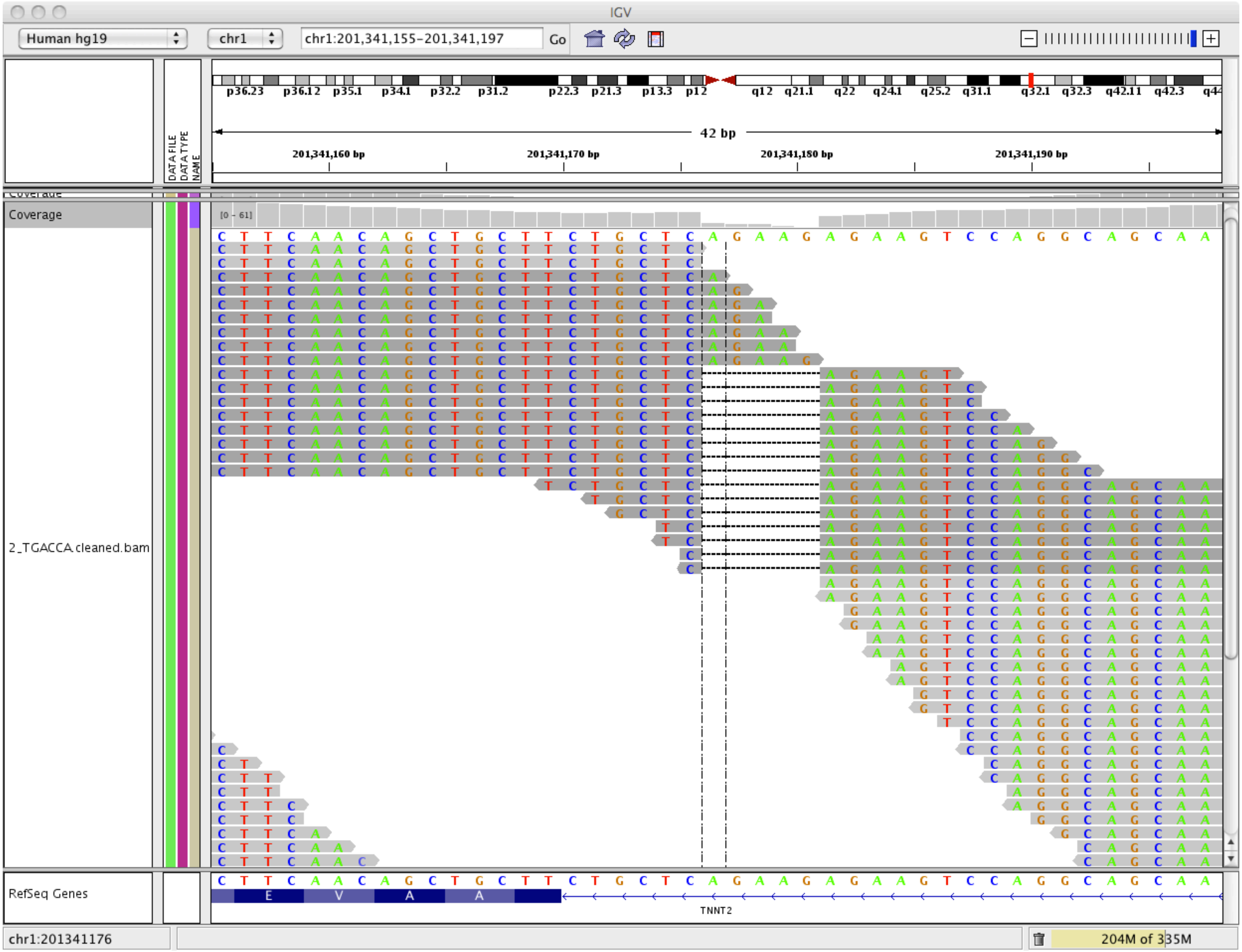


Human hg19

chr1

chr1:201,341,155-201,341,197

IGV
Go
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3 q44
42 bp
201,341,160 bp
201,341,170 bp
201,341,180 bp
201,341,190 bp
DATA FILE
DATA TYPE
NAME
Coverage
[0 - 61]
2_TGACCA.bam
RefSeq Genes
C T T C A A C A G C T G C T T C T G C T C A G A A G A G A A G T C C A G G C A G C A A
E V A A
TNNT2
chr1:201341176
203M of 335M


IGV
Human hg19
chr1
chr1:201,341,155-201,341,197
Go
p36.23 p36.12 p35.1 p34.1 p32.2 p31.2 p22.3 p21.3 p13.3 p12 q12 q21.1 q22 q24.1 q25.2 q31.1 q32.1 q32.3 q42.11 q42.3 q44
42 bp
201,341,160 bp
201,341,170 bp
201,341,180 bp
201,341,190 bp
DATA FILE
DATA TYPE
NAME
Coverage
Coverage
[0 - 61]
C T T C A A C A G C T G C T T C T G C T C A G A A G A G A A G T C C A G G C A G C A A
2_TGACCA.cleaned.bam
RefSeq Genes
C T T C A A C A G C T G C T T C T G C T C A G A A G A G A A G T C C A G G C A G C A A
E V A A
TNNT2
chr1:201341176
204M of 335M

Steps of cancer genome analysis with next-generation sequencing

Getting started

Data quality control

Alignment

Variant calling

Visualization

Artifact removal

Significance analysis

Analysis of public data sets

Variant calling: mutation detection

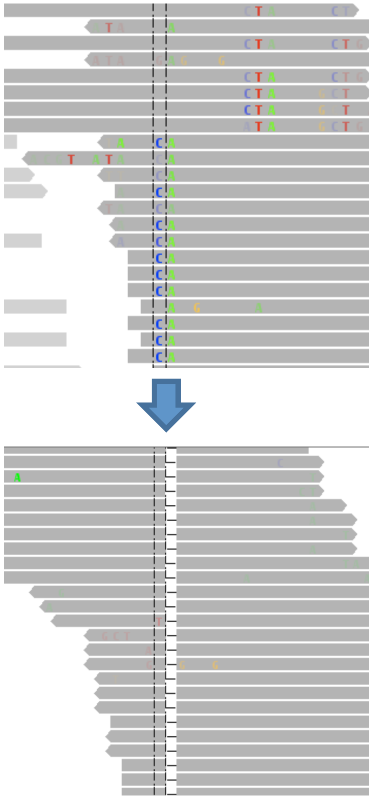
Kristian Cibulskis

Gad Getz

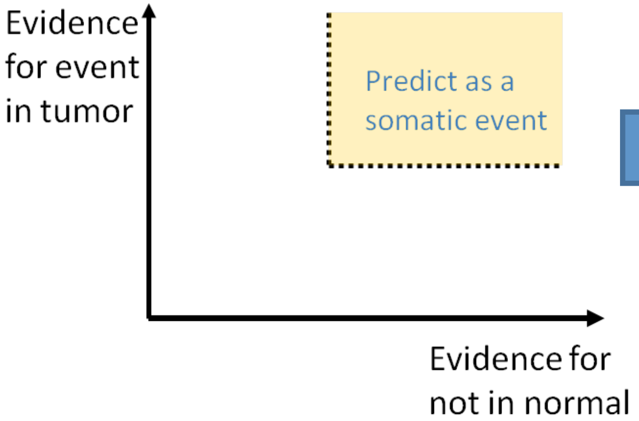
MuTector: Approach

Pre-processing

- Remove duplicate reads
- Calibrate quality scores
- Remove noisy reads
- Local realign



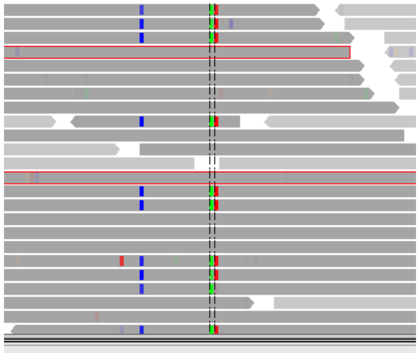
Statistical analysis


Evidence for event in tumor
Predict as a somatic event
Evidence for not in normal

Post-processing

Artifact filtering:

- Misaligned reads
- Events observed only in one direction



Bayesian classifier

Tumor

$$\frac{\text{Prob (Tumor is mutated | Data)}}{\text{Prob (Tumor is reference| Data)}} > \boxed{W}_T$$

Normal

$$\frac{\text{Prob (Normal is reference | Data)}}{\text{Prob (Normal is non-reference | Data)}} > \boxed{W}_N$$

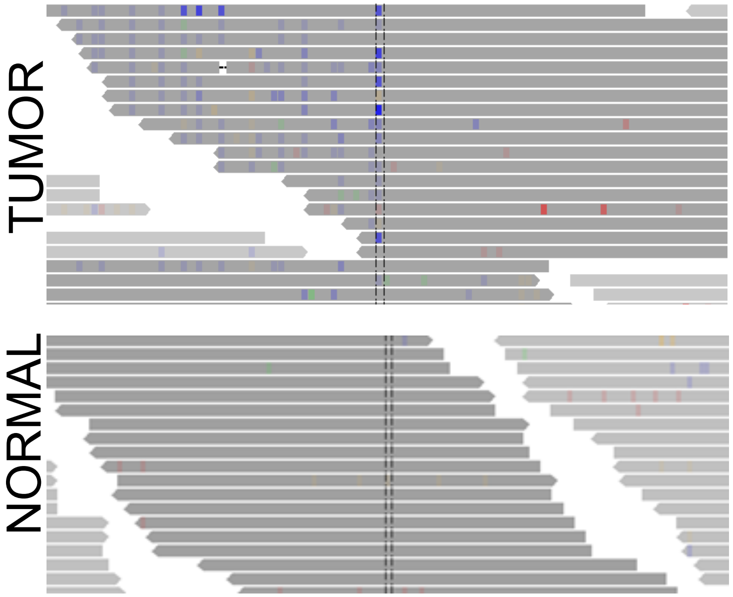
MuTector: Control low rate of two types of false positives

Signal: ~1 somatic mutation per Mb.

Need error rate << signal rate (<< 10^{-6} errors/base)!

Noise: Two types of false positives

NO EVENT



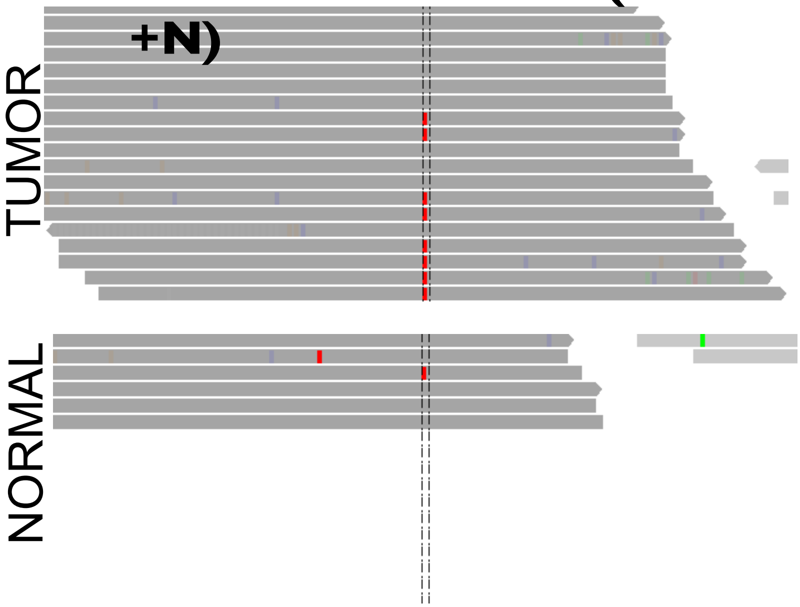
At risk: Every base

Source: Misread bases

Sequencing artifacts

Misaligned reads

GERMLINE EVENT (in T +N)



At risk: ~1000 germline variants / Mb (dbSNP)

~50 rare germline variants / Mb (not in dbSNP)

Source: Low coverage in normal (sampling noise)

Misaligned or unaligned reads (indels)

Variant detection: non-human sequences

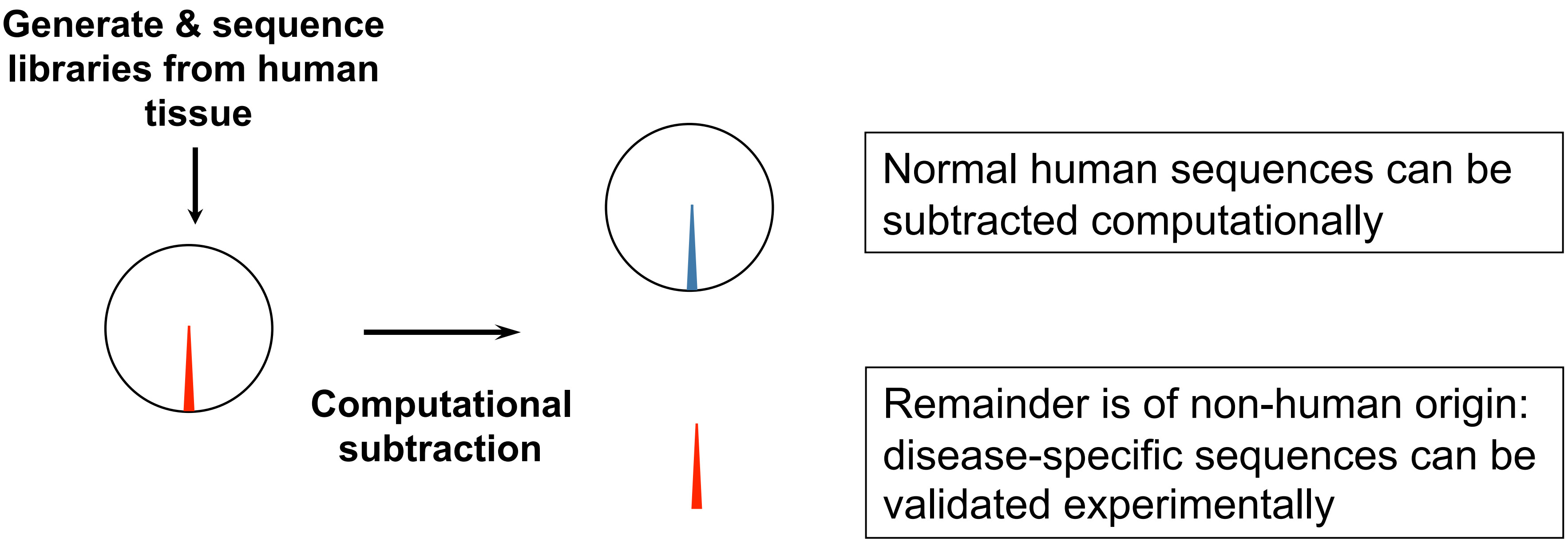
Alex Kostic
Chandra Pedamallu
Akin Ojesina
Joonil Jung

Sequence-based computational subtraction for pathogen discovery

Principle

The human genome sequence is nearly complete

Infected tissues contain human and microbial RNA and DNA



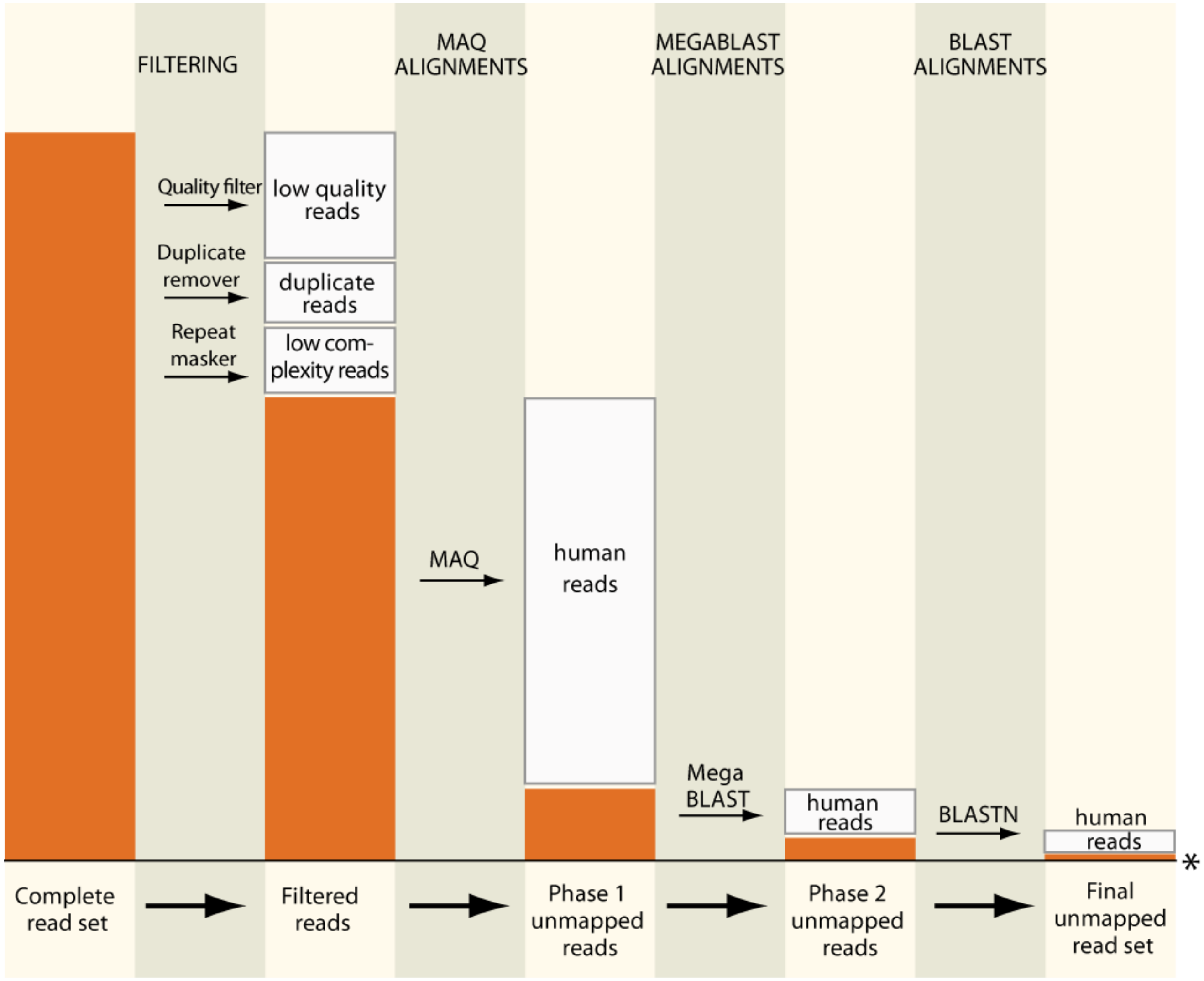
Generate & sequence libraries from human tissue

Computational subtraction

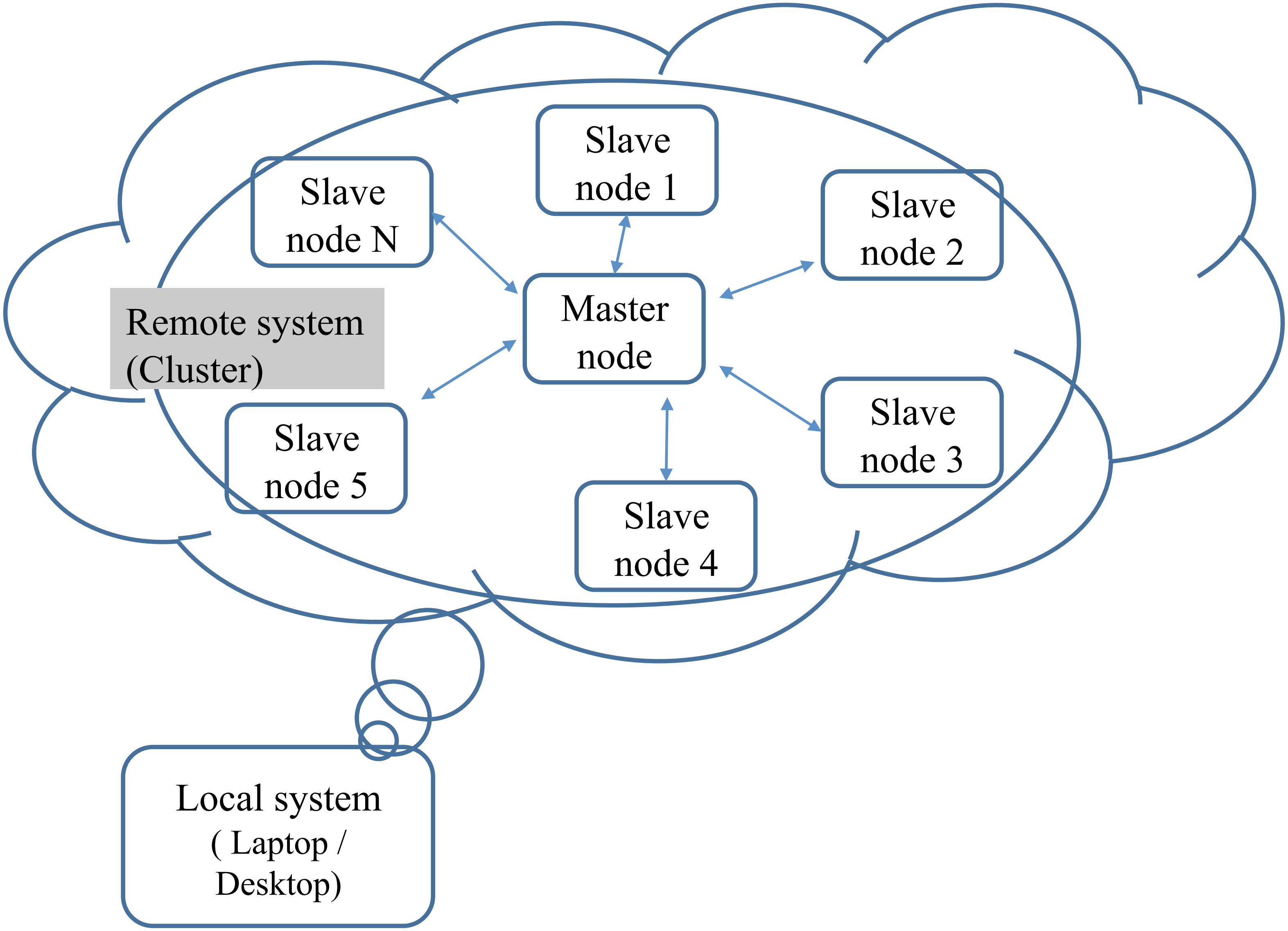
Normal human sequences can be subtracted computationally

Remainder is of non-human origin: disease-specific sequences can be validated experimentally

PathSeq: Computational Subtraction Workflow

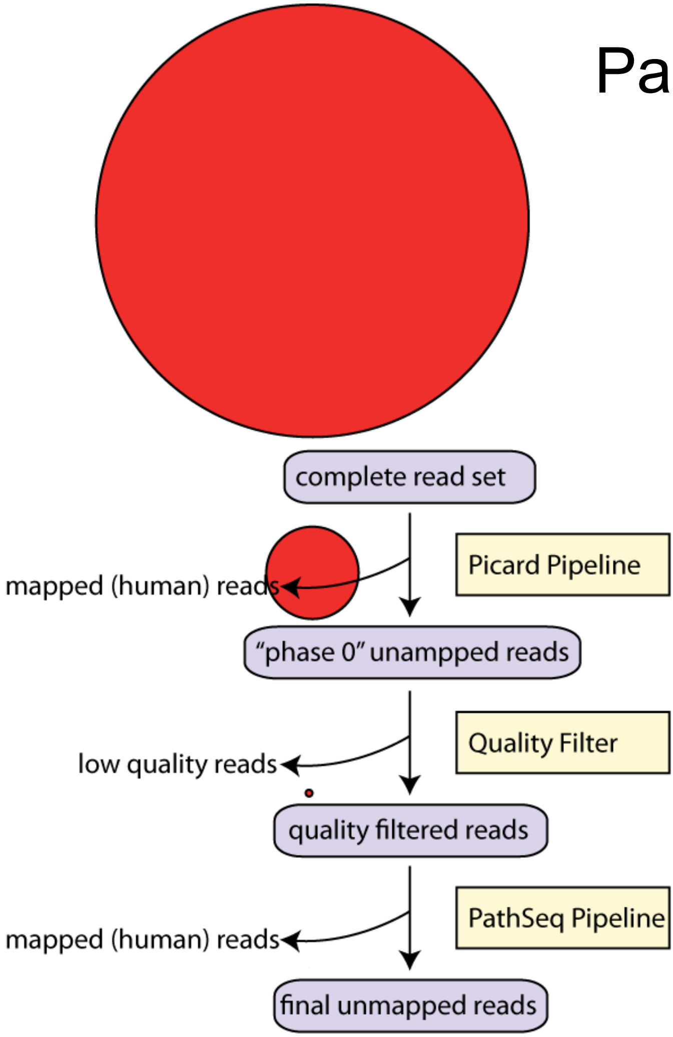


PathSeq implemented on cloud computing



PathSeq: Subtraction efficiency > 1 / 15 million

PathSeq analysis of
ovarian cancer
genome data



complete read set

Picard Pipeline

mapped (human) reads

"phase 0" unampped reads

Quality Filter

low quality reads

quality filtered reads

PathSeq Pipeline

mapped (human) reads

final unmapped reads

	Number of Reads (101bp)
complete read set	1,340,765,698
"phase 0" unampped reads	260,459,260
quality filtered reads	45,057,752
final unmapped reads	778

Variant detection: absolute allele-level copy number calling

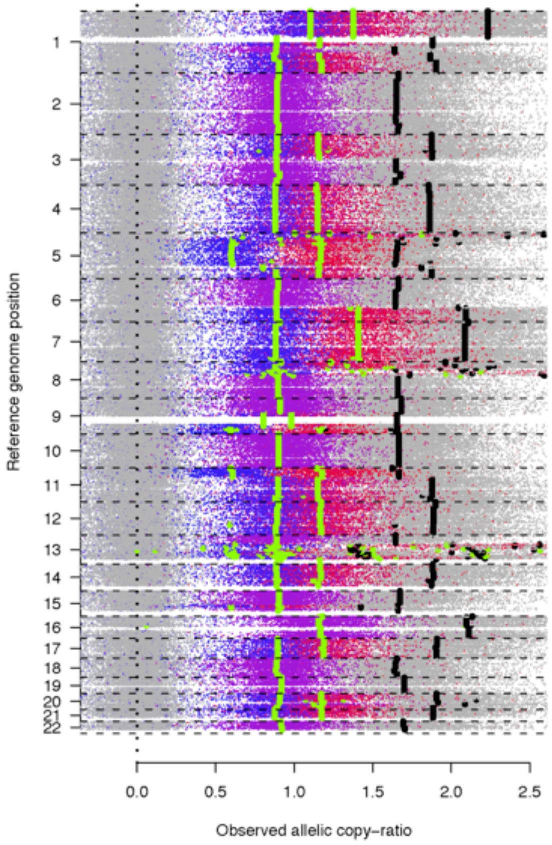
Scott Carter

Gad Getz

Allelic copy-ratio histograms are the basis for purity / ploidy determination

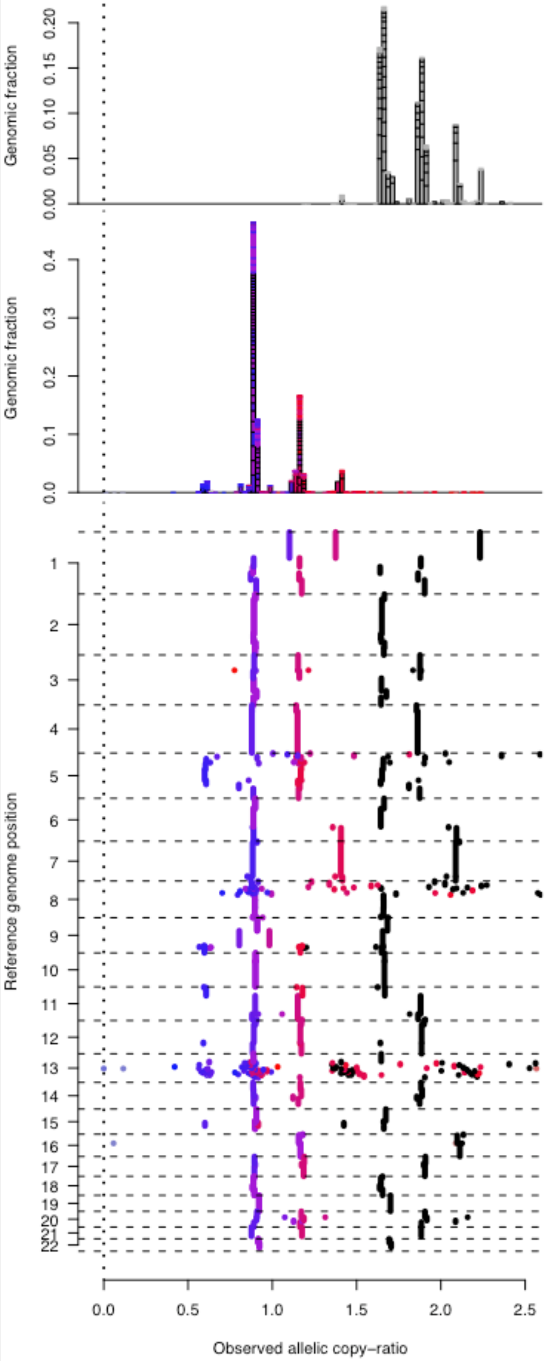
- Haplotype-specific copy histograms must be inferred from allele-specific SNP measurements

Fit with SNP-array error-model

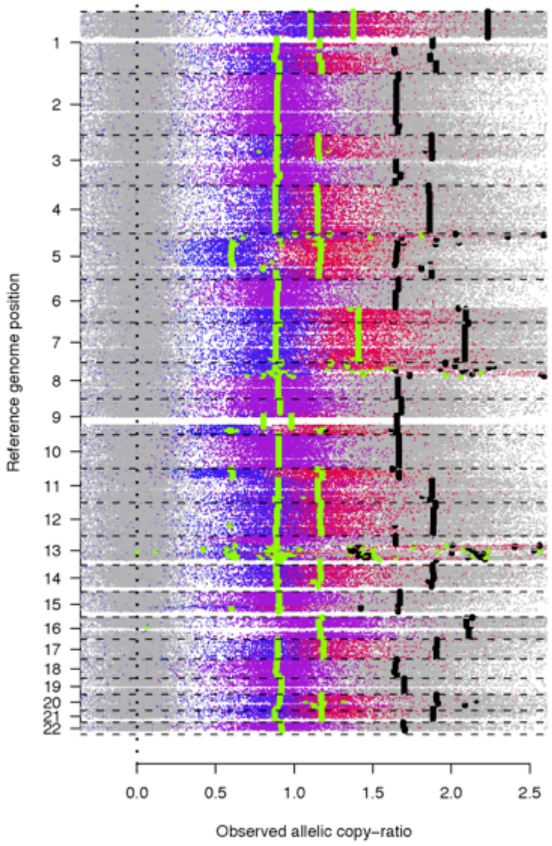


Colored SNPs are germline-heterozygous

Collapse to segment-means for each haplotype



Allelic copy-ratio histograms are the basis for purity / ploidy determination



Total copy

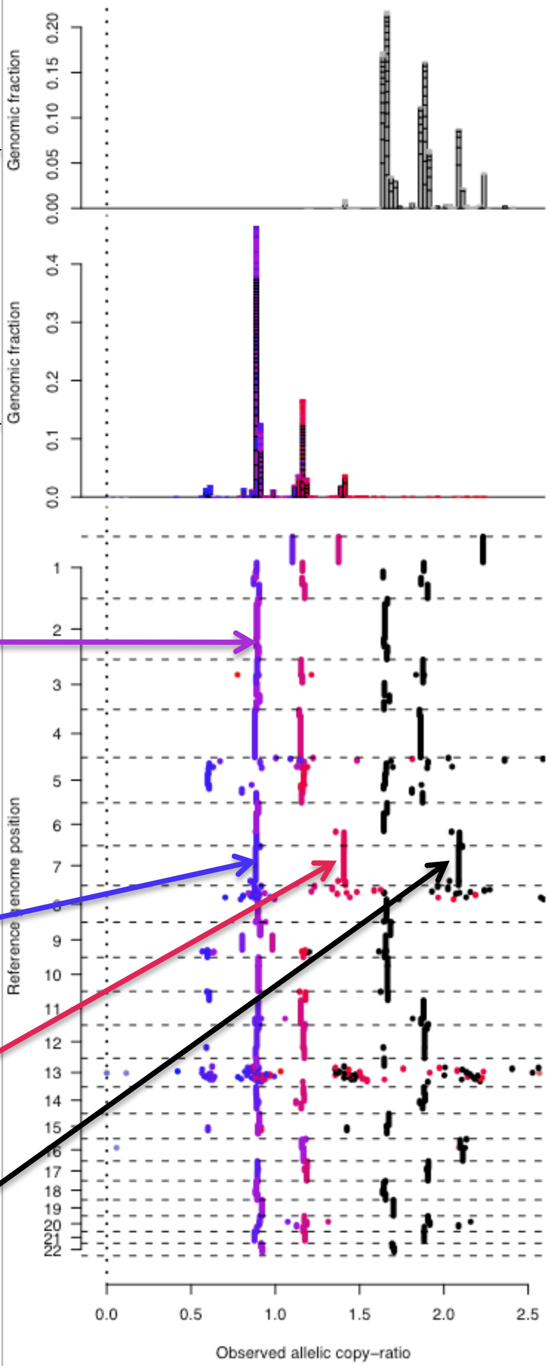
Haplotype-specific copy

Region at allelic-balance (unphased)

Lower-copy haplotype

Higher-copy haplotype

Total copy



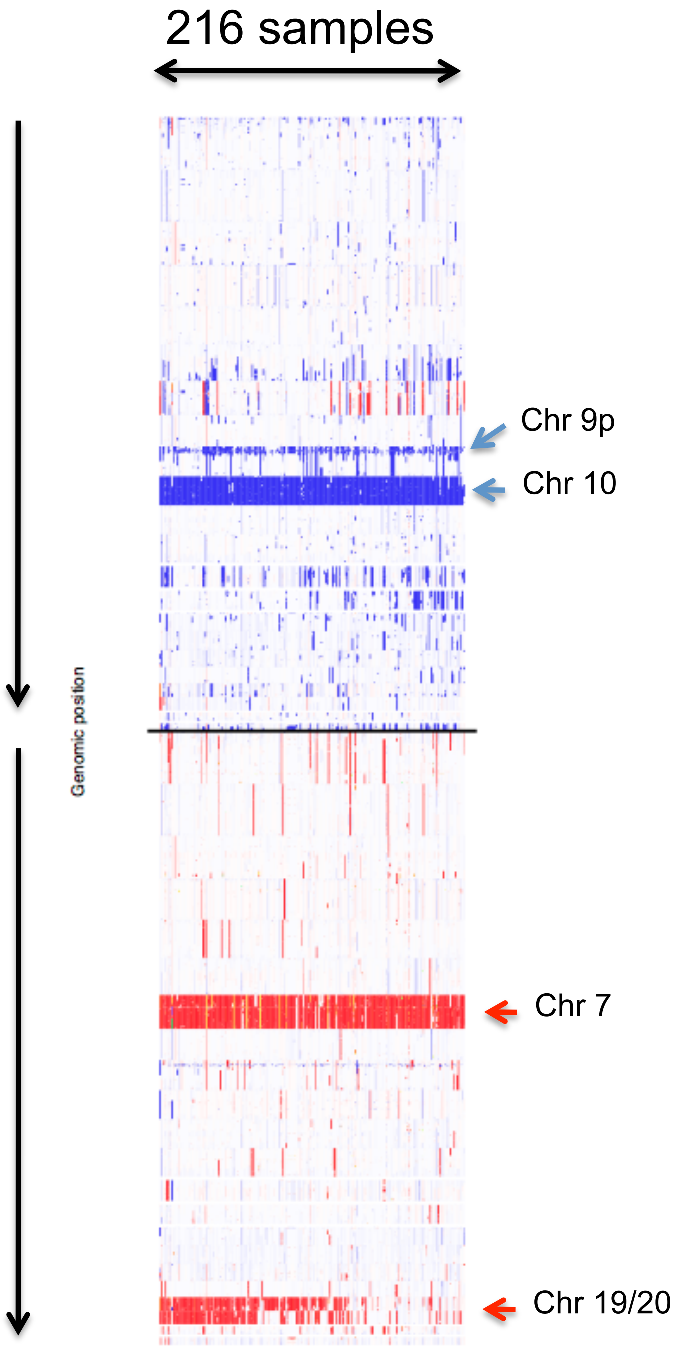
Visualizing absolute allelic copy-number data:

Glioblastoma multiforme (GBM)

- 🟦 0 copies
- □ 1 copy (“neutral”)
- 🟥 2 copies
- 🟩 3 copies
- 🟨 4 copies

Genome order –
Low-copy haplotypes

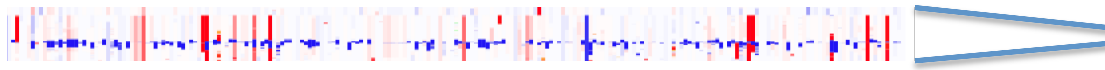
Genome order –
High-copy haplotypes

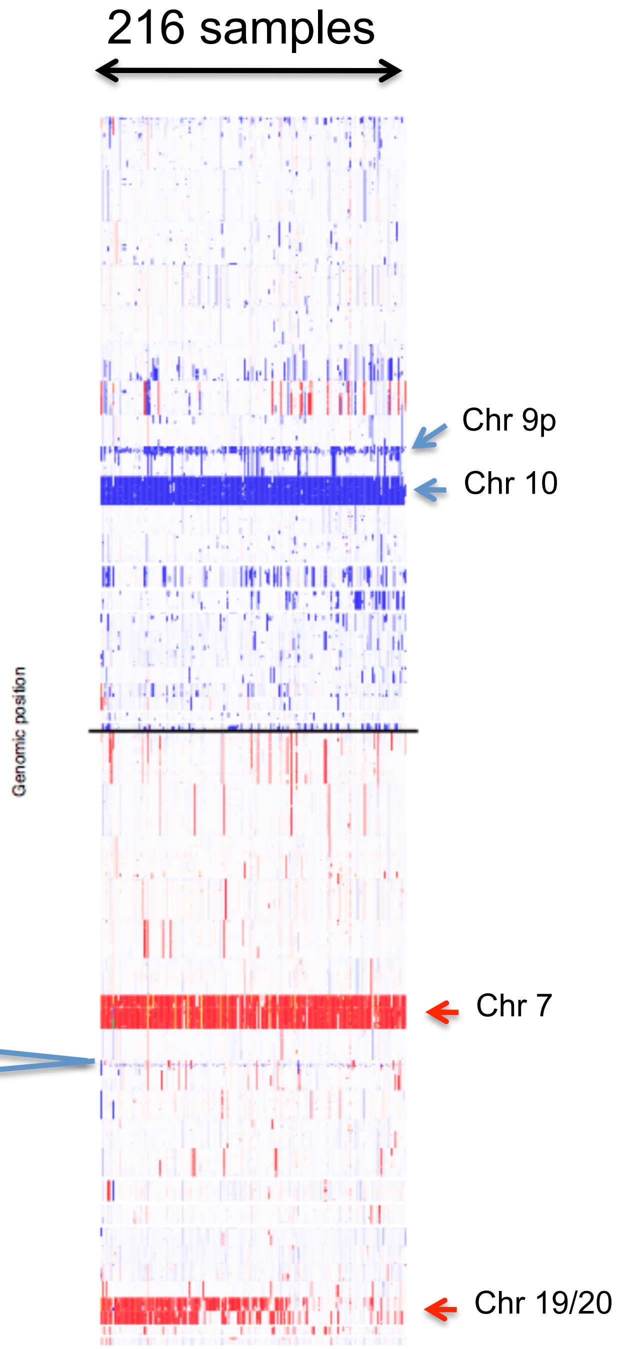


Visualizing absolute allelic copy-number data:

Glioblastoma multiforme (GBM)

Frequent homozygous deletion of *CDKN2A/B* on chr 9p





Steps of cancer genome analysis with next-generation sequencing

Getting started

Data quality control

Alignment

Variant calling

Visualization

Artifact removal

Significance analysis

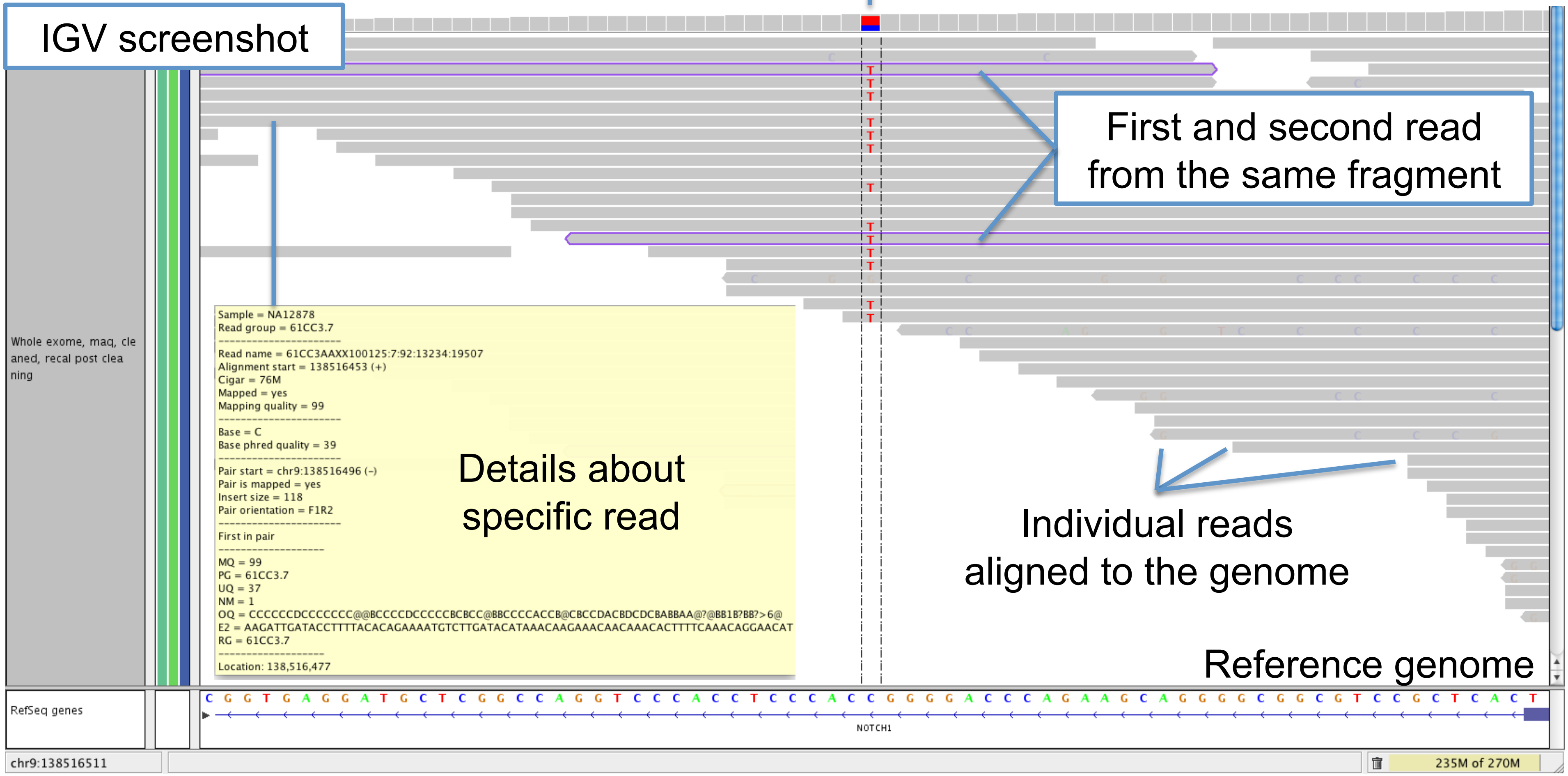
Analysis of public data sets

Visualizing next-generation sequencing data: the Integrated Genome Viewer (IGV)

Non-reference bases are colored; reference bases are grey

Clean C/T heterozygote

Depth of coverage



IGV screenshot

First and second read from the same fragment

Details about specific read

Individual reads aligned to the genome

Reference genome

```
Sample = NA12878
Read group = 61CC3.7
----------------------
Read name = 61CC3AAXX100125:7:92:13234:19507
Alignment start = 138516453 (+)
Cigar = 76M
Mapped = yes
Mapping quality = 99
----------------------
Base = C
Base phred quality = 39
----------------------
Pair start = chr9:138516496 (-)
Pair is mapped = yes
Insert size = 118
Pair orientation = F1R2
----------------------
First in pair
-------------------
MQ = 99
PG = 61CC3.7
UQ = 37
NM = 1
OQ = CCCCCCDCCCCCCC@@BCCCCDCCCCCBCBCC@BBCCCCACCB@CBCCDACBDCDCBABBAA@?@BB1B?BB?>6@
E2 = AAGATTGATACCTTTTACACAGAAAATGTCTTGATACATAAACAAGAAACAACAAACACTTTTCAAACAGGAACAT
RG = 61CC3.7
-------------------
Location: 138,516,477
```

Whole exome, maq, cleaned, recal post cleaning
T
RefSeq genes
C G G T G A G G A T G C T C G G C C A G G T C C C A C C T C C C A C C G G G G A C C C A G A A G C A G G G G C G G C G T C C G C T C A C T
NOTCH1
chr9:138516511
235M of 270M

Steps of cancer genome analysis with next-generation sequencing

Getting started

Data quality control

Alignment

Variant calling

Visualization

Artifact removal

Significance analysis

Analysis of public data sets

Artifact removal: “If it’s interesting, it’s probably an artifact!”

Alignment problems

- Genes with close homologs and pseudogenes
- Alignment of insertions and deletions

Whole genome amplification

Stochastic errors

Read quality problems

Read duplication from excess PCR

How to find them: look for an interesting result and then try to understand why it happened

Steps of cancer genome analysis with next-generation sequencing

Getting started

Data quality control

Alignment

Variant calling

Visualization

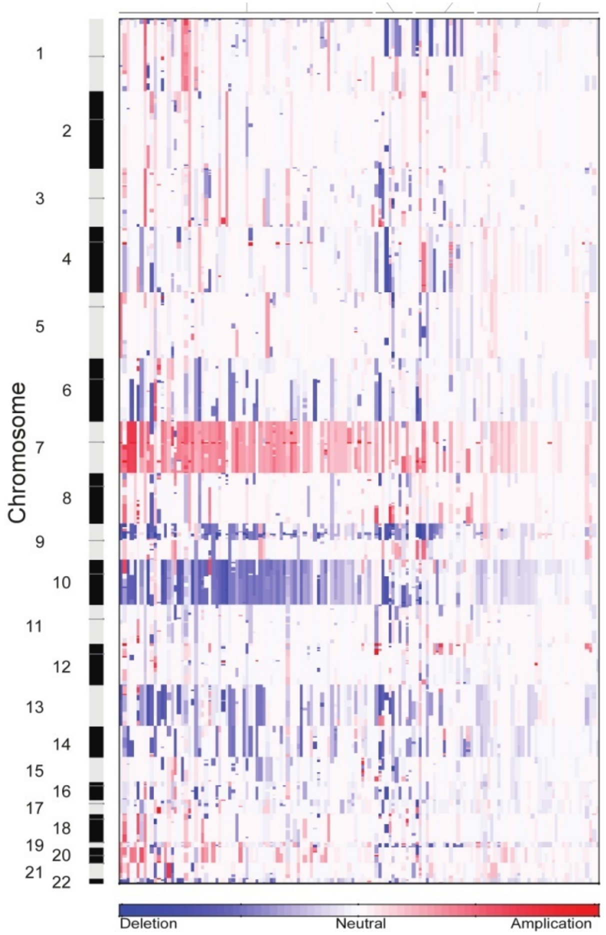
Artifact removal

Significance analysis

Analysis of public data sets

The Fundamental Challenge of Cancer Genome Analysis: Distinguishing “driver” from “passenger” alterations

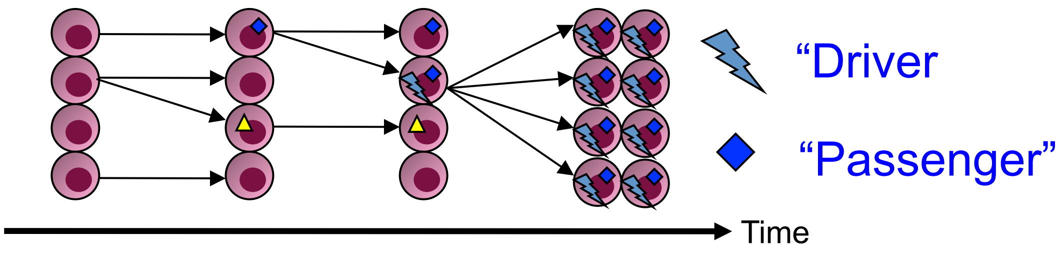
141 glioblastoma samples



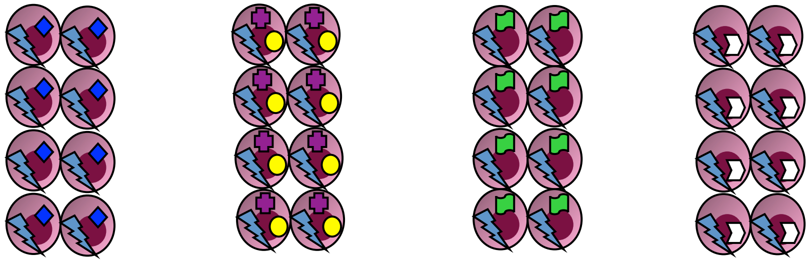
Nearly every region is altered in at least one tumor

Beroukhim, Getz et al, *PNAS*, 104(50) 2007-12, 2007.

Only some of aberrations present in a tumor clone are related to cancer growth (“drivers”)



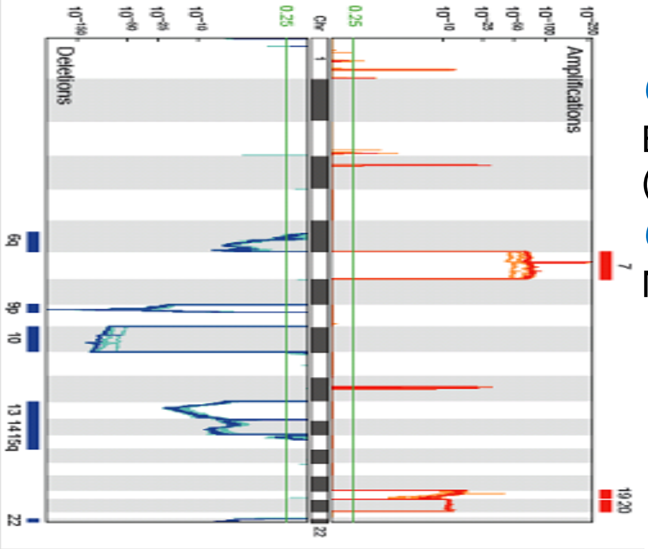
Can be distinguished by studying many samples and identifying aberrations that occur more frequently than expected by chance



For SCNAs, an additional challenge is identifying which of the many affected genes are actually being targeted

Tools for detecting cancer genes / regions / pathways

COPY NUMBER



GISTIC 1.0

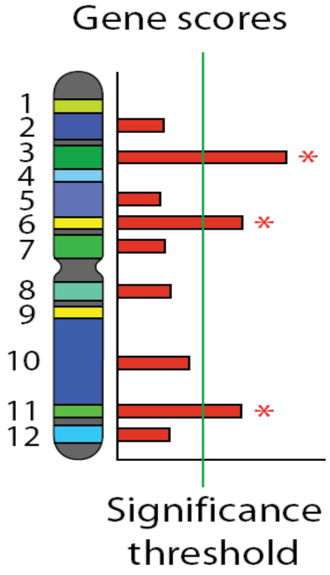
Beroukhim et al. PNAS (2007)

GISTIC 2.0

Mermel et al. *submitted*

Uses: Frequency and amplitude of events
Separates broad and focal gains and losses

MUTATIONS



MutSig

Getz et al. Science (2007)
Lawrence et al. in development

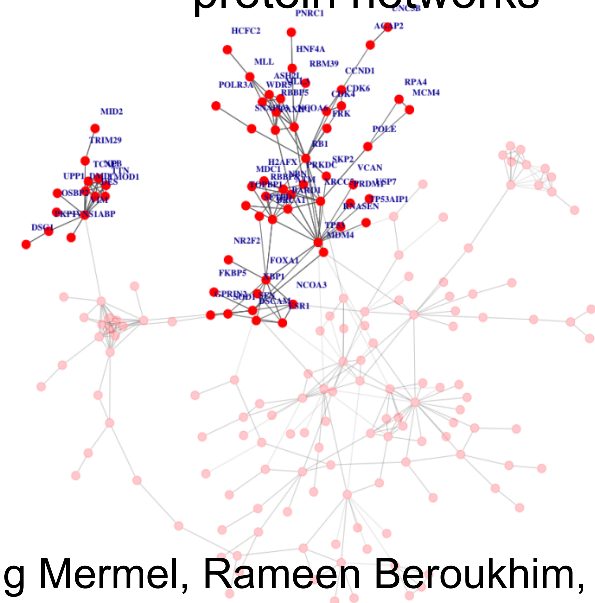
Uses: Number and types of mutations: CpG, C or G, A or T, indel, null
Works on genes, genesets and conserved regions (intervals on the genome)

ALL MODALITIES

NetSig (in development)

Zou et al., in development

Uses: all types of alterations to identify clusters of mutated genes in protein-protein networks



Craig Mermel, Rameen Beroukhim, Steve Schumacher, Mike Lawrence, Lihua Zou, Alex Ramos, Gregory Kryukov, Petar Stojanov

Steps of cancer genome analysis with next-generation sequencing

Getting started

Data quality control

Alignment

Variant calling

Visualization

Artifact removal

Significance analysis

Analysis of public data sets: see The Cancer Genome Atlas and the International Cancer Genome Consortium

Summary: next-generation analysis of cancer is powerful and do-able

Acknowledgements

Analysis Team
Mike Lawrence
Kristian Cibulskis
Andrey Sivachenko
Craig Mermel
Scott Carter
Yotam Drier
Gordon Saksena
Doug Voet
Wendy Winckler
Alex Ramos
Trevor Pugh
Mike Berger
Mike Chapman
Aaron McKenna
Petar Stojanov
Gregory Kryukov
Alex Kostic
Peter Carr
Mike Noble
Nicolas Stransky
Joonil Jung
Derek Chiang
Roel Verhaak

THE CANCER GENOME ATLAS

NHGRI NCI SIGMA

Stacey Gabriel

Levi Garraway

Lynda Chin

Gad Getz

Todd Golub

Eric Lander

Broad Institute of Harvard and MIT

Dana Farber Cancer Institute

Project Management
Carrie Sougnez
Erica Shefler
Daniel Auclair
Marisa Cortes
Kristin Thompson

Jill Mesirov
Jim Robinson
Helga Thorvaldsdottir
Marc-Danie Nazaire

Broad Institute Sequencing Program and Platform
Robb Onofrio
Brendan Blumenstiel
Huy Nguyen
Mellisa Parkin
Wendy Winckler

Tim Fennell
Lauren Ambrogio
Sheila Fisher
Joshua Levin
Xian Adiconis
Andreas Gnirke
Toby Bloom
Chad Nusbaum

Broad Institute Biological Samples Platform

Kristin Ardlie

David Jaffe

Mark DePristo
Eric Banks
Kiran Gam