Cold Spring Harbor Laboratories
Advanced Sequencing Technologies and Applications

Introduction to NGS data analysis

Sorana Morrissy, PhD

Charbonneau Cancer Institute

Alberta Children's Hospital Research Institute

Dept. of Biochemistry and Molecular Biology

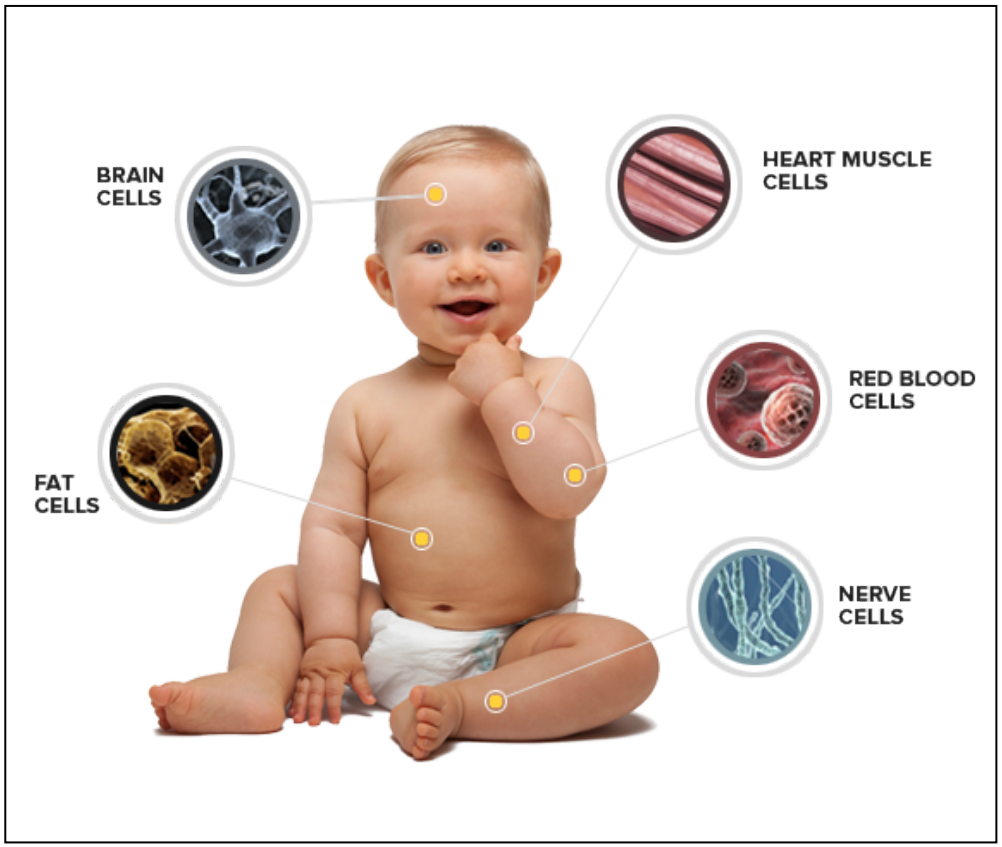
Cumming School of Medicine

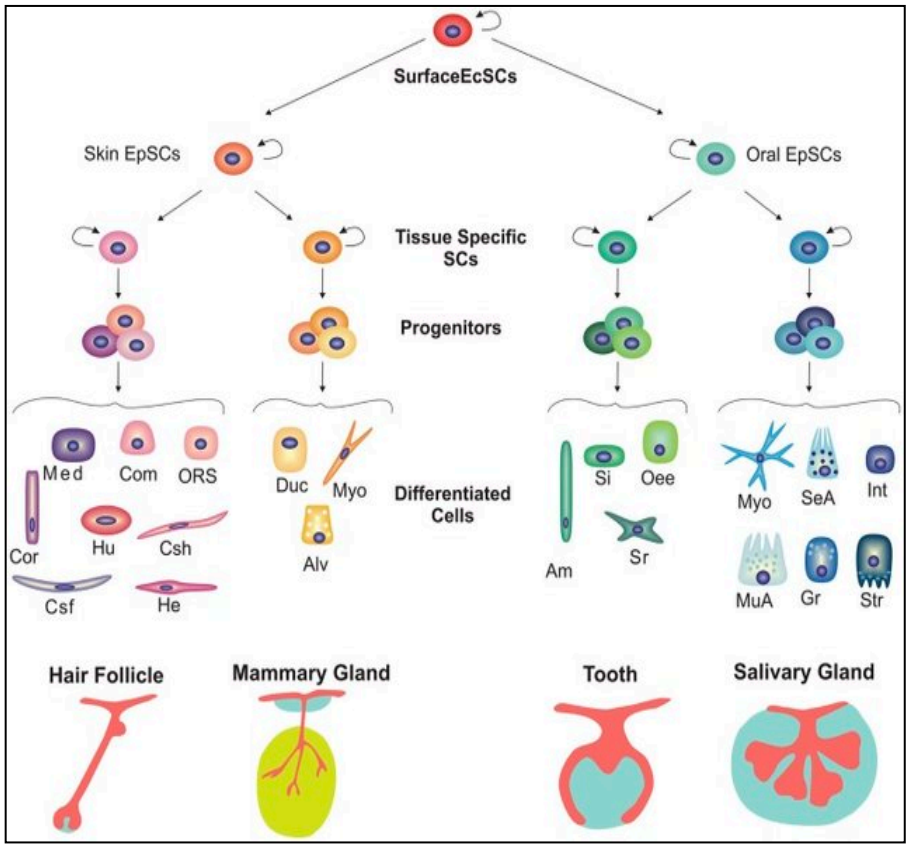
University of Calgary, Canada

Outline

- NGS data and resources
  - Understanding the biology of normal and disease states
- NGS data analysis
  - Single Nucleotide Variants
  - Structural Variants
  - scRNAseq analysis

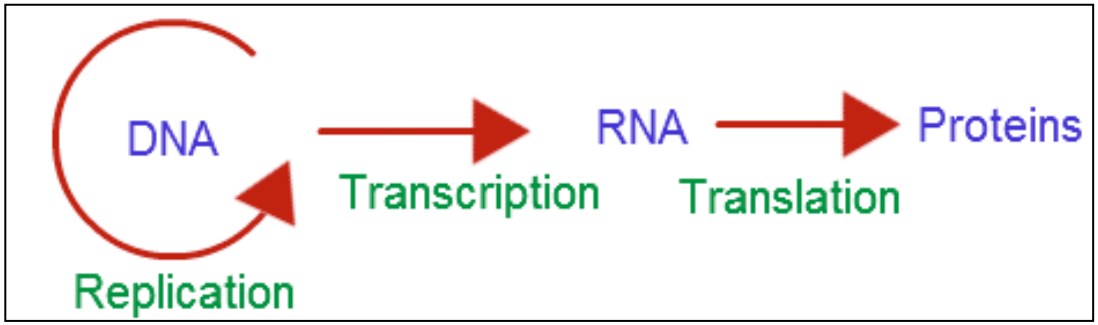
We are all made of cells



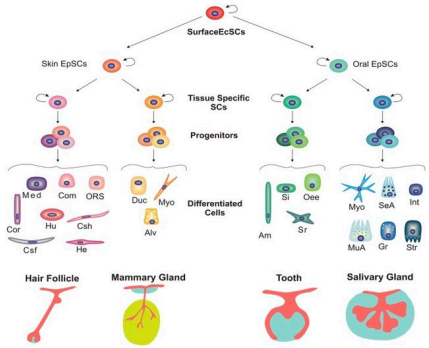


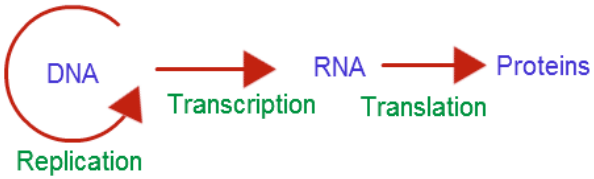
https://genographic.nationalgeographic.com/science-behind/genetics-overview/

DOI: 10.3389/fphys.2012.00107



Data Collections



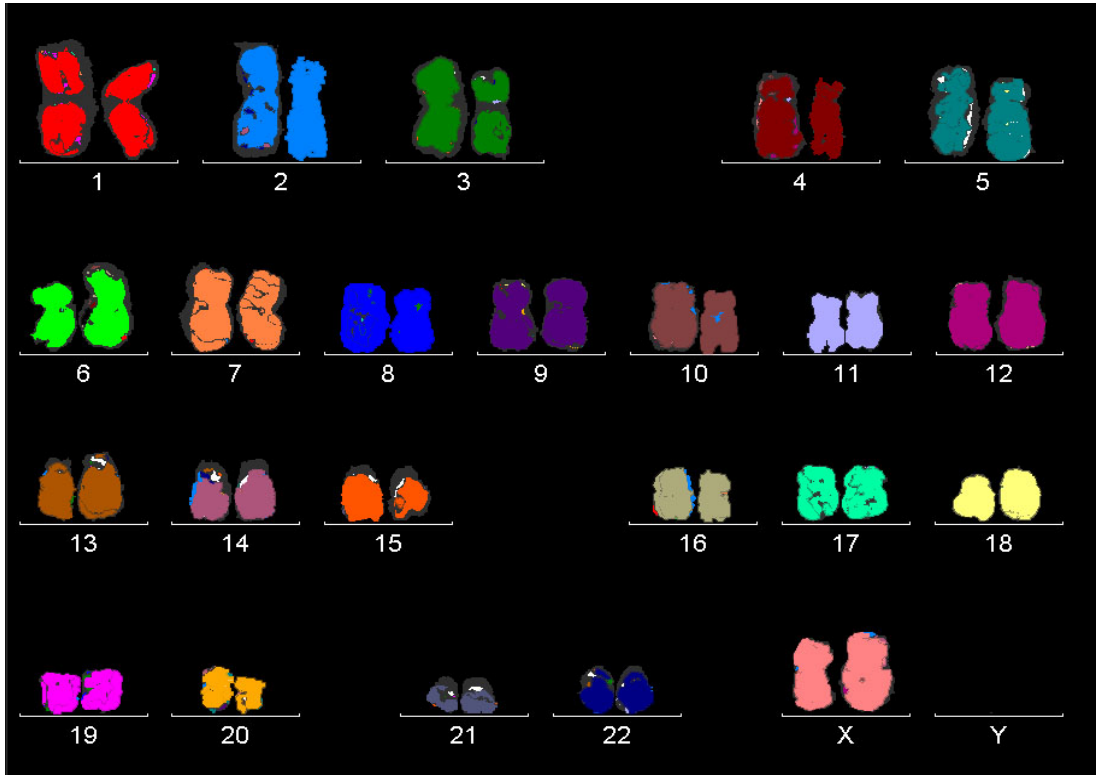


Normal States

Disease states

Germline variation Expression regulation	Germline alterations	Somatic alterations
Variation: dbSNP, DGV, UKBioBank, gnomAD, MVP **Tissues**: GTEx **Cell types / states**: Human Cell Atlas, Tabula Muris	**OMIM**: Catalog of Human Genes and Genetic Disorders	**TCGA**: The Cancer Genome Atlas **ICGC**: International Cancer Genome Consortium
WGS: short and long reads DNA, epigenetic modifications	**RNAseq**: bulk, single cell	**Proteomics**: Ab, global

A normal human karyotype



Genome Reference Consortium

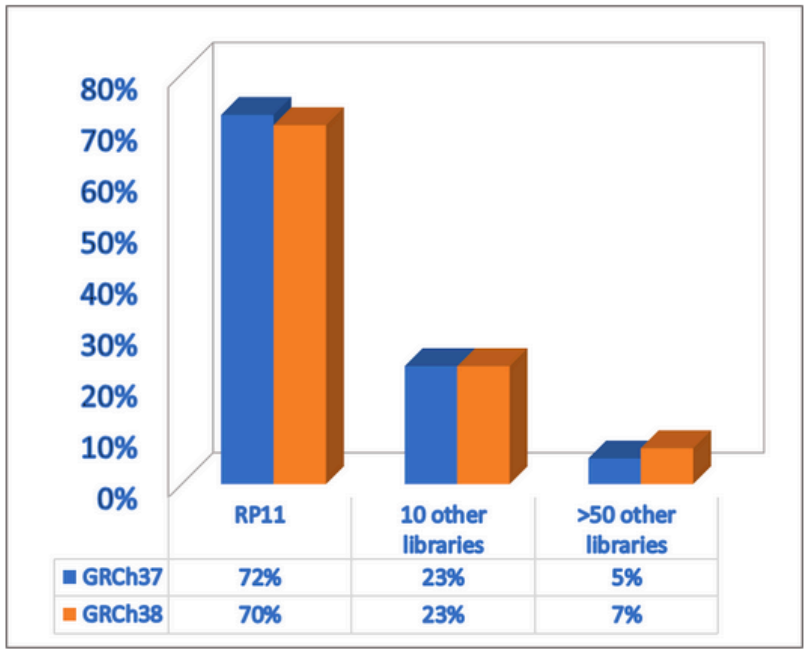
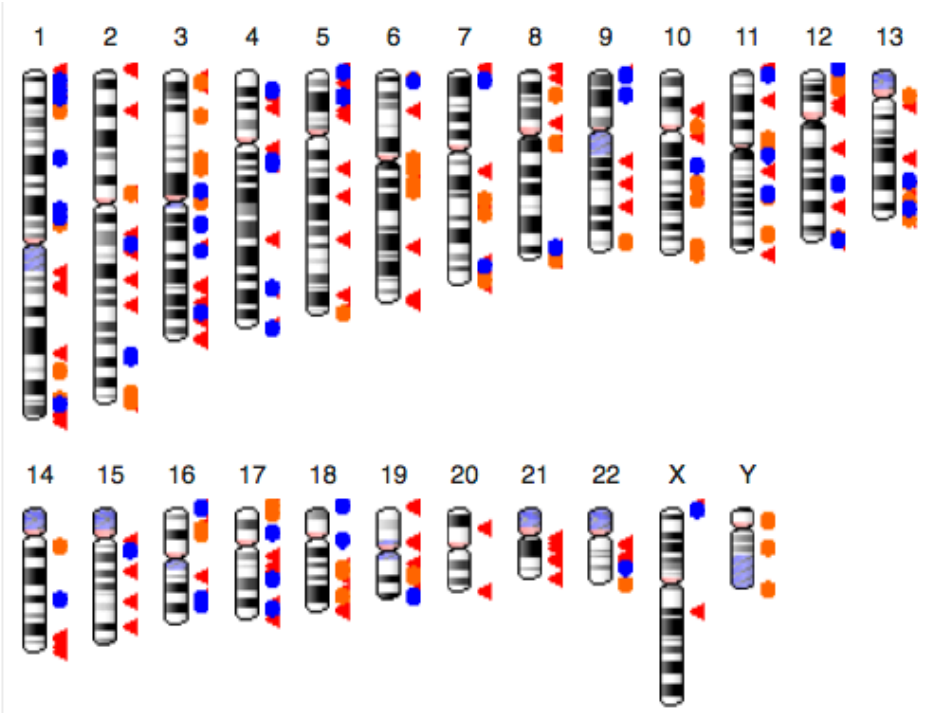


Figure 1. Contribution of genomic libraries to GRCh37 and GRCh38.

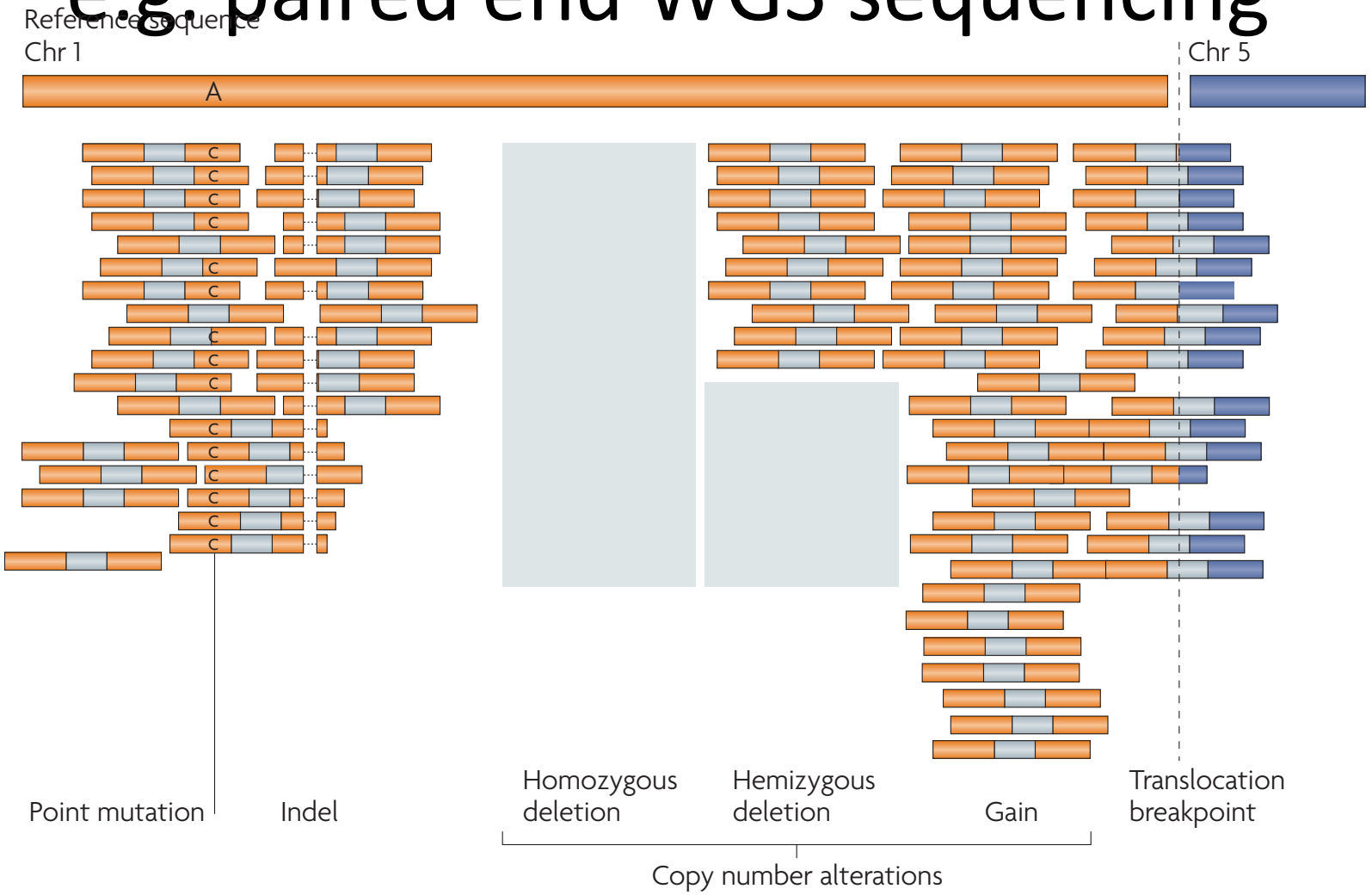


Region containing alternate loci
Region containing fix patches
Region containing novel patches

Ideogram of the latest human assembly, GRCh38.p12

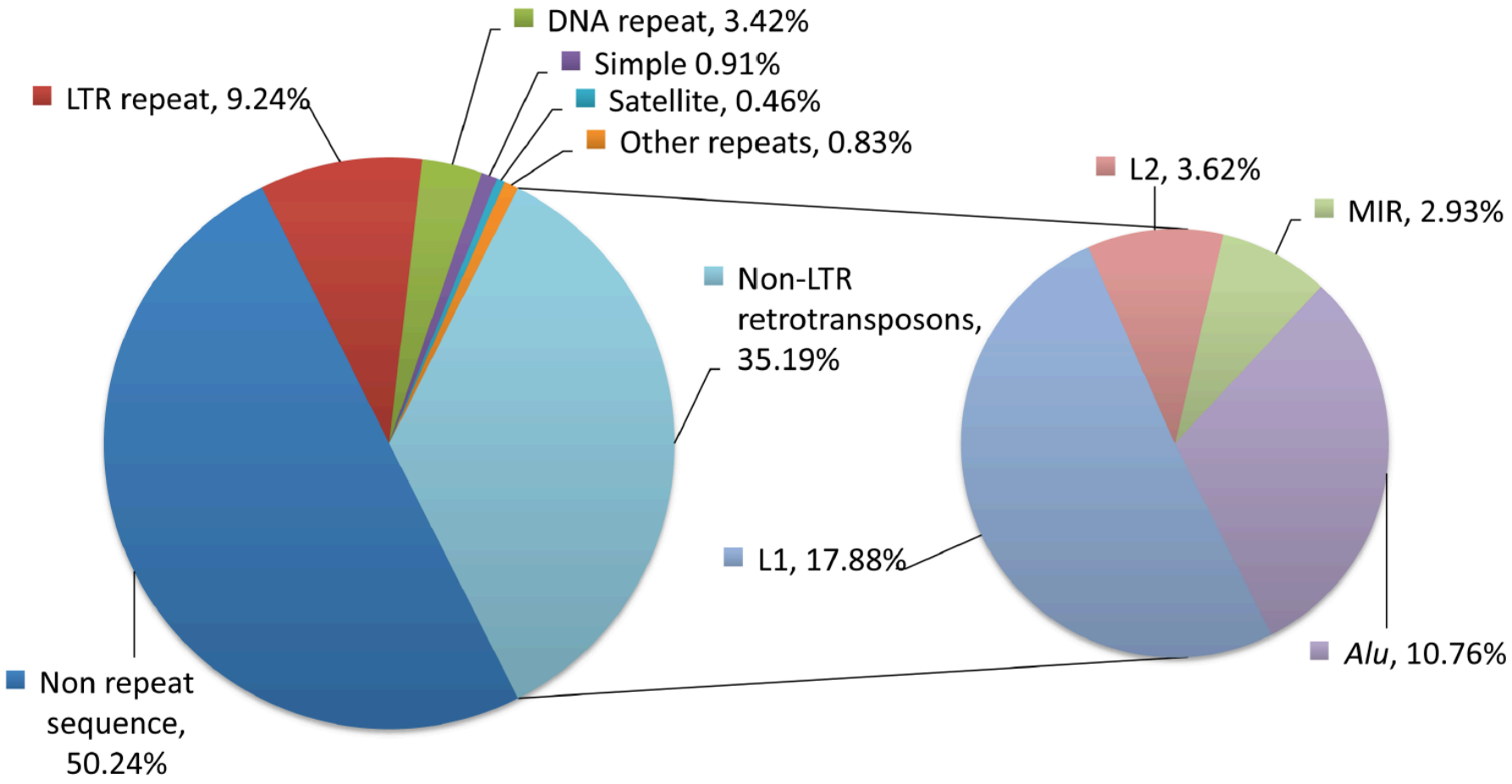
First draft in 2001
GRCh38.p12 (2017-12-21)
RP11 - African-European male
Haploid with alt loci
3,095,978,931 bases
Total regions: 317
Regions with alt loci: 178
Alternate loci: 261

Next Generation Sequencing
e.g. paired end WGS sequencing



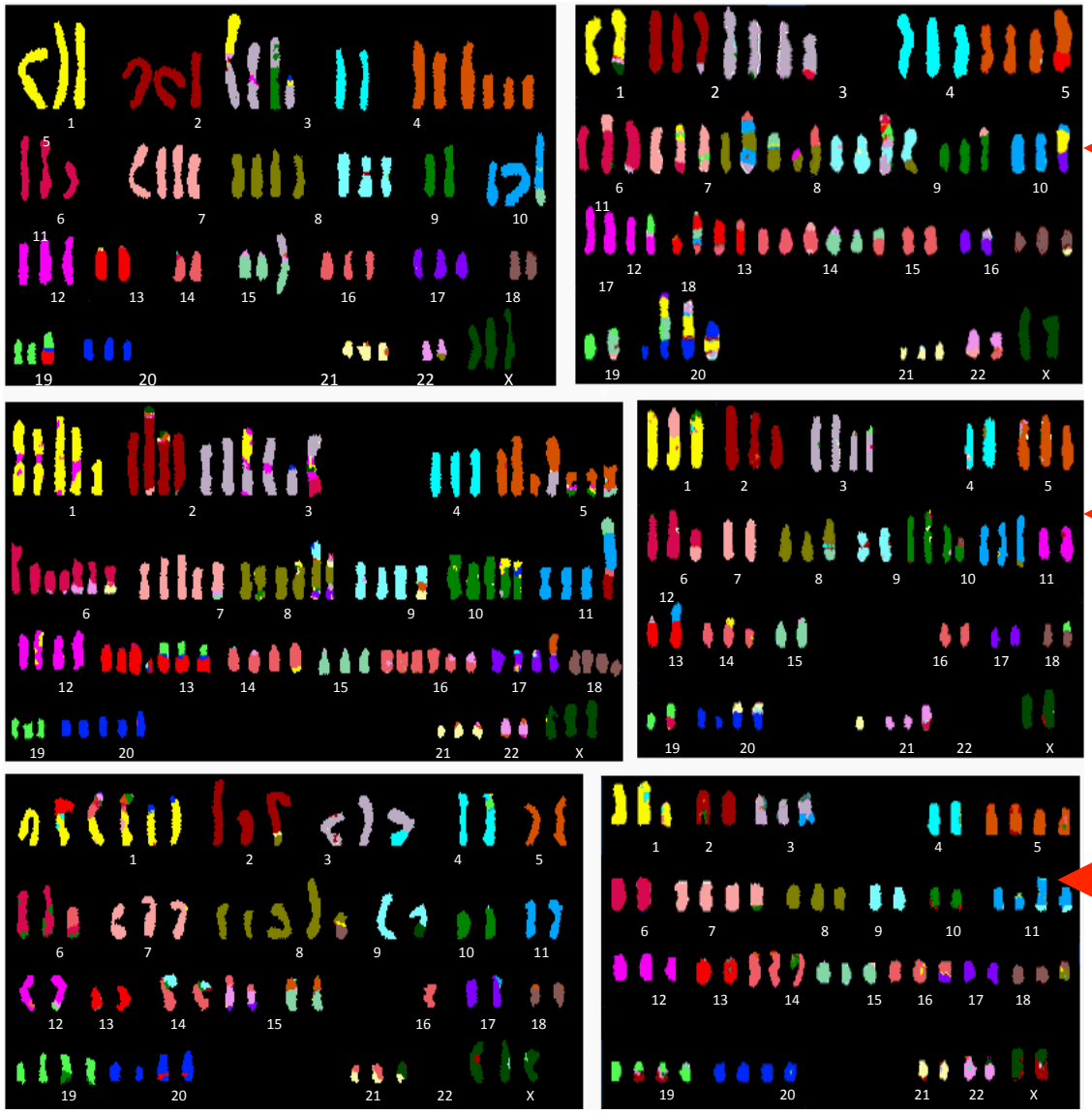
Meyerson and Getz, 2010, Nat Rev Genet.

Composition of the human genome



http://dx.doi.org/10.1371/journal.pgen.1003402

Cancers exhibit disrupted karyotypes



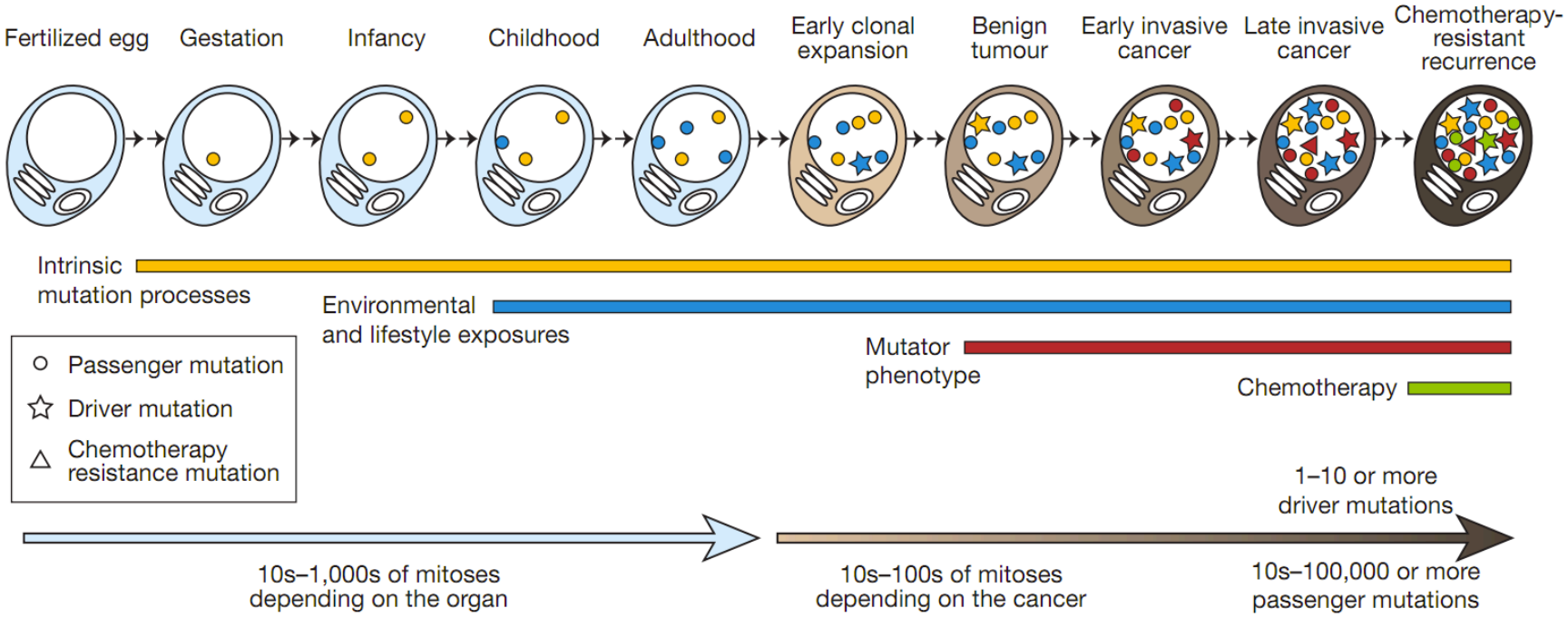
translocations

High ploidy

Broad events vs focal events

David Huntsman

Cancer cells accumulate somatic alterations over time



Mutation frequency depends on cancer type (childhood vs adult)
External forces (e.g. drug treatment) can select for specific clones (e.g. cells with resistance mutations)

Stratton MR, Campbell PJ, Futreal PA. *Nature*. 2009 Apr 9;458(7239):719-24. Review.

Next-generation sequencing of cancer

33 tumor sites | 31,000 patients | 10 rare cancers

*Genomic DNA or RNA **

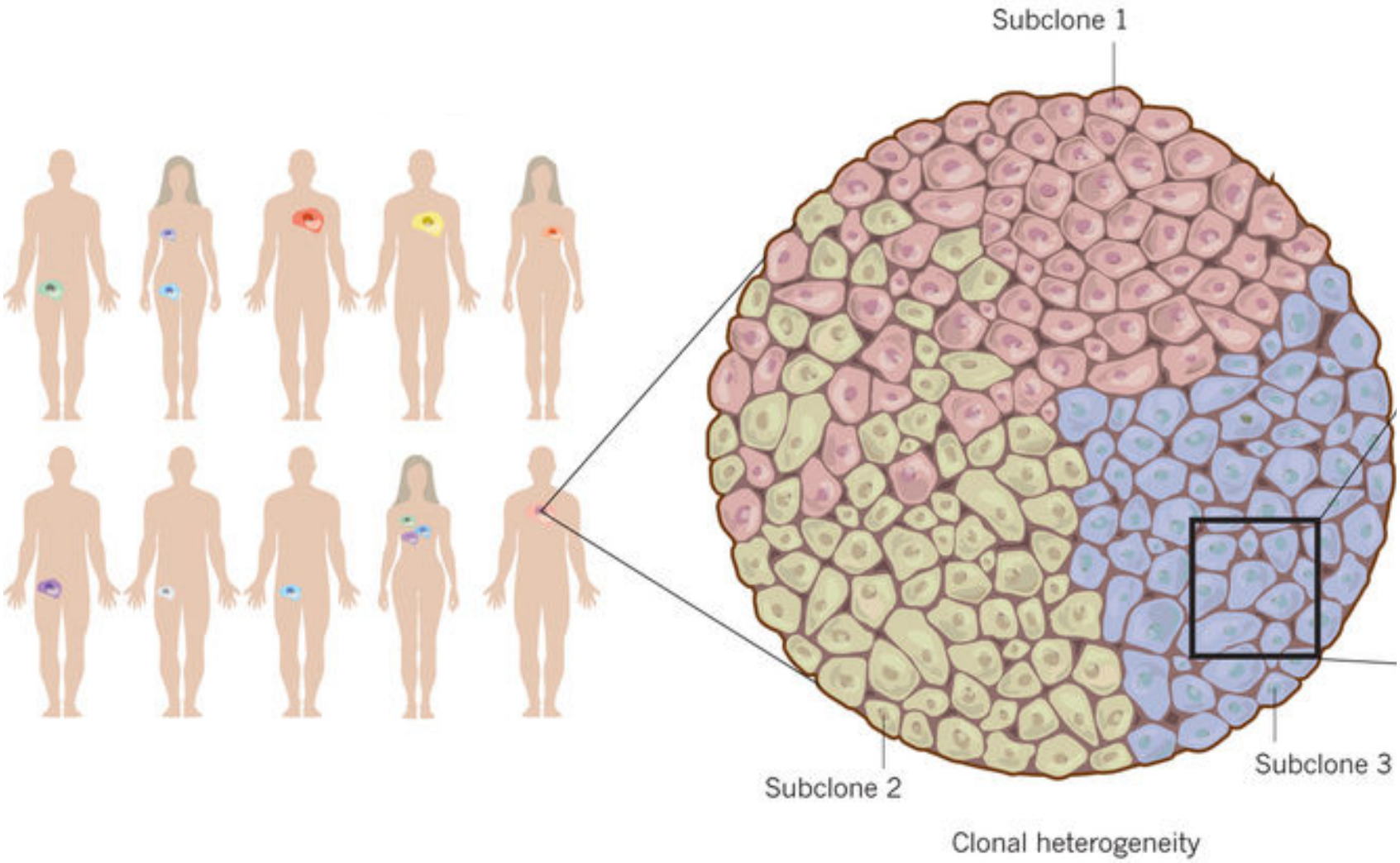
Library of DNA fragments

Sequencing device

Computational analysis

** Proteomics*

Inter- and Intra-Tumoral Heterogeneity



Burrell et al., Nature, 2013

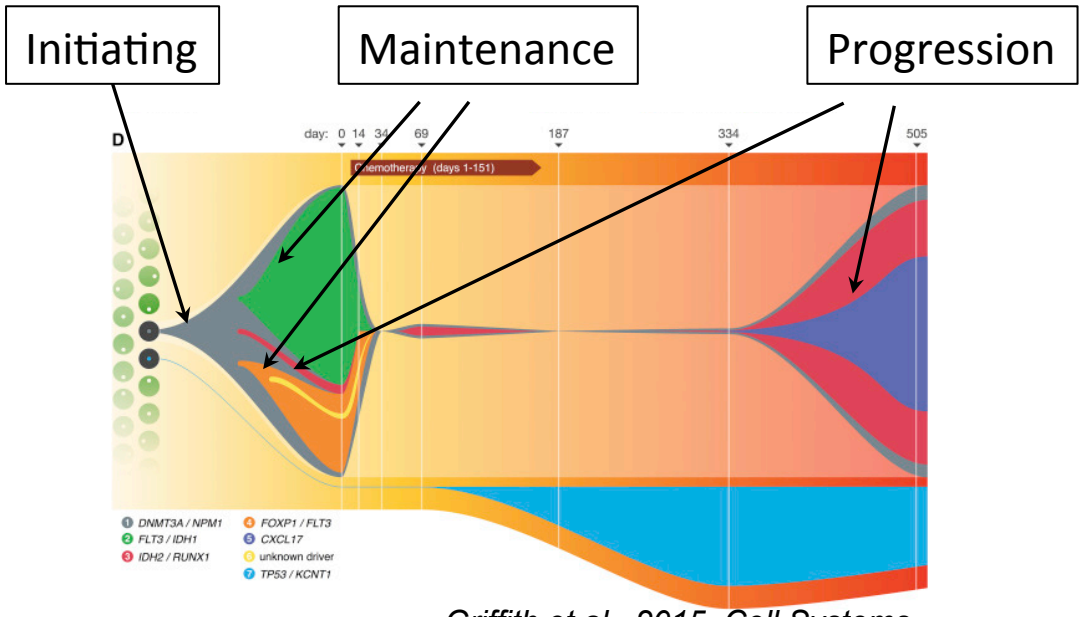
Cancer is an Evolutionary Process

Driver mutation: directly/indirectly confers a selective growth advantage to the cell

Selective growth advantage (*s*): difference between birth and death in a cell population. In normal adult cells in the absence of injury, *s* = 0.000000

Passenger mutation: no direct/indirect effect on the selective growth advantage of the cell

Subclonal mutation: exists in only a subset of the neoplastic cells within a tumor



Griffith et al., 2015, Cell Systems

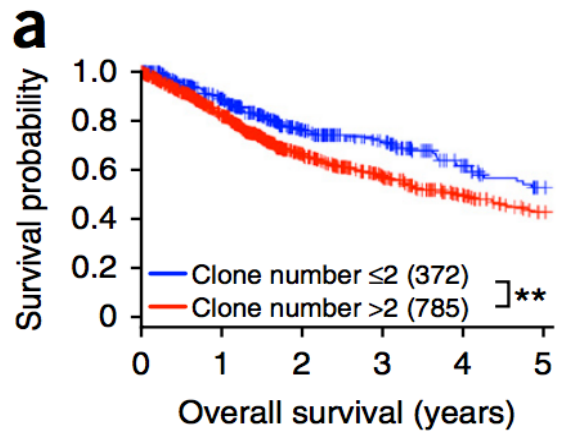
ITH is clinically relevant in multiple cancers

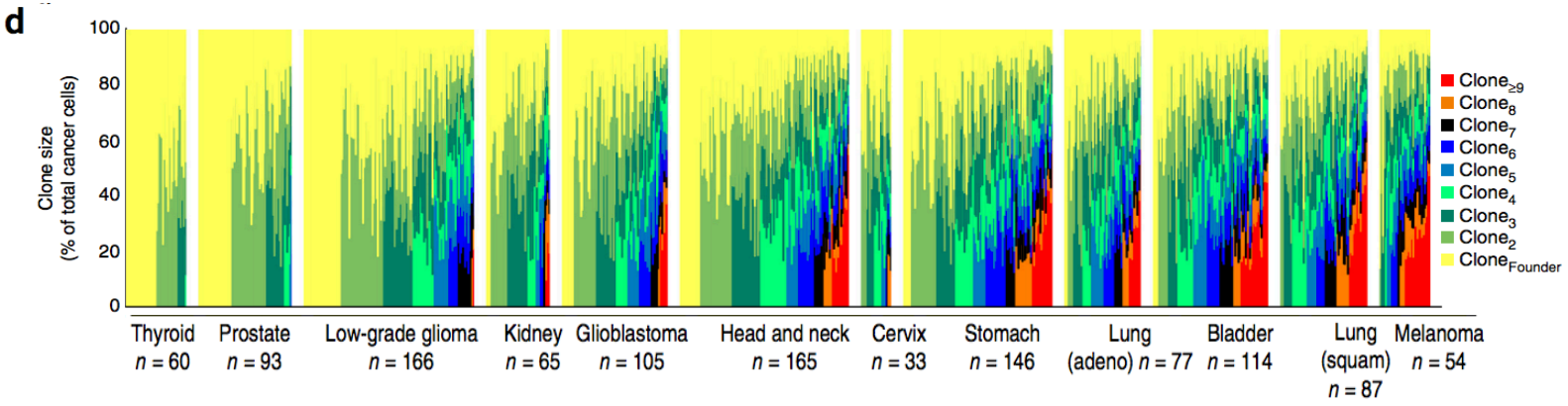
ANALYSIS

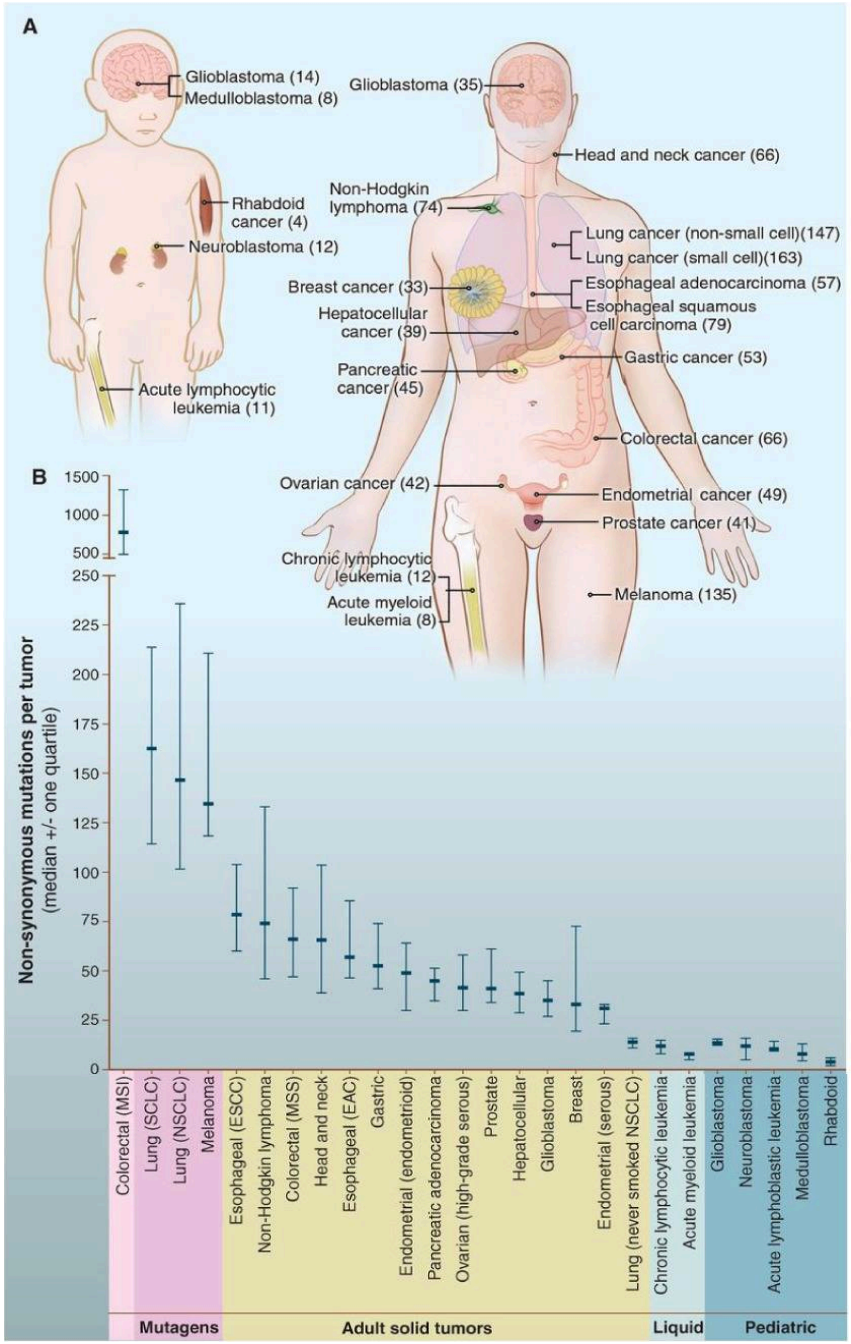
nature medicine

Pan-cancer analysis of the extent and consequences of intratumor heterogeneity

Noemi Andor[1,2], Trevor A Graham[3], Marnix Jansen[3], Li C Xia[1], C Athena Aktipis[4,5], Claudia Petritsch[6–8], Hanlee P Ji[1,9,12] & Carlo C Maley[4,10–12]







Cancer genome landscapes

Typical tumors arise from **2-8 sequentially acquired "driver" genes** conferring a selective growth advantage

- 120-140 genes with driver mutations

Average tumor has 33-66 genes with protein coding changes (drivers + passengers)

Notable outliers

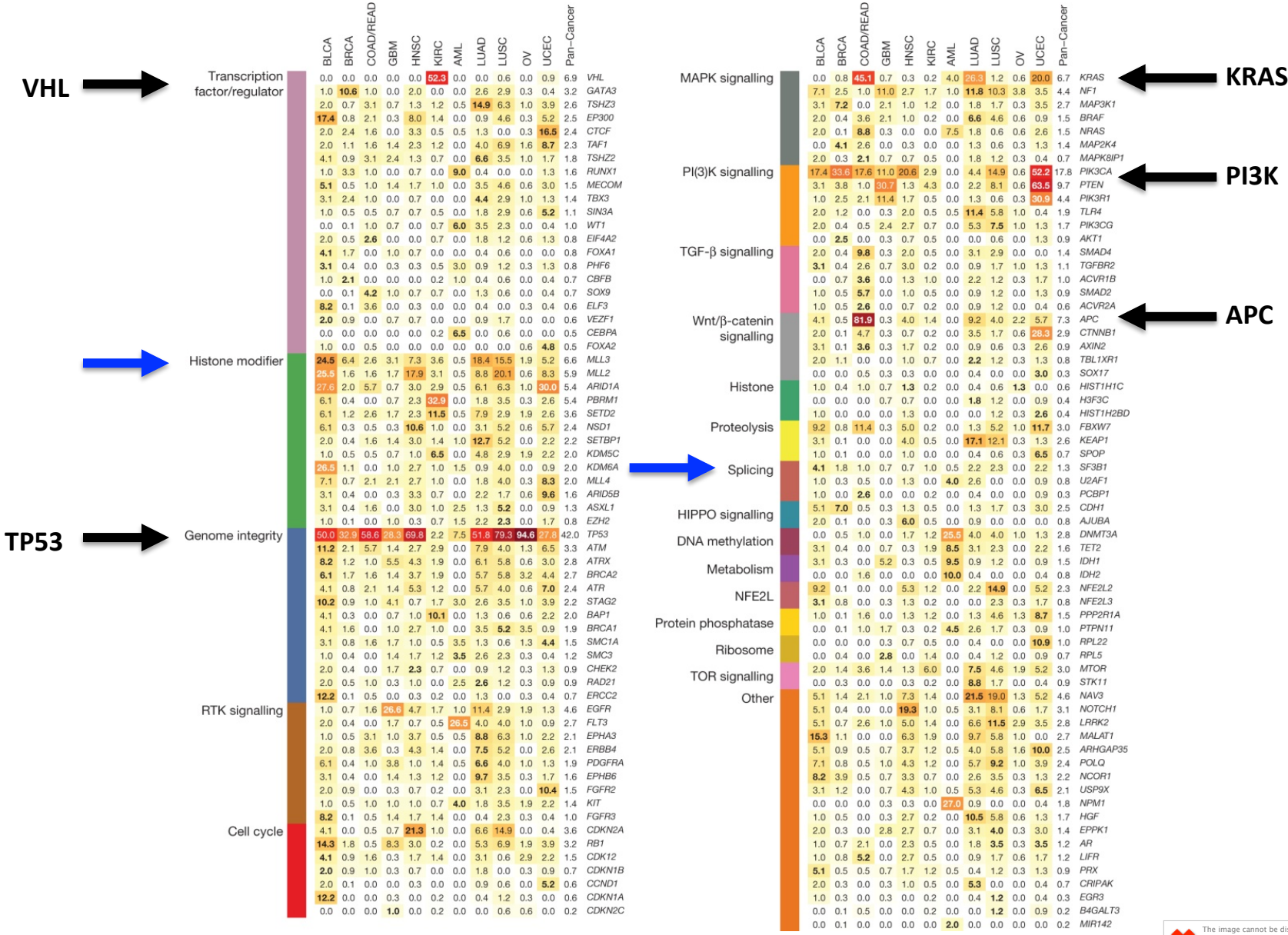
- Melanoma and lung: involvement of potent mutagens (UV light and smoking) – **100's**
- Tumors with defects in DNA repair (mismatch repair: CRC, bMMRD brain tumors) – **1,000s**
- Paediatric tumors - **<10**

Tumor suppressor gene **(TSG)-inactivating mutations predominate** over oncogene-activating mutations

- Few individual tumors contain more than one oncogenic mutation
- Oncogenes are easier to drug than TSGs
  - Target pathways instead

Vogelstein et al., 2013, Science

127 Significantly mutated genes from 20 cellular processes in cancer identified in 12 cancer types.



C Kandoth *et al. Nature* **502**, 333-339 (2013) doi:10.1038/nature12634

Considerations for measuring variant allelic distributions from NGS cancer data

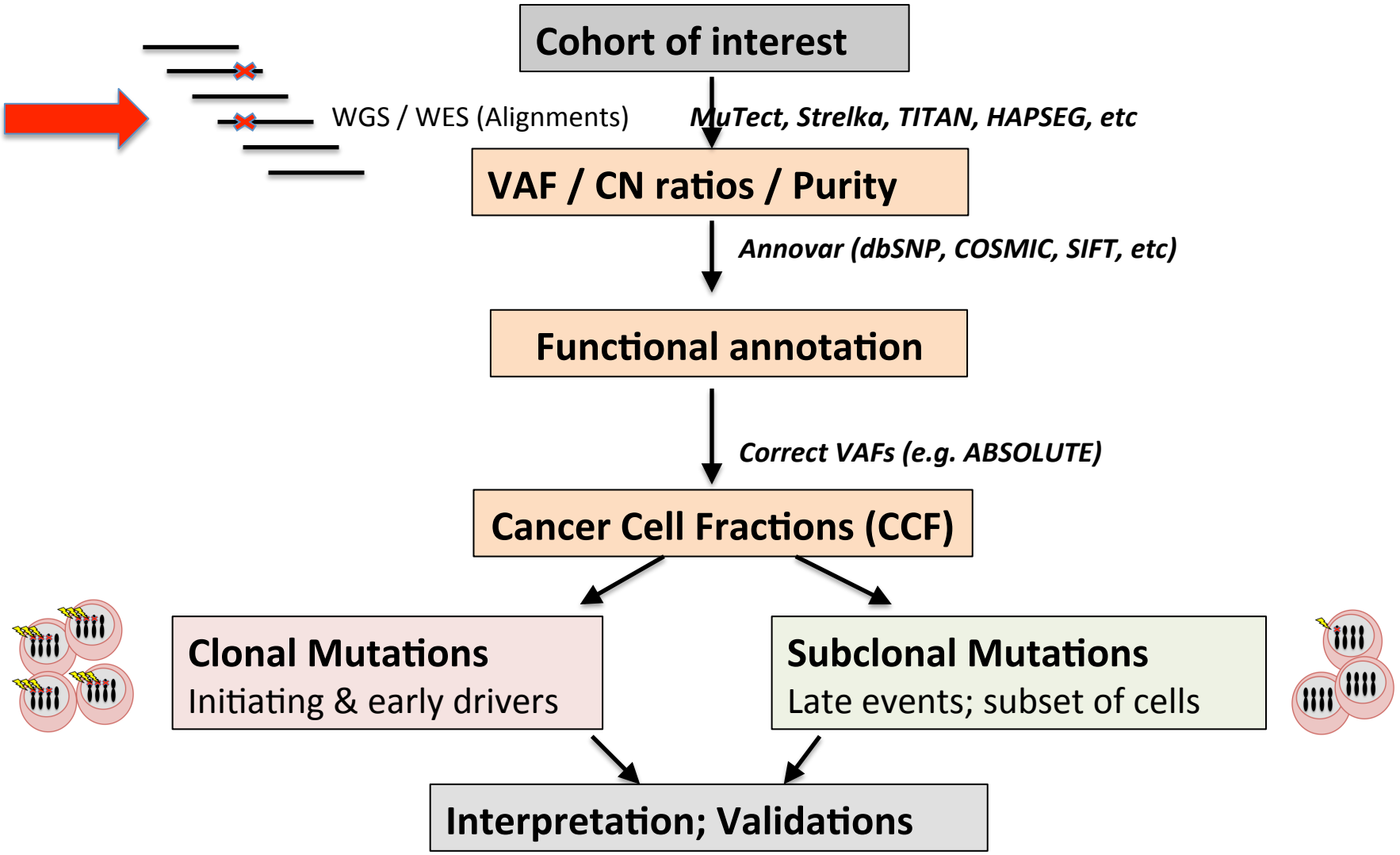
Cancer genomes have specific properties that warrant specialized analytical strategies

- **Tumor/normal admixture**
  - Tumour DNA is often contaminated with DNA from non-malignant cells
  - May dilute important biological signals

- **Intra-tumoural heterogeneity**
  - Cancer is often a mosaic of cellular populations that are genomically distinct

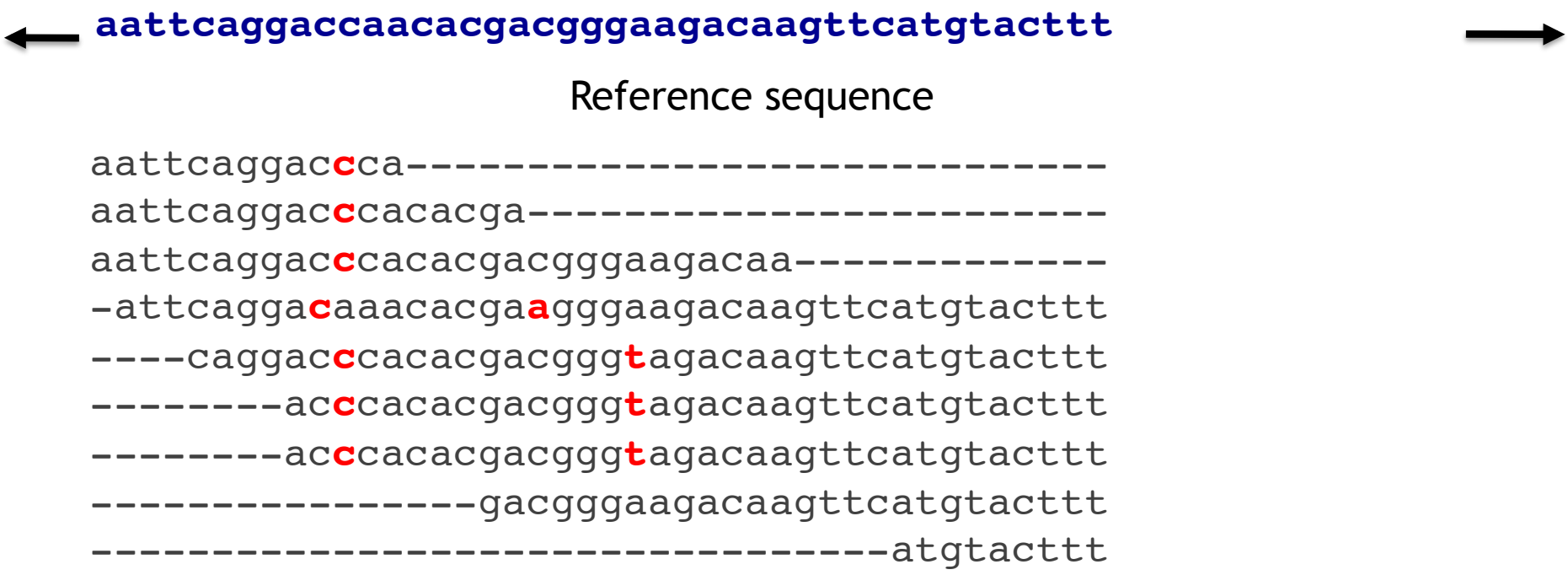
- **Genomic instability**
  - Copy number changes, loss of heterozygosity and genomic rearrangements will distort expected allelic distributions

Analytic approach

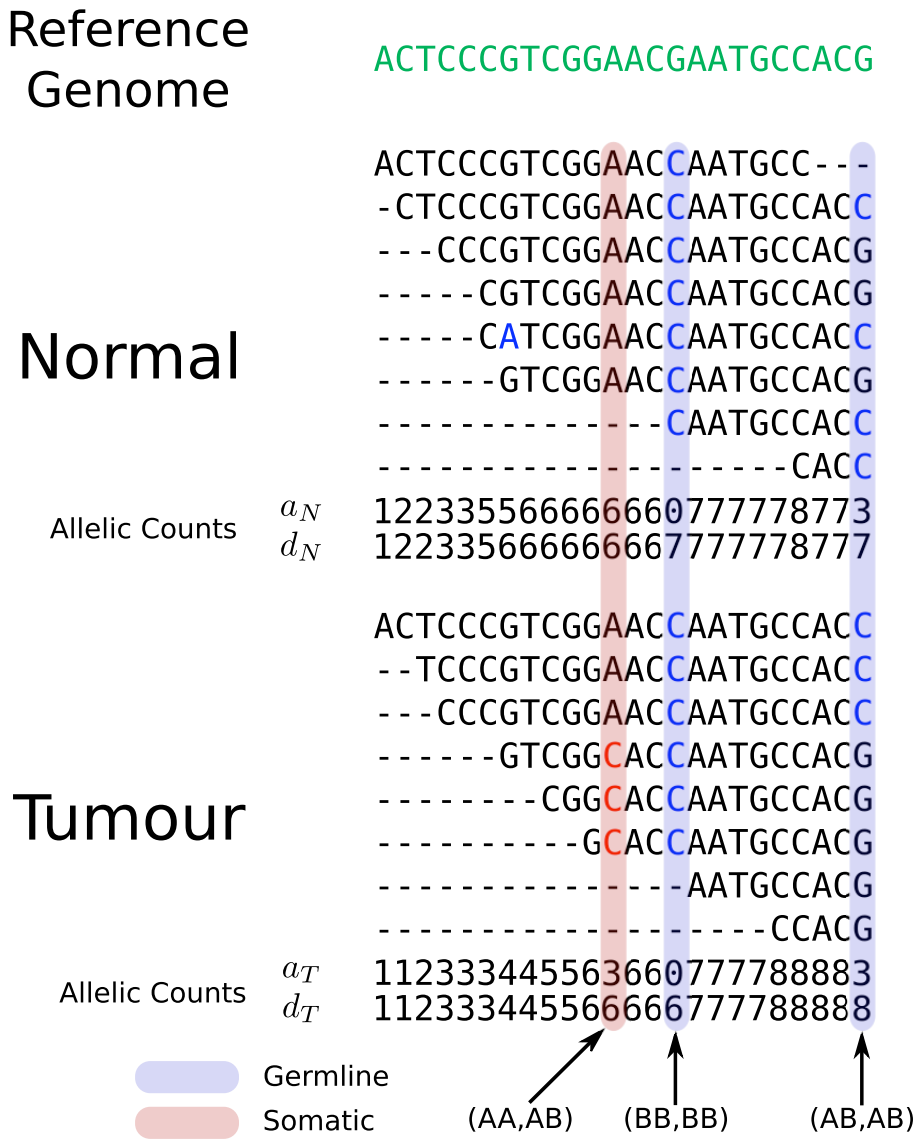


Aligning billions of short reads to the genome

- MAQ, BWA, SOAP, SHRiMP, Mosaik, BowTie, Rmap, ELAND, BWA-Mem
- Chopping, hashing and indexing the genome + string matching with mismatch tolerance



Tumour-Normal allelic count data



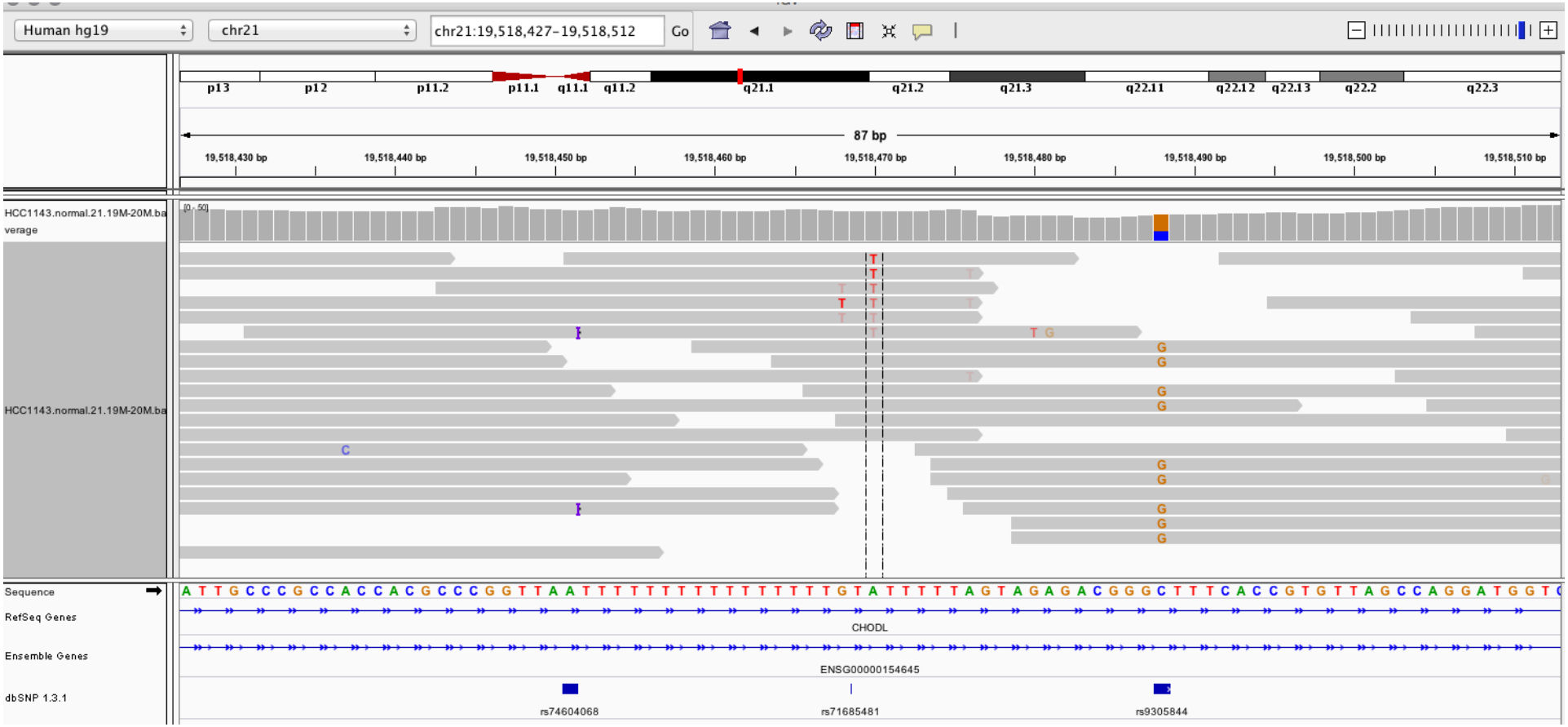
$$P(G^i_{(g_N,g_T)} = 1)$$

$g_N \backslash g_T$	AA	AB	BB
AA	0.01	0.95	0.00
AB	0.00	0.04	0.00
BB	0.00	0.00	0.00

Roth et al Bioinformatics 2012

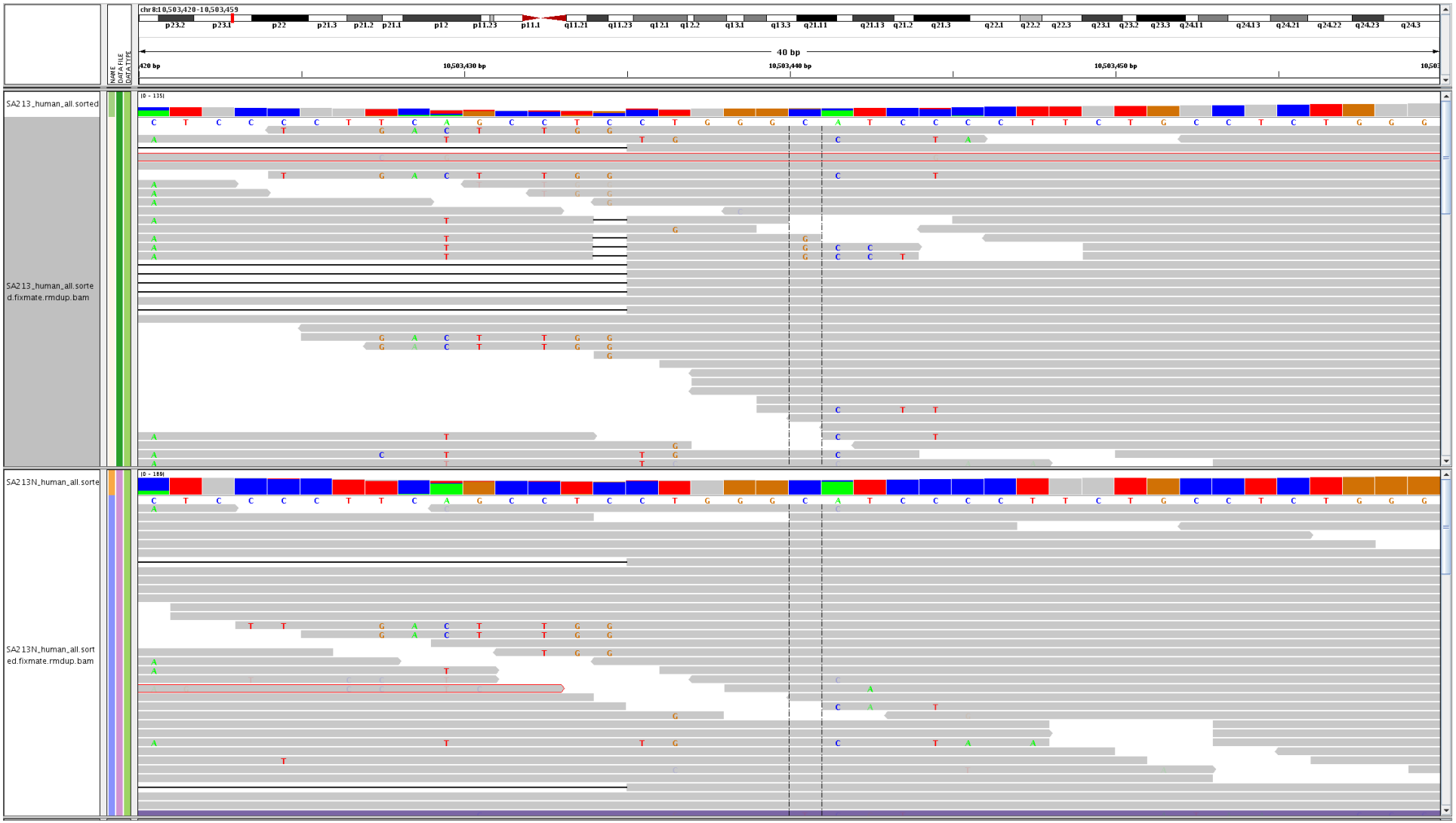
Example artifacts that induce false positives

Sequencing errors



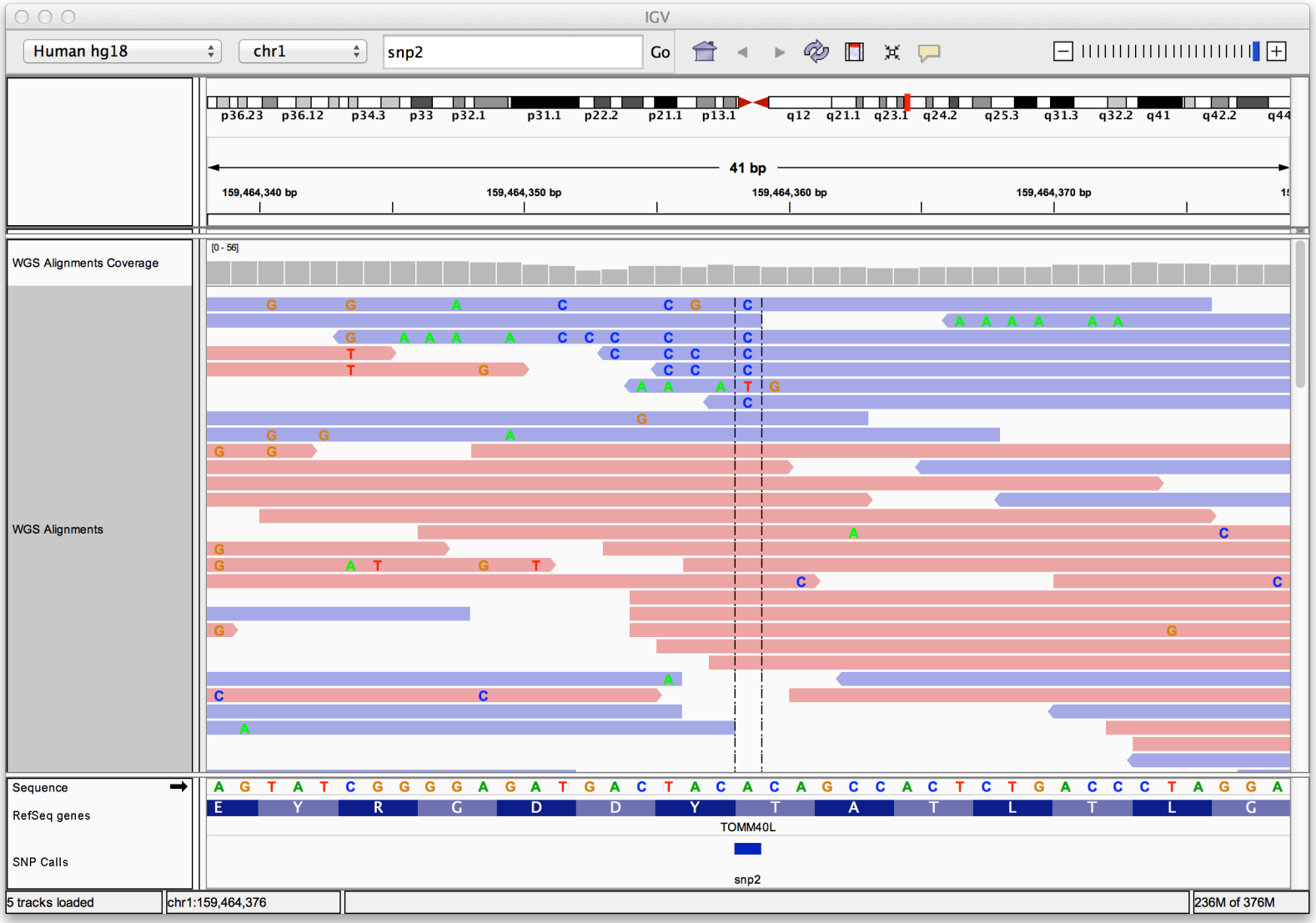
Example artifacts that induce false positives

Indels – these regions typically undergo a local realignment step

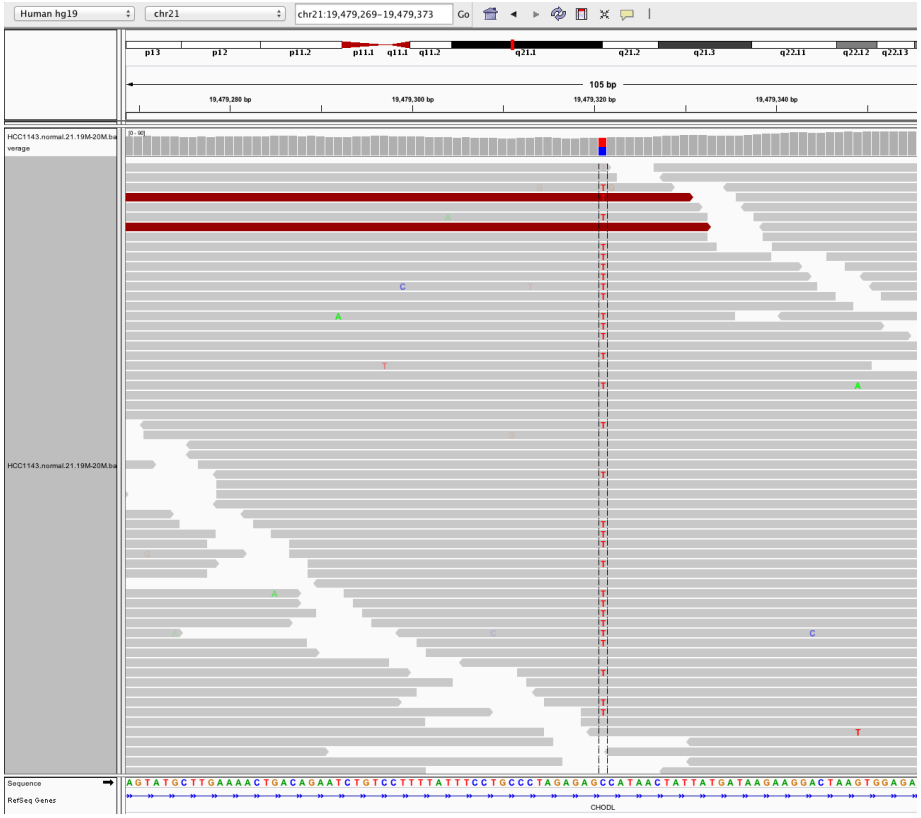


Example artifacts that induce false positives

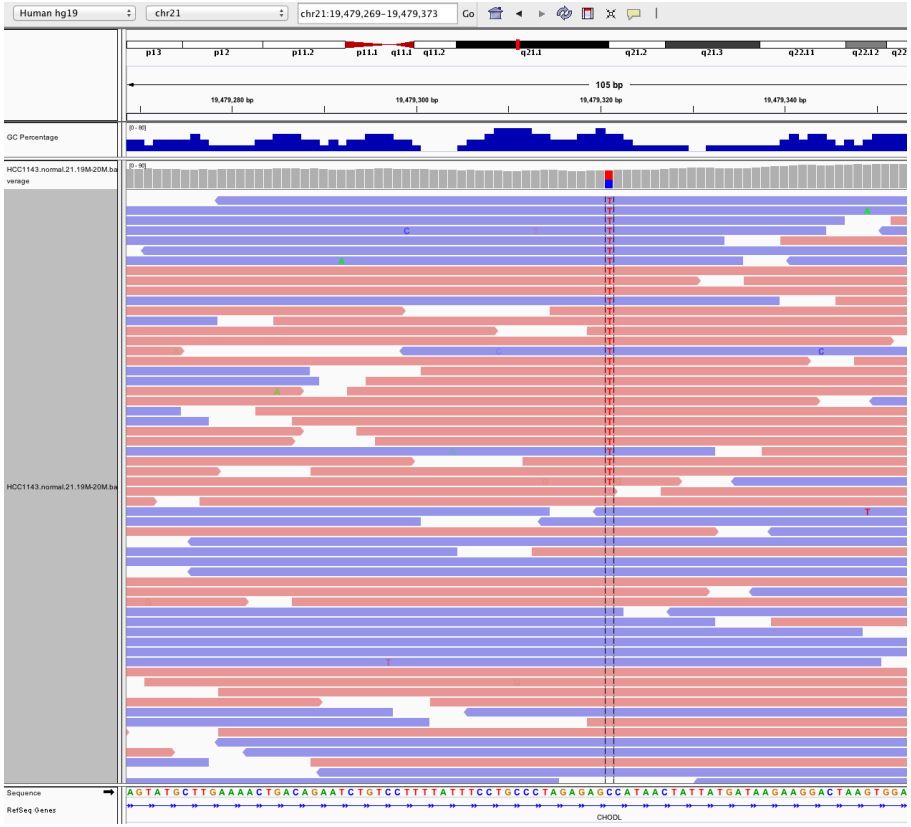
All reads from the same strand



True positive examples

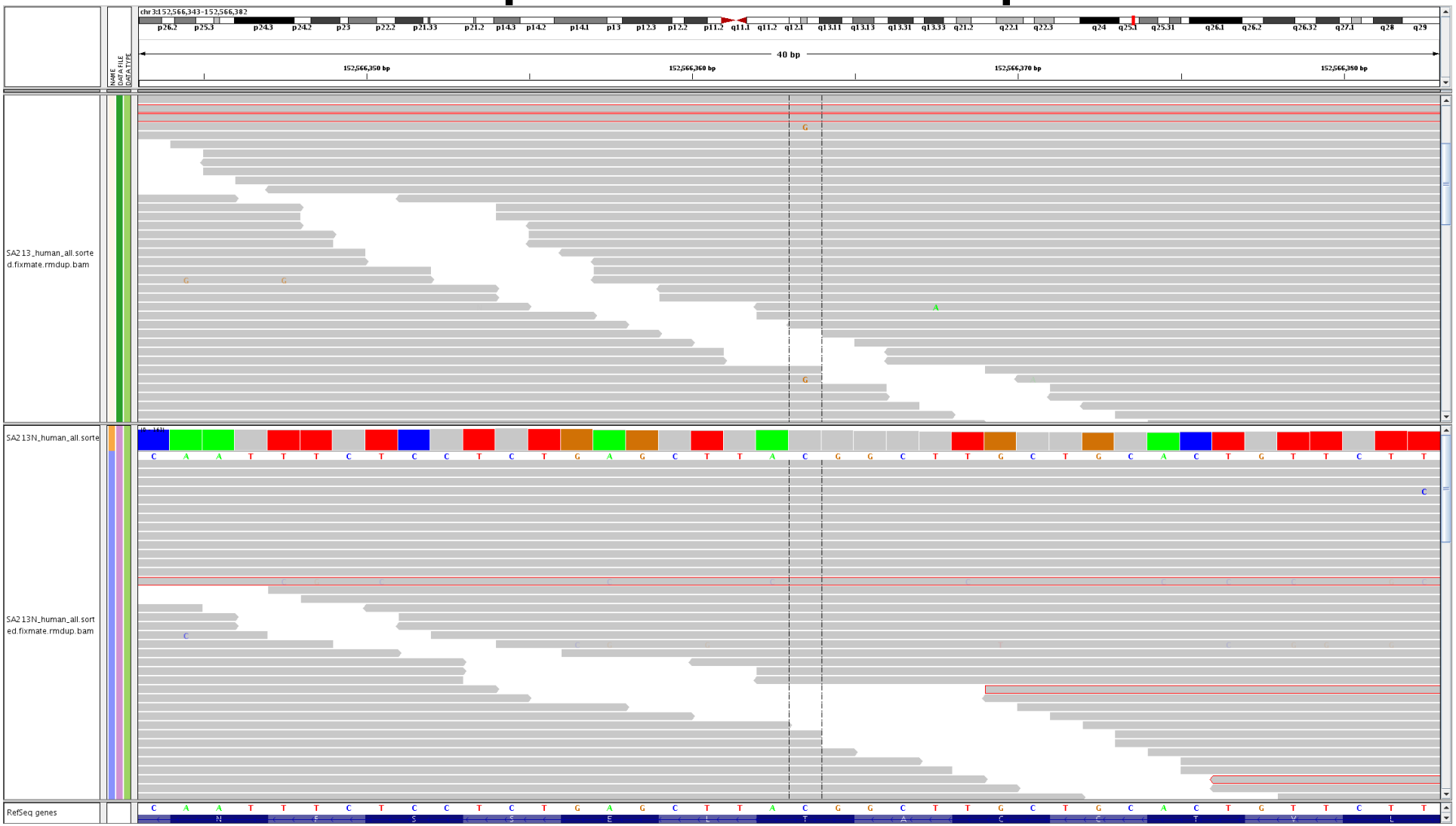


Sort alignments by “start location”

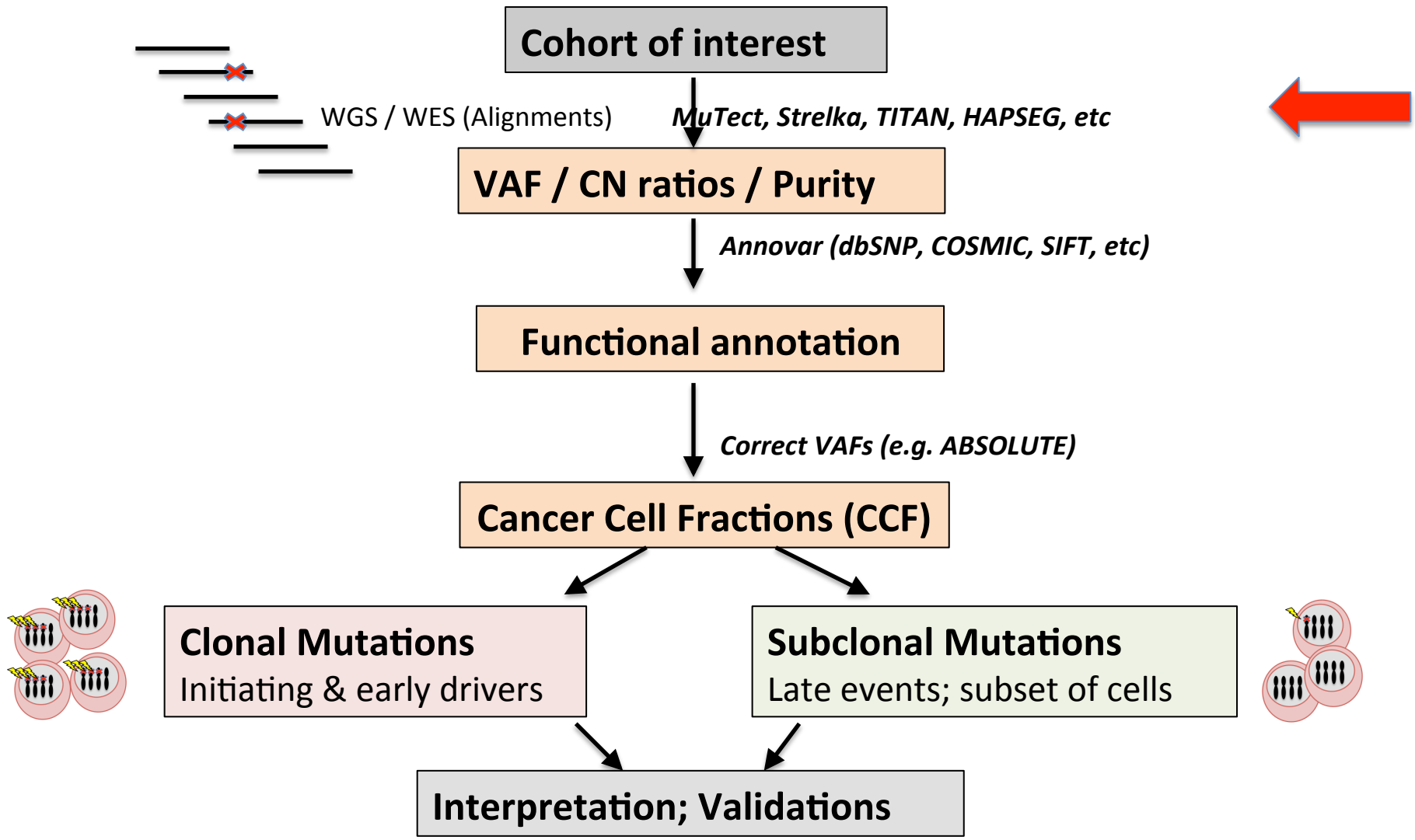


Sort alignments by “base”
Color alignments by “read strand”

True positive examples



Analytic approach



Available tools for somatic mutation calling and visualization

Available tool suite: Samtools

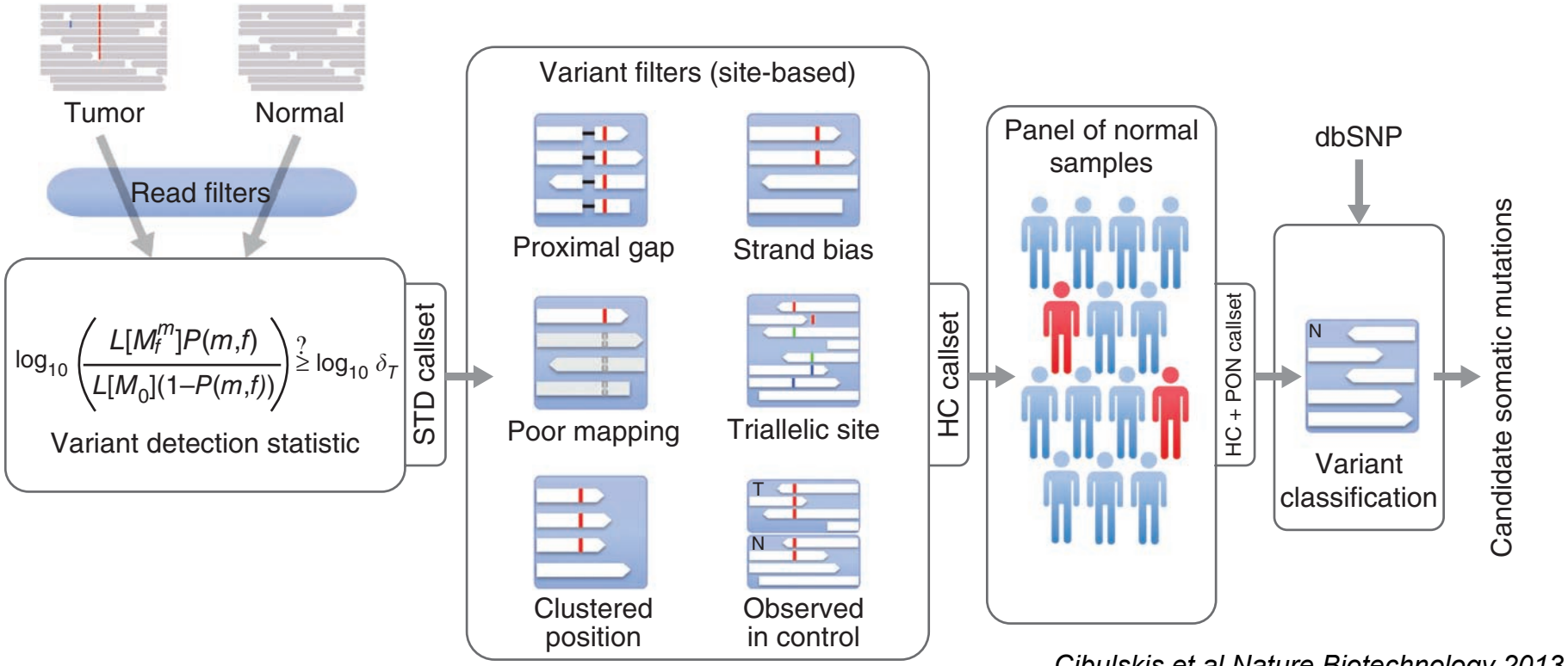
- Suite of tools for working with alignment files in the community standard sam/bam/cram format
  - http://samtools.sourceforge.net/SAM1.pdf
- Implemented in C
  - Fast and memory efficient

Available tool suite: GATK

– Widely used pipeline for variant calling – Broad best practices

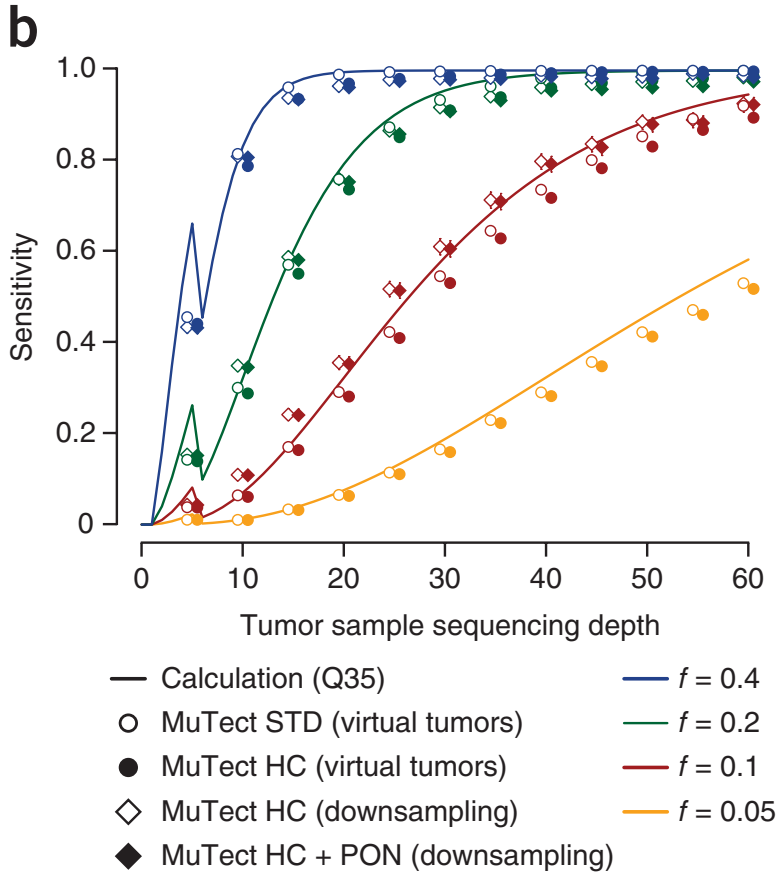
Available tools: MuTect2

- MuTect: http://www.broadinstitute.org/software/cprg/?q=node/34
- **MuTect2** released 2017, also calls indels



Cibulskis et al Nature Biotechnology 2013

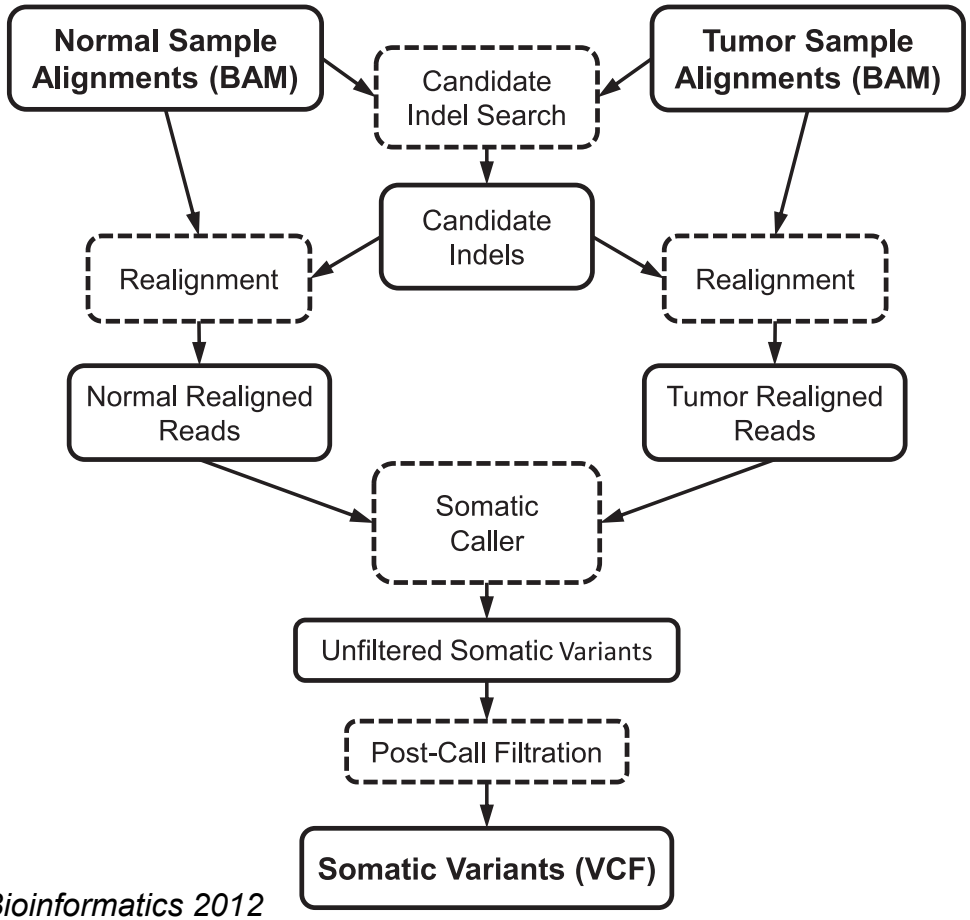
MuTect2: sensitivity to low-frequency mutations



Cibulskis et al Nature Biotechnology 2013

Available tools: Strelka (high specificity)

Strelka (from Illumina): https://sites.google.com/site/strelkasomaticvariantcaller/



Saunders et al Bioinformatics 2012

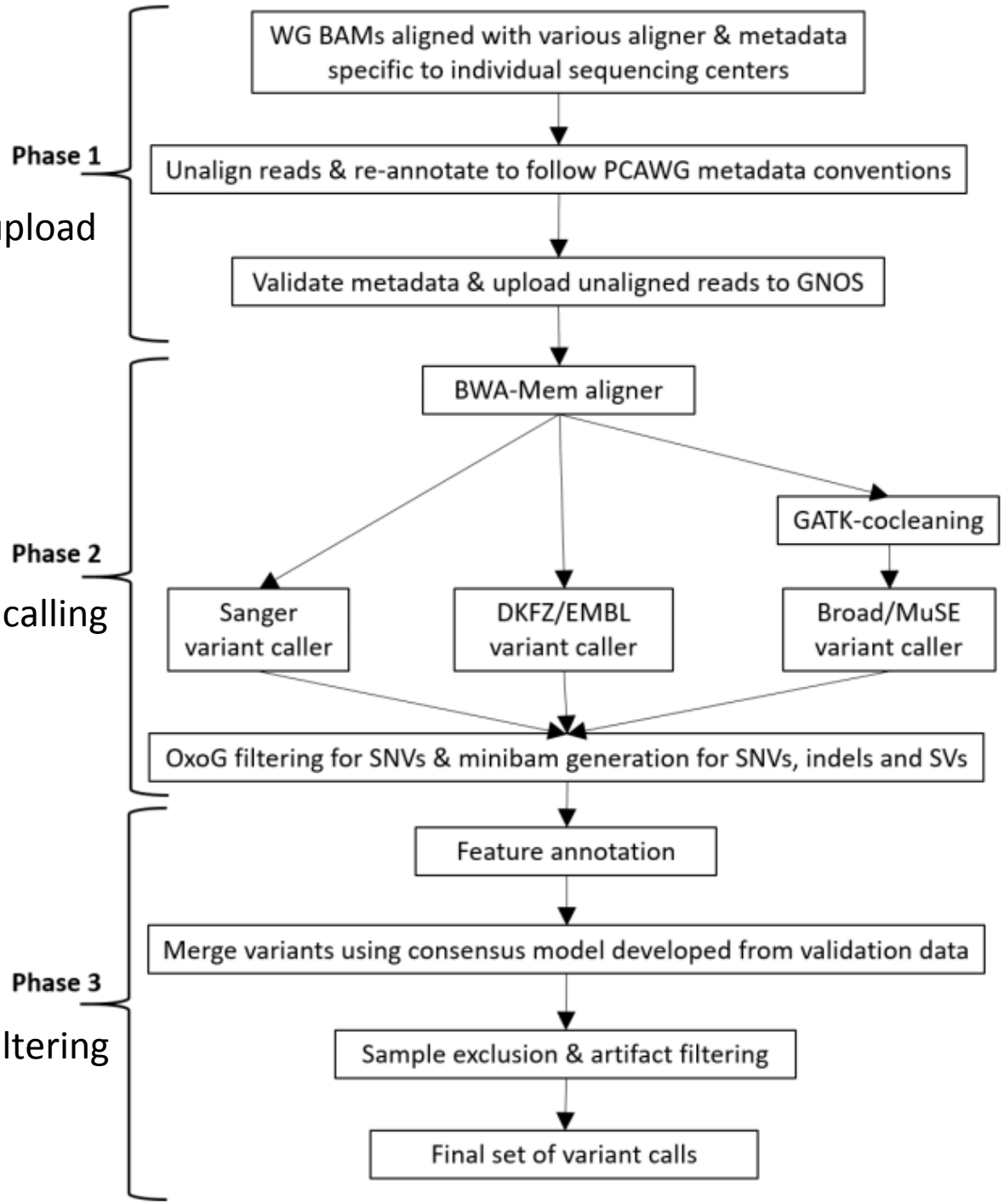
PCAWG-TCGA

pan-cancer analysis of whole genomes

Data marshaling and upload

Sequence alignment and variant calling

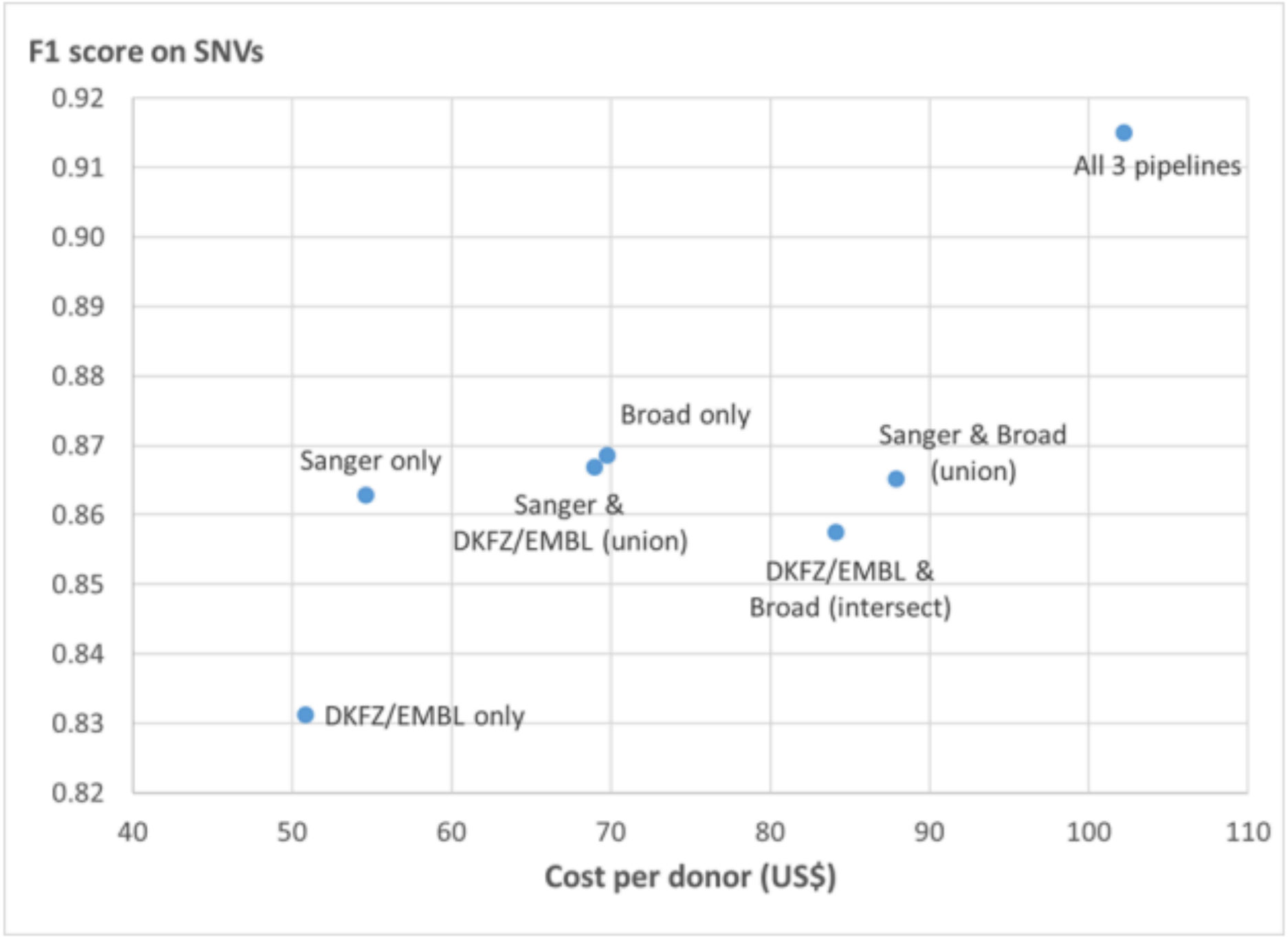
Variant Merging and Filtering



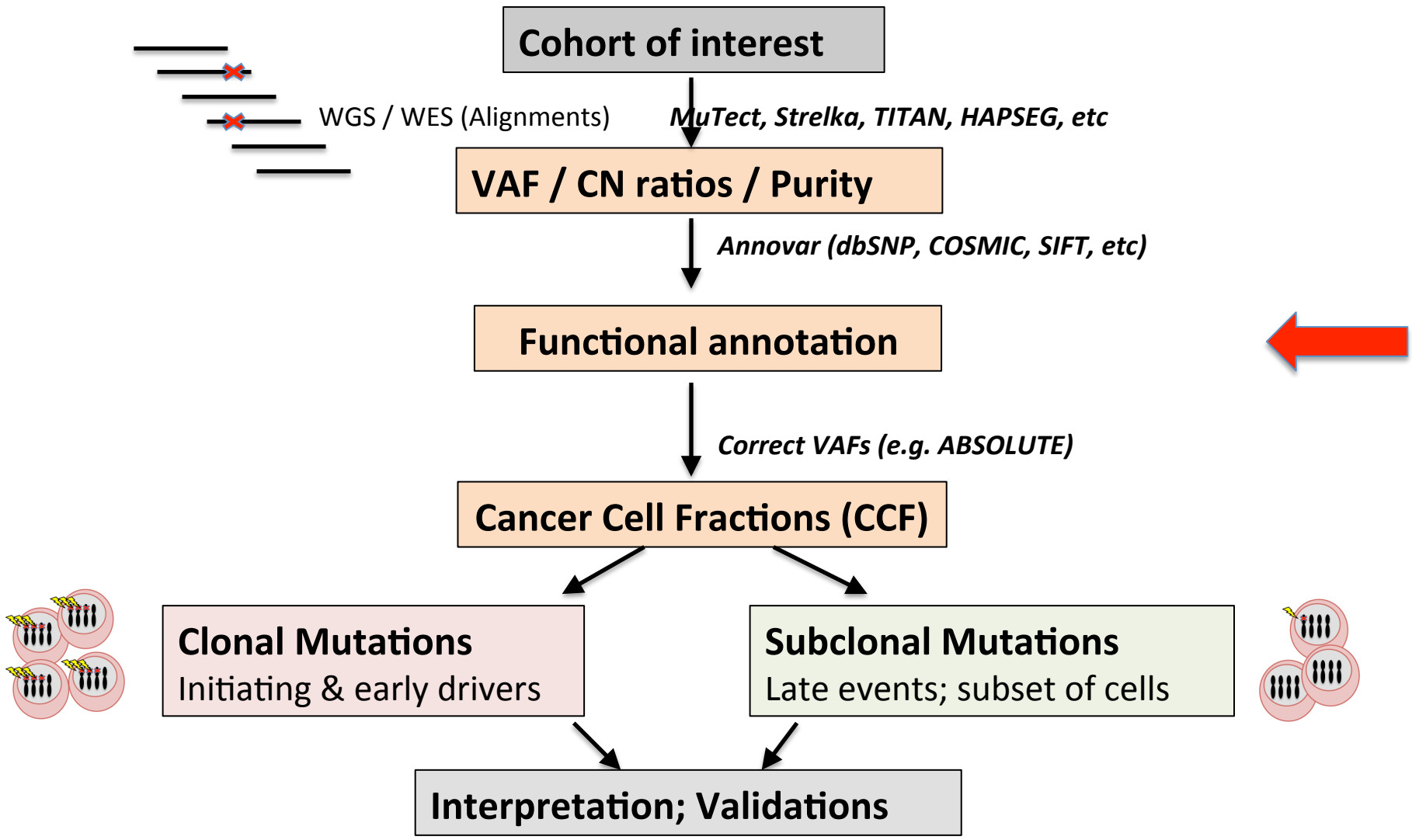
PCAWG – Phase 2 (T:N pair)

(1) SNVs, (2) indels, (3) SVs and (4) SCNAs

	BWA	Sanger	DKFZ/EMBL	Broad	OxoG
Analytical components in workflow	BWA-Mem Picard Biobambam samtools	CaVEMan[1] cgpPindel[2] BRASS[3] ascatNgs[4]	dkfz_snv[1] Platypus[2] DELLY[3] ACE-seq[4]	GATK cocleaning MuTect[1] MuSE[1,#] Snowman[2,3] dRanger[3]	OxoG VariantBam
Workflow controller	SeqWare	SeqWare	Roddy, SeqWare	Galaxy	SeqWare
Recommended compute requirements	4 cores, 15GB RAM	16 cores, 4.5GB RAM/core	16 cores, 64GB RAM	32 cores, 244GB RAM	8 cores, 64GB RAM
Average runtime across all compute environments	2.0 +/- 1.7 days	5.3 +/- 5.5 days	3.2 +/- 1.7 days	5.1 +/- 2.2 days	2.6 +/- 1.3 hours
Benchmark on AWS	5.8 days on 4-core m1.xlarge	2.2 days on 32-core r3.8xlarge	1.7 days on 32-core r3.8xlarge	3.7 days on 32-core r3.8xlarge	4 hours on 8-core m2.4xlarge
Core hours per run	557	1690	1306	2842	32
Output files per run	120GB	2 GB	5 GB	35 GB	1.5 GB


F1 score on SNVs
0.92
0.91
0.90
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0.82
40
50
60
70
80
90
100
110
Cost per donor (US$)
All 3 pipelines
Broad only
Sanger & Broad (union)
Sanger only
Sanger & DKFZ/EMBL (union)
DKFZ/EMBL & Broad (intersect)
DKFZ/EMBL only

Analytic approach

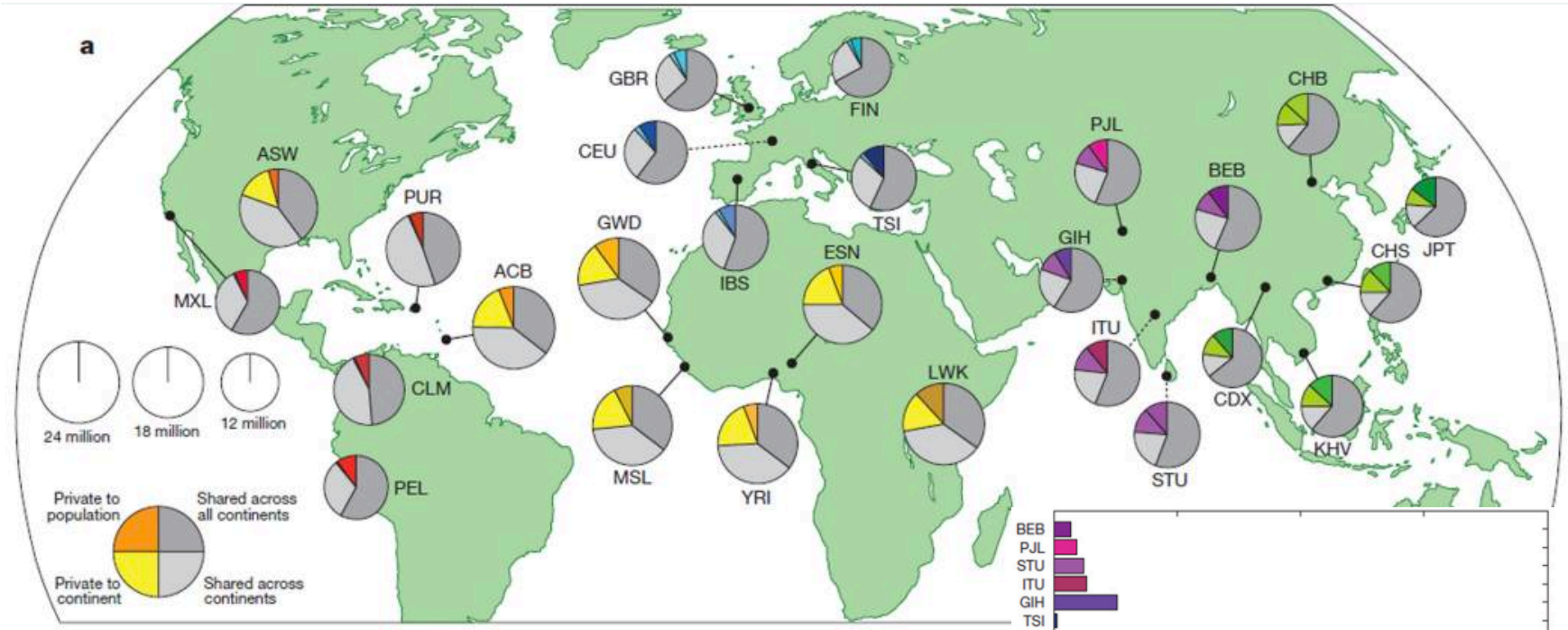


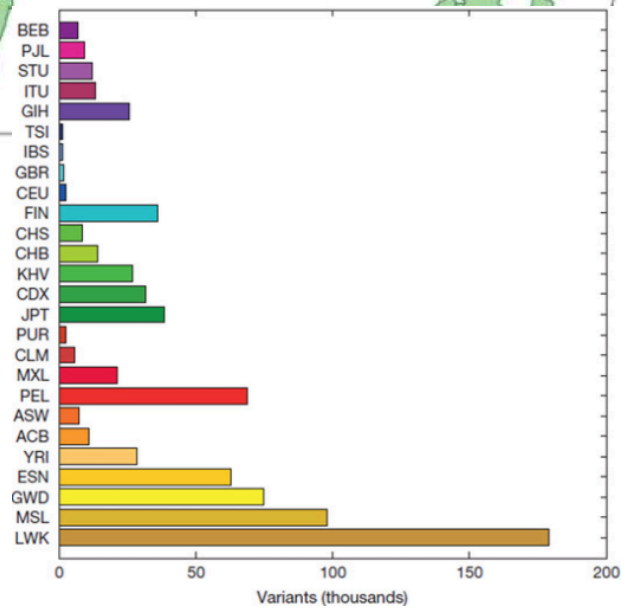
Mutations vs polymorphisms

- Single Nucleotide Polymorphisms (**SNPs**)
  - Common mutations present in >1% of the population
  - If deleterious to fitness, selected against (rare)
  - If advantageous, selected for (prevalent)
  - May be associated with disease susceptibility, drug responses
  - Germline

- Single Nucleotide Variants (**SNVs**)
  - Infrequent and potentially harmful variation, usually associated with disease
  - <1% of the population
  - Germline occurrence leads to disease predisposition (TSG)

Databases of variants: 1000 genomes project





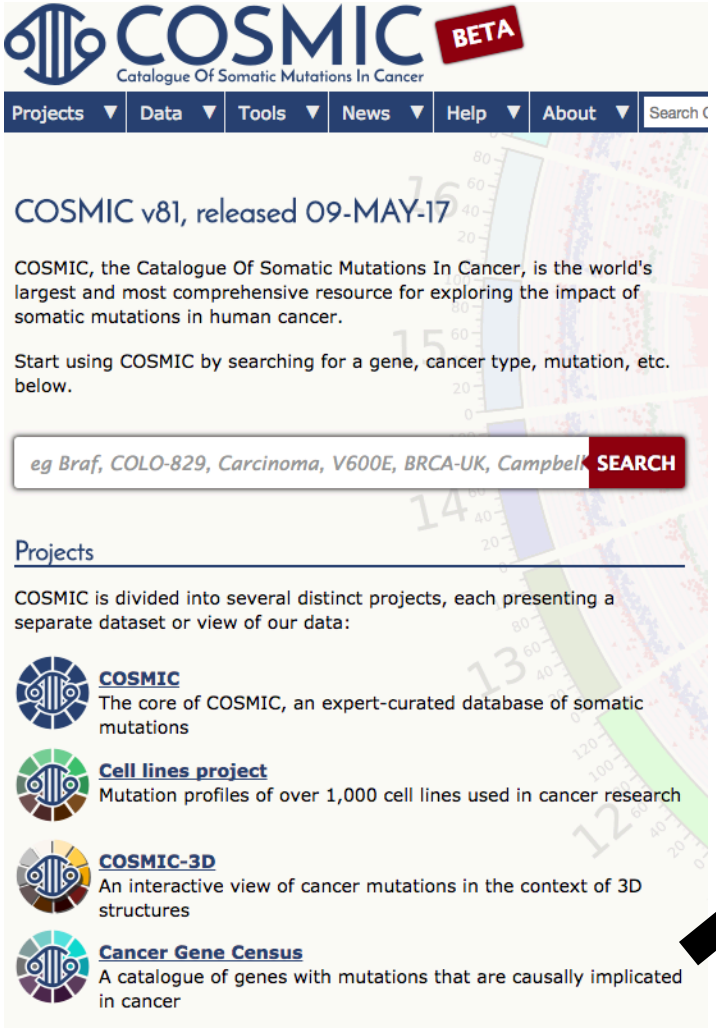
On average, each person carries around 250-300 loss-of-function variants in annotated genes and 50-100 variants previously implicated in inherited disorders.

Estimated rate of de novo mutation is ~ 0.1 – 1 mutation per cell cycle

Databases of variants: dbSNP

- dbSNP150: ~130 million SNPs with a known frequency in human populations
- Up to 8% of SNPs may be false positives due to related sequences (e.g. SNP is “found” by PCR assay that uses primers which bind to paralogs)
- High-quality calls from population re-sequencing efforts (1000g)
- snp128**NonFlagged**:
  - flagged SNPs
    - SNPs < 1% minor allele frequency (MAF) (or unknown)
    - "clinically associated”

Databases of variants: COSMIC



Sorted By	
Amplifications	18
Chromosome	616
Frameshift Mutations	151
Gene Symbol	616
Germline Mutations	101
Large Deletions	41
Missense Mutations	243
Nonsense Mutations	153
Other Mutations	37
Somatic Mutations	575
Splicing Mutations	76
Translocations	360

cancer.sanger.ac.uk/cosmic

Database of variants: COSMIC

COSMIC BETA
Catalogue Of Somatic Mutations In Cancer

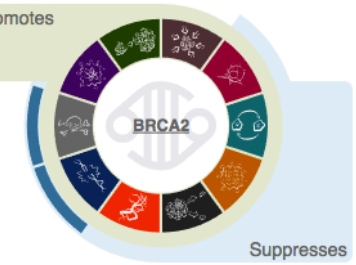
Hallmarks of Cancer

Projects | Data | Tools | News | Help | About | Search COSMIC... SEARCH | Login

BRCA2

switch view

familial breast/ovarian cancer gene 2



Function summary

DNA damage repair protein [Pubmed]

genome instability and mutations

critical for stabilization of stalled replication forks [Pubmed]; stimulates RAD51-dependent homologous recombination during mitosis [Pubmed]

escaping programmed cell death

silencing promotes resistance to anoikis [Pubmed]

Role in cancer	TSG
Cell division control	critical for meiotic recombination [Pubmed]
Differentiation and development	deficiency leads to meiotic impairment and infertility in mice [Pubmed]
Types of alteration in cancer	recurrent missense mutations: N372H [Pubmed]
Clinical impact	germline variant N372H is associated with higher risk of ovarian cancer, non-Hodgkin lymphoma [Pubmed]

Colour	Mutation type	Number of samples (%)
	Nonsense substitution	79 (9.91%)
	Missense substitution	490 (61.48%)
	Synonymous substitution	140 (17.57%)
	Inframe insertion	1 (0.13%)
	Frameshift insertion	30 (3.76%)
	Inframe deletion	4 (0.50%)
	Frameshift deletion	74 (9.28%)
	Complex mutation	0 (0.00%)
	Other	7 (0.88%)
	Total unique samples	797

Annotation tools

- **ANNOVAR**: Functional annotation of genetic variants from high-throughput sequencing data

- **SNPEFF**: Genetic variant annotation and effect prediction toolbox.

- **VEP**: determines effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions

- **GEMINI**: a flexible framework for exploring genome variation, prioritizing genetic variants in various disease contexts based on genome annotation, sample genotypes, and sample relationships

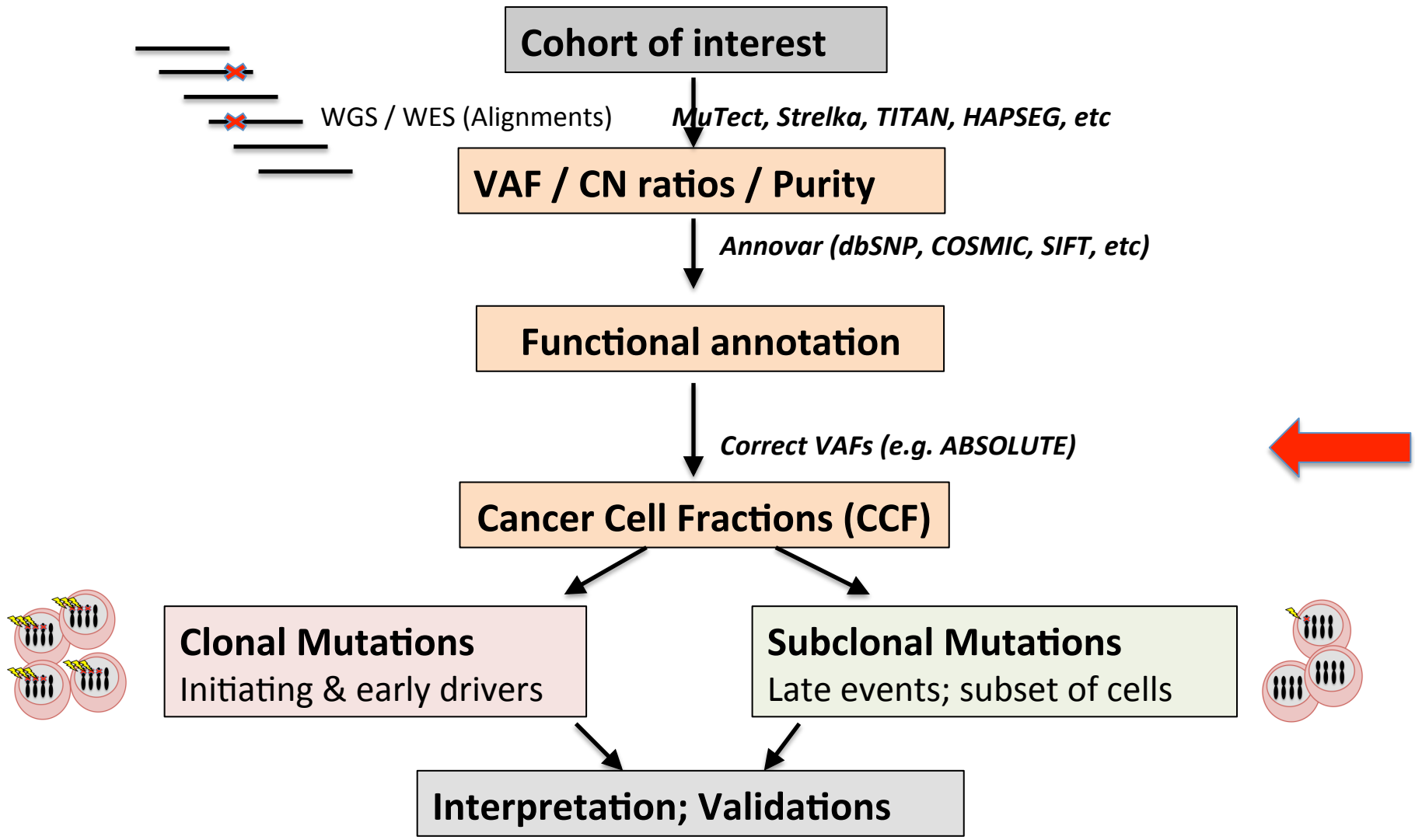
Annotations: Annovar

- Functional annotation of variants:
  - **Gene-based**: protein coding changes, amino acids affected
    - RefSeq genes, UCSC genes, ENSEMBL genes, GENCODE genes
  - **Region-based**: conserved regions, TFBS, GWAS hits, seg duplication regions, etc
  - **Filter-based**: dbSNP, 1000g, gnomAD, etc
    - mutation effect and impact predictions: SIFT/PolyPhen/LRT/MutationTaster/ MutationAssessor/FATHMM/MetaSVM/MetaLR

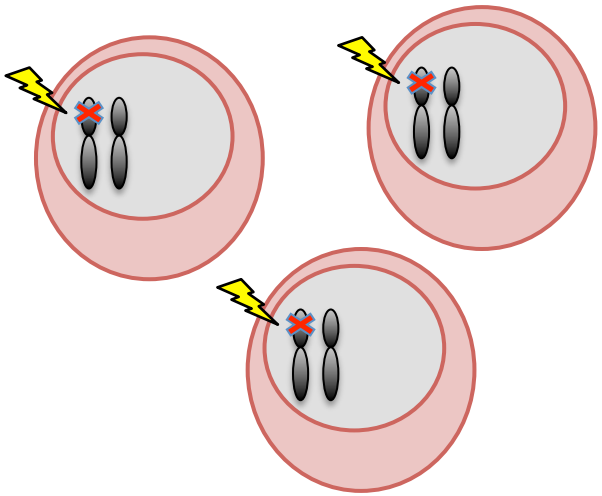
- Available databases
  - More on GRCh37 (hg19) than GRCh38

http://doc-openbio.readthedocs.io/projects/annovar/en/latest/user-guide/download/?highlight=database

Analytic approach

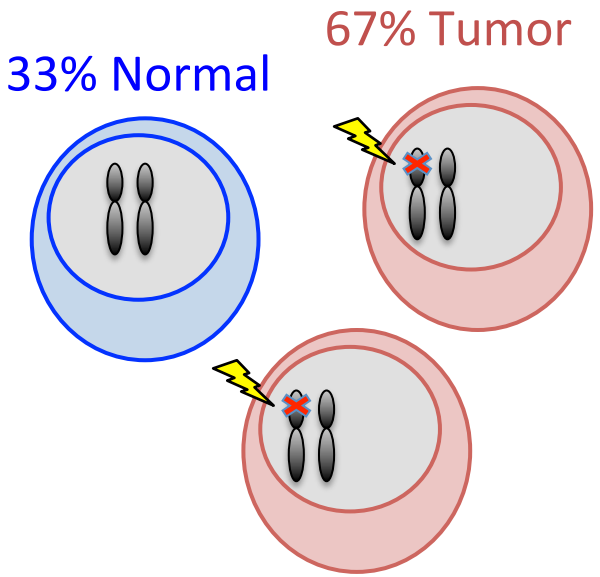


Metrics for reporting mutations: VAF, CCF, multiplicity



Purity = 100%
Ploidy = 2n
Mutation multiplicity (copies/cell) = 1

- Variant Allele Fraction = 3/6 = **0.5**
- **Cancer Cell Fraction = 1**

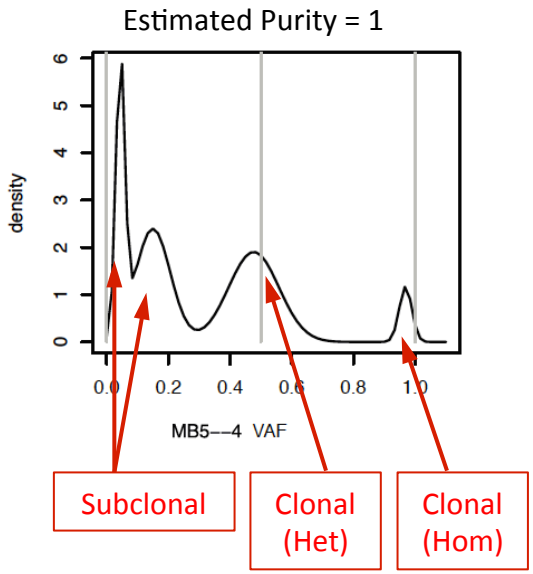


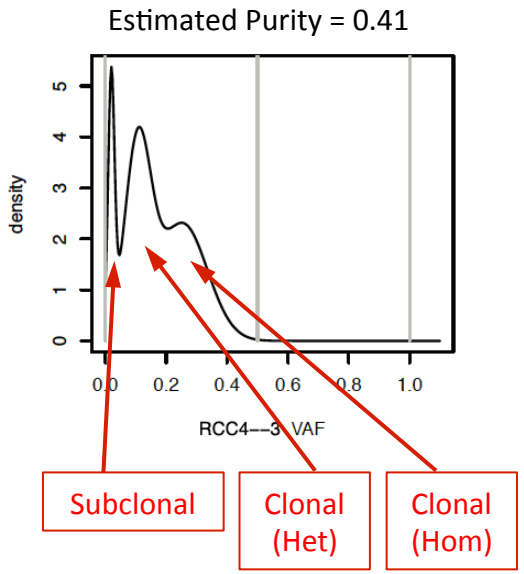
Purity = 67%
Ploidy = 2n
Mutation multiplicity (copies/cell) = 1

- Variant Allele Fraction = 2/6 = **0.33**
- **Cancer Cell Fraction = 1**

Metrics for reporting mutations: VAF, CCF, multiplicity

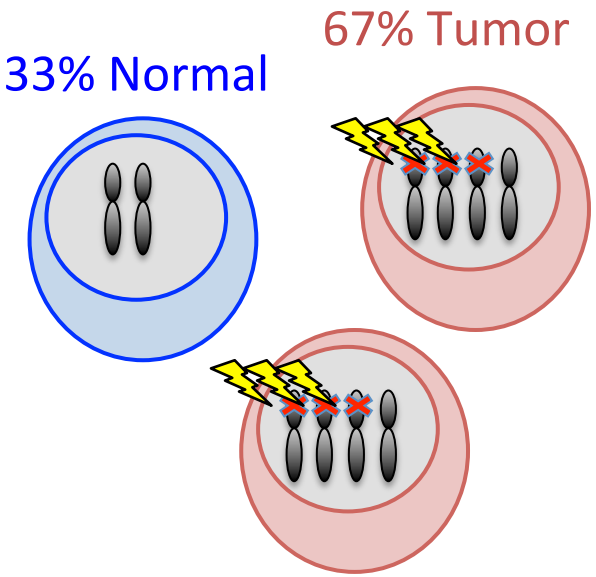
Low purity "pushes" the VAF towards the left





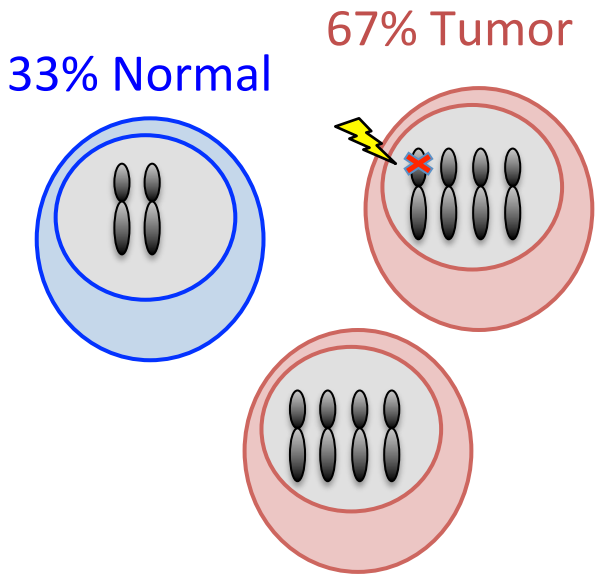
Morrissy et al., 2017, Nature Genetics

Metrics for reporting mutations: VAF, CCF, multiplicity



Purity = 67%
Ploidy = 4n
Mutation multiplicity (copies/cell) = 3

- Variant Allele Fraction = 6/10 = 0.6
- **Cancer Cell Fraction = 1**

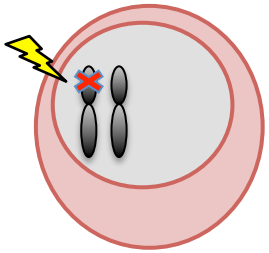


Purity = 67%
Ploidy = 4n
Mutation multiplicity (copies/cell) = 1

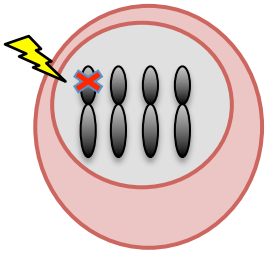
- Allelic Fraction = 1/10 = 0.1
- **Cancer Cell Fraction = 0.5**

Multiplicity and CCF are associated with timing of mutation

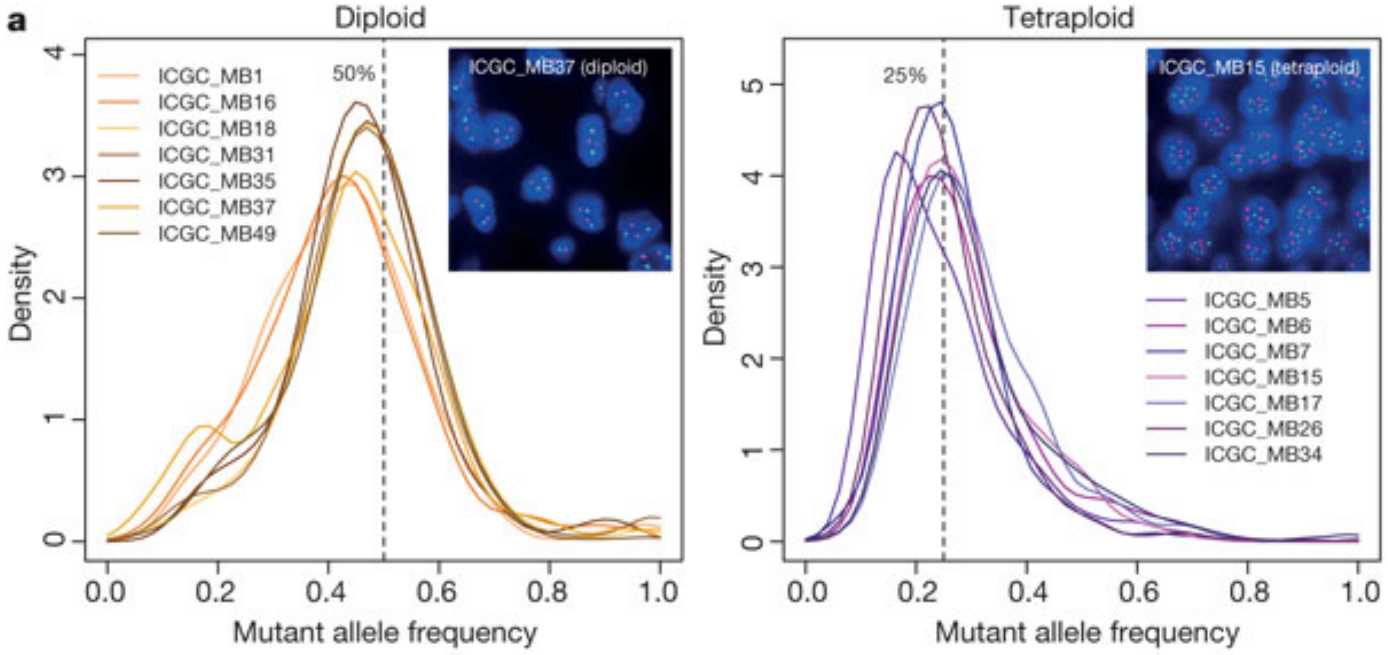
~100% Tumor



~100% Tumor

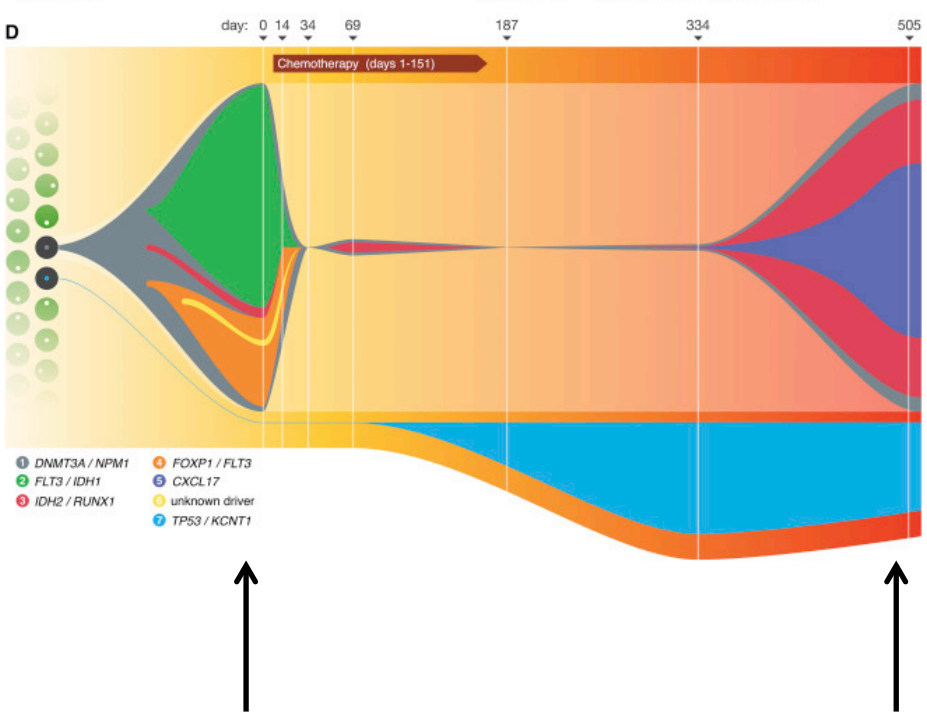


Mutation occurs after genome duplication



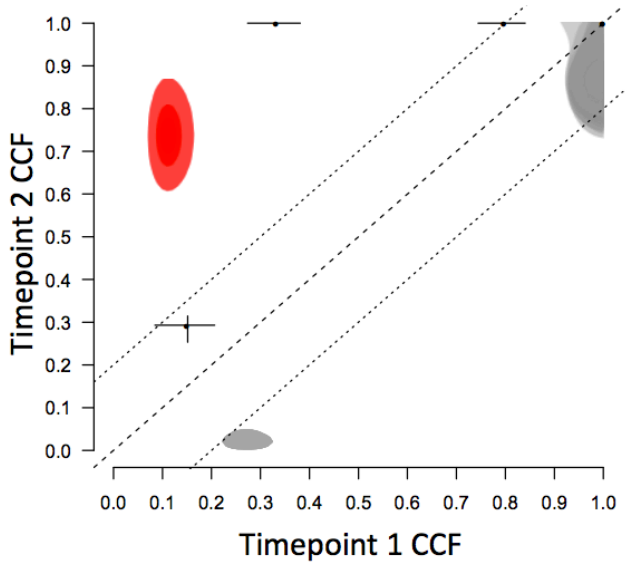
Jones et al., 2012, Nature

Using CCF to infer clonal dynamics

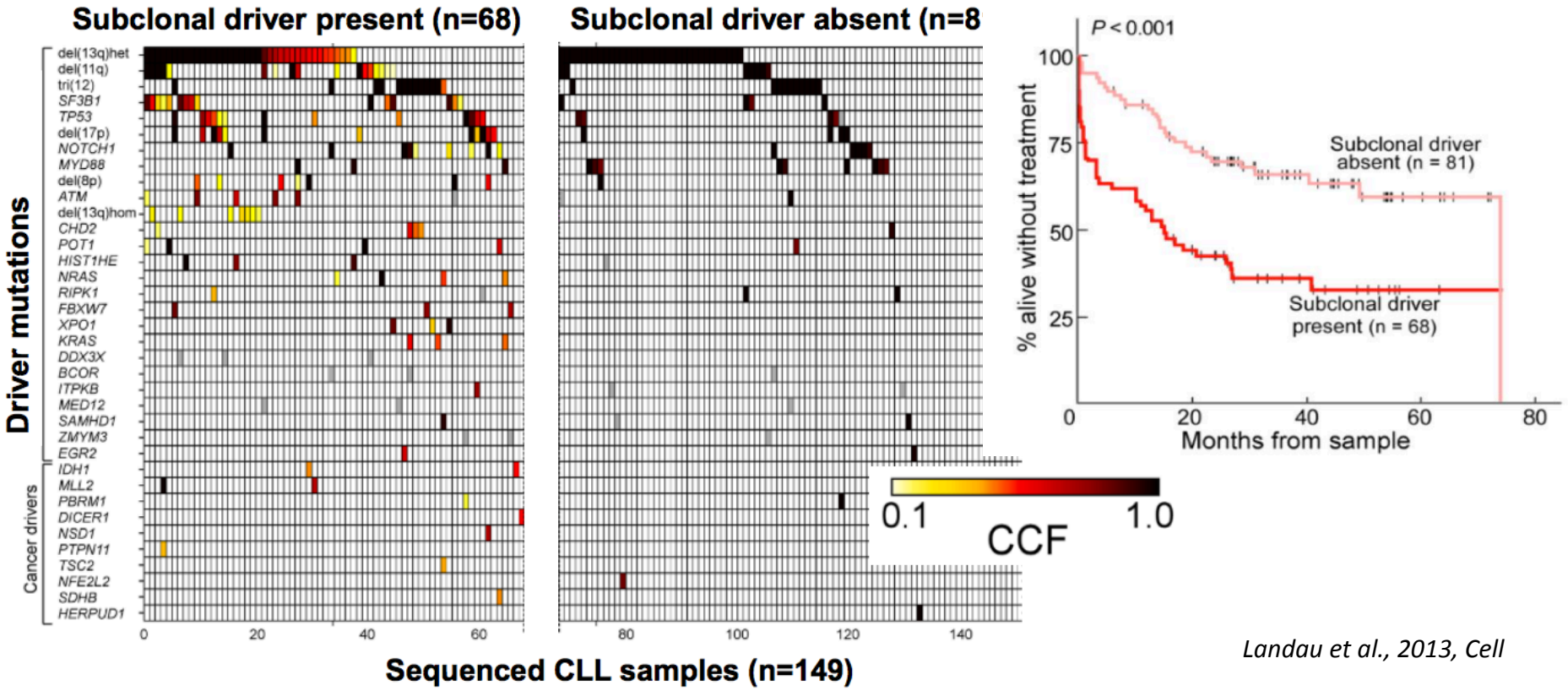


Timepoint 1

Timepoint 2



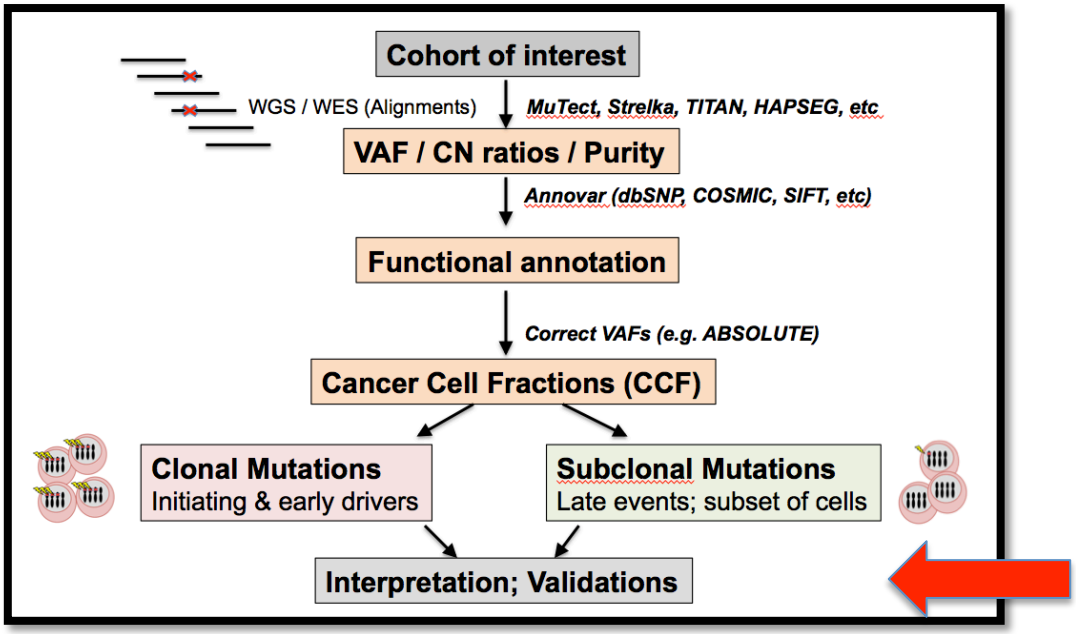
The presence of subclonal drivers adversely impacts patient outcome



Landau et al., 2013, Cell

Many examples in the literature highlight the clinical relevance of subclonal events in a multitude of cancers

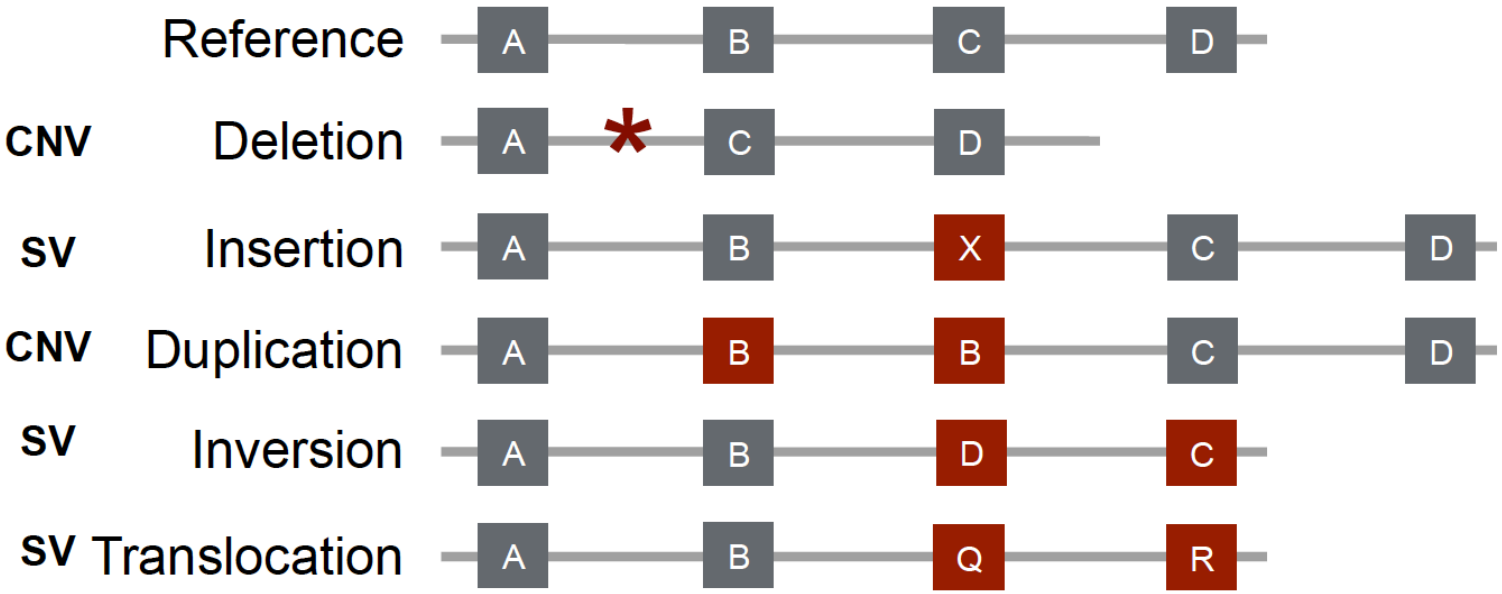
Interpretation of somatic mutations



- Recurrence / Significance analysis (drivers vs passengers)
- Patterns of clonal evolution
- Mutational mechanisms
- Association with clinical variables (subtype, survival, met)

Copy number and structural variants (CNVs, SVs)

Copy number and structural variants



SV is a superset of copy number variation (CNV). Not all structural changes affect copy number (e.g., inversions)!

A.Quinlan

Why are CNVs and SVs important?

- Common in the population (DGV). Humans differ by:
  - A few hundred inversions
  - A few hundred duplications
  - About 3,000 deletions (>=500bp)
  - Tens of retrotransposon insertions
- Affect a large fraction of the genome
- Genetic basis of traits
  - Gene dosage effects
  - Neuropsychiatric disease (e.g. autism, SZ)
- Major driver of genome evolution
  - Speciation driven by rapid changes in genome architecture
  - Genome instability and aneuploidy are hallmarks of solid tumor evolution

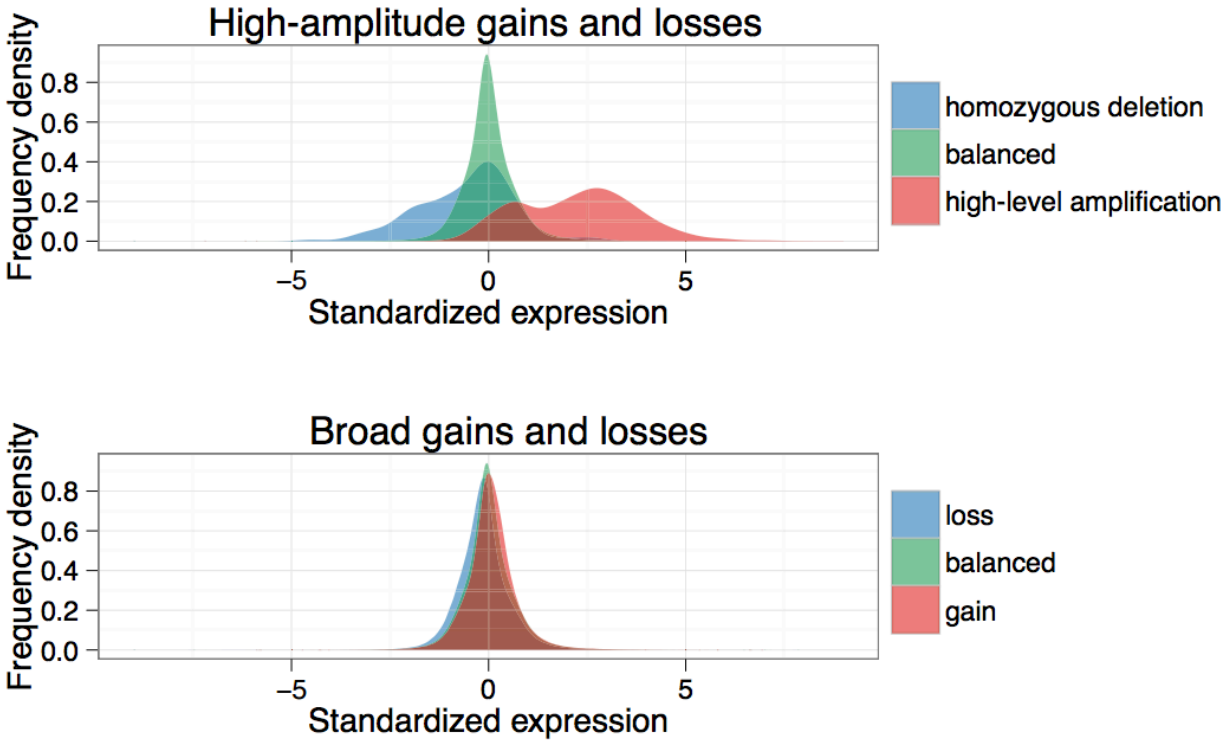
SVs and human phenotypes

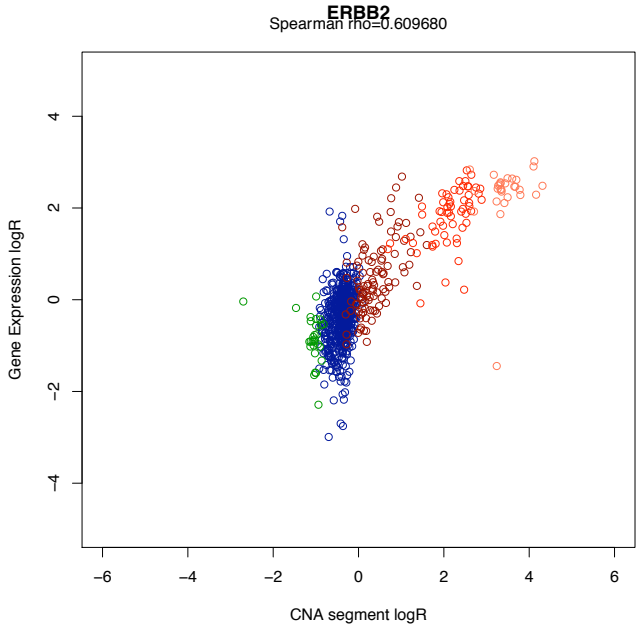
Table 2 Examples of copy number variations (CNVs) and conveyed genomic disorders[a]

Phenotype	OMIM	Locus	CNV
Mendelian (autosomal dominant)[b]			
Williams-Beuren syndrome	194050	7q11.23	del
7q11.23 duplication syndrome	609757	7q11.23	dup
Spinocerebellar ataxia type 20	608687	11q12	dup
Smith-Magenis syndrome	182290	17p11.2/*RAI1*	del
Potocki-Lupski syndrome	610883	17p11.2	dup
HNPP	162500	17p12/*PMP22*	del
CMT1A	118220	17p12/*PMP22*	dup
Miller-Dieker lissencephaly syndrome	247200	17p13.3/*LIS1*	del
Mental retardation	601545	17p13.3/*LIS1*	dup
DGS/VCFS	188400/192430	22q11.2/*TBX1*	del
Microduplication 22q11.2	608363	22q11.2	dup
Adult-onset leukodystrophy	169500	*LMNB1*	dup
Mendelian (autosomal recessive)			
Familial juvenile nephronophthisis	256100	2q13/*NPHP1*	del
Gaucher disease	230800	1q21/*GBA*	del
Pituitary dwarfism	262400	17q24/*GH1*	del
Spinal muscular atrophy	253300	5q13/*SMN1*	del
beta-thalassemia	141900	11p15/*beta-globin*	del
alpha-thalassemia	141750	16p13.3/*HBA*	del

A.Quinlan

Focal CNAs alter expression of the genes they harbour





17q11.2-q12

Shih et al., 2012, Nature

Actionable gene-based copy number alterations

Table 1. Categories of Genomic Alteration and Exemplary Cancer Genes

Category of Genomic Alteration	Exemplary Cancer Gene	Type of Cancer	Targeted Therapeutic Agent
Translocation	*BCR-ABL*	Chronic myelogenous leukemia	Imatinib
	PML-RARα	Acute promyelocytic leukemia	All-*trans*-retinoic acid
	EML4-ALK	Breast, colorectal, lung	ALK inhibitor
	ETS gene fusions	Prostate	—
	Other	Leukemias, lymphomas, sarcomas	—
Amplification	*EGFR*	Lung, colorectal, glioblastoma, pancreatic	Cetuximab, gefitinib, erlotinib, panitumumab, lapatinib
	ERBB2	Breast, ovarian	Trastuzumab, lapatinib
	KIT, PDGFR	GISTs, glioma, HCC, RCC, CML	Imatinib, nilotinib, sunitinib, sorafenib
	MYC	Brain, colon, leukemia, lung	—
	SRC	Sarcoma, CML, ALL	Dasatinib
	PIK3CA	Breast, ovarian, colorectal, endometrial	PI3-kinase inhibitors
Point mutation	*EGFR*	Lung, glioblastoma	Cetuximab, gefitinib, erlotinib, panitumumab, lapatinib
	KIT, PDGFR	GISTs, glioma, HCC, RCC, CML	Imatinib, nilotinib, sunitinib, sorafenib
	PIK3CA	Breast, ovarian, colorectal, endometrial	PI3-kinase inhibitors
	BRAF	Melanoma, pediatric astrocytoma	RAF inhibitor
	KRAS	Colorectal, pancreatic, GI tract, lung	Resistance to erlotinib, cetuximab (colorectal)

Abbreviations: ALK, anaplastic lymphoma kinase; GIST, GI stromal tumor; HCC, hepatocellular carcinoma; RCC, renal cell carcinoma; CML, chronic myelogenous leukemia; ALL, acute lymphoblastic leukemia; PI3, phosphatidylinositol-3.

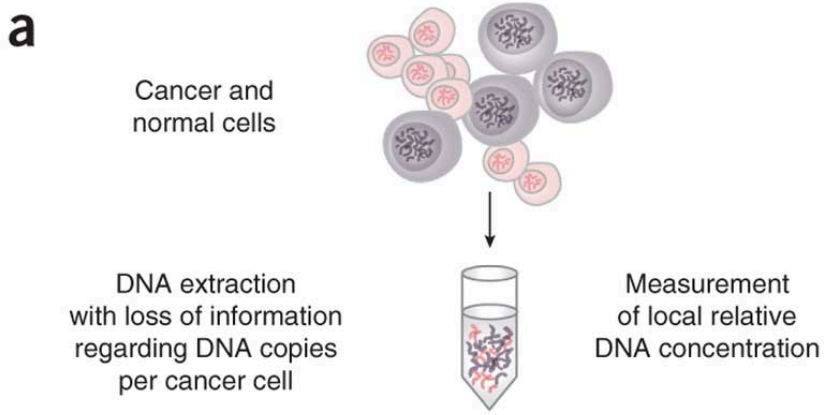
MacConaill and Garraway. Jour Clin Oncology (2010)

Analysis strategy depends on key factors:

samples: germline vs somatic
data: read length
strategy: alignment vs assembly

Tools: ascatNGS, ABSOLUTE, Control-Freec, Delly, BRASS, ACE-seq, dRanger

Purity and ploidy



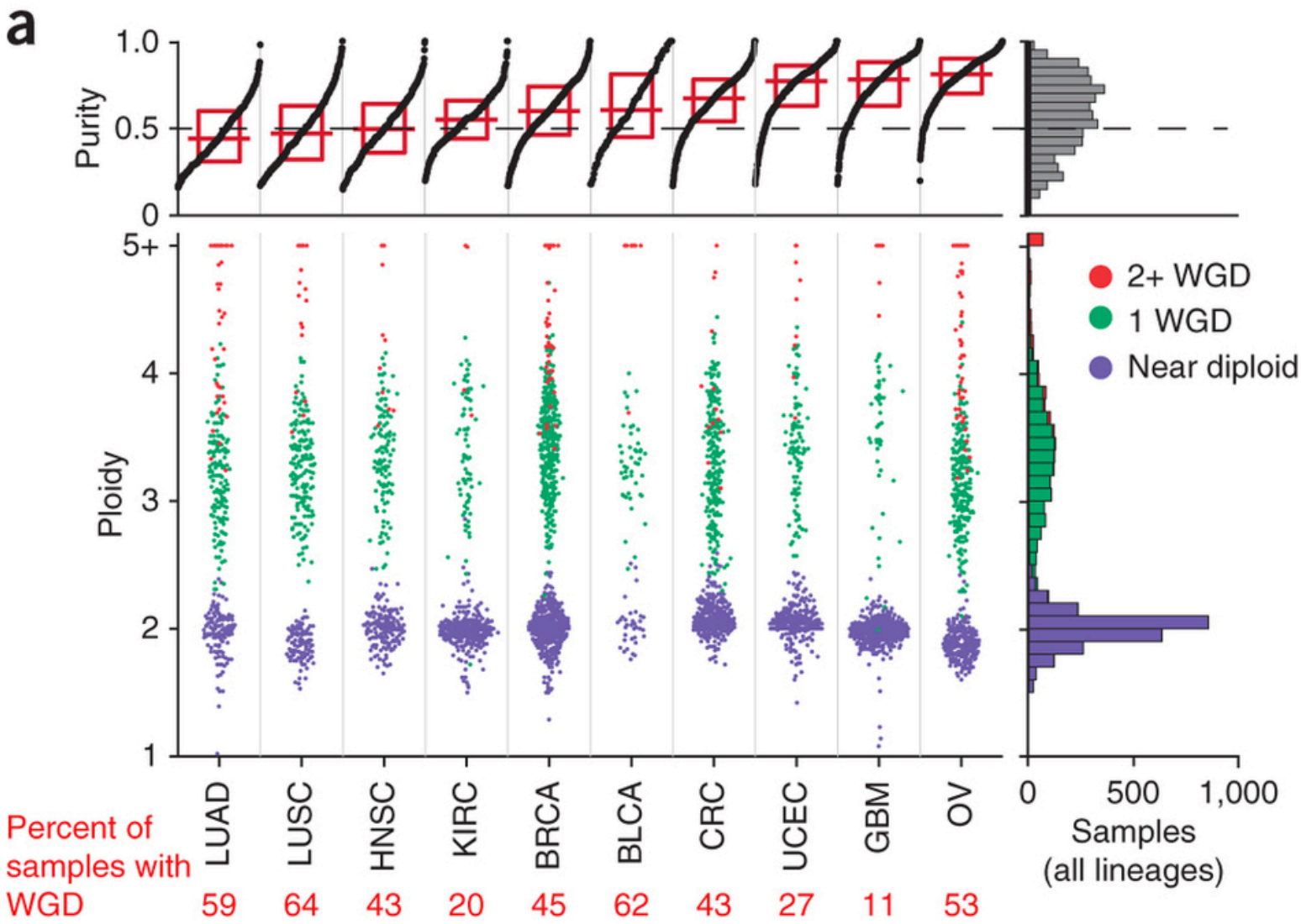
(1) Purity: fraction of all cancerous cells within a heterogeneous tumor sample

(2) Ploidy: baseline copy number of genomic segments or entire chromosomes

"Identifiability problem"

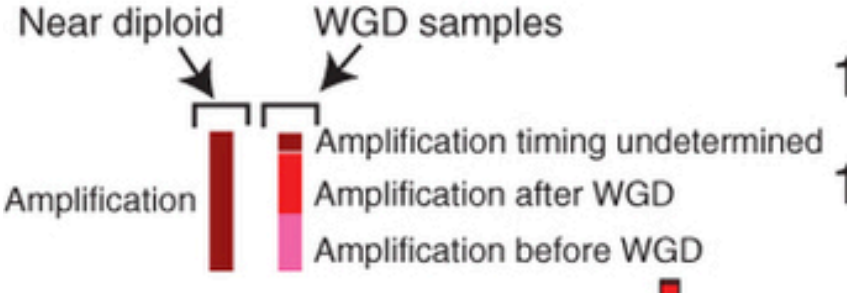
- *Equivalent Copy Number*:
  - Homozygous deletion combined with 30% tumor purity
    - Tumor (0*0.3) + Normal (2 * 0.6) = 1.2
  - Heterozygous deletion combined with 60% tumor purity
    - Tumor (1*0.6) + Normal (2 * 0.3) = 1.2
- *Equivalent Copy Number and BAF*:
  - One copy gain in a diploid tumor: AB -> AAB
  - One copy loss in a tetraploid tumor: AABB -> AAB

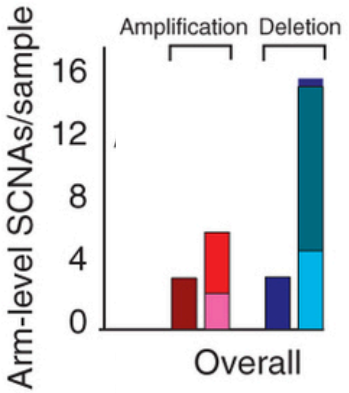
Pan-cancer: purity and ploidy estimates



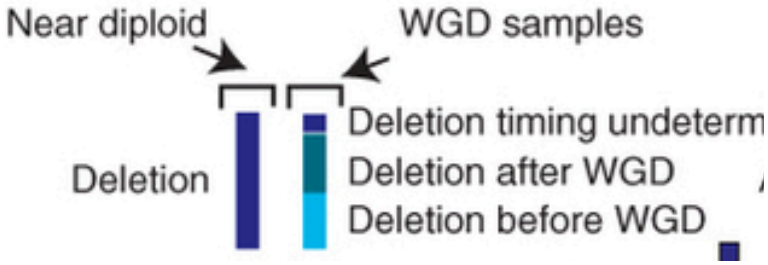
Zack et al., 2013, Nature Genetics

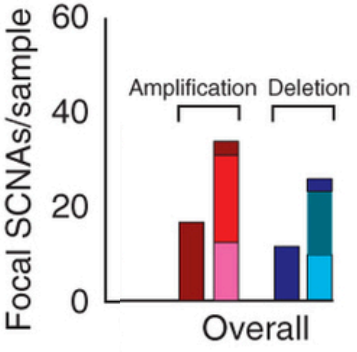
Genome doubling (GD) is an early event in genomic instability





Broad events



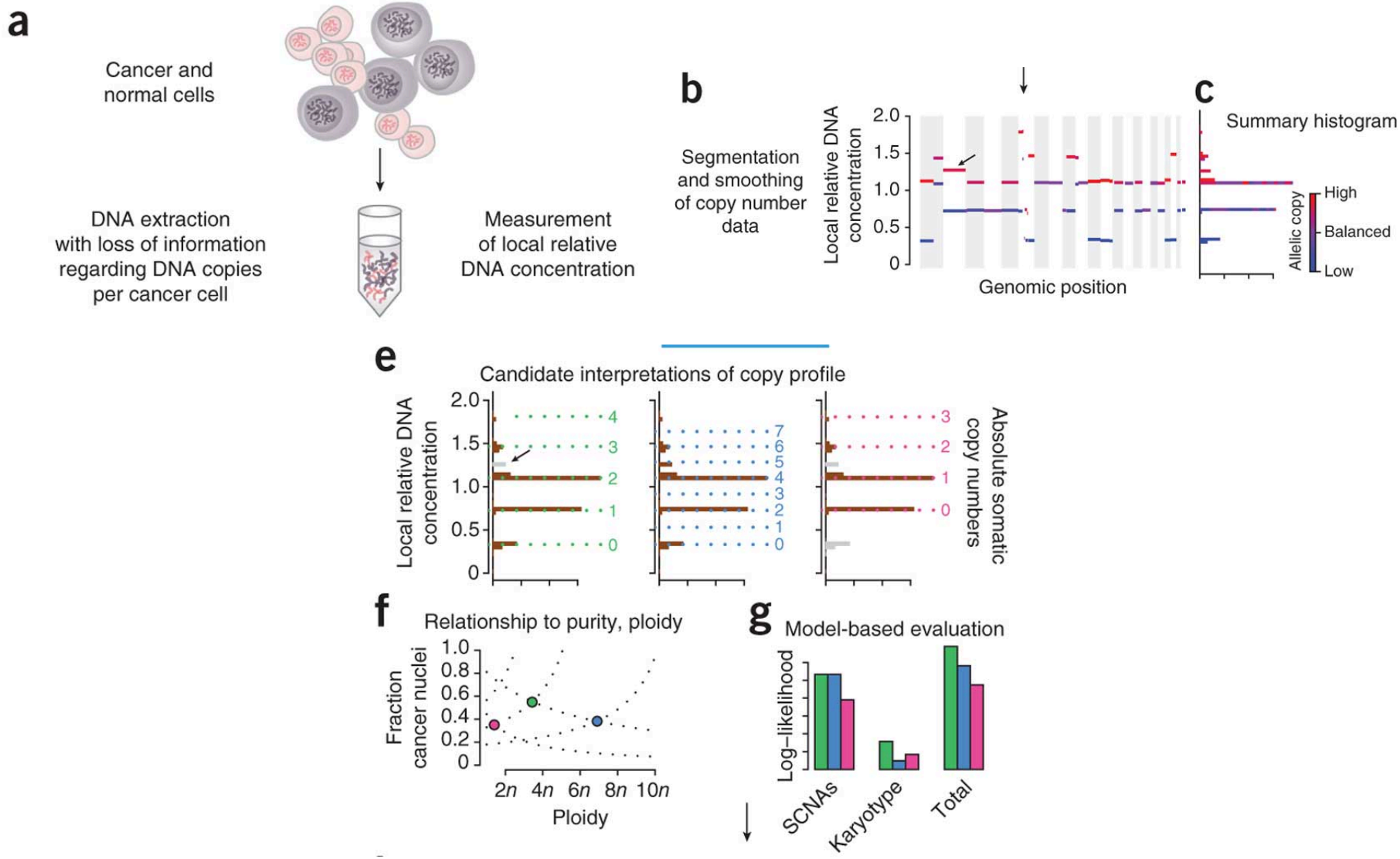


Focal events

More CNAs occur after GD than before
After GD, deletions outnumber gains

Zack et al., 2013, Nature Genetics

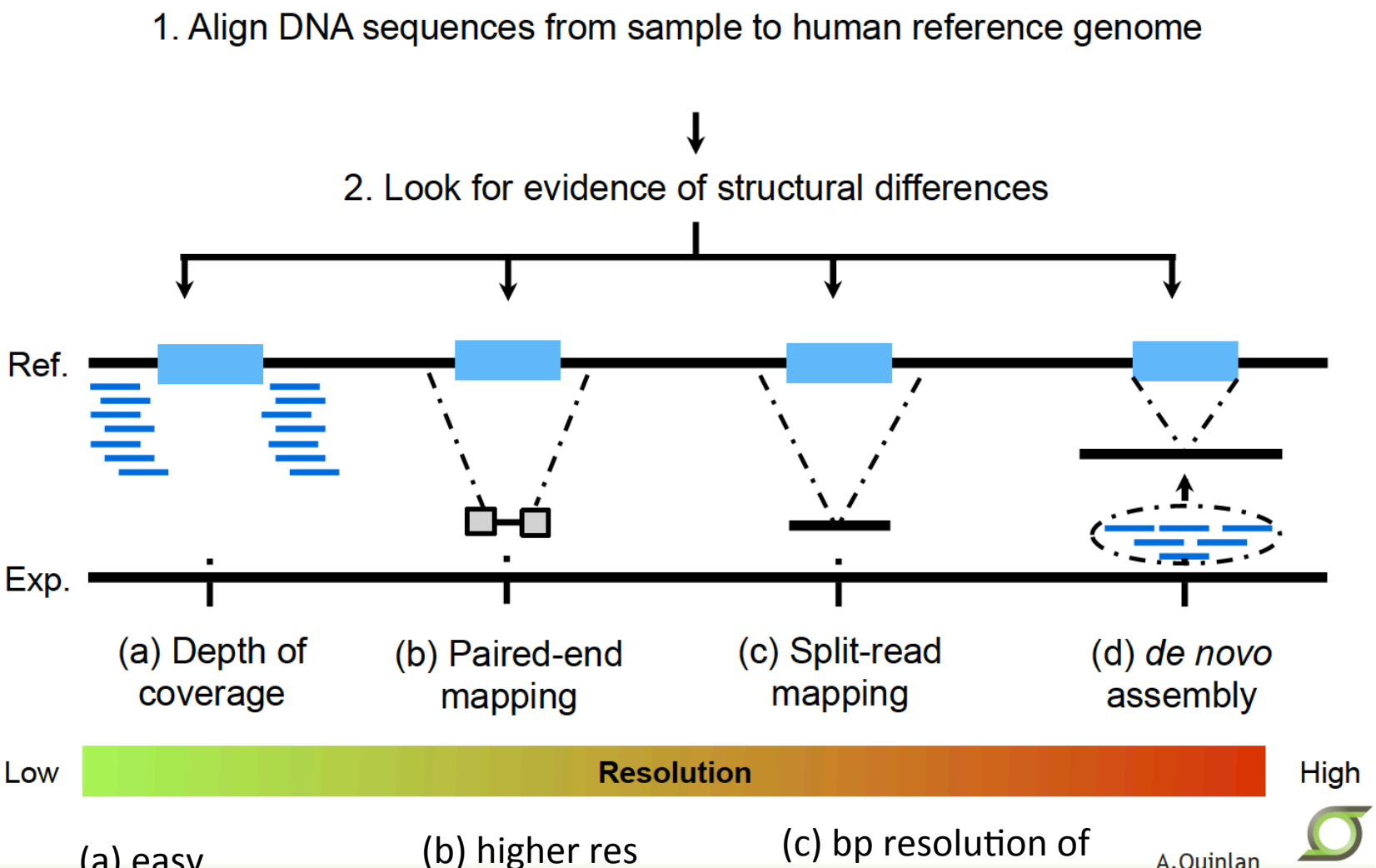
Purity and ploidy



ABSOLUTE: *Carter et al., 2012, Nature Biotechnology*

Signals of structural variation

Sequence alignment "signals" for structural variation



(a) easy amp/del 2-5kb res

(b) higher res no bp cov FP chimeric molecules

(c) bp resolution of breakpoint Misalignments in low-complexity regions

A.Quinlan

Viewing Structural Events in IGV

Insert size

Pair orientation

BROAD INSTITUTE

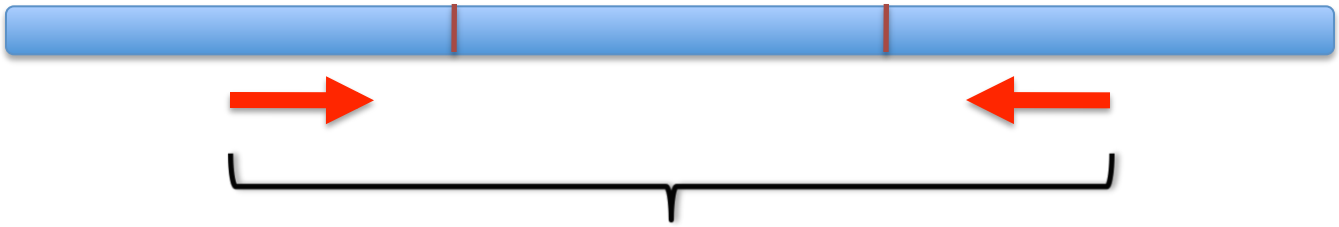
Integrative Genomics Viewer

Deletion

Integrative Genomics Viewer

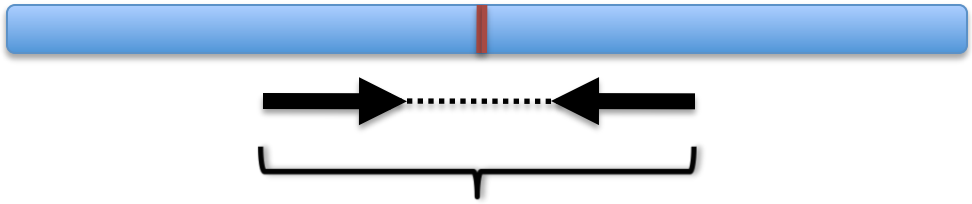
Inferred insert size is > expected value

Reference Genome



inferred insert size

Subject



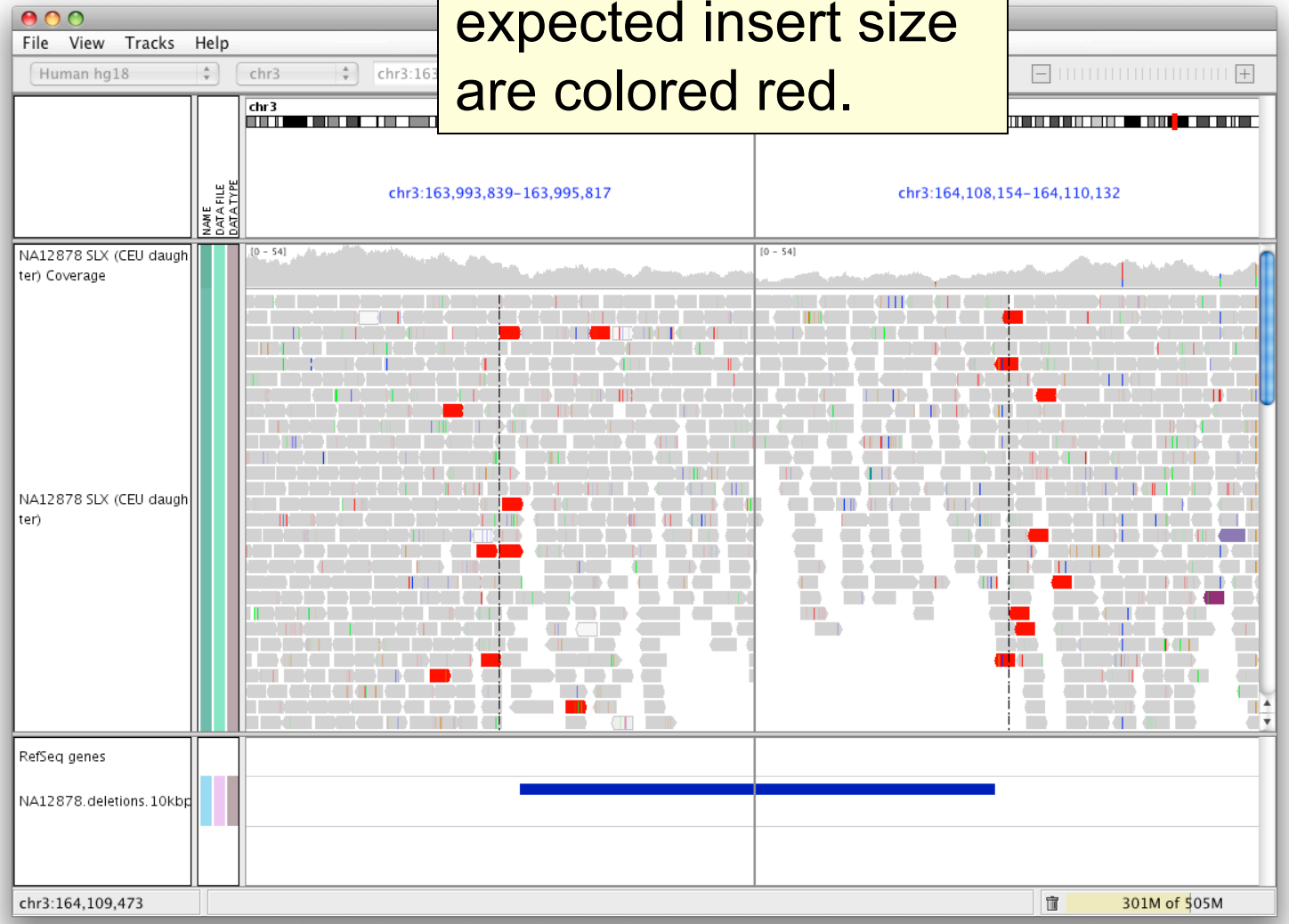
expected insert size

BROAD INSTITUTE

Deletion

Integrative Genomics Viewer

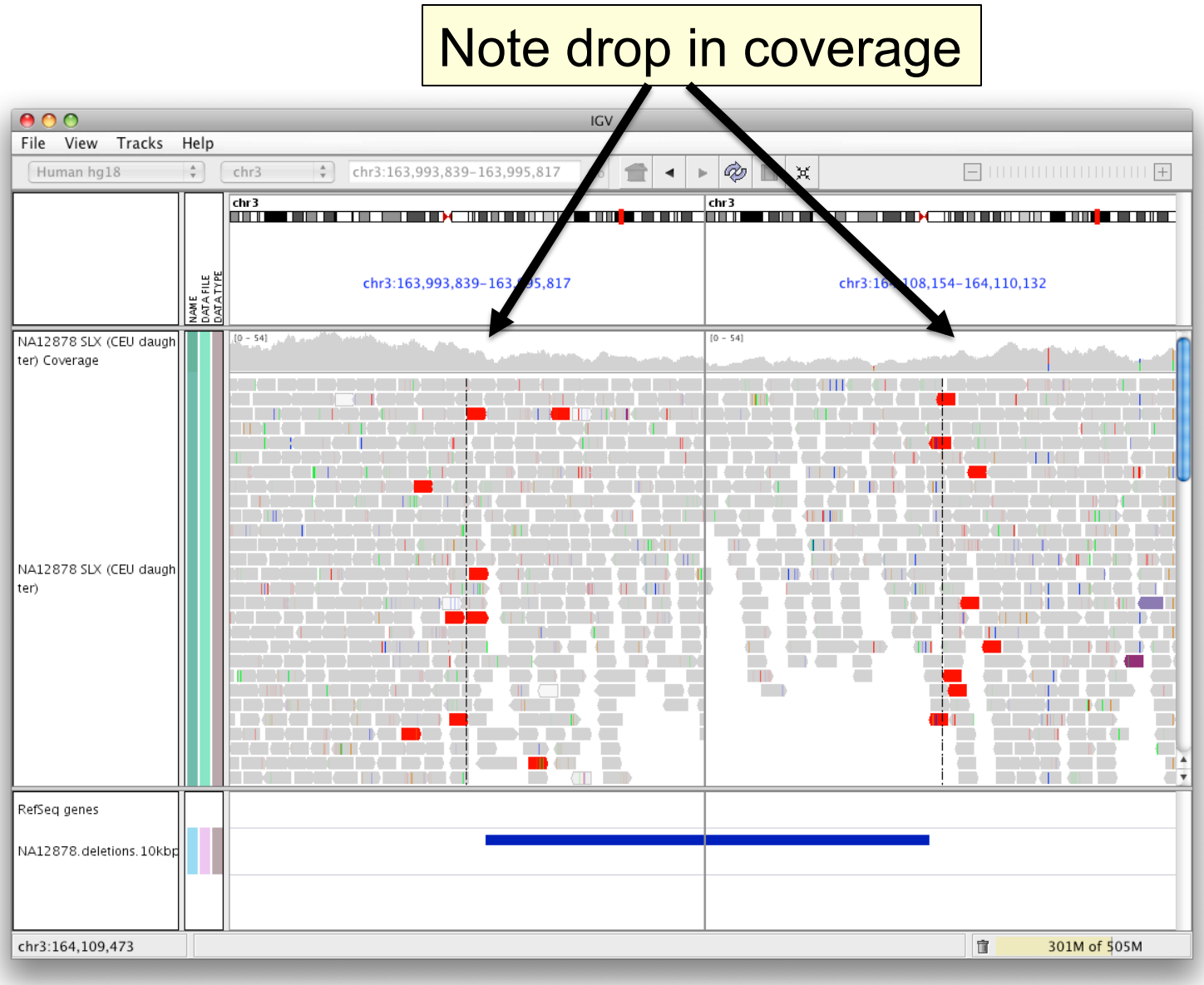
Pairs with larger than expected insert size are colored red.



BROAD INSTITUTE

Deletion

Integrative Genomics Viewer



BROAD INSTITUTE

Insert size color scheme

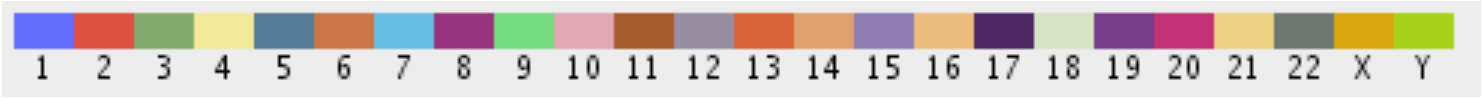
Integrative Genomics Viewer

- Smaller than expected insert size:

- Larger than expected insert size:

- Pairs on different chromosomes

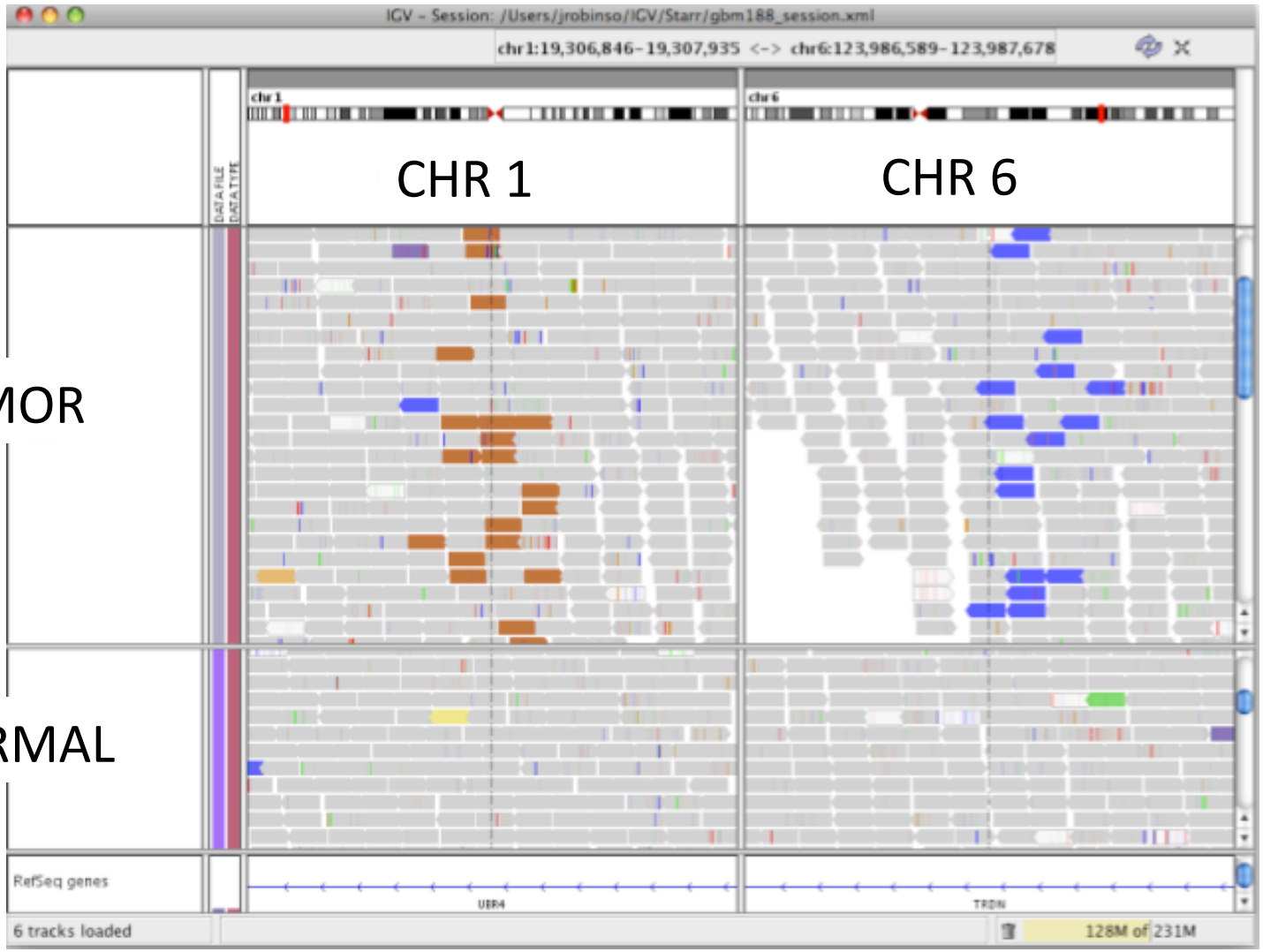
Each end colored by chromosome of its mate



BROAD INSTITUTE

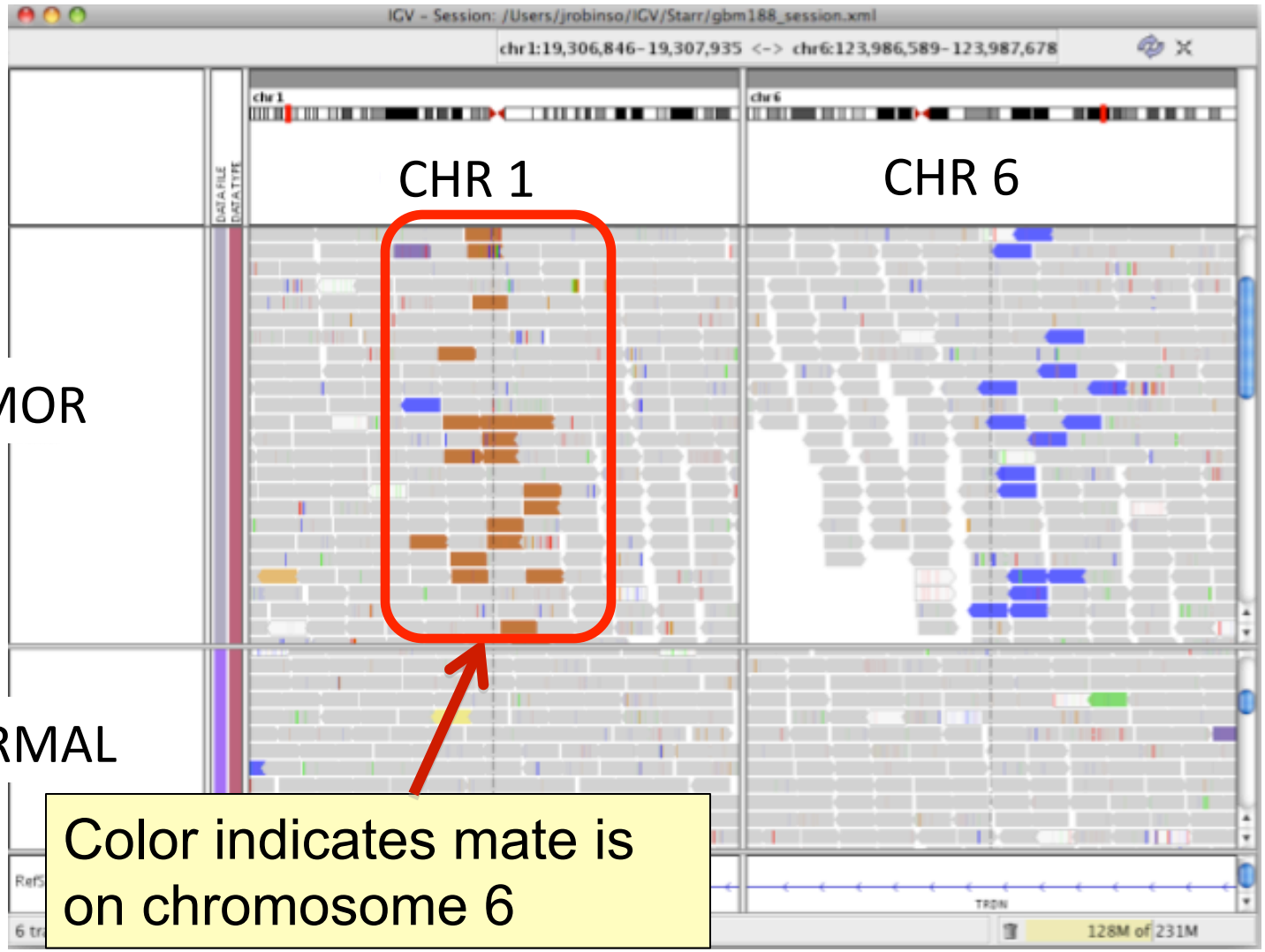
Rearrangement

Integrative Genomics Viewer



BROAD INSTITUTE

Rearrangement

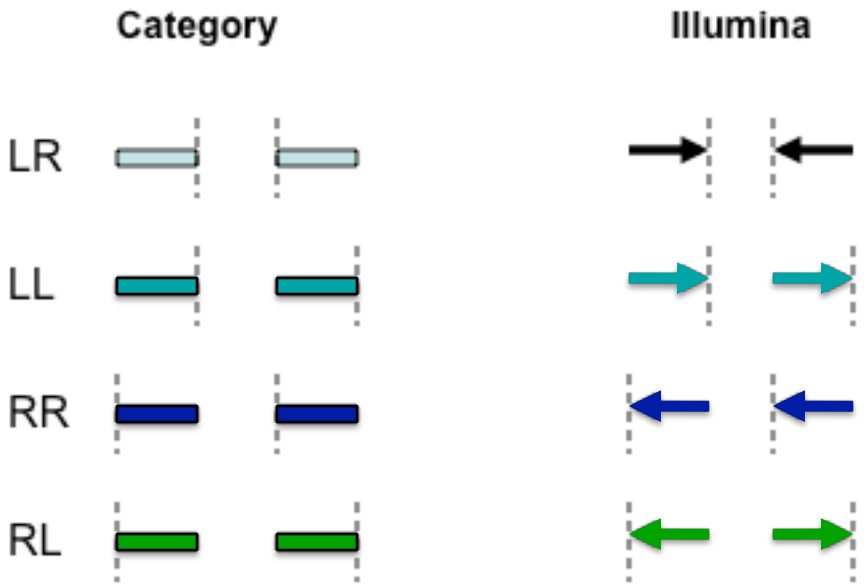


Interpreting Pair Orientations

BROAD INSTITUTE

Integrative Genomics Viewer

Interpretation of read pair orientations



LR	Normal reads. The reads are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome.
LL,RR	Implies inversion in sequenced DNA with respect to reference.
RL	Implies duplication or translocation with respect to reference.

These categories only apply to reads where both mates map to the same chromosome.

Figure courtesy of Bob Handsaker

Inversion

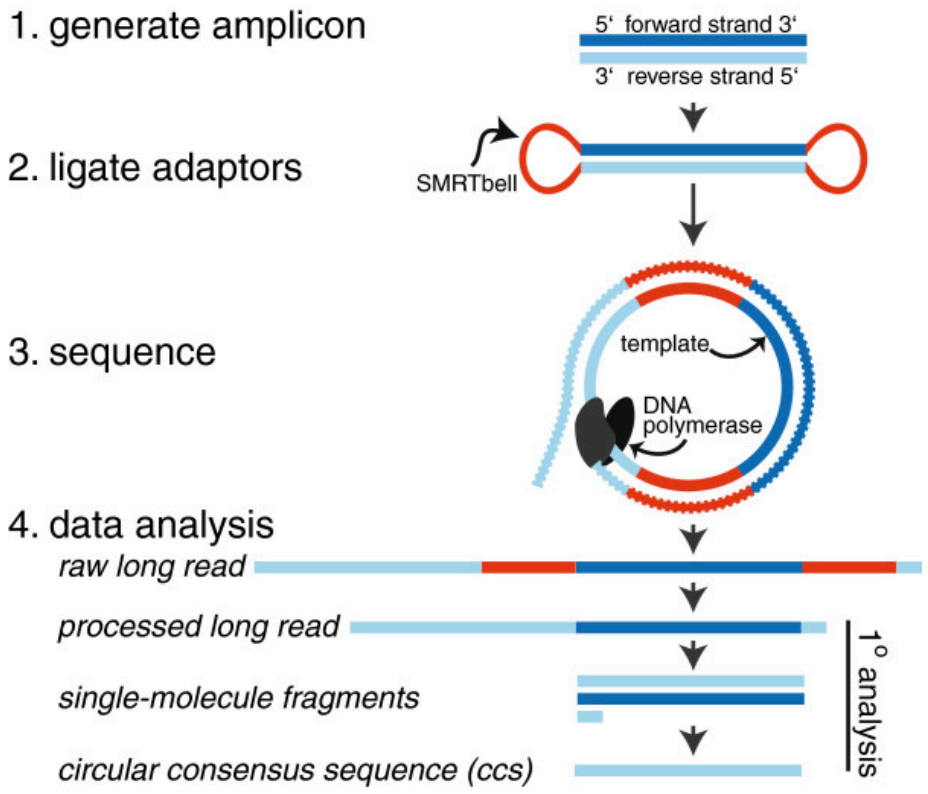
Integrative Genomics Viewer

BROAD INSTITUTE

Challenges and Solutions

- Many false positives
  - Short reads, heuristic alignment, repetitive genome
  - PCR artifacts (amplification)
  - Ref genome errors (gaps, mis-assemblies)
  - Need stringent filtering!
- Many false negatives (up to 30%)
  - Low/moderate physical coverage due to small insert size
  - Breakpoints enriched in repetitive DNA
  - Filtering out of true positives
  - *FNs in reference data (DGV, SNP arrays, short reads)
- Solutions
  - Long read sequencing (**PacBio**, Oxford Nanopore)
  - **10X Genomics linked-reads**

PacBio



Physical and sequence coverage over long ranges spanning multiple breakpoints and repetitive elements

Germline vs somatic ?

10X Genomics Linked Reads

Genotoul GeT

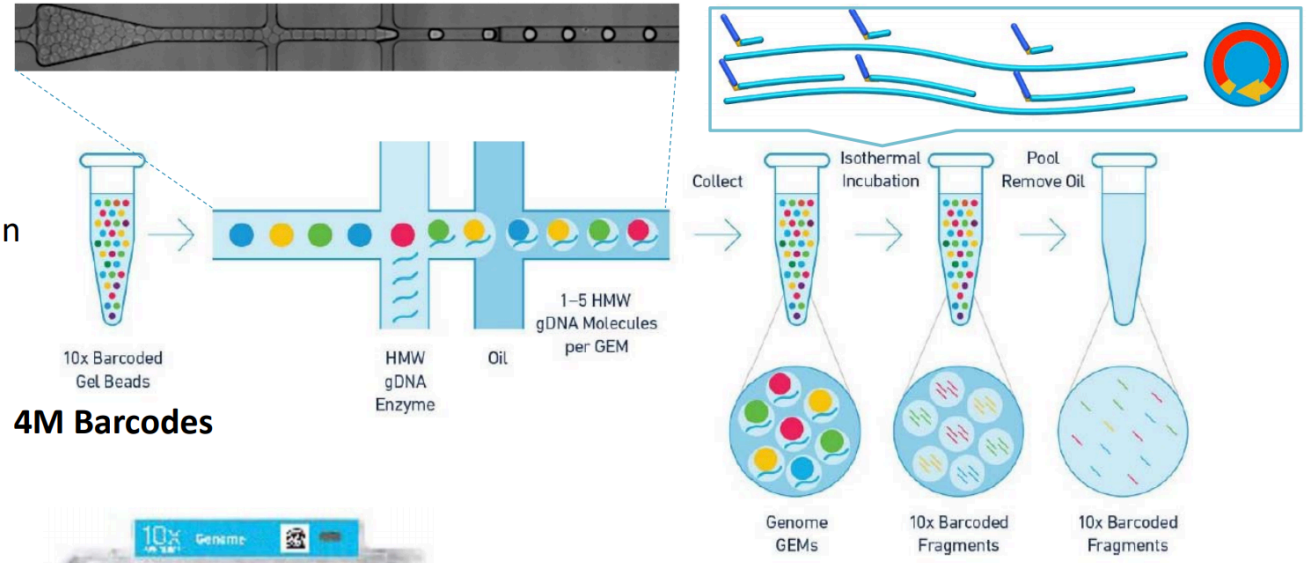
Chromium 10XGENOMICS

10X GENOMICS

How does it work ?

Complete wetlab workflow: 2 days

GEM (Gel Bead in Emulsion)

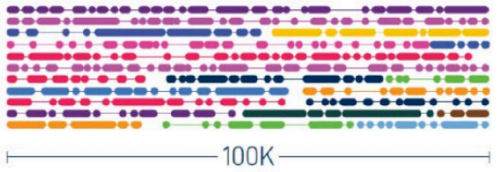


4M Barcodes

Input: ~1 ng HMW gDNA

Linked-Reads

Sequencing on HiSeq3000 2x150pb



http://get.genotoul.fr

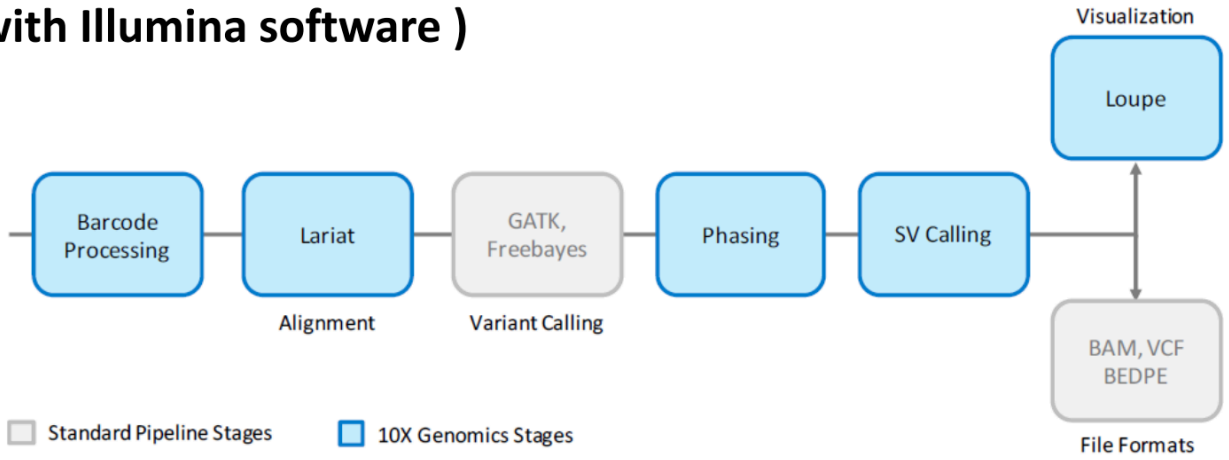
10X Genomics Linked reads

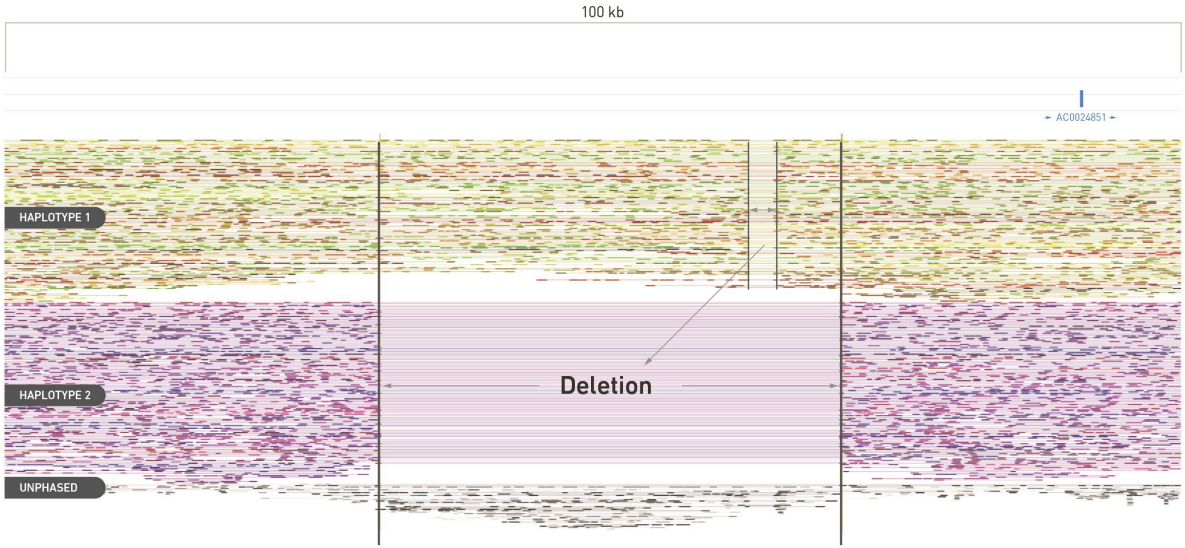
Genotoul GeT

LongRanger software

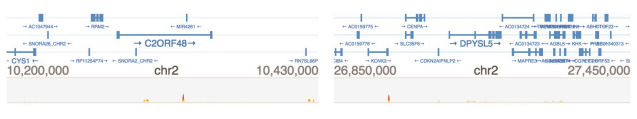
Analysis pipelines that perform :

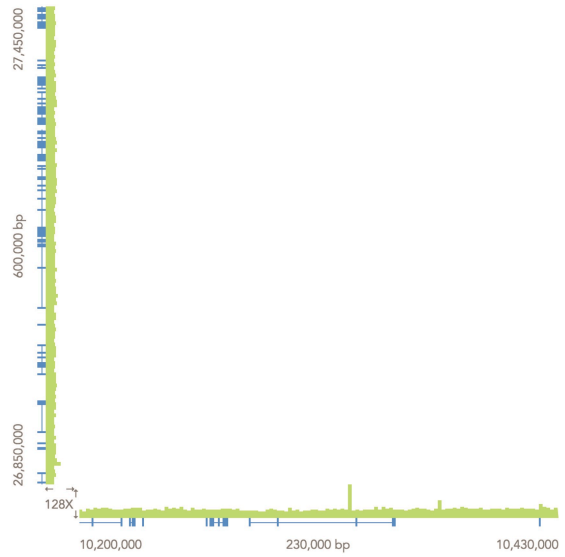
- **Sample demultiplexing (with Illumina software)**
- **Barcode processing**
- **Alignment (Lariat)**
- **Quality control**
- **Variant calling**
- **Phasing**
- **Structural variant calling**



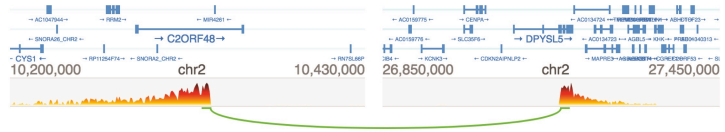


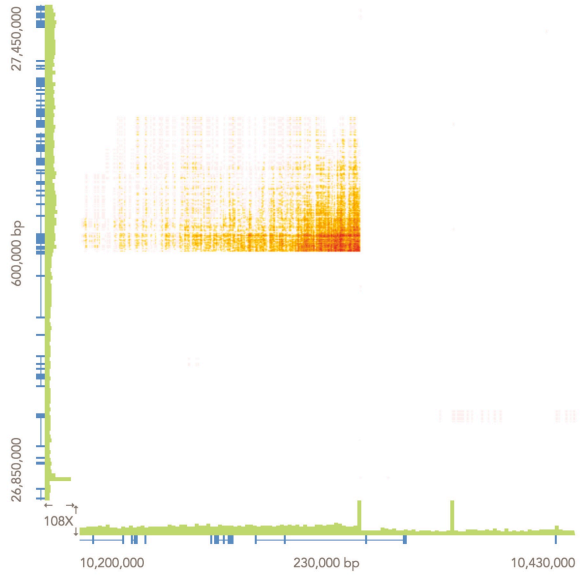
Sample with no SV





Sample with a large deletion on Chr 2



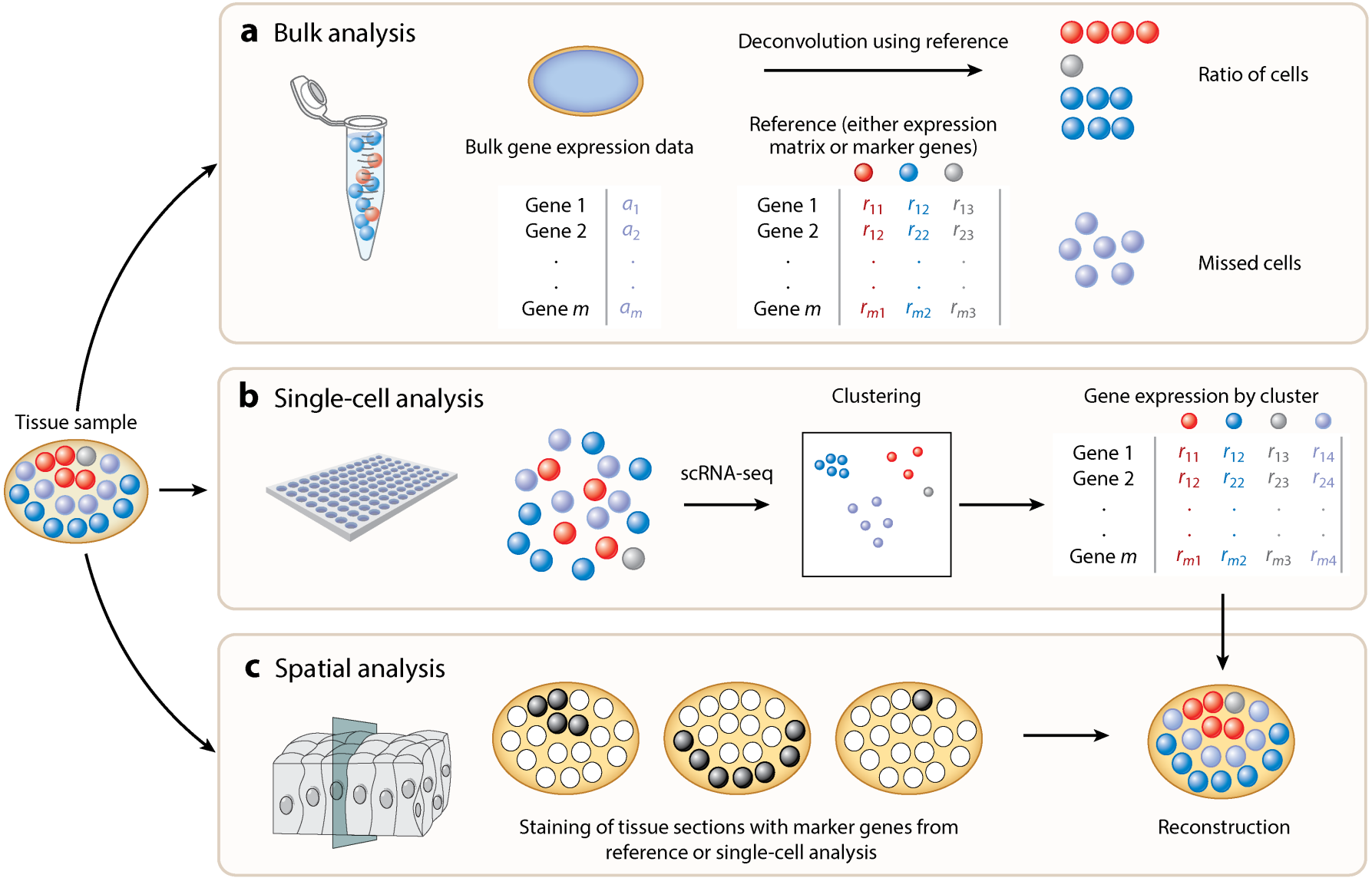


Long read technologies will improve SV detection

Need for population controls (DGV-style) in order to interpret pathogenic variants appropriately

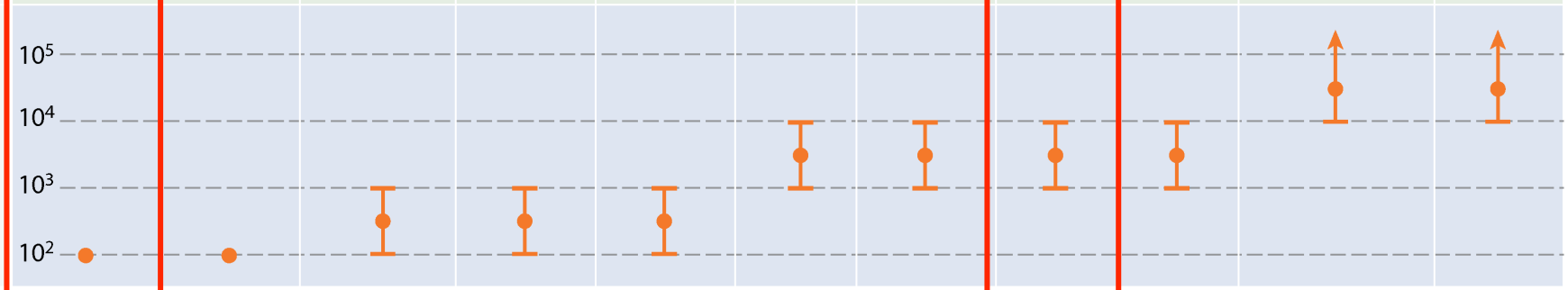
Single cell sequencing

scRNAseq



Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

scRNAseq

	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3'	5'	3'	3'	3'	3'	3'	3'	3'	3'
Number of cells per assay											

Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

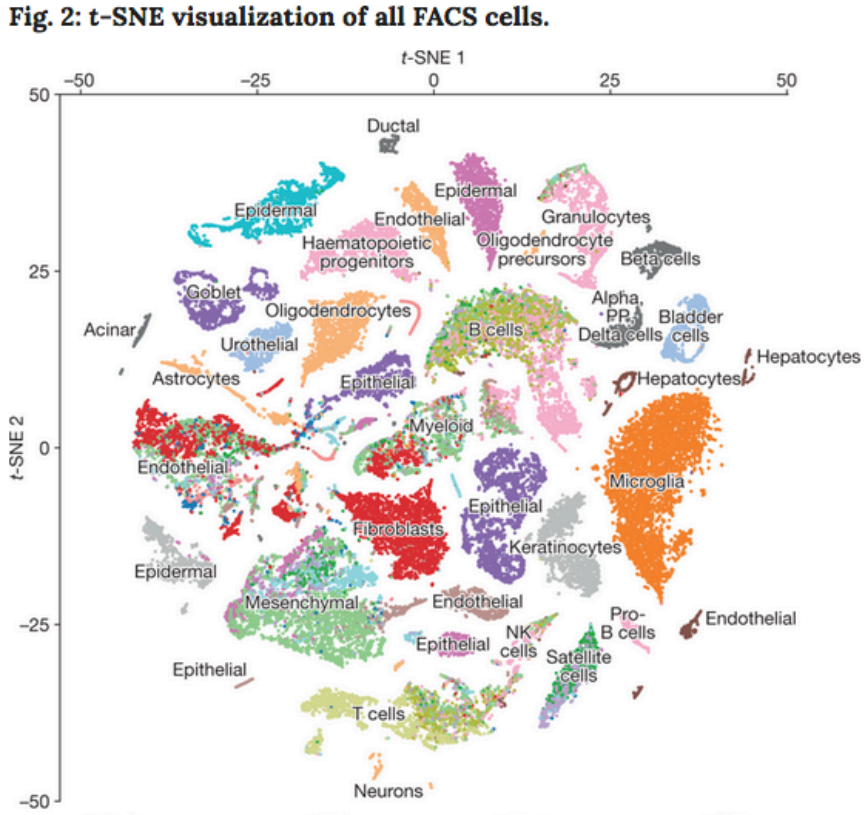
scRNAseq – experimental choices (Tabula Muris)

- **SMART-seq2:**
  - generates a "typical" high-resolution RNAseq library from each cell
    - Splicing, fusions, mutations,
    - low expression genes (TFs), co-regulated genes
  - low throughput, high capture rate
  - standard transcriptome analysis methods

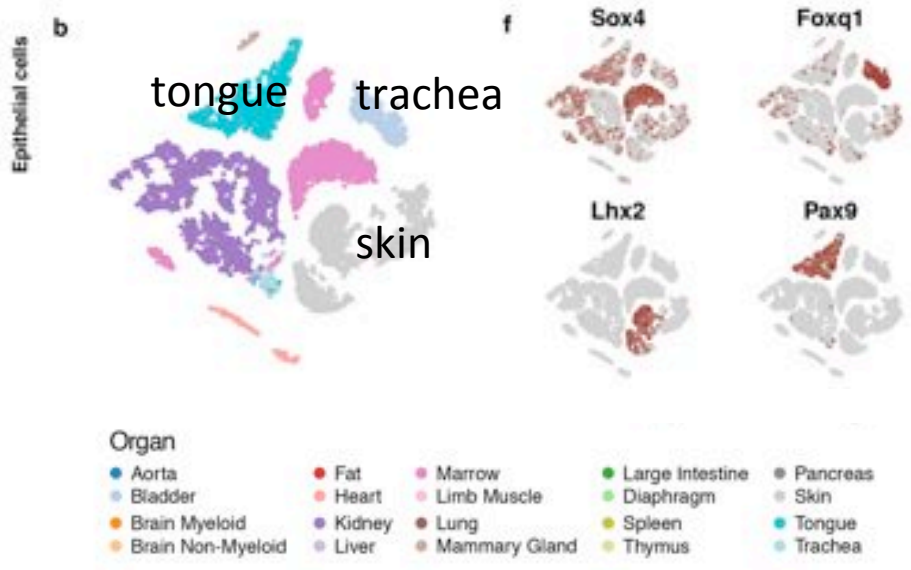
- **10X Genomics Chromium scRNAseq:**
  - approximation of a transcriptome (few genes, 3' tag)
  - high throughput, lower capture rate
  - identification of rare cell types
  - "dropout" of expression (impute missing data)
  - novel analytic approaches evolving

Tabula Muris:

20 organs from four male and three female mice



Cell types can be cleanly clustered using TFs
Cell identities can be defined though their underlying regulatory networks



General analytic pipelines

Loupe (10XG), Seurat, Monocle2/3, (many others)

- **QC to remove low quality cells**
  - poor viability, low efficiency cDNA production, etc
  - Relative library size, nr detected genes, spike in RNA, doublets
- **Normalization and Dimensionality reduction**
  - Normalization (diff lib sizes and mRNA captured)
  - PCA to reduce dimensions (1000's -> 10-100)
  - tSNE embedding for 2D visualization of cells based on PCA
- **Clustering into subpopulations**
  - biologically meaningful trends in the data (functional similarity, developmental relationships)
- ***Cluster cell-type annotation***
- **Infer trajectory of cells**
  - linear differentiation, multipronged fate decisions
- **Identify differentially expressed genes**
  - across clusters / trajectories

10X Genomics: cell_ranger to Loupe Cell Browser

Estimated Number of Cells

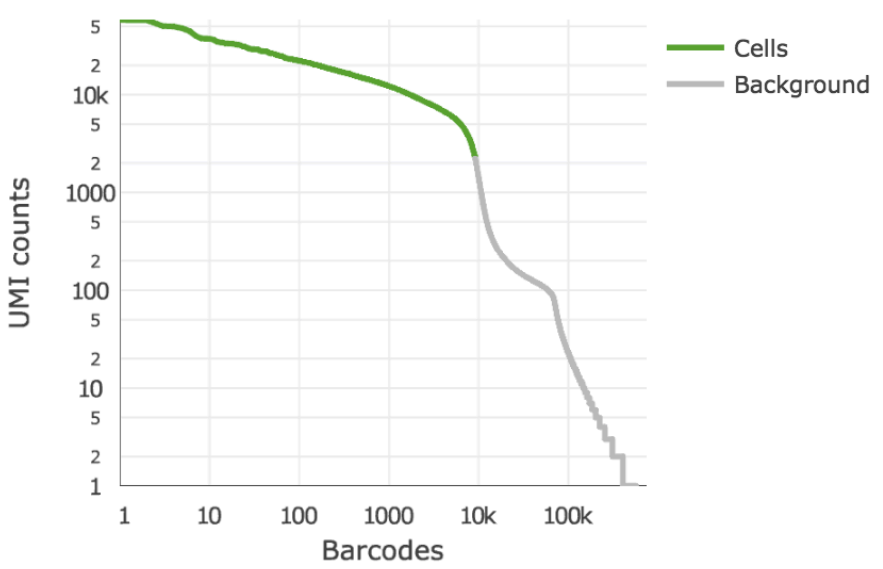
9,128

Mean Reads per Cell	Median Genes per Cell
41,998	**2,508**

Sequencing

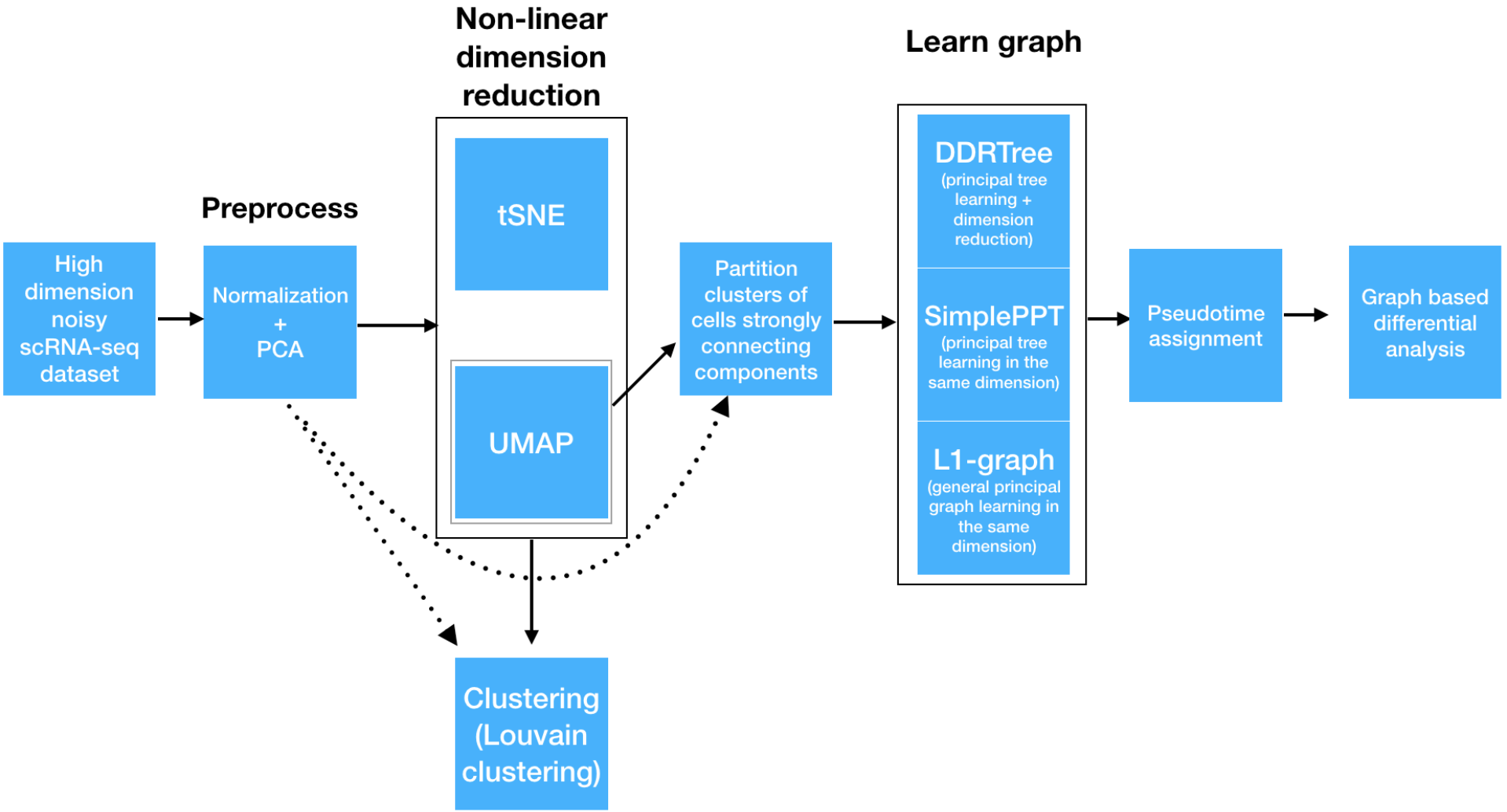
Number of Reads	383,366,284
Valid Barcodes	98.3%
Sequencing Saturation	65.2%
Q30 Bases in Barcode	98.3%
Q30 Bases in RNA Read	81.6%
Q30 Bases in Sample Index	95.8%
Q30 Bases in UMI	98.2%

Cells



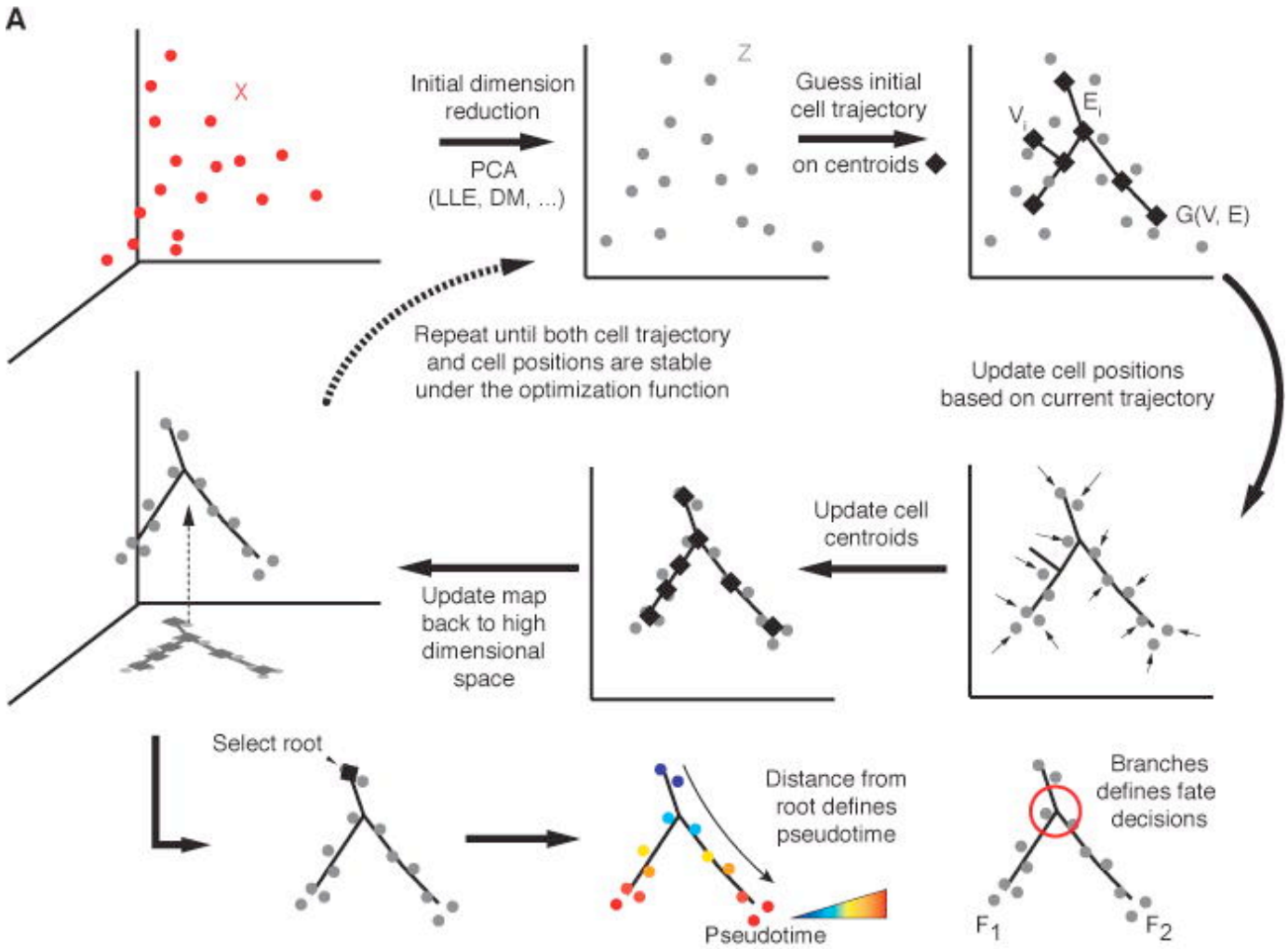
Estimated Number of Cells	9,128
Fraction Reads in Cells	83.2%
Mean Reads per Cell	41,998
Median Genes per Cell	2,508
Total Genes Detected	19,205
Median UMI Counts per Cell	6,360

Monocle3


High dimension noisy scRNA-seq dataset
Preprocess
Normalization + PCA
Non-linear dimension reduction
tSNE
UMAP
Clustering (Louvain clustering)
Partition clusters of cells strongly connecting components
Learn graph
DDRTree
(principal tree learning + dimension reduction)
SimplePPT
(principal tree learning in the same dimension)
L1-graph
(general principal graph learning in the same dimension)
Pseudotime assignment
Graph based differential analysis

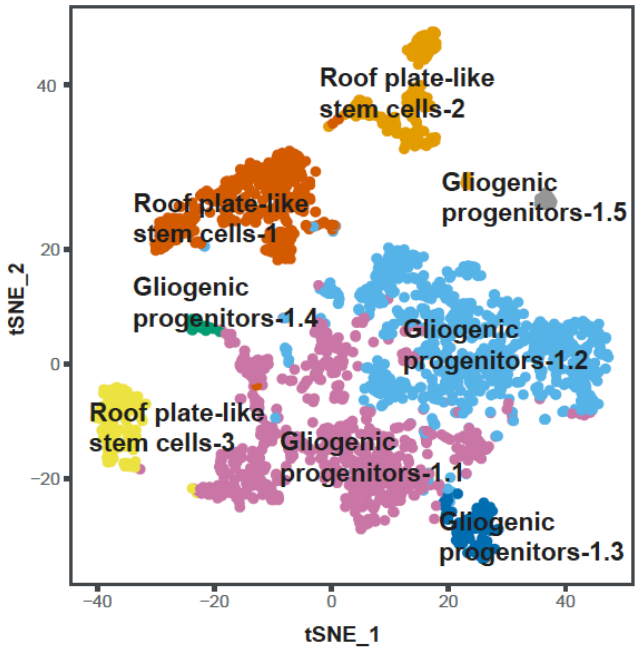
Pseudotime

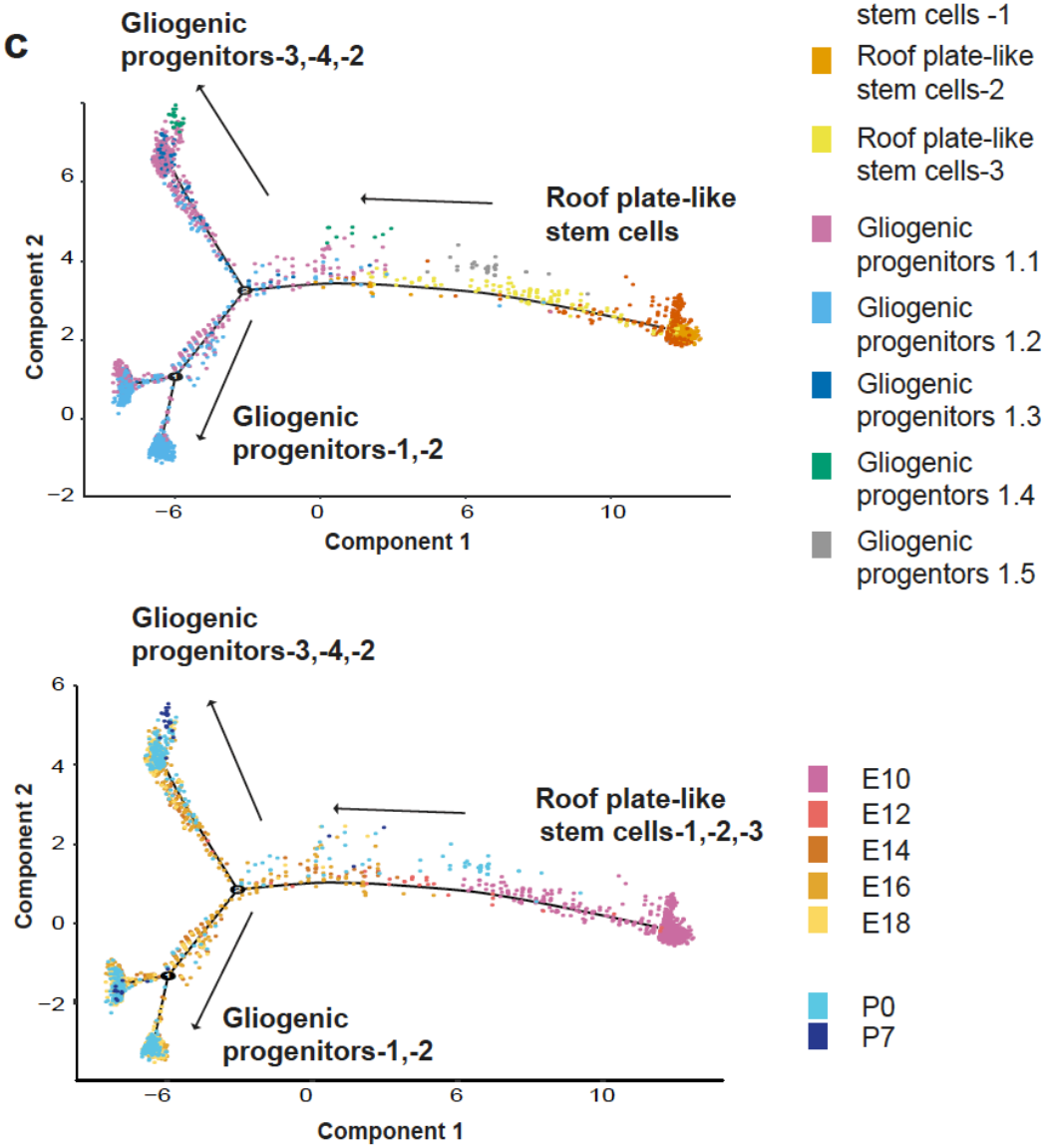
Pseudotime: an abstract unit / measure of progress that an individual cell has made through a biological process such as cell differentiation. (assuming somewhat smooth transitions).



Single cell RNAseq of mouse cerebellum

Subset of cell clusters





Vladoiu et al., 2018, *bioRxiv*

Medulloblastoma Cell of Origin

Mouse CB: 60,000 single cells | 9 timepoints

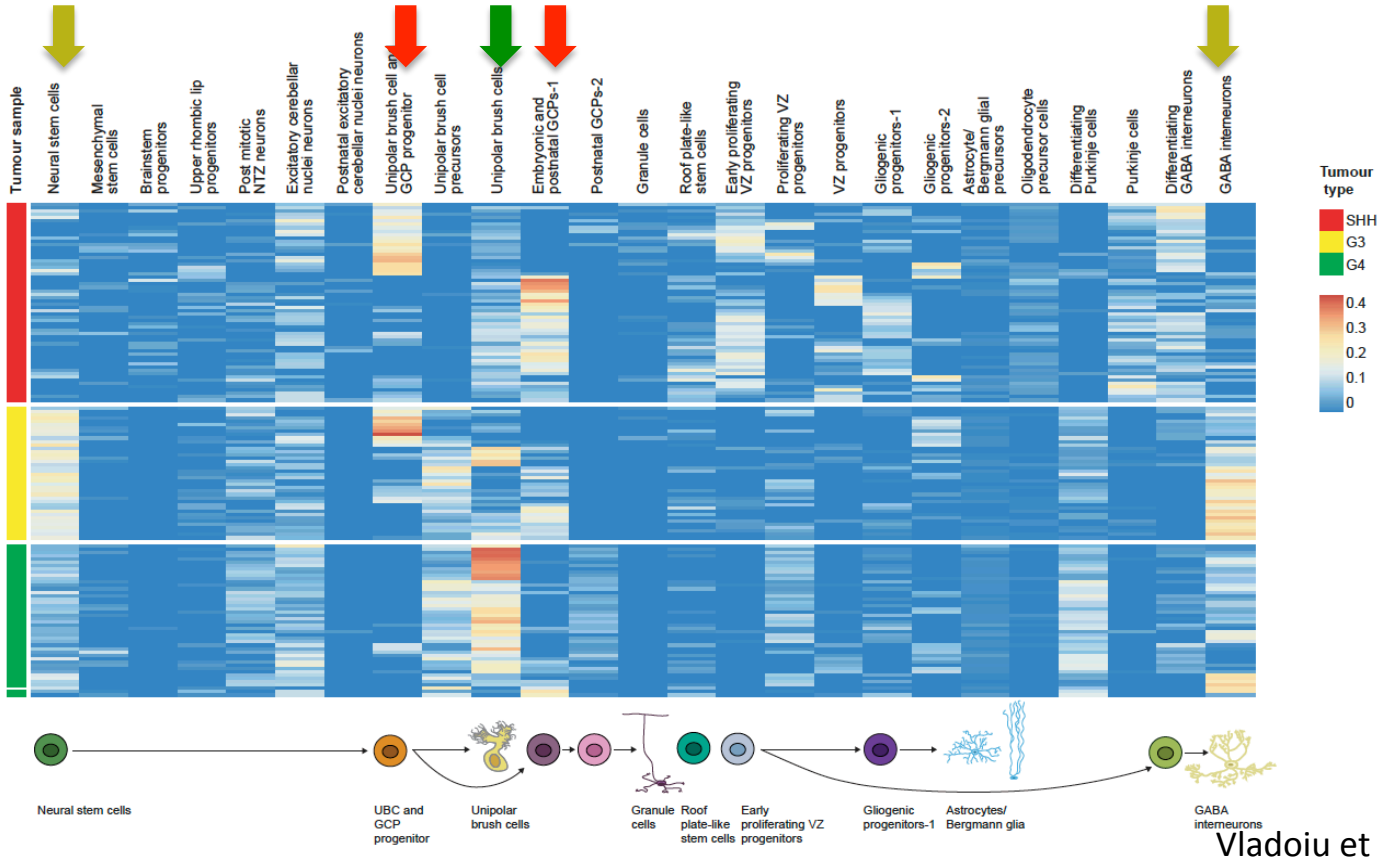
-> distinct, temporally restricted cell lineages

Human tumors (n=145):

SHH: granule cell hierarchy (pre vs postnatal origin ~ prognosis)

Group3: Nestin +ve stem cells (early neural stem-like cells & multi-lineage differentiation)

Group4: unipolar brush cells (E14-E17, gone by P0)



Vladoiu et al., bioRxiv, 2018

We are on a break