Structural Variants and Copy-Number Variants

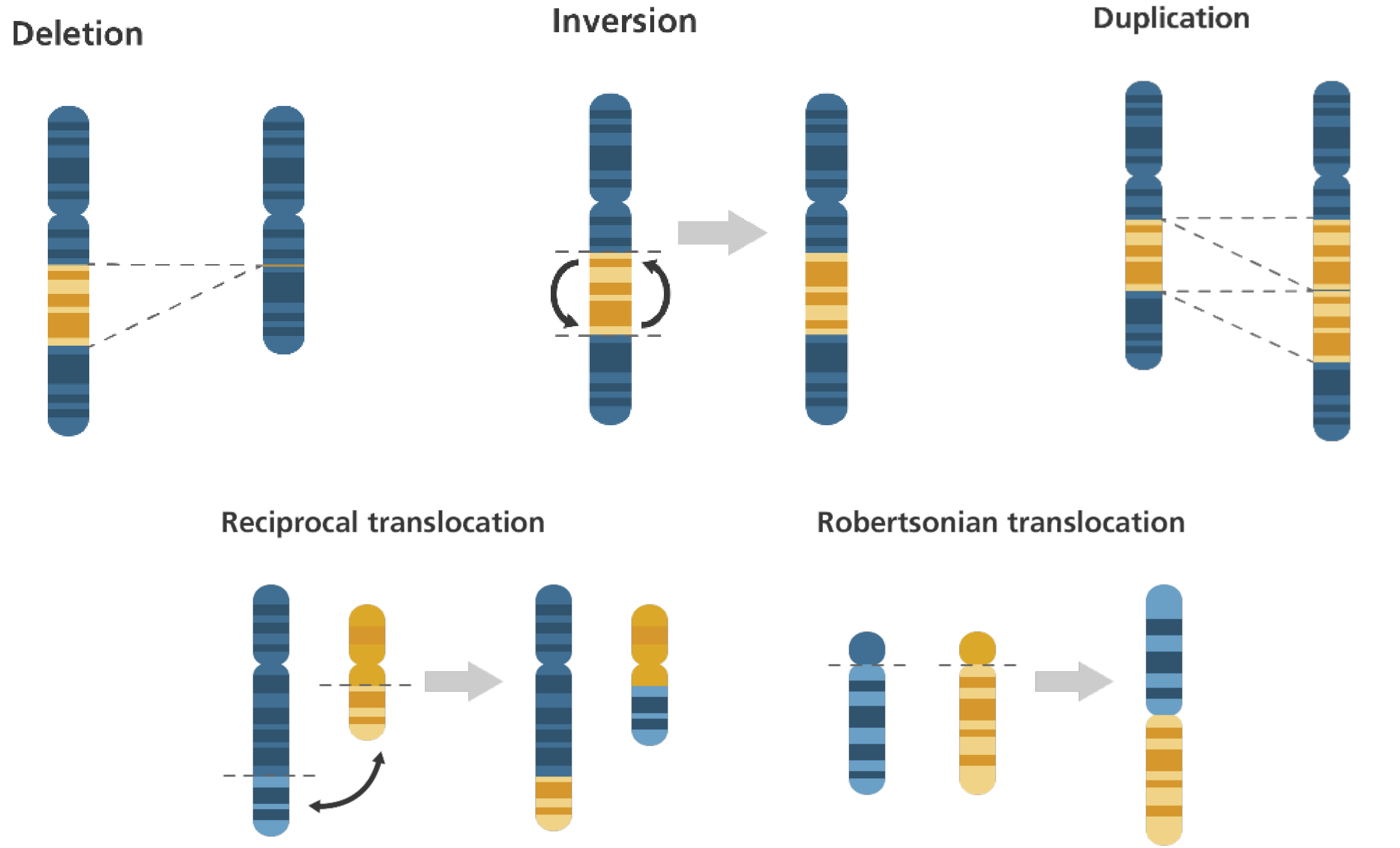
Tobias Rausch

European Molecular Biology Laboratory (EMBL)

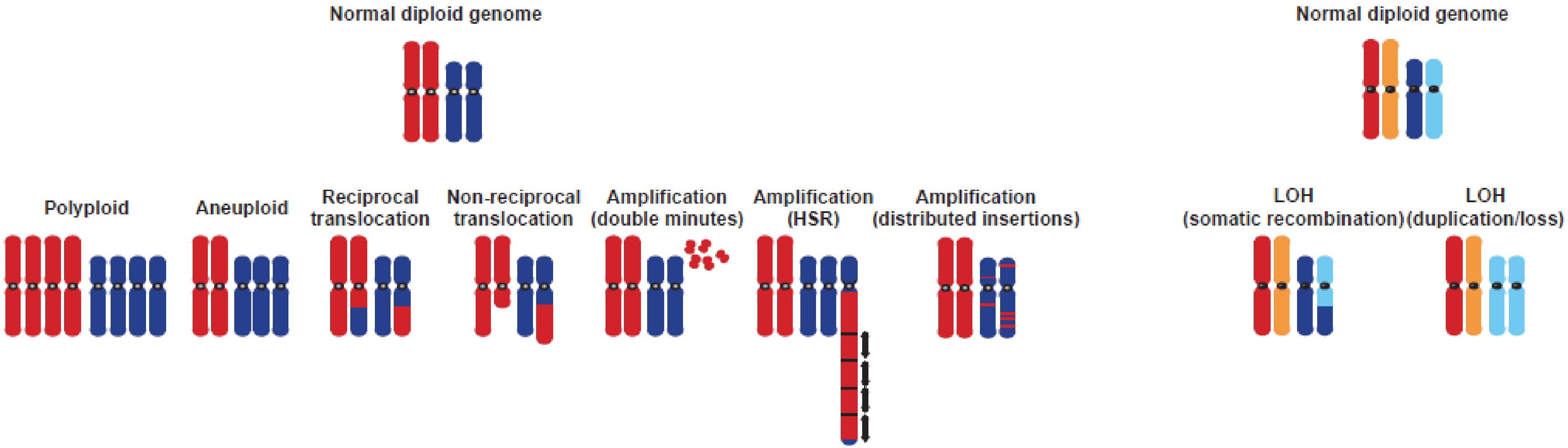
25 June 2024

EMBL

Somatic and Germline Structural Variants (SVs)



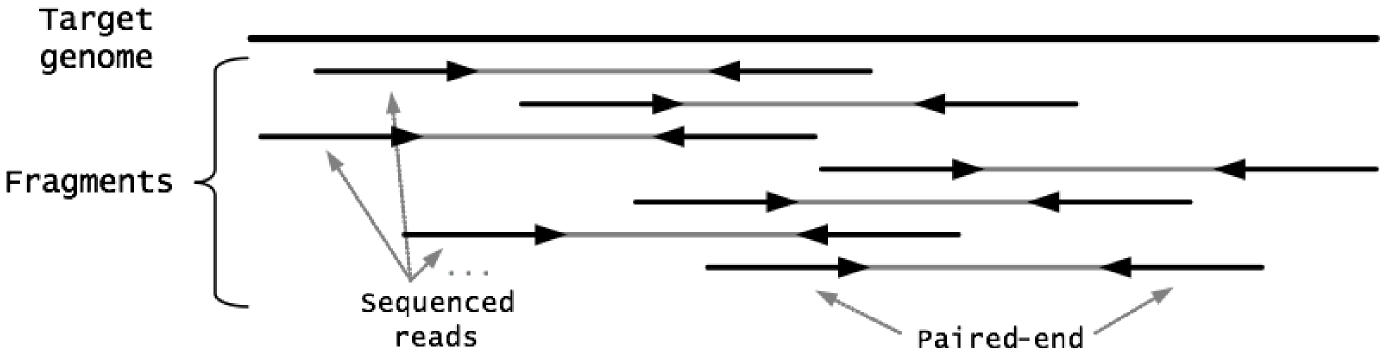
Cancers harbor a wide Range of Chromosome Abberrations



Source: Chromosome aberrations in solid tumors. Albertson et al., Nat Genet. 2003 Aug;34(4):369-76.

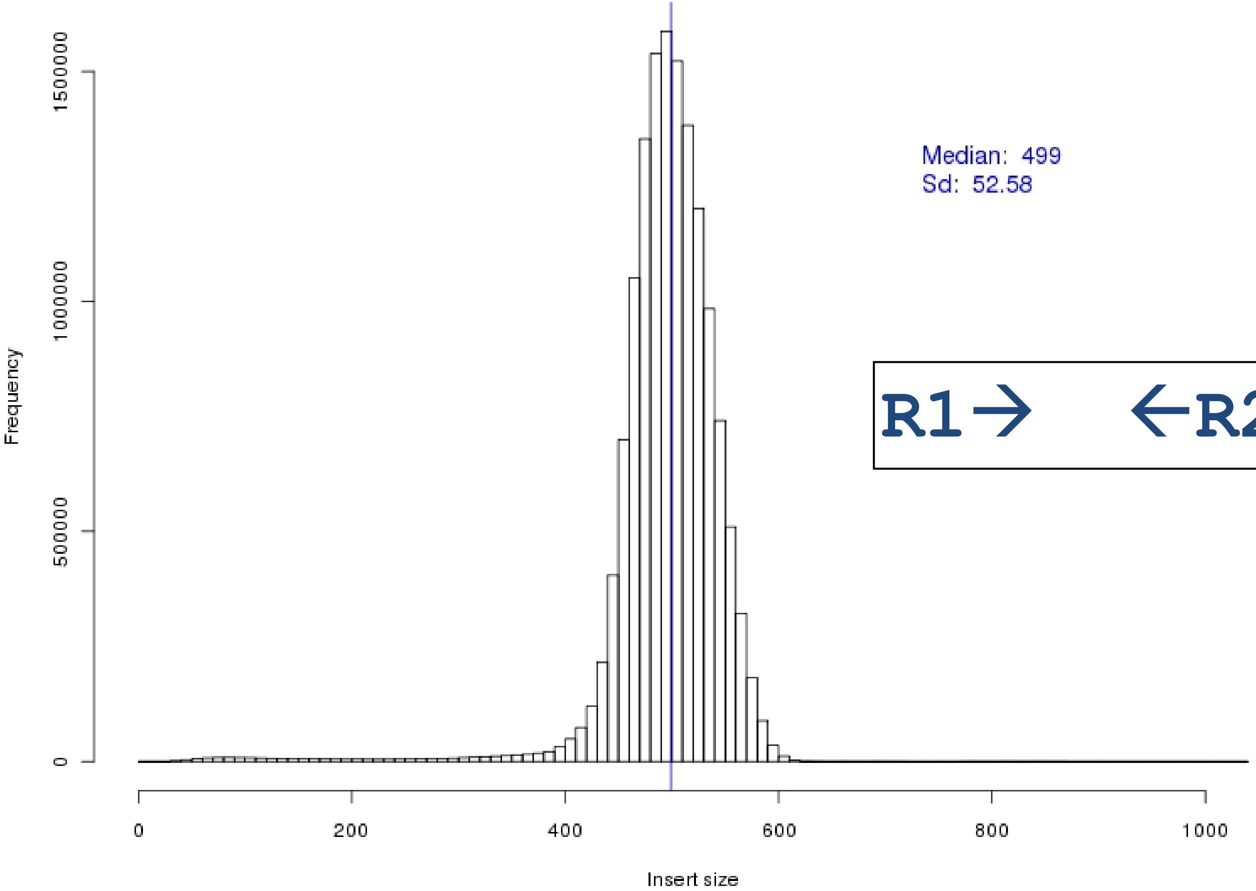
Structural Variants (SVs)

Paired-End Sequencing



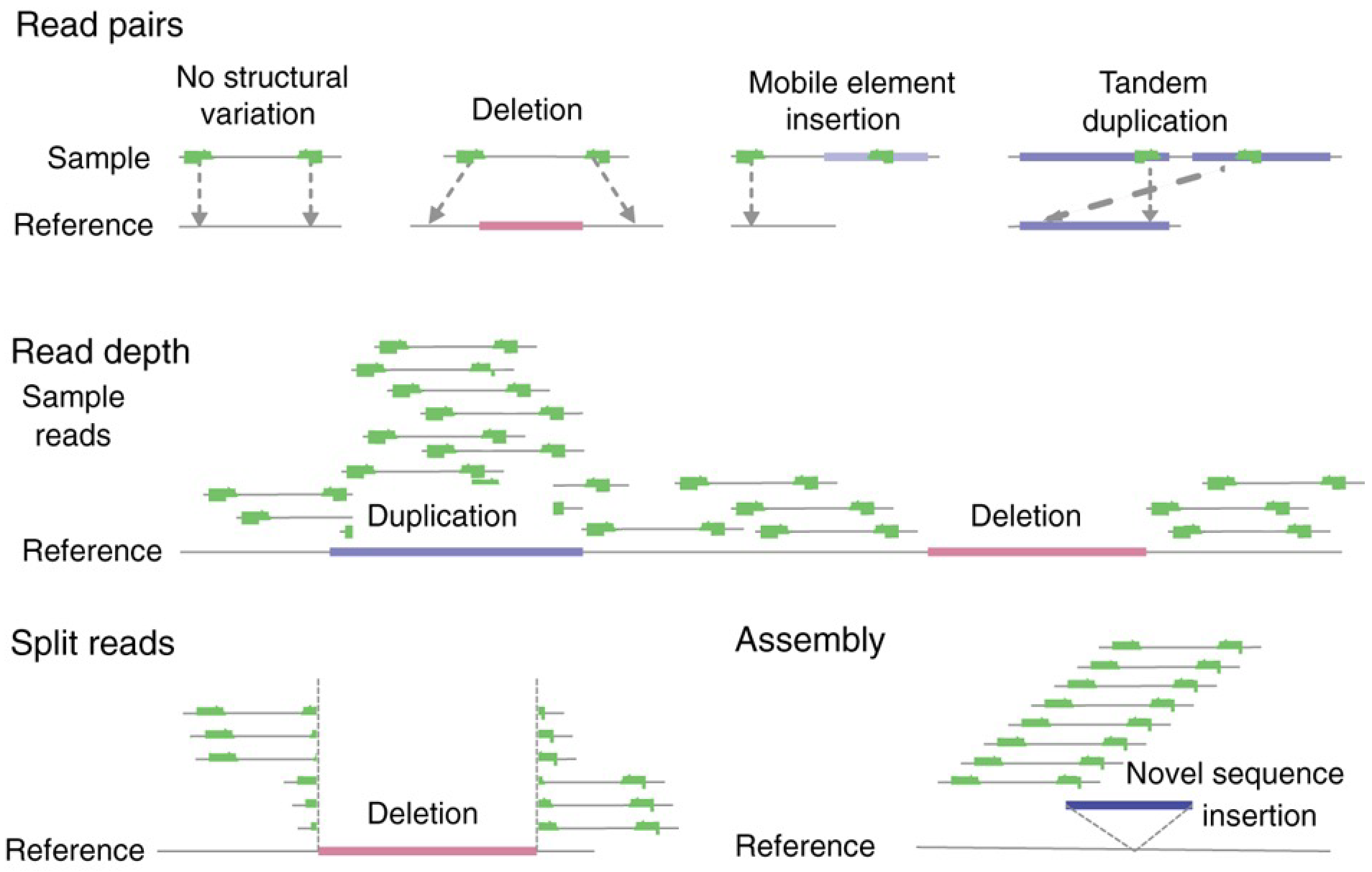
EMBL

Paired-End Libraries



R1→ ←R2

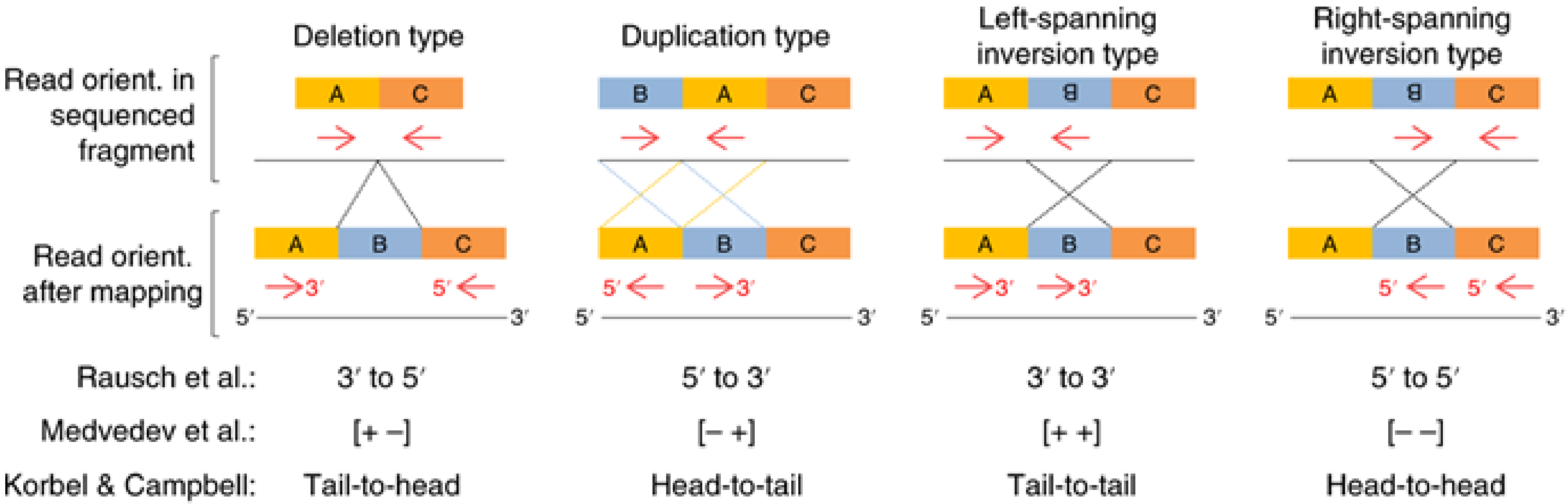
SV Discovery Approaches



EMBL

Source: Structural variation: the genome's hidden architecture. Baker M. Nat Methods. 2012 Jan 30;9(2):133-7.

Paired-end mapping

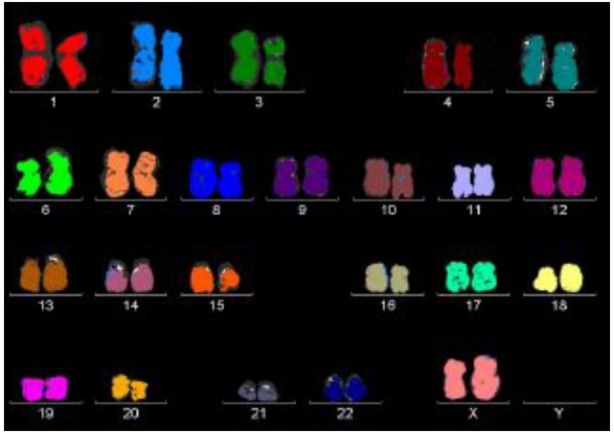


Source: https://doi.org/10.1038/s41596-019-0234-7

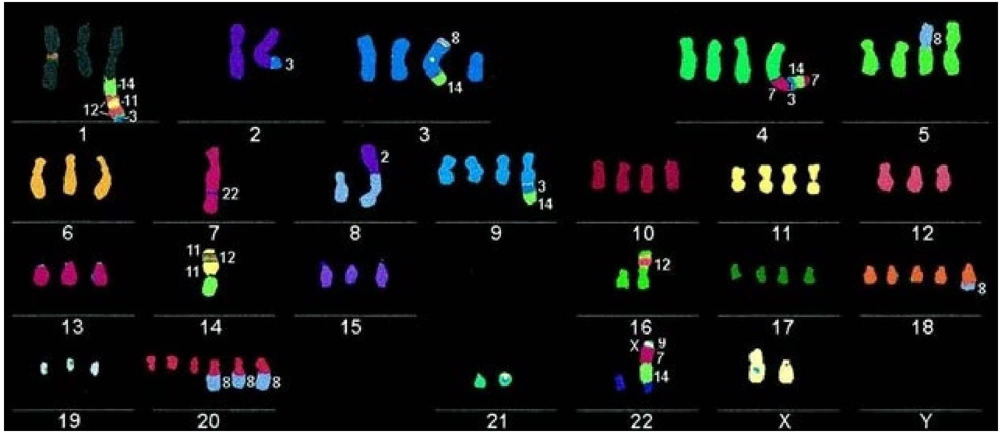
Copy-Number Variants (CNVs)

Human karyotype

Normal Karyotype



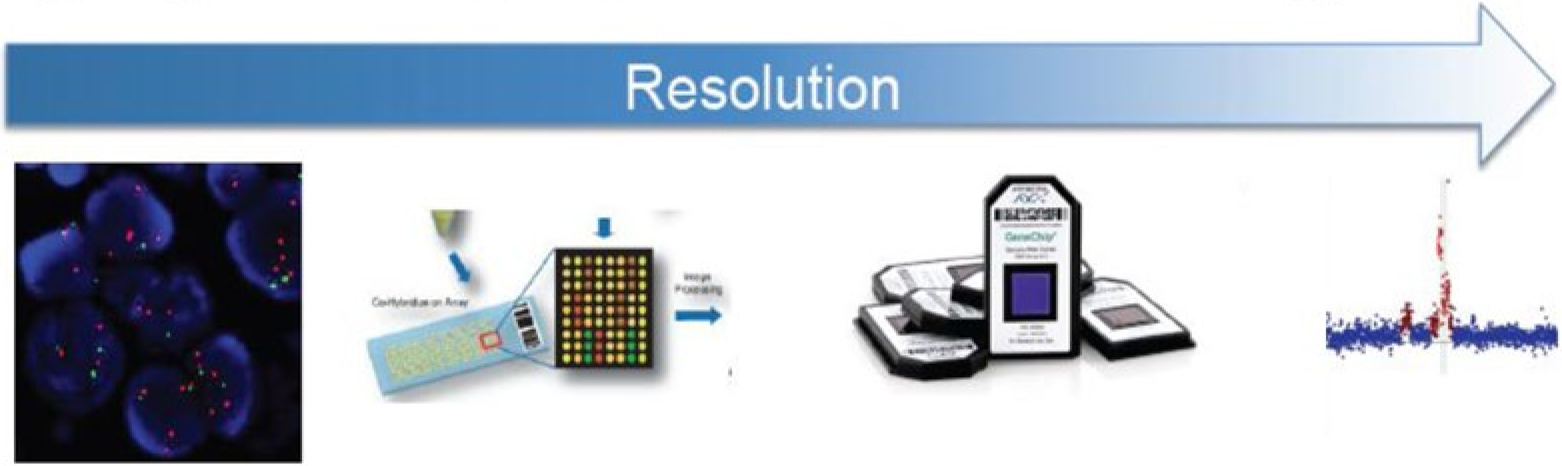
Cancer Karyotype (NSCLC cell line D117)



EMBL

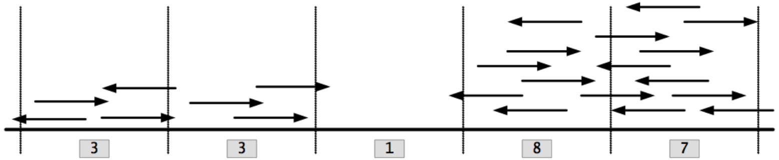
CNV detection technologies

Tech:	FISH	Array CGH	Genotype arrays	WGS
#:	<10	30-100K	100K-2M	3G!



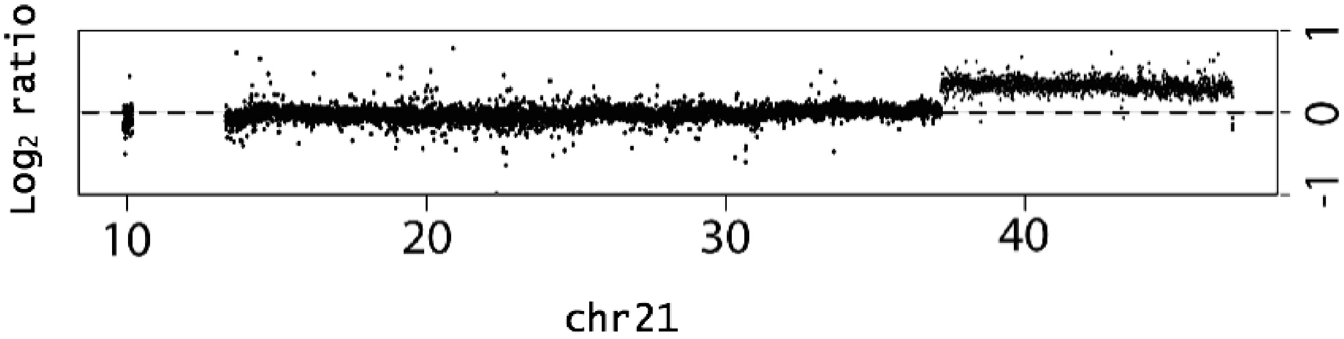
Tumor / Normal Read-Depth Ratio

- Read counting in windows for tumor and normal data



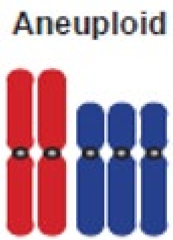
- Log2 ratio for each window
- Chromosome-wide plot

$$\log_2 \frac{\#\text{Reads}_{\mathit{Disease}}}{\#\text{Reads}_{\mathit{Normal}}}$$

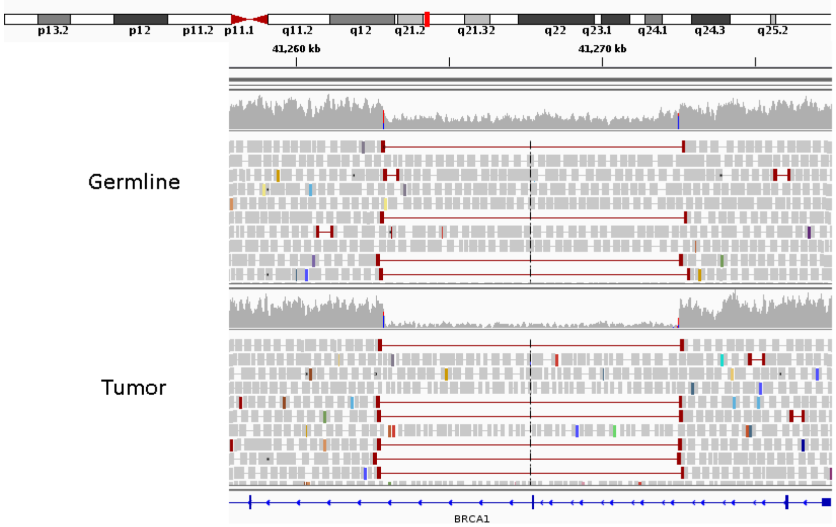


Copy-Number Variants

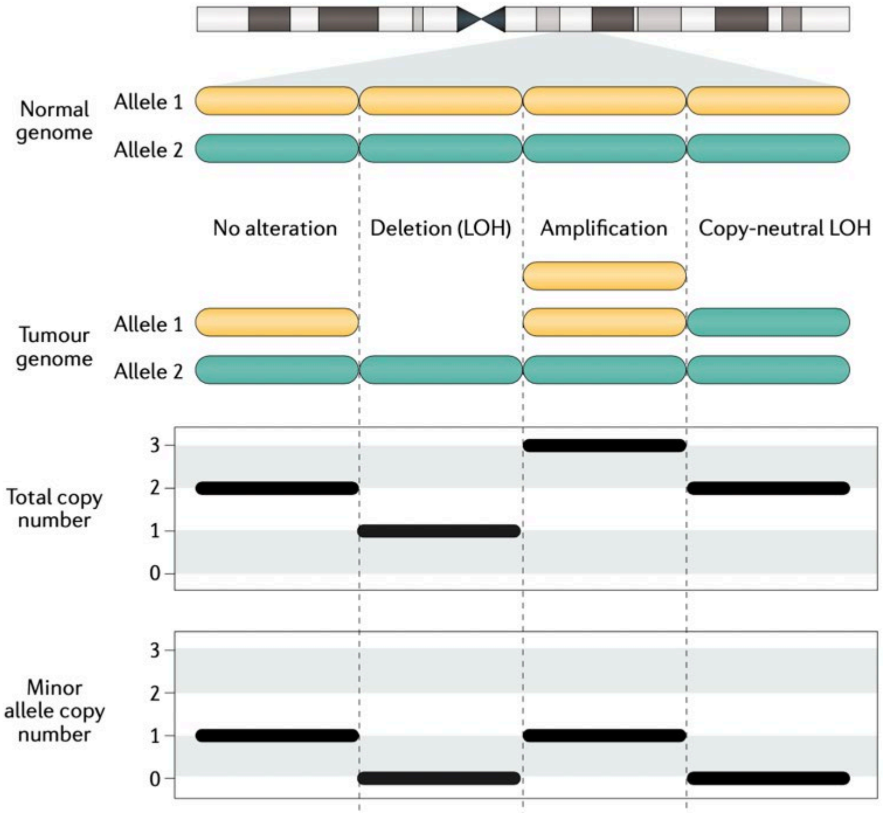
- Can vary the gene dosage of a tumor suppressor or oncogene
- Aneuploidy or non-reciprocal translocations are one form of CNV
- Rare pathogenic germline CNVs can affect known cancer predisposition genes
- Recurrent deletions or duplications indicate a selective advantage

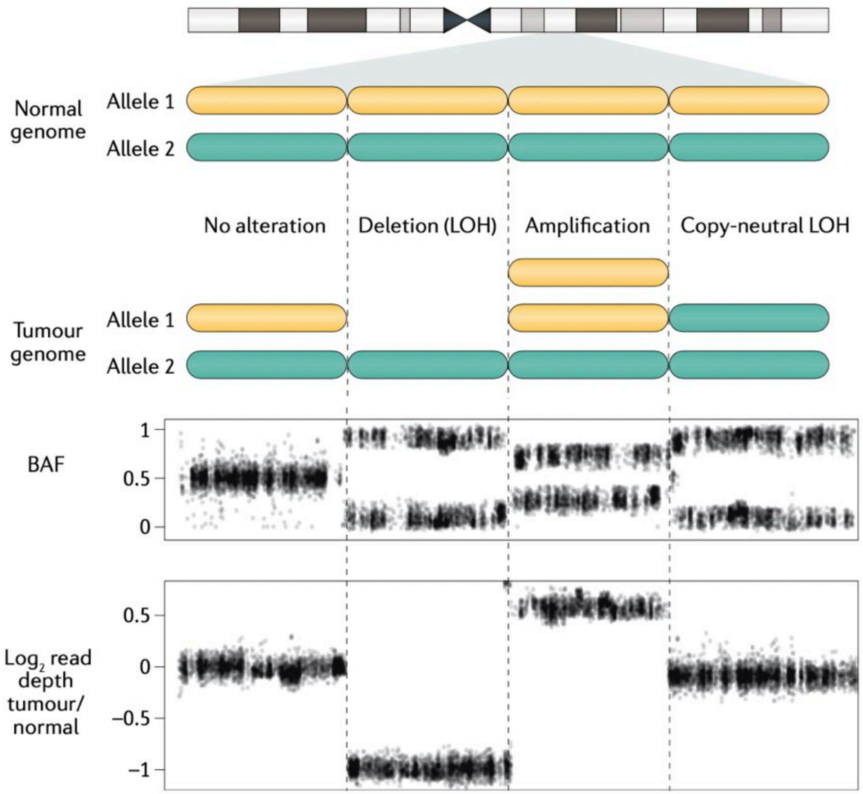


BRCA1 Deletion chr17:41,262,855-41,272,521



Copy-Number Variation

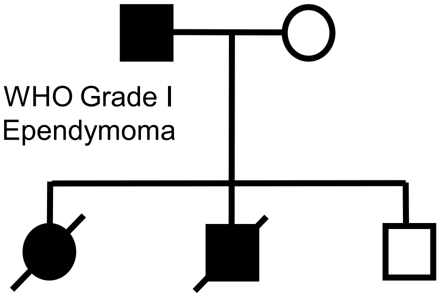




Somatic Structural Variants

Childhood Brain Tumor
Medulloblastoma

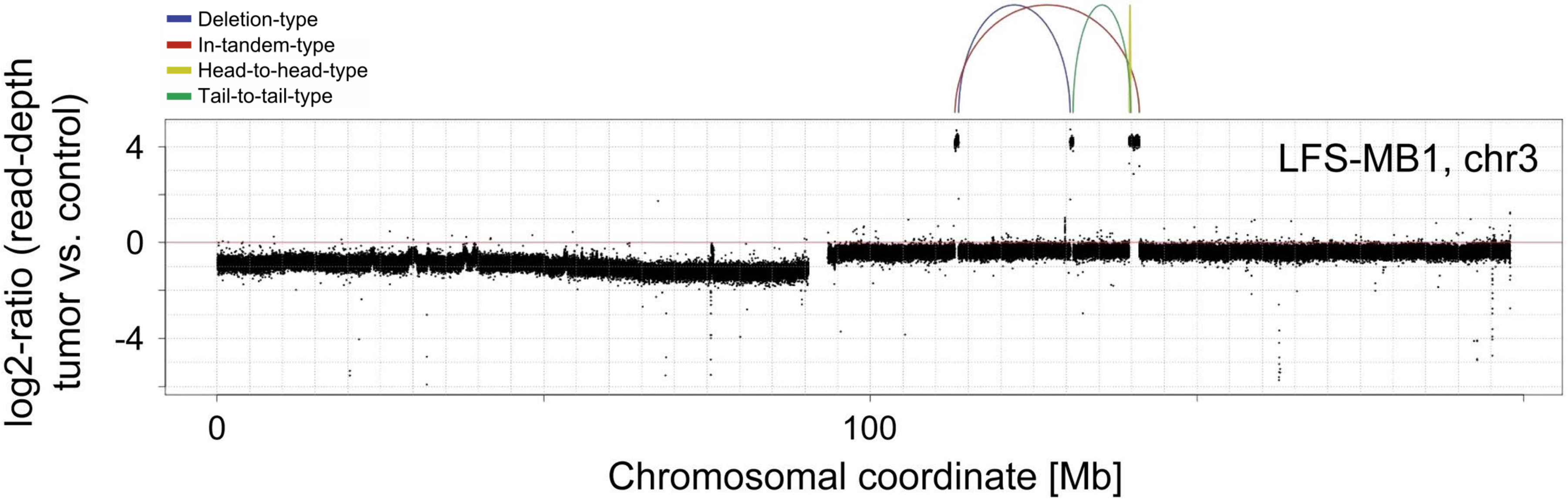
- Li-Fraumeni syndrome
  - Germline TP53 mutation



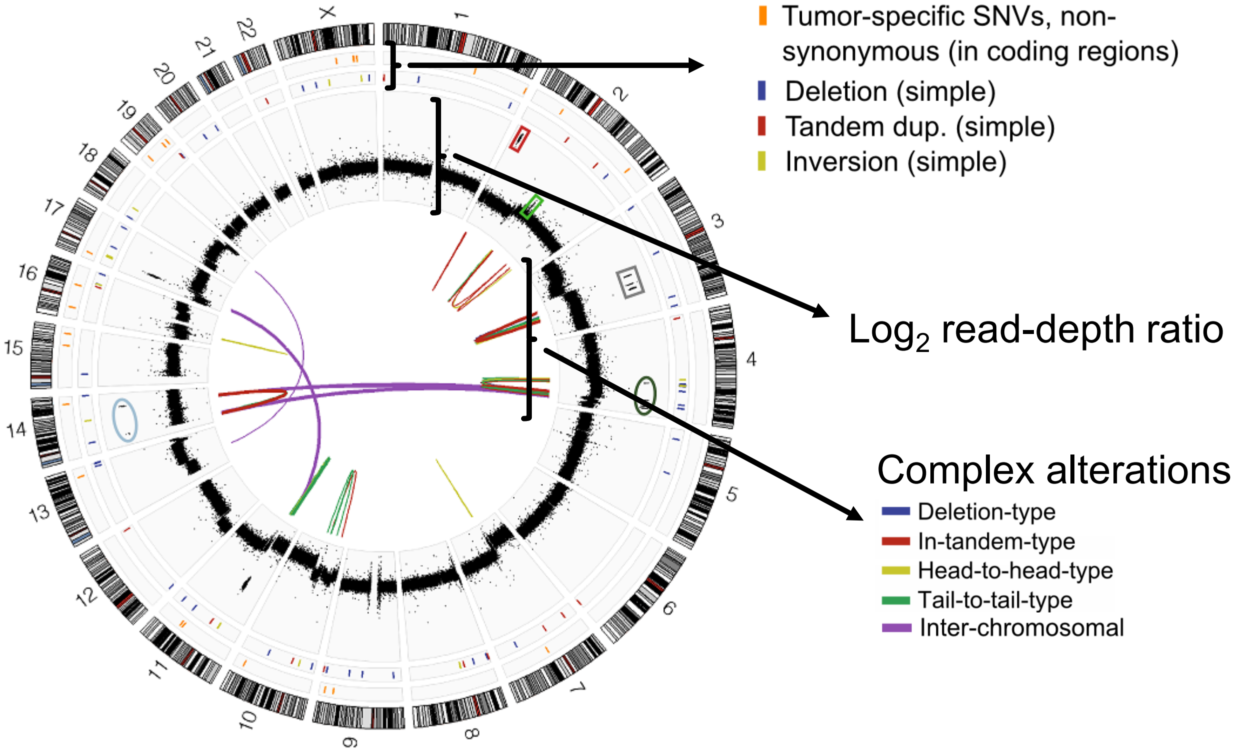
- WHO Grade I Ependymoma

- SHH-medulloblastoma
- Myelodysplastic syndrome (MDS)

- Choroid plexus carcinoma (CPC)



Somatic DNA alterations



Tumor-specific SNVs, non-synonymous (in coding regions)

Deletion (simple)

Tandem dup. (simple)

Inversion (simple)

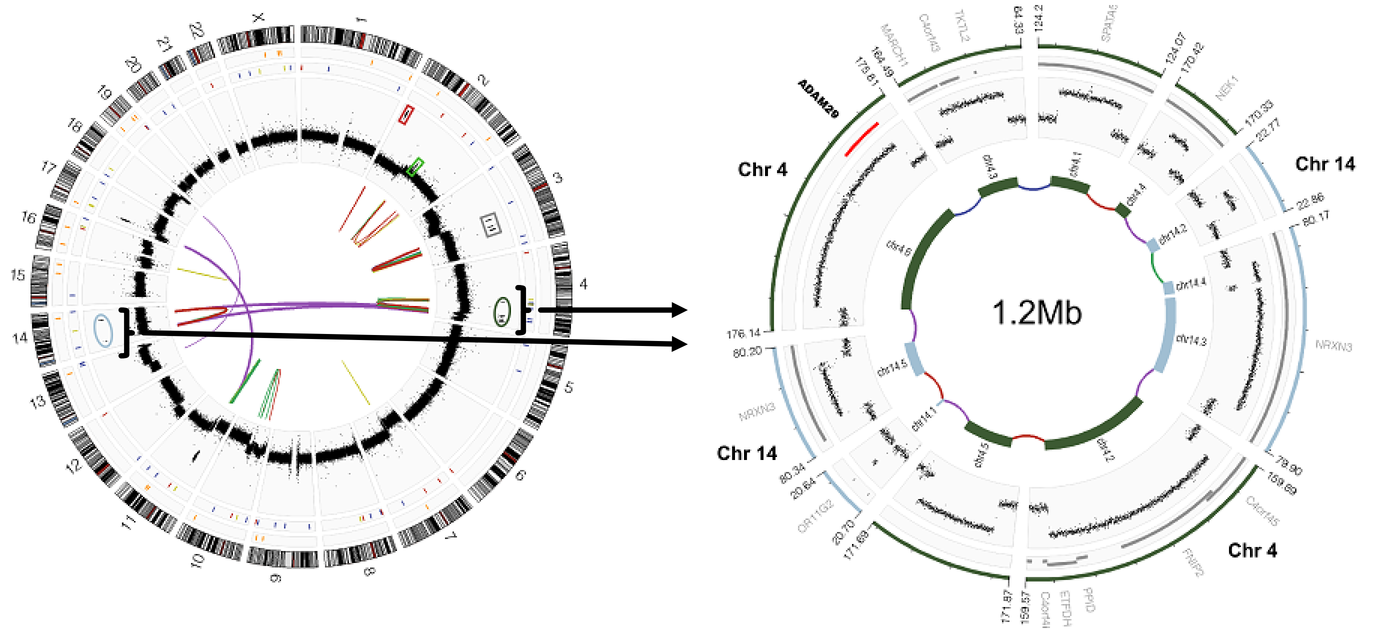
Log$_2$ read-depth ratio

Complex alterations

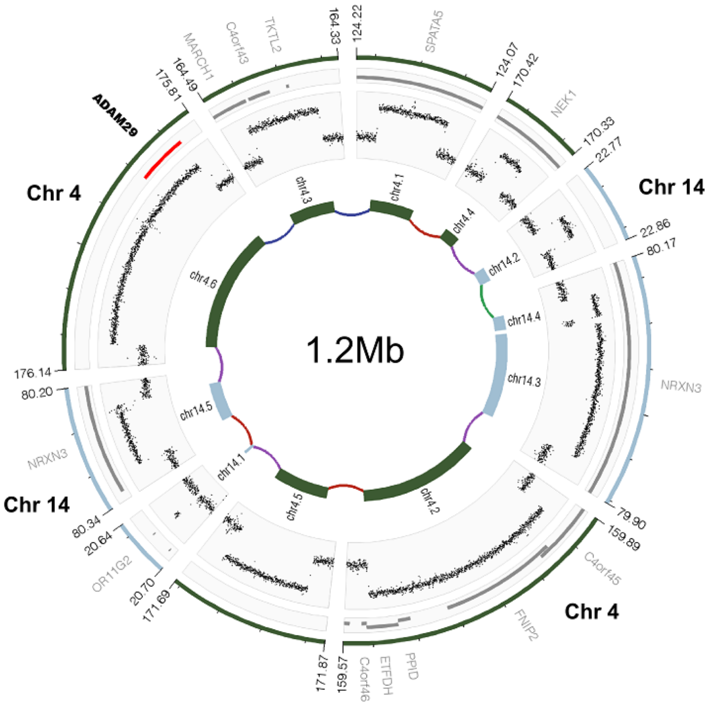
- Deletion-type
- In-tandem-type
- Head-to-head-type
- Tail-to-tail-type
- Inter-chromosomal

Source: Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Rausch et al., Cell. 2012 Jan 20;148(1-2):59-71.

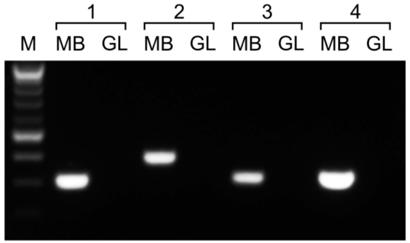
Complex DNA alterations forming double-minute chromosome



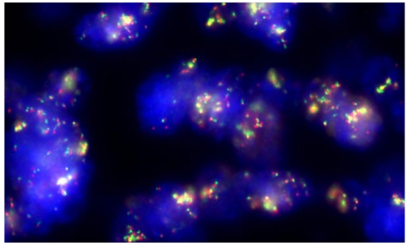
Validation of double-minute and co-localization of distant segments



Inter-chromosomal connections validated by PCR

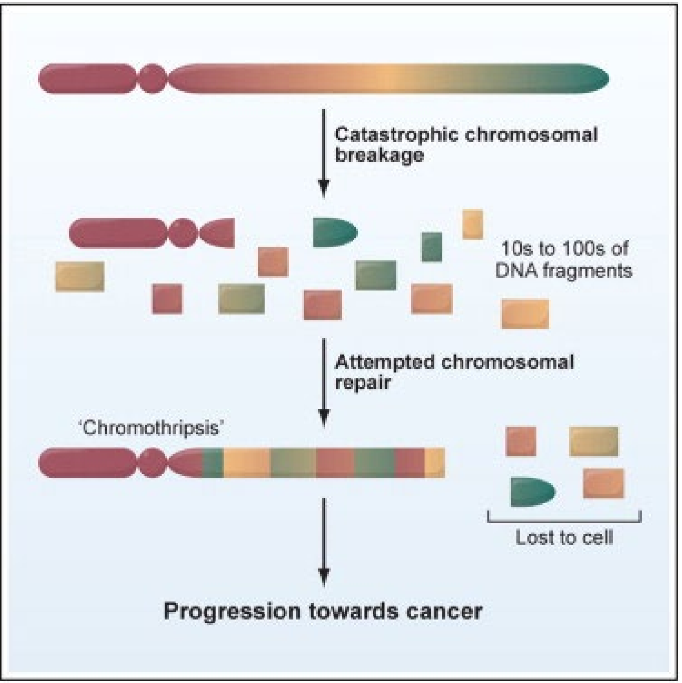


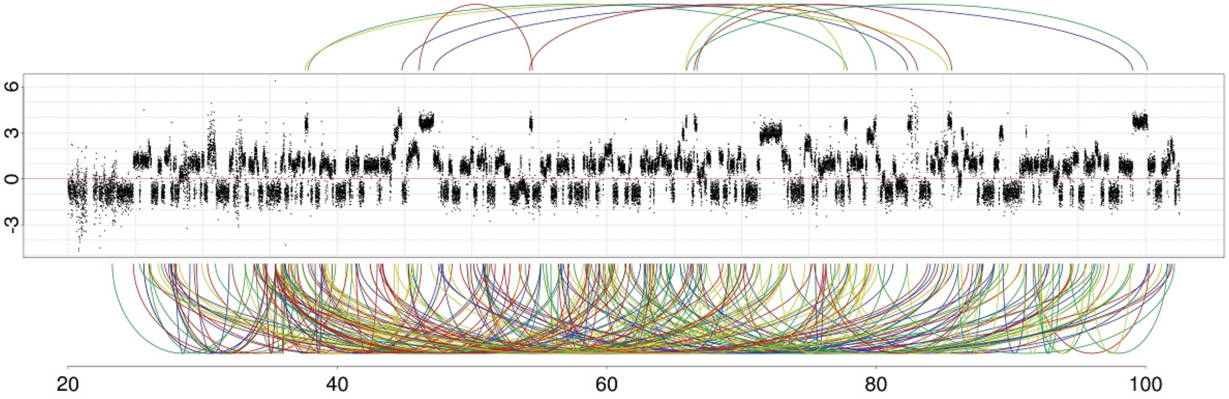
Co-localization of distal segments on chr3 by FISH



EMBL

Chromothripsis


Catastrophic chromosomal breakage
10s to 100s of DNA fragments
Attempted chromosomal repair
'Chromothripsis'
Lost to cell
Progression towards cancer



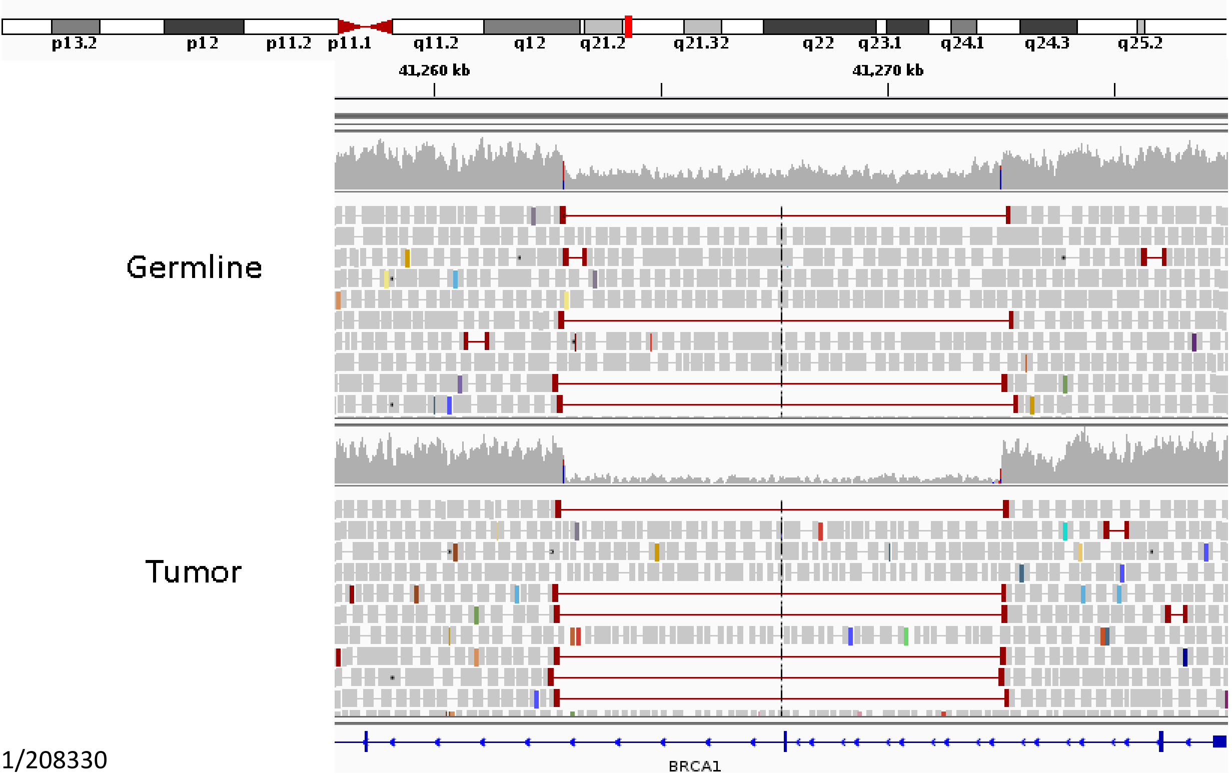
Source: Stephens et al., Cell, 2011; Rausch et al., Cell, 2012

Predisposing Germline Structural Variants

Cancer Predisposition

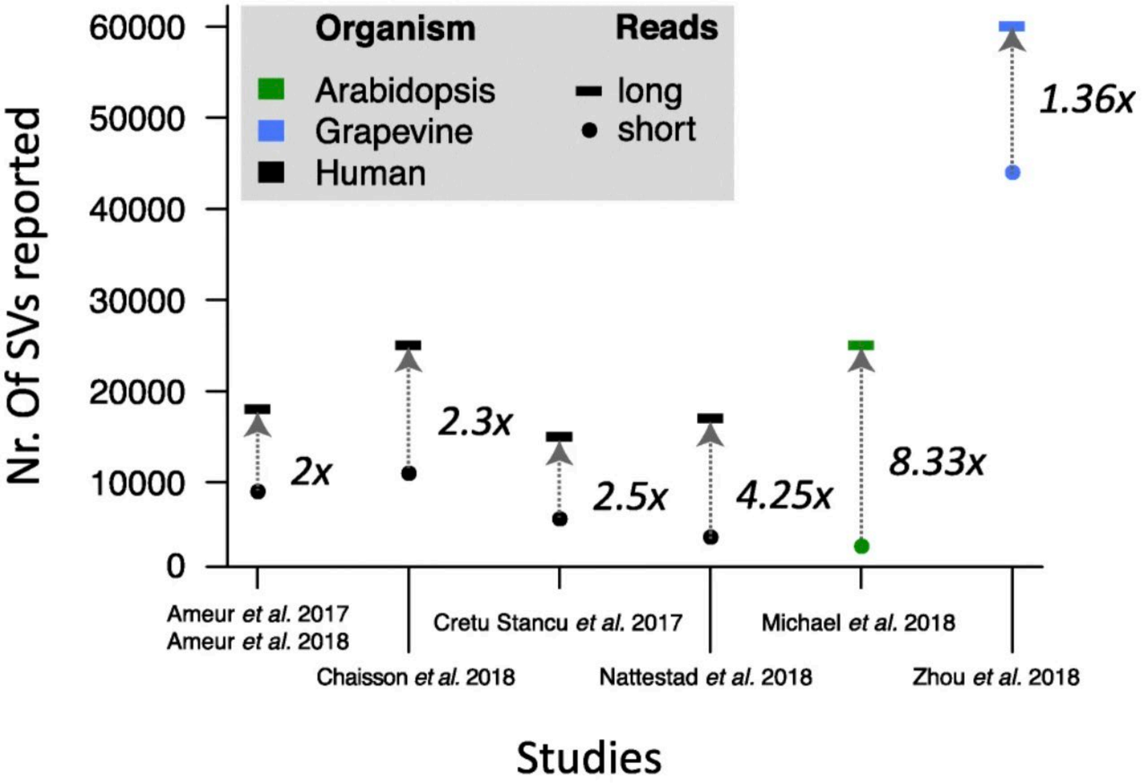
- Heterozygous germline deletion
- Loss of wildtype copy in the tumor

BRCA1 Deletion chr17:41,262,855-41,272,521

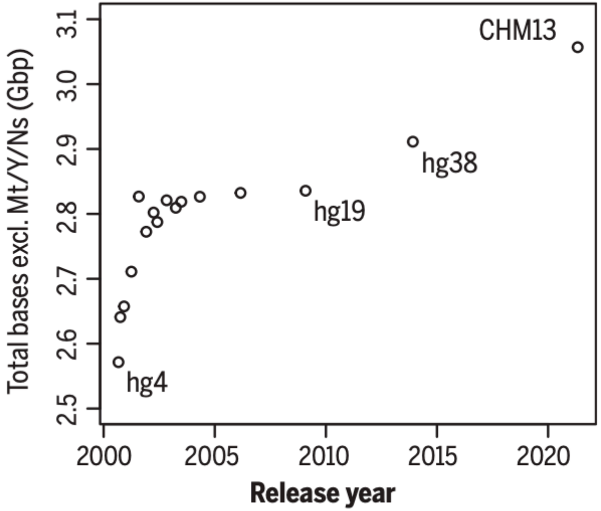


Source: https://www.biorxiv.org/content/early/2017/11/01/208330

Germline SV detection using short-reads is largely incomplete!



Telomere-to-Telomere (T2T) Assembly – CHM13



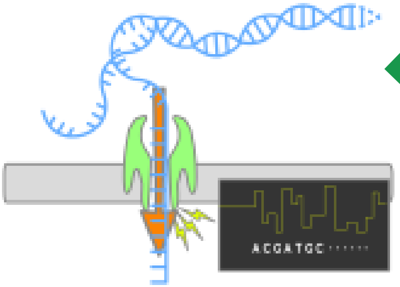
EMBL

Source: https://doi.org/10.1186/s13059-019-1828-7, https://doi.org/10.1126/science.abj6987

Long-reads and T2T references for SV discovery

Short-reads: 100bp-300bp

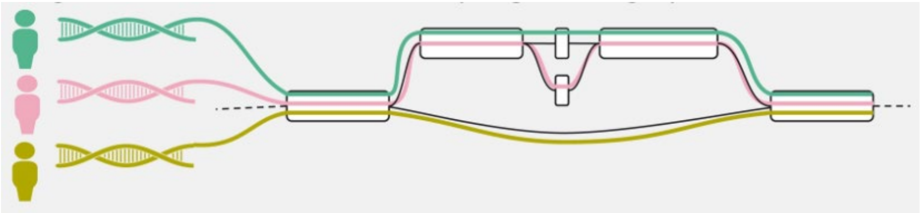
Long-reads: 1,000bp-20,000Kbp, few >>20Kbp



Nanopore sequencing

Linear reference genome (GRCh38)

Graph pan-genome

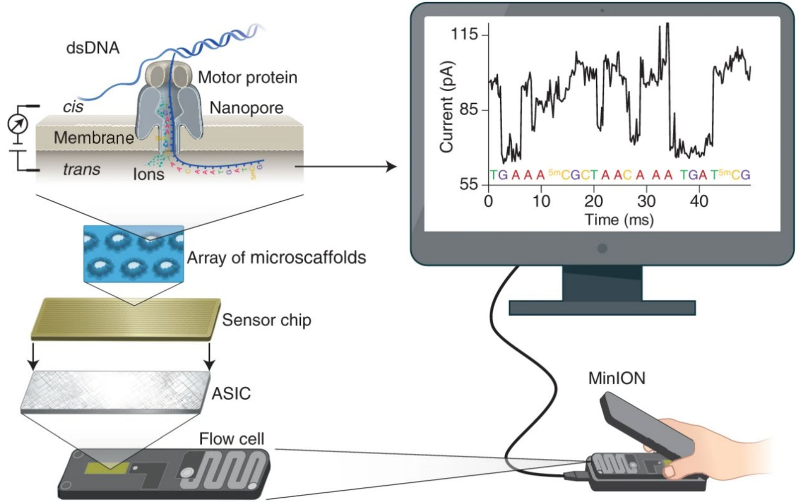


Human Pangenome Reference Consortium

?

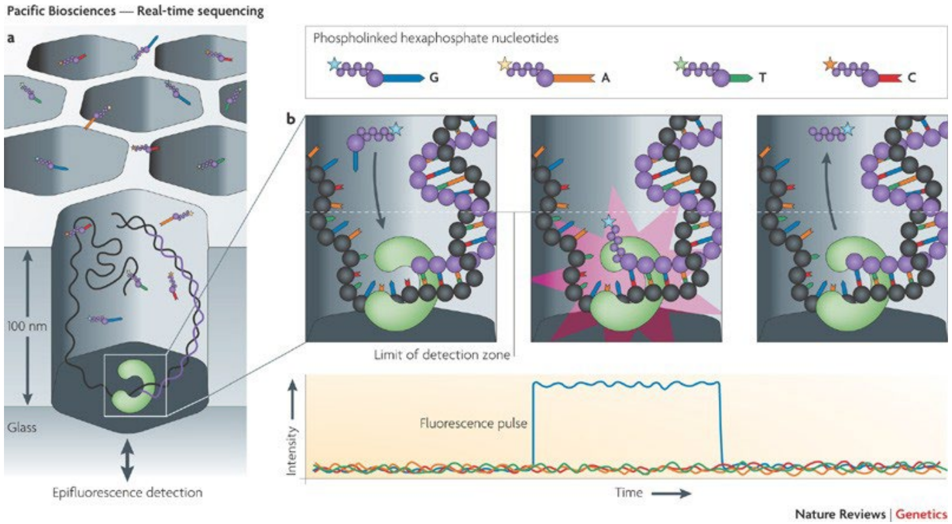
Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio)

Oxford Nanopore Sequencing



1,000bp – 20,000bp reads but some >>20Kbp
~1 error in 100 bases

Pacific Biosciences Sequencing

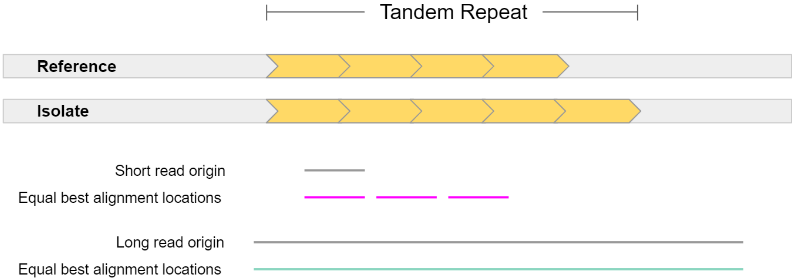

Pacific Biosciences — Real-time sequencing
a
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
100 nm
Glass
Epifluorescence detection
Limit of detection zone
Fluorescence pulse
Intensity
Time
Nature Reviews | Genetics

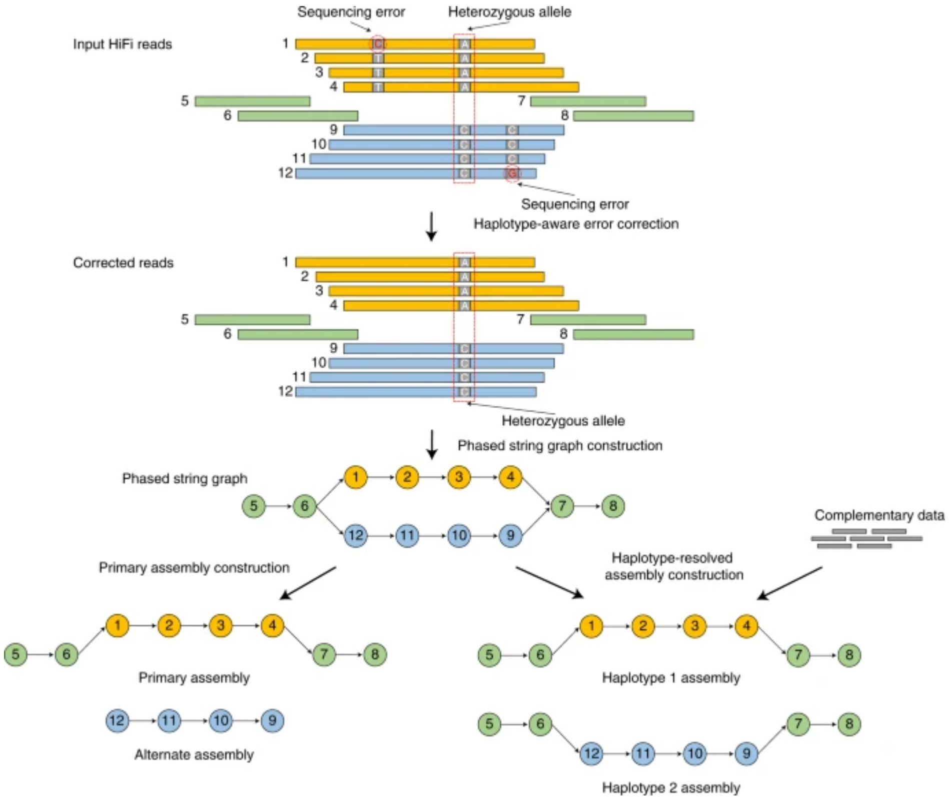
1,000bp – 20,000bp reads
~1 error in 10,000 bases

Sources: https://doi.org/10.1007/s11825-019-0249-z, https://doi.org/10.1038/s41587-021-01108-x, https://doi.org/10.1038/nrg2626

Long read applications

- *De novo* genome Assembly
- Haplotype-resolved genome analysis
- Structural variant (SV) discovery
  - Repetitive SVs
  - Complex SVs
- Resolving genome structure
  - Derivative chromosomes in cancer





EMBL

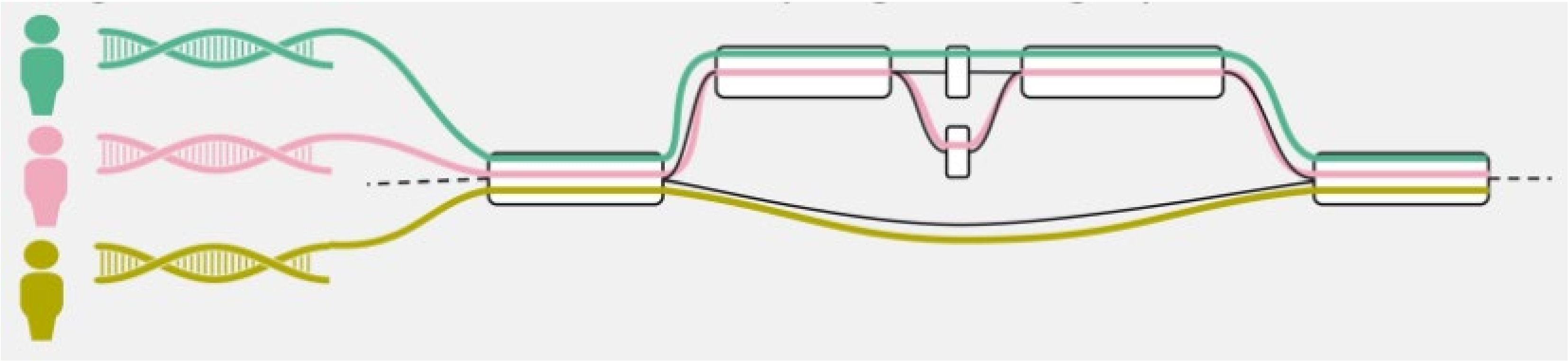
Source: https://www.nature.com/articles/s41592-020-01056-5

Pan-genome graphs

- A succinct representation of a set of reference genomes

Haplotype-resolved human assemblies

Pan-genome graph



Human Pangenome Reference Consortium (HPRC)

- 44 samples (88 haplotypes) + GRCh38 + CHM13

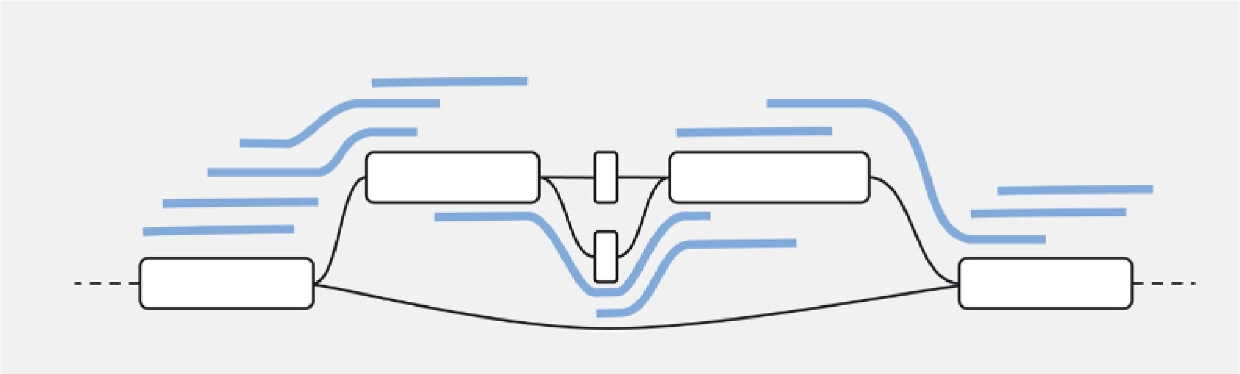
Source: https://doi.org/10.1038/s41586-023-05896-x, https://doi.org/10.1126/science.abg8871

Pan-genome graphs

- HPRC pan-genome graph: 90 haplotypes (44 samples, GRCh38, CHM13)

- How to incorporate all types of variation?
  - Coarse-grained pangenome graph (structural variants only)
    - 751M on disk: **391,950 segments** (S); 566,204 links (L); 3,198,196,033bp
  - Fine-grained pangenome graph (including small variants)
    - 8.6G on disk: **81,415,956 segments** (S); 112,955,105 links (L); 3,287,932,785bp

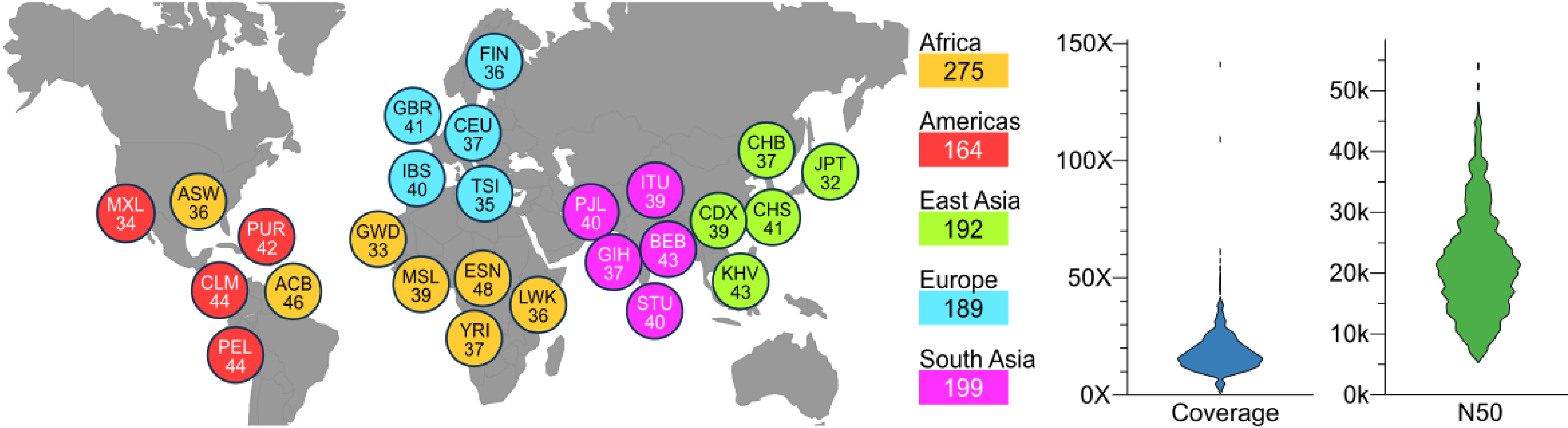
Alignment to pan-genome graph



Source: https://doi.org/10.1126/science.abg8871, https://github.com/lh3/minigraph

Genome variation discovery using long reads and graph genomes

1000 Genomes ONT Project

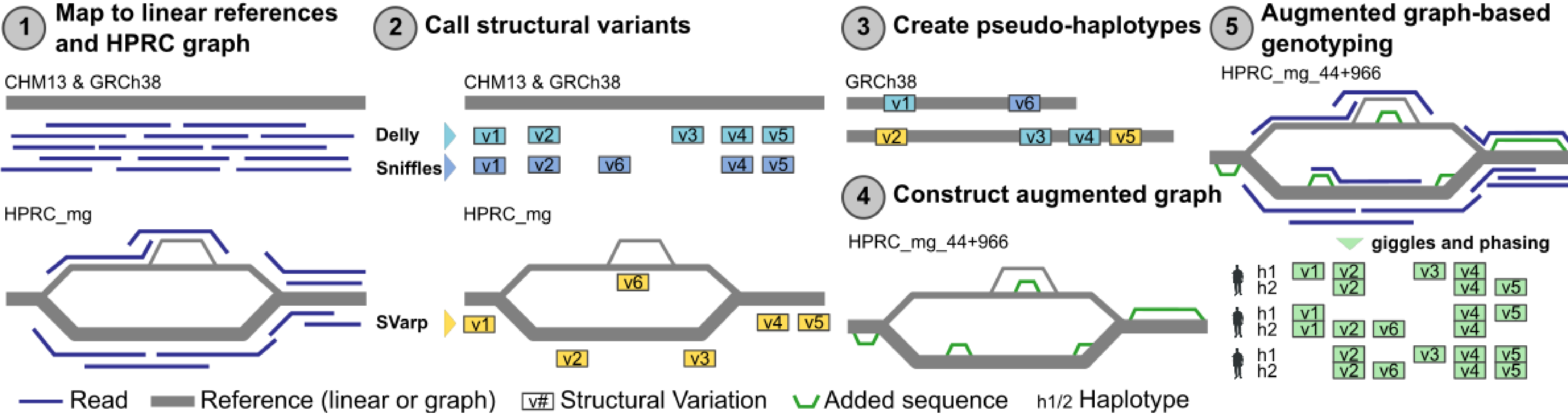


1,019 samples sequenced with ONT

- ~15x coverage
- **Structural variant calling** using pan-genome graphs

bioRxiv preprint: https://www.biorxiv.org/content/10.1101/2024.04.18.590093v1

Variant Calling Strategy



1 Map to linear references and HPRC graph

2 Call structural variants

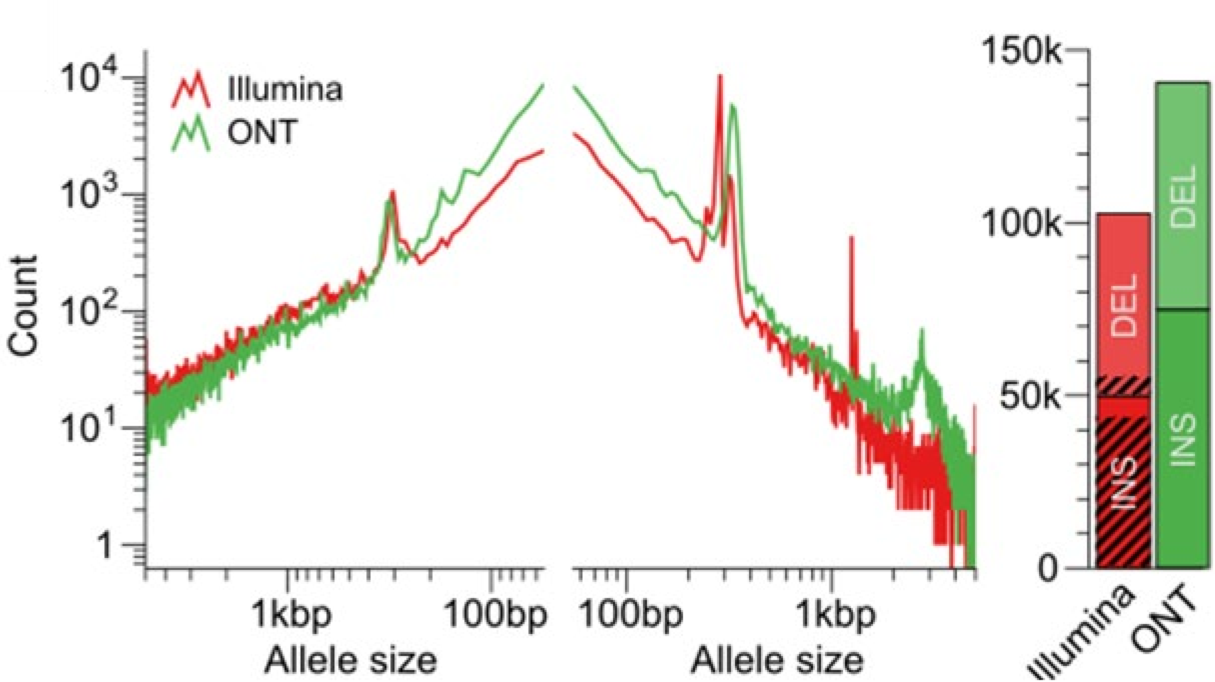
3 Create pseudo-haplotypes

4 Construct augmented graph

5 Augmented graph-based genotyping

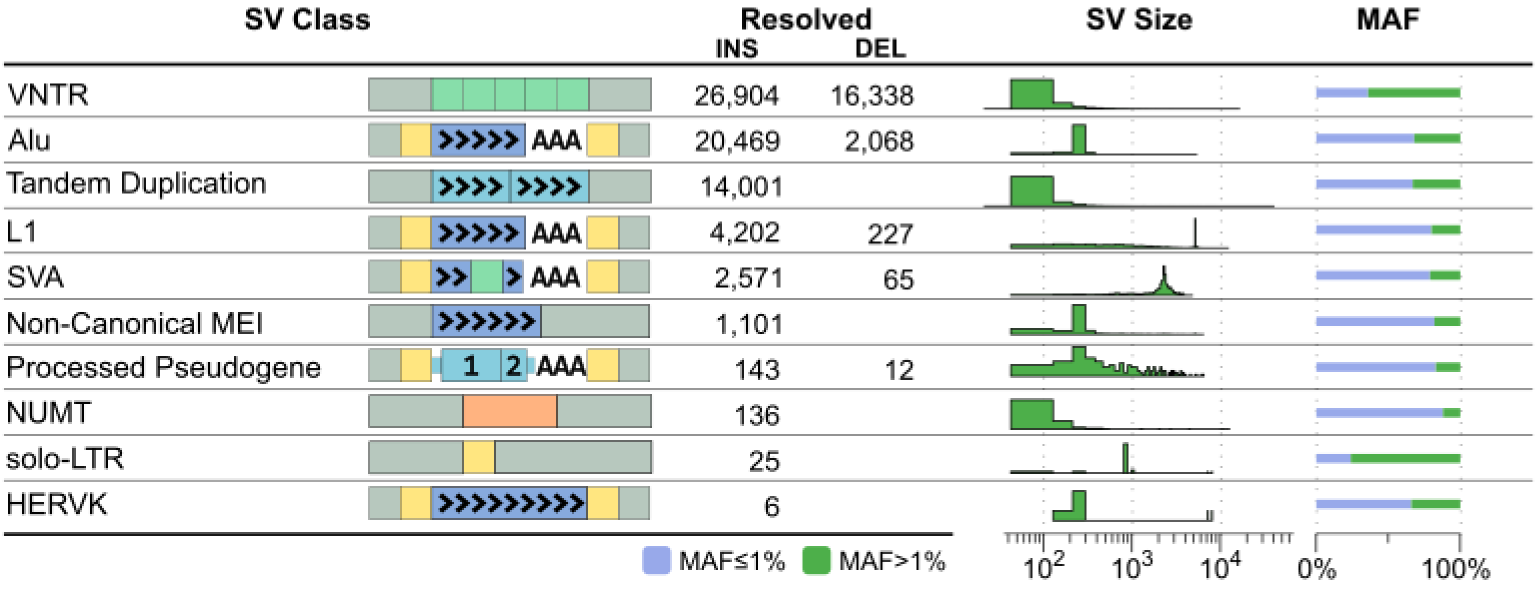
bioRxiv preprint: https://www.biorxiv.org/content/10.1101/2024.04.18.590093v1

Long-reads facilitate the discovery of sequence-resolved insertions



EMBL

Insertion SV classes: VNTR variation is abundant!

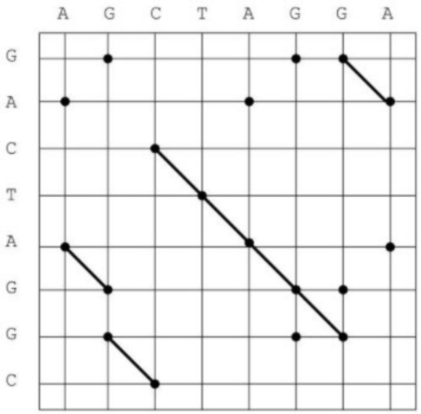


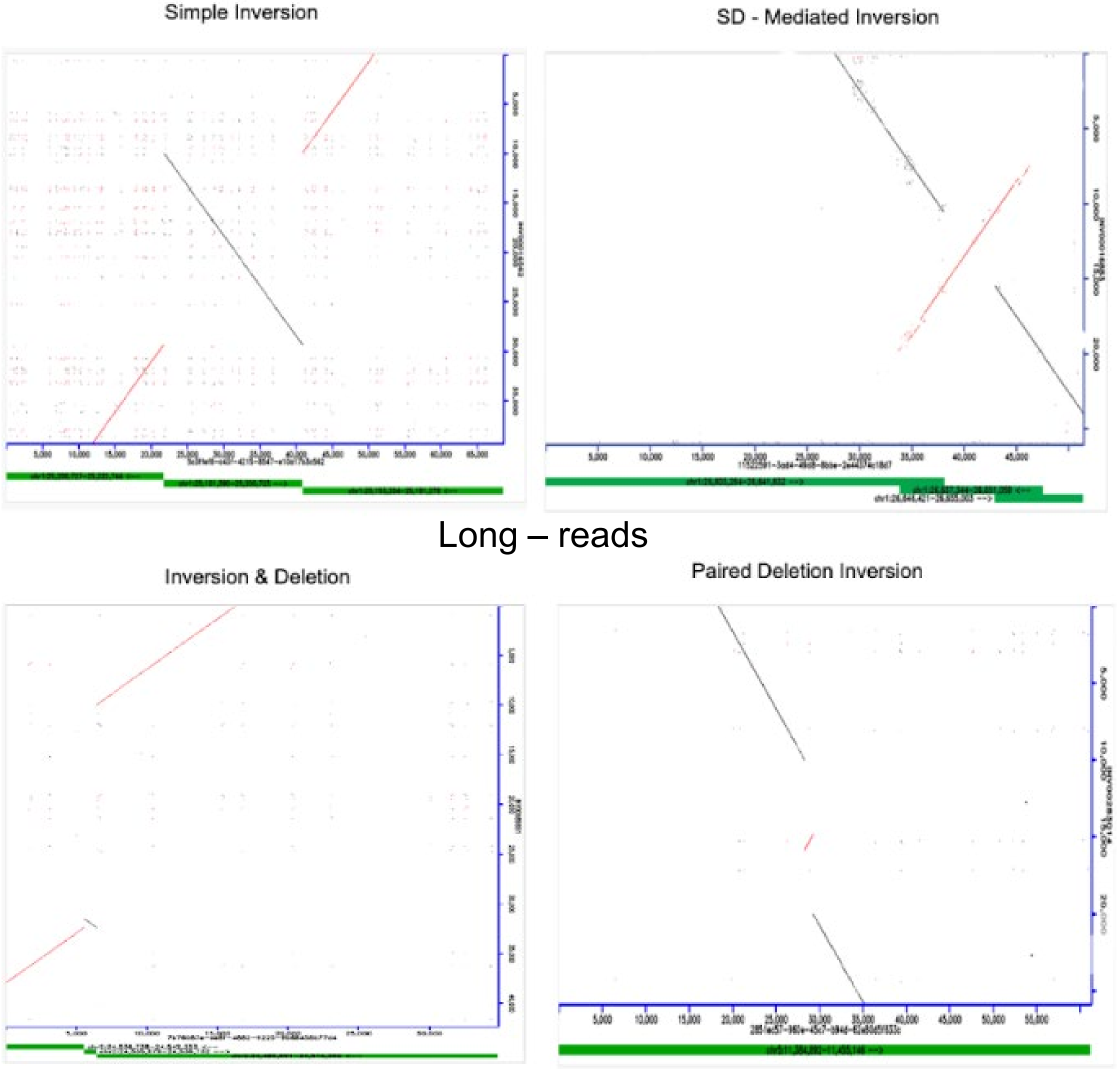
SV Class	Resolved	
	INS	DEL
VNTR	26,904	16,338
Alu	20,469	2,068
Tandem Duplication	14,001	
L1	4,202	227
SVA	2,571	65
Non-Canonical MEI	1,101	
Processed Pseudogene	143	12
NUMT	136	
solo-LTR	25	
HERVK	6	

bioRxiv preprint: https://www.biorxiv.org/content/10.1101/2024.04.18.590093v1

Improved resolution of complex SVs and Inversions

Dotplot




Simple Inversion

SD - Mediated Inversion

Long – reads

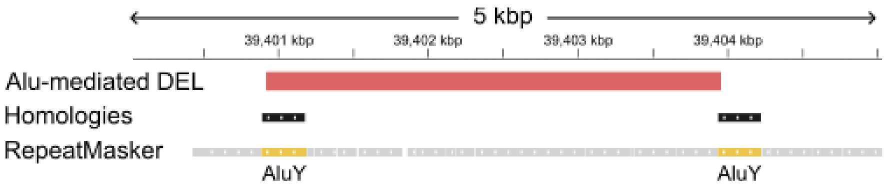
Inversion & Deletion

Paired Deletion Inversion

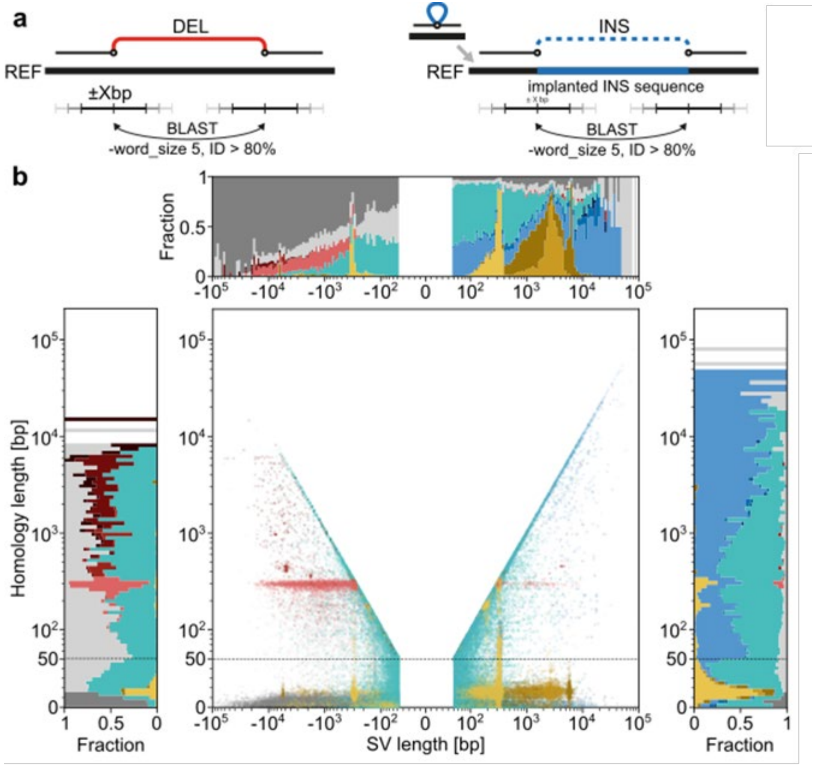
Reference genome

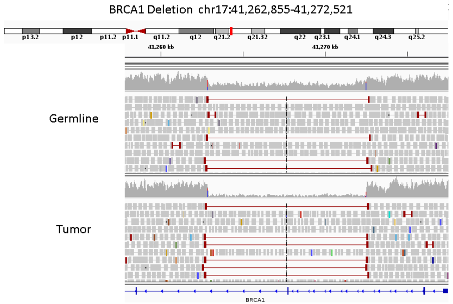
Repeat-mediated SVs

- A large fraction of deletions (other than VNTR) is repeat-mediated (35%)



Repeat-mediated
● Alu-mediated
● LTR-mediated
● L1-mediated
● SD-mediated
Duplications
● Tandem
● Interspersed
● Complex
● Inverted
Mobile Elements
● Alu
● SVA
● L1
● VNTR
● NHEJ
● Other
All





Cancer Predisposing SVs

Hypothetical Example: Deletion that affects geneX

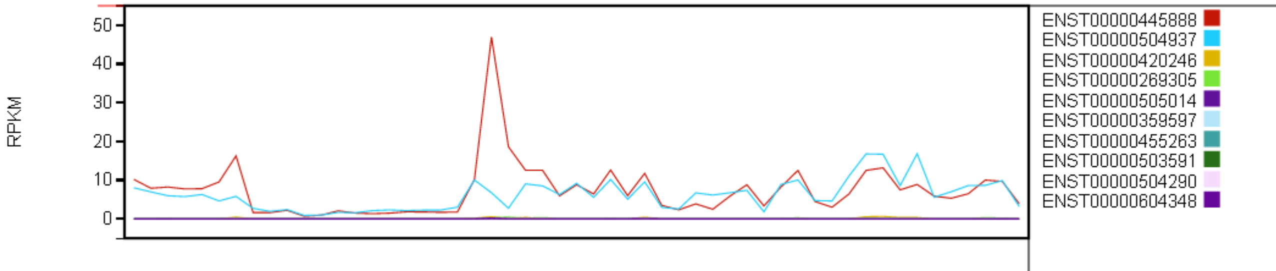
- Pan-cancer cohort (e.g. 300 breast cancer samples)
  - 18 out of 300 samples are a carrier
  - Allele frequency: ~6%

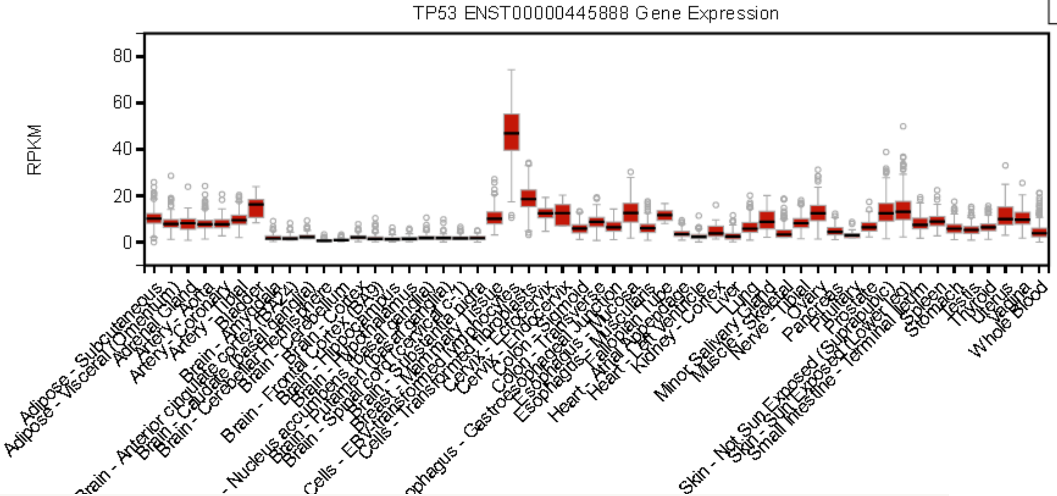
- 1000 Genomes cohort (2504 samples)
  - 5 out of 2504 samples are a carrier
  - Allele frequency: ~0.2%

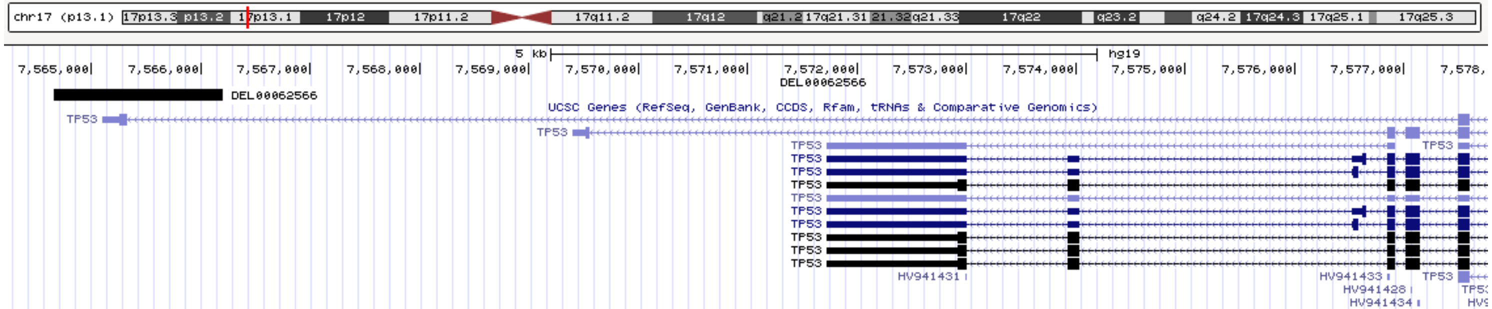
- SV may confer a higher risk for breast cancer but be aware of many possible confounders!
  - Sex, Related individuals, Population structure, …
    - All 5 carriers have European ancestry and the cohort of Europeans is much smaller than 2504 samples
  - Technical confounders: Low vs. high-coverage, different insert size, error rate

Structural variants affecting genes

GTEx Portal:
Genotype-Tissue Expression
https://gtexportal.org/home/



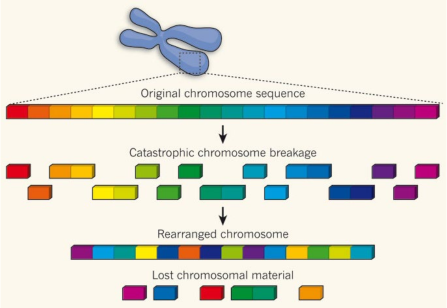




EMBL

Summary - Challenges in SV and CNV Calling

- Comprehensive detection usually requires long-reads
- Breakpoints from short-reads are often imprecise (e.g., is a fusion gene in-frame?)
- Copy-number baseline in cancer is not necessarily copy-number 2
- Incomplete copy-number and SV polymorphism map
- Very incomplete understanding of SV mechanisms and structure
  - Repeat-mediated SVs
  - Role of centromeres and telomeres?
- Short-read methods tend to have a very high false positive rate
  - Long-read methods are still being developed
- Complex rearrangements are difficult to disentangle with short reads
- Assembling a cancer genome is currently NOT possible


Original chromosome sequence
Catastrophic chromosome breakage
Rearranged chromosome
Lost chromosomal material

Thank you for your attention!

EMBL