

**03.11.2023**


# **CLINICAL GENOMIC TESTING**

**Matti Kankainen, PhD**


**Laboratory of Genetics, Diagnostic centre, Helsinki and Uusimaa hospital district**

**Translational Immunology Program, Faculty of Medicine, University of Helsinki**

# OVERVIEW

- **Basics**
  - Clinical genome-wide methods
  - Structural variants
  - Germline variant calling in rare diseases
  - Future
- 

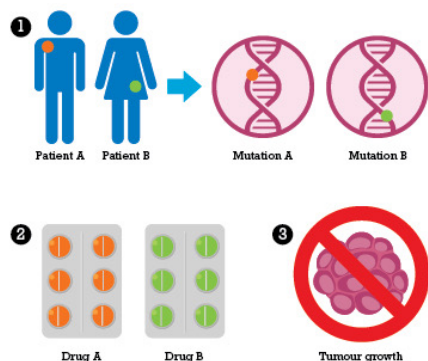
# GENETIC STUDY

- The study of a sample of DNA to identify genetic variants at chromosomal or nucleotide level
  - Used to specify and confirm diagnoses, understand pathogenesis, choose therapies, monitor treatment success, monitor disease progress, identify patients at risk of developing a disease, etc
  - Long lasting / Permanent consequences on patients and their relatives
  - Key role in precision medicine
- 

# APPLICATION AREAS

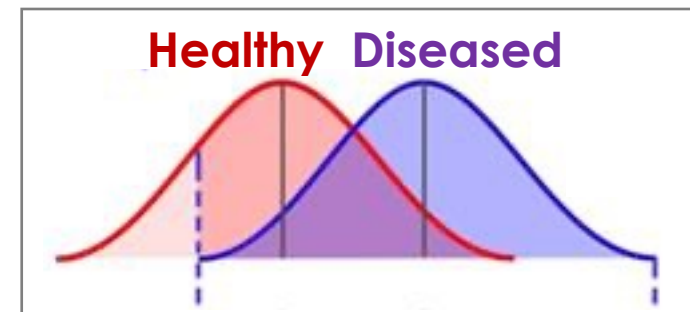
**Rare diseases:** diseases that are individually rare (1/2000) but collectively very frequent (8% of the population). 80% have genetic cause

**Separately (1/2000)**      **Collectively (1/17)**



**Cancers:** 40% of people will have a cancer during their lifetime. Genetic data important in identifying individuals at risk, choosing therapy, monitoring treatment outcomes, and confirming diagnosis

**Screening of individuals at risk of developing endemic diseases:** Reduction of diabetes incidence by 20% would save in 743M€ in healthcare expenses in Finland every year



# GOOD GENETIC TEST

**Agile** - Usable to different sample types and diseases

**Rapid** - Results must be delivered rapidly to the clinic (days/weeks)

**Sensitivity** - True pathogenic variants identified correctly

**Specificity** - False predictions a major cause of delay

**Cost efficient** - Unclever to waist limited financial resources


**Validated** - All clinical tests must be to validated by the laboratory

**Reproducible** - Re-analysis produces concordant results

**Reliability** - Instruments / Algorithms don't get broken / crash

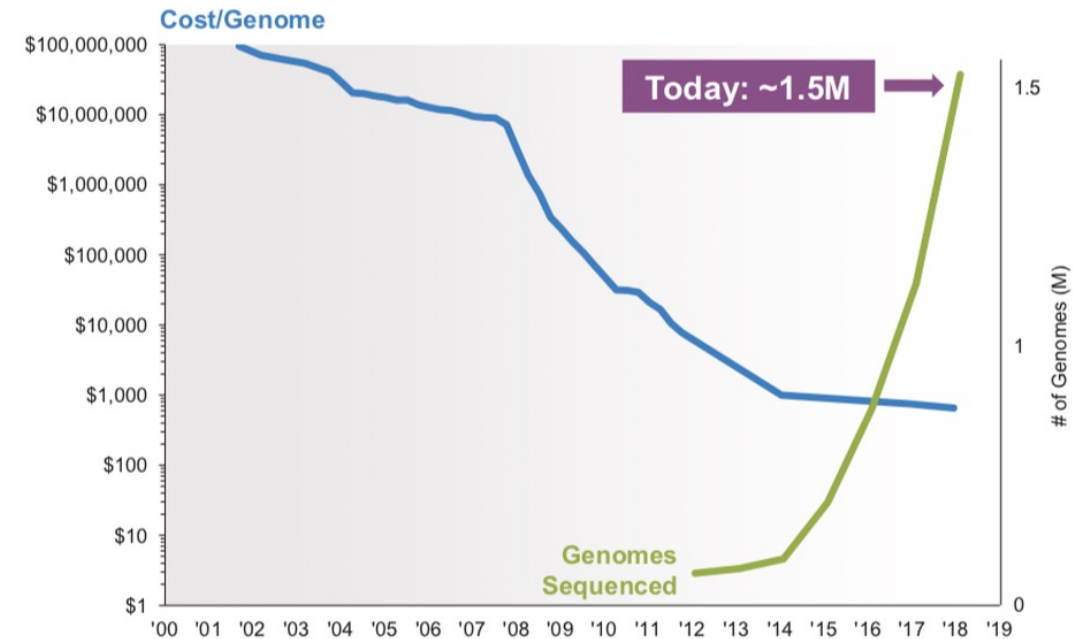
**Standardized** - Same protocols / algorithms / parameters applied

# OVERVIEW

- Basics
  - **Clinical genome-wide methods**
  - Structural variants
  - Germline variant calling in rare diseases
  - Future
- 

# HIGH-THROUGHPUT SEQUENCING

- High-throughput sequencing is the process of identifying the sequences of vast numbers of short DNA fragments in parallel. Used in increasing levels also in clinic
- High output. Barcoding allows to analysis of hundreds of samples per analysis run
- Highly accurate. Multiple independent interrogations made for each region of interest
- Relative rapid turnaround time (2-14d). Costs ~300 € per sample



# SEQUENCING INSTRUMENTS USED IN CLINIC

**Illumina  
Novaseq™ 6000**



**High yield (250-3000 Gb)  
2x150 bp  
20-500 WES/WGS  
Rapid (24h)**

**Ion Torrent S5  
prime**



**Low yield (10-50 Gb)  
200-400 bp  
~100 TS (20-400 genes)  
Rapid (24h)**

**Oxford  
Nanopore**



**Low yield (42 Gb)  
20 kbp  
Single molecule  
Runtime data-analysis  
Direct RNA-sequencing  
DNA/RNA modifications  
Rapid (16-72h)**

**Pacbio  
RS II**

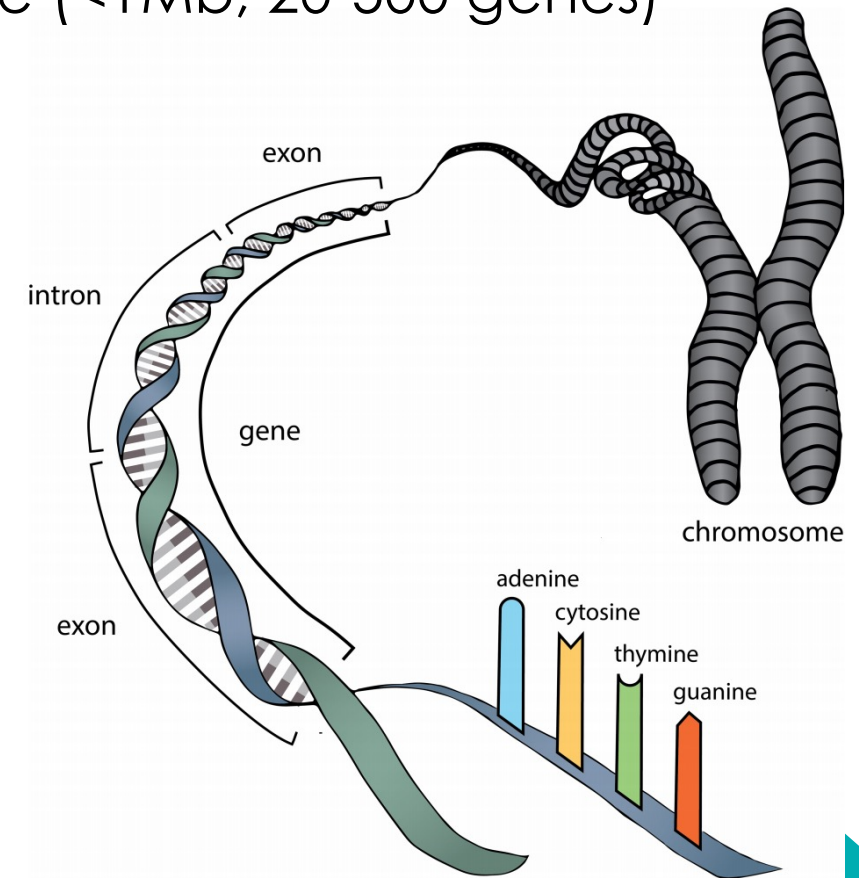


**Low yield (5 Gb)  
10-16 kbp  
Single molecule  
Rapid (10h)**



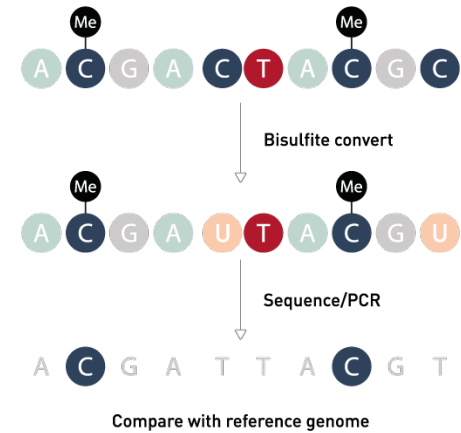
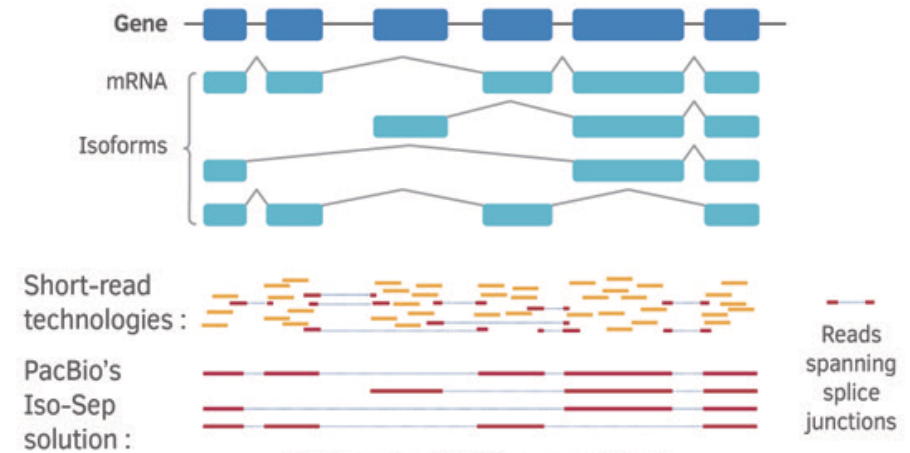
# CLINICALLY RELEVANT SEQUENCING STRATEGIES

- Gene-panel sequencing (TS)
  - Defined set of regions associated with a disease (<1Mb, 20-500 genes)
  - Up to ultra-high depth (~5000X), price ~500€
- Whole-exome sequencing (WES)
  - Protein coding regions (~60Mb / 22,000 genes)
  - Cover ~85% of known pathogenic variants
  - Intermediate depth (~30-500X), price ~1500 €
- Whole-genome sequencing (WGS)
  - Whole genome (~3,000Mb / 60,000 genes)
  - Low depth (~30X), price ~2500 €



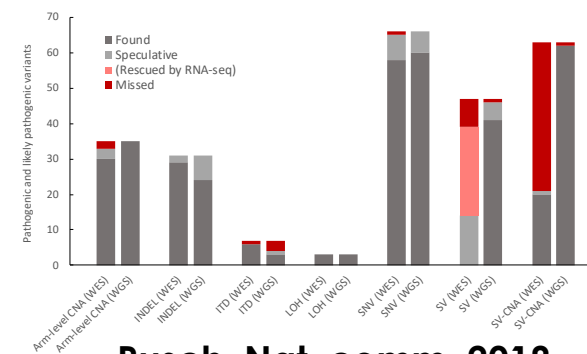
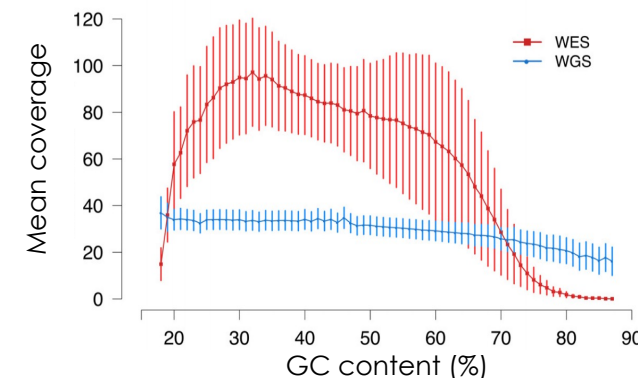
# CLINICALLY RELEVANT SEQUENCING STRATEGIES

- Transcriptome sequencing (TAS / RNA-seq)
  - mRNA + lincRNA, mRNA, or total-RNA
  - Fusion genes, differential expression
  - High depth (200-1000X), price 300 €
  
- Methylation sequencing
  - Methylated regions / Whole genome
  - Hyper- and hypomethylation cytosines
  - Intermediate depth (~30-500X), price ~500 €



# WHOLE-GENOME SEQUENCING

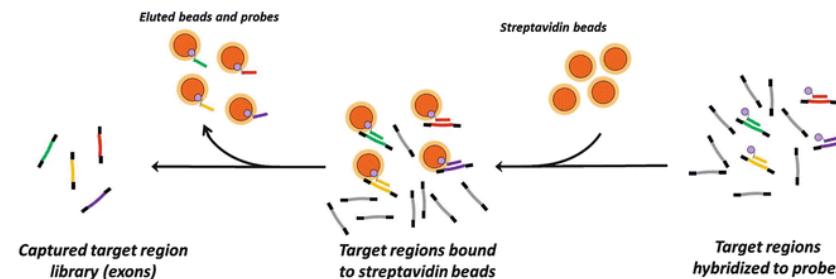
- WGS provides most homogenous and stable coverage across the genome. Most suitable for understanding extreme GC-regions
- Superior in identifying structural variants and large-scale genome instabilities like chromothripsis, kataegis, and chromoplexy
- At similar read depths WES and WGS identify SNVs and indels equally accurately. Price still limits the use of WGS in detection of (somatic) short variants (e.g. 100X = 2000€)



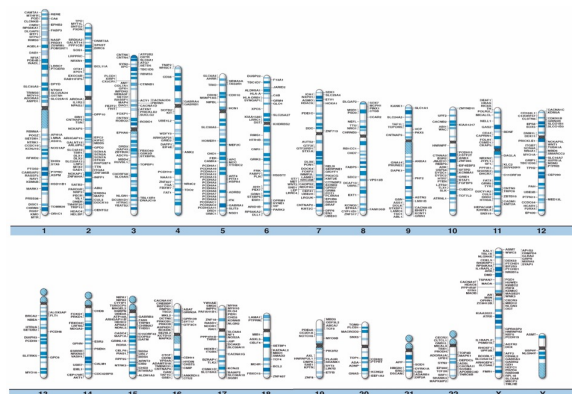
Rusch, Nat. comm, 2018

# WHOLE-EXOME SEQUENCING

Methodologically as WGS, but protocol involves enrichment of coding and regulative regions (1-2% of the genome) by hybridization

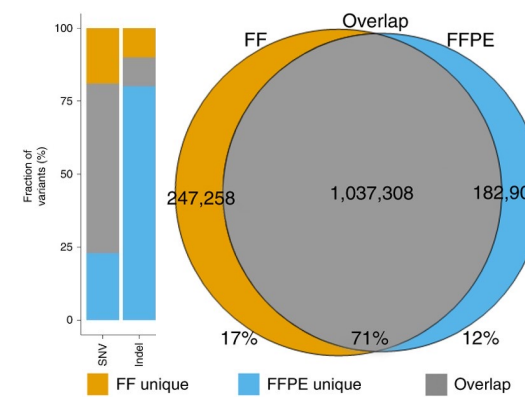


792 autism spectrum disorder genes



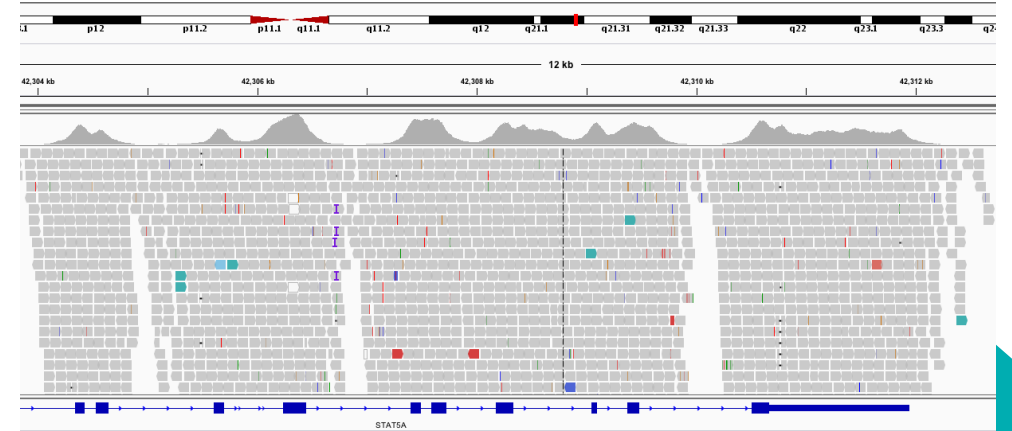
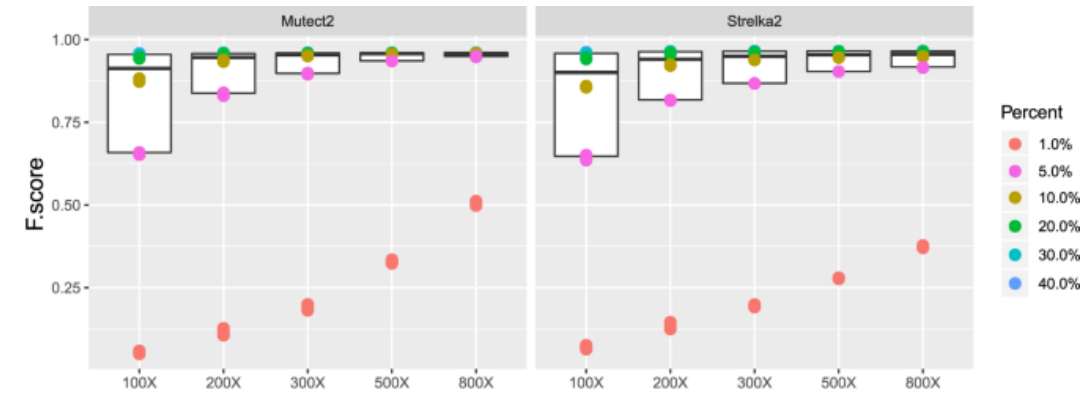
Captures coding and regulative regions of almost all genes and thereby preclude time consuming and laborious preselection of target genes

Suitable also to FFPE and other poor-quality samples common in clinic



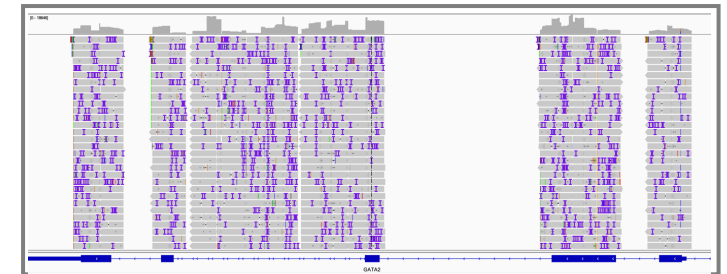
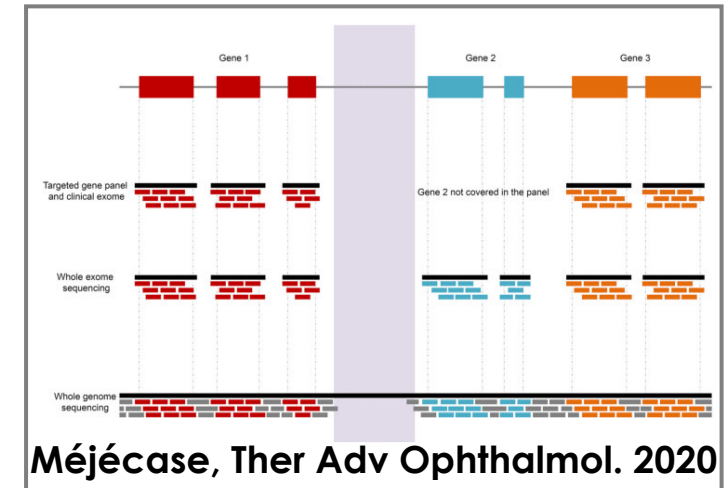
# WHOLE-EXOME SEQUENCING

- The cost-efficiency of WES entails use of greater sequencing depth that improves accuracy and detection of variants present only in a fraction of cells in comparison to WGS
- Coverage variation and absence of reads spanning structural variant (SV) breakpoints make identification of SVs other than copy-number variants (CNV) challenging



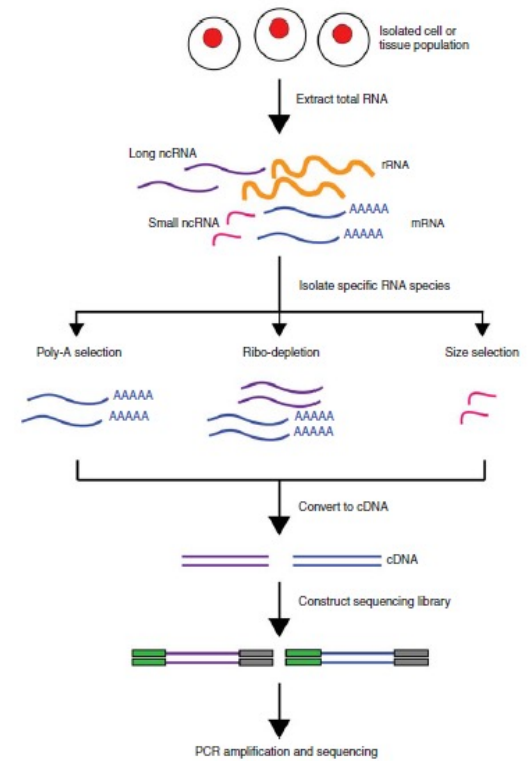
# GENE-PANEL SEQUENCING

- Focuses on preselected clinically relevant target regions. Enrichment by hybridization and/or PCR-amplification
- Most panels include exons of 50-500 genes and are disease and/or disease-set specific
- Superior in identifying somatic and mosaic variants present in a fraction of cells. Limited ability to detect SVs and chromosomal aneuploidies
- Most suitable to FFPE and other poor-quality samples common in clinic




# TRANSCRIPTOME SEQUENCING

- RNA is the source material. Includes often selection of poly-A mRNA, depletion of rRNA and/or depletion of globin RNA
- Samples should be preserved immediately using products like RNAlater, PAXgene, etc
- Used to detect gene fusions, splice variants, deep intronic mutations, and allele specific expression of genes
- Transcriptome is dynamic and varies due to tissue, cellular conditions, and environment etc. Use of patient-matched controls or control-sets (>40 subjects) recommendable



# OVERVIEW

- Basics
  - Clinical genome-wide methods
  - **Structural variants**
  - Germline variant calling in rare diseases
  - Future
- 



# STRUCTURAL VARIANTS

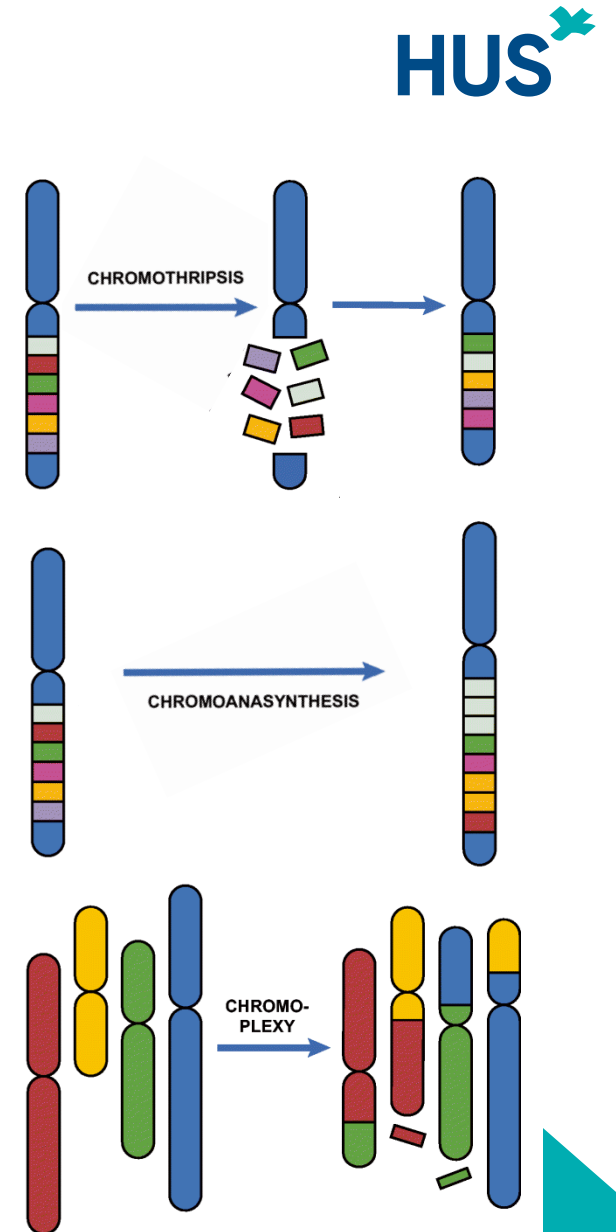
- Structural variation is generally defined as a DNA variant ~1 kb or more in size
- Includes inversions, deletions, duplications, translocations and insertions within and between chromosomes
- Duplications and deletions can also occur at genome, chromosome or chromosome arm level

	Assembly	Short Reads Mapping	Long Reads Mapping
Deletion			
Duplication			
Inversion			
Insertion			
Translocation			

Paired end read   Unmapped read   Split reads on the reference indicating SV type by its directions   Long read   Split long read

# CHROMOSOMAL ABNORMALITIES

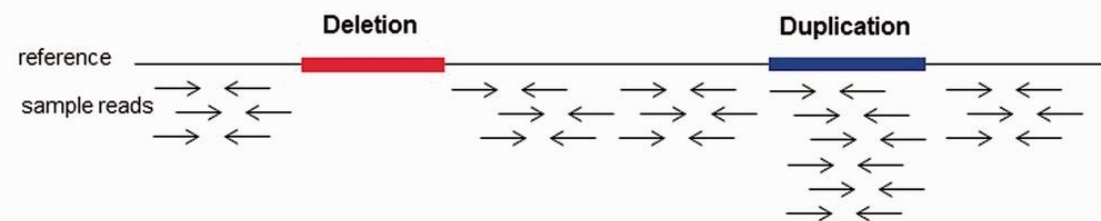
- In complex rearrangements multiple mutations occur in a single catastrophic event and result in a scrambled genome
- Include kataegis, chromoplexy, chromothripsis and chromoanasythesis
- Typically affects cancer cells. Prevalence across cancers varies from ~0% (e.g. leukaemia) to >50% (e.g. soft tissue cancers)



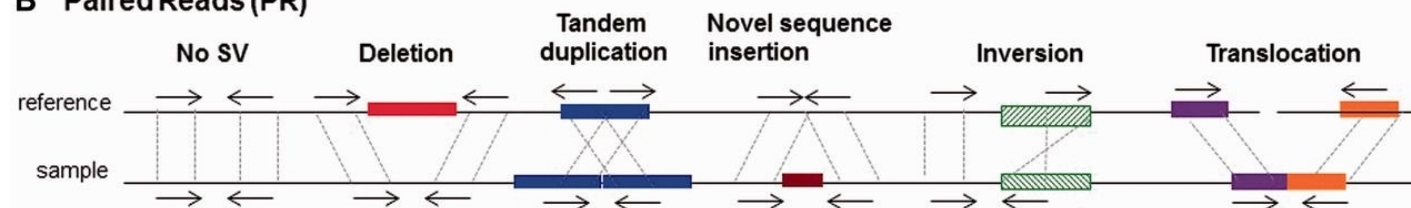
# STRUCTURAL VARIANT BIOINFORMATICS

- Four major analysis strategies available for SV detection
- Read-depth most useful for WES and TS. Paired reads, split reads, and *de novo* assembly mainly applicable to WGS data
- Solutions used in clinical diagnosis typically combine methods relying on different strategies

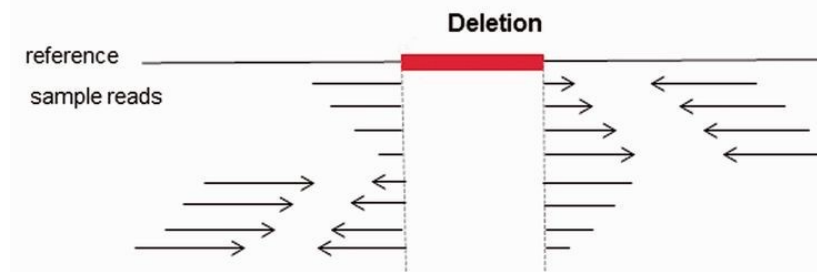
**A Read Depth (RD)**



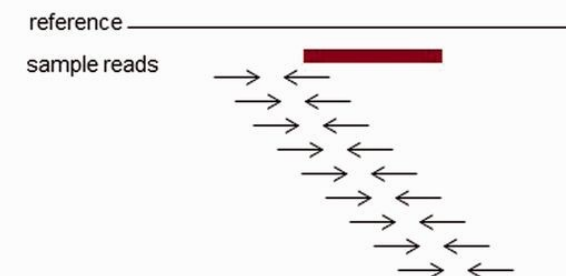
**B Paired Reads (PR)**



**C Split Reads (SR)**

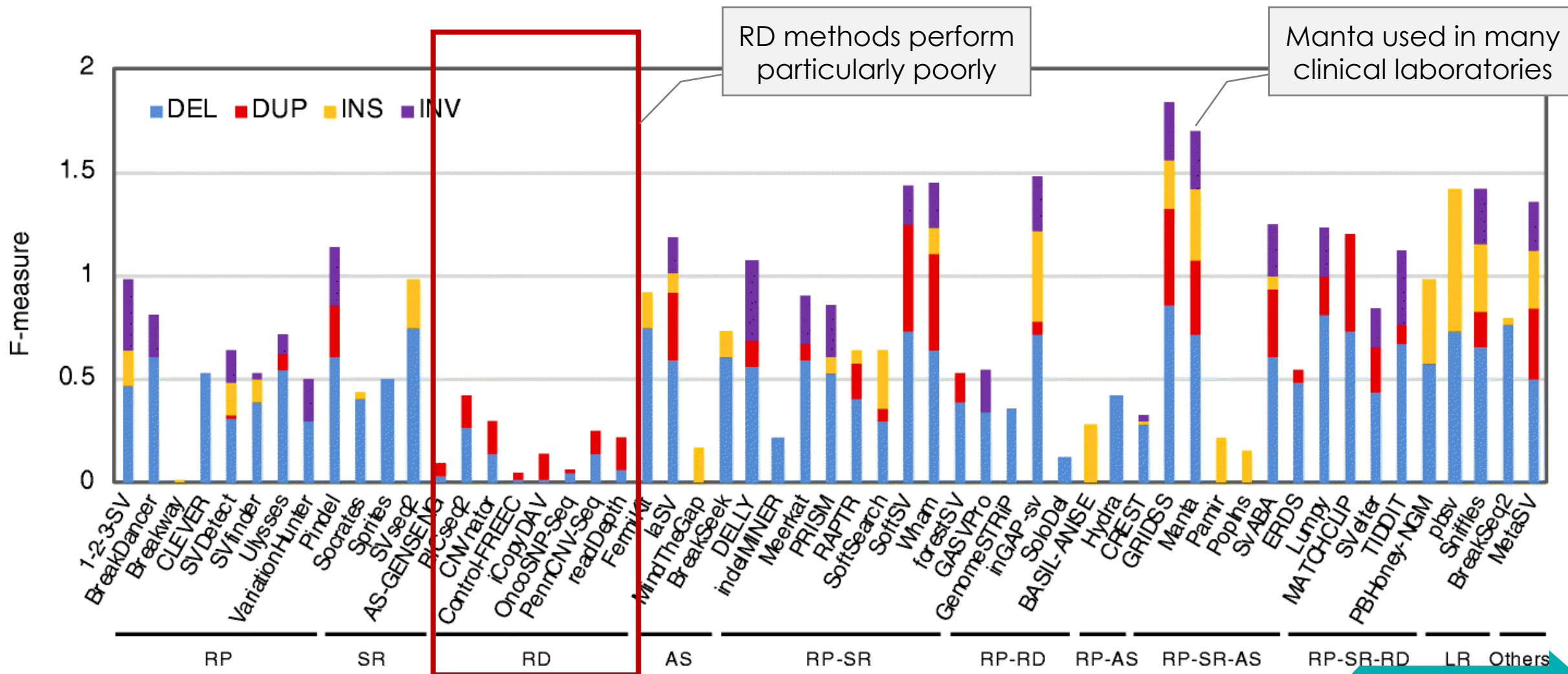


**D. De Novo Assembly (AS)**



Escaramís, Brief Funct Genomics, 2015,

# STRUCTURAL VARIANT CALLER PERFORMANCE

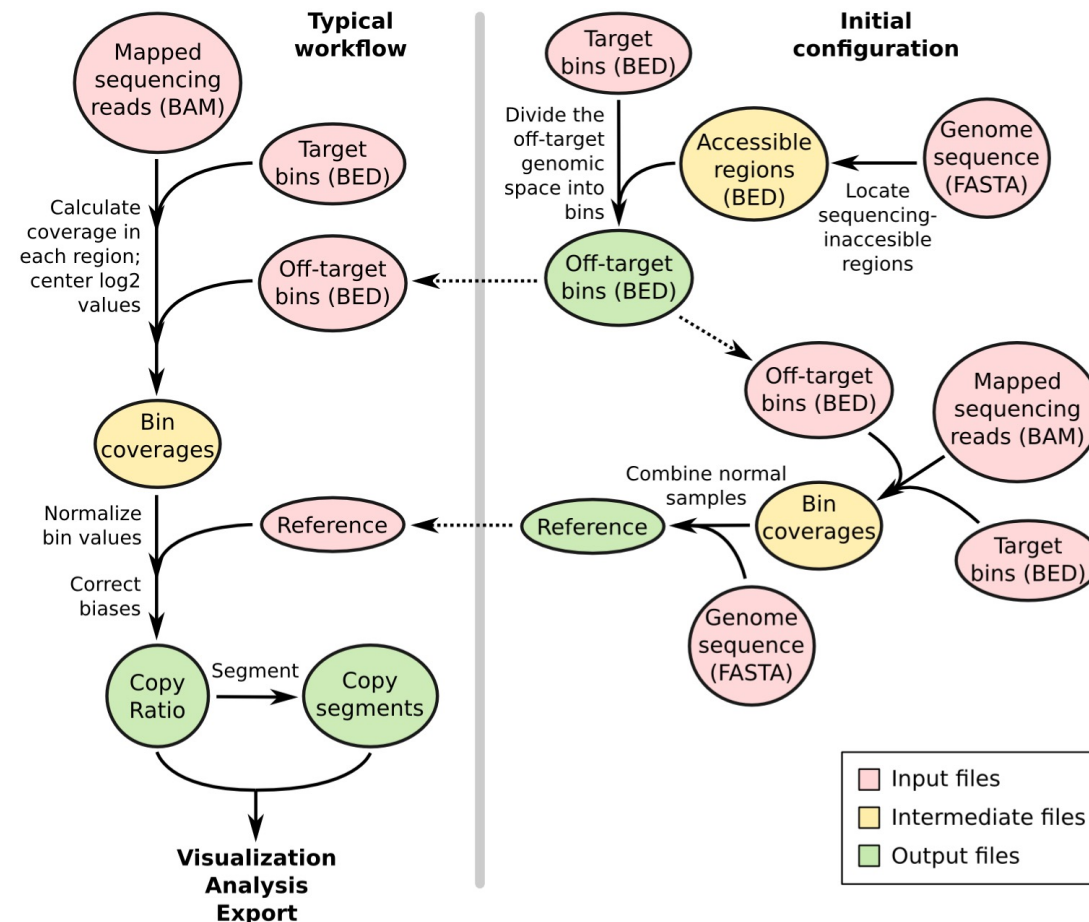


# READ DEPTH APPROACH

- The read-depth method allows to detect deletions and duplications. Based on the hypothesis of a correlation between the depth of coverage of a genomic region and the copy number of the region
- Works better on large-sized CNVs (>3 exons)
- Sensitivity and precision varies by coverage, platform, assay and tissue-type. Results confirmed at clinic using additional laboratory methods
- Most algorithms require large-size control background sample sets. Control samples must have been generated using the same platform and assays

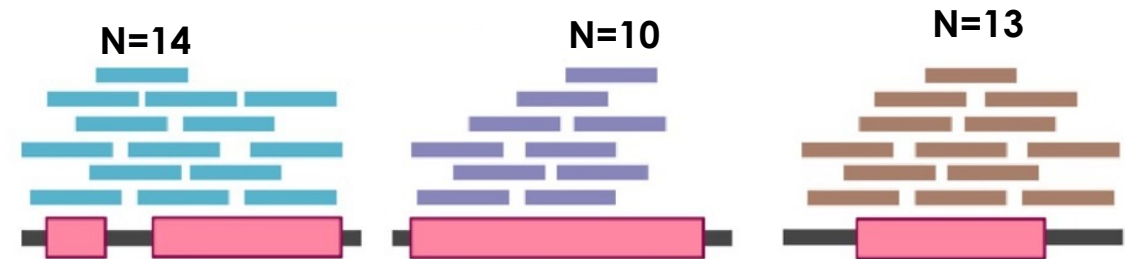
# SUMMARY OF READ DEPTH APPROACH

1. Map reads to the genome
2. Count number of alignments in each target interval
3. Remove biases (introduced by GC content, sample quality, total depth, mappability, exon capture efficiency etc) and latent systemic artifacts
4. Detect CNV boundaries by segmentation and output copy number estimates



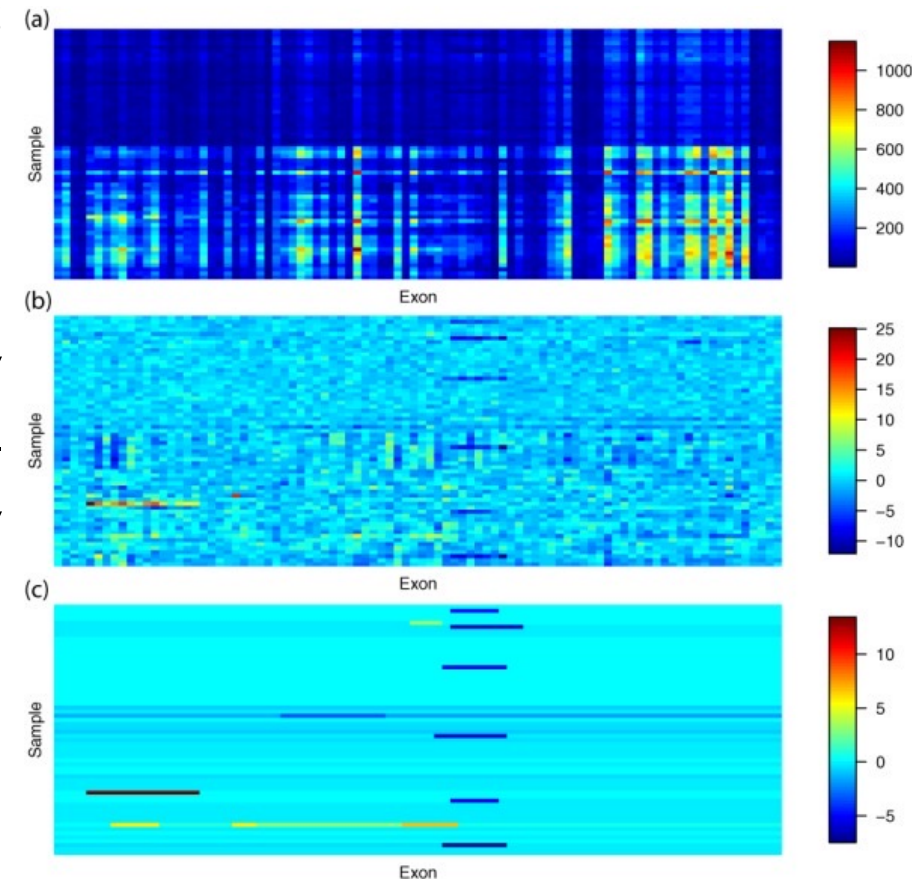
# READ BINNING

- Process of determining the number of reads at certain genomic loci
- Regions captured by the assay (with some padding) used typically in WES
- In WGS analyses, the reference is divided in larger (often equally sized and non-overlapping) bins
- Typically only unambiguously mapped reads used



# NORMALIZATION

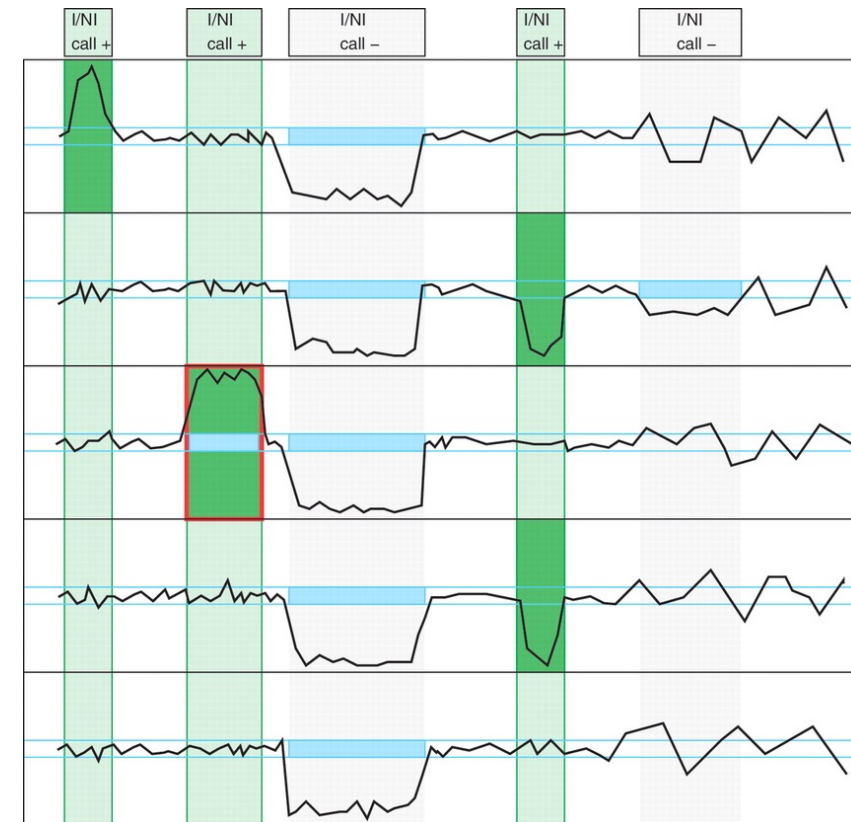
- Process of removing artifacts and biases introduced by GC content, sample quality, total depth, mappability, capture efficiency *etc*
- Achieved often with the help of a set of healthy reference samples using binomial or Poisson log-linear model. Identification of latent factors rely on singular value decomposition, PCA *etc*
- Methods not relying on reference samples make loess, polynomial *etc* fits between counts and genomic features






# SEGMENTATION

- Detection of variation along a chromosome to define CNV start and end sites
- Achieved in most methods using circular binary segmentation (CBS). Alternatives include Poisson likelihood-based segmentation.
- Algorithms not relying on reference samples make loess or polynomial fits between counts and genomic features



# OVERVIEW

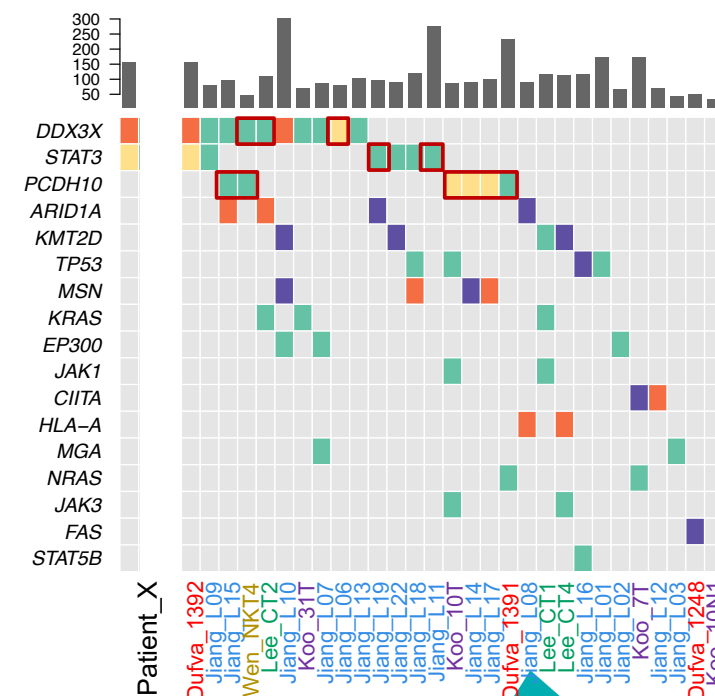
- Basics
  - Clinical genome-wide methods
  - Structural variants
  - **Germline variant calling in rare diseases**
  - Future
- 

# WES ANALYSIS OF RARE DISEASES

- Diagnosis of rare diseases often done with WES given the large number of potential candidate disease-genes and high contribution of exonic and splice-site (point) mutations
- WES analyses focus on SNVs and Indels Diagnosis achieved in ~30-70% of monogenic cases (+40% in comparison to conventional methods)
- SV/CNV analysis increases diagnostic yield 10-20%. External methods (e.g. MLPA or DDPCR) used to confirm SV/CNV findings
- May reveal incidental genetic findings unrelated to the initial indication. Current standard is to report those in 81 medically actionable genes

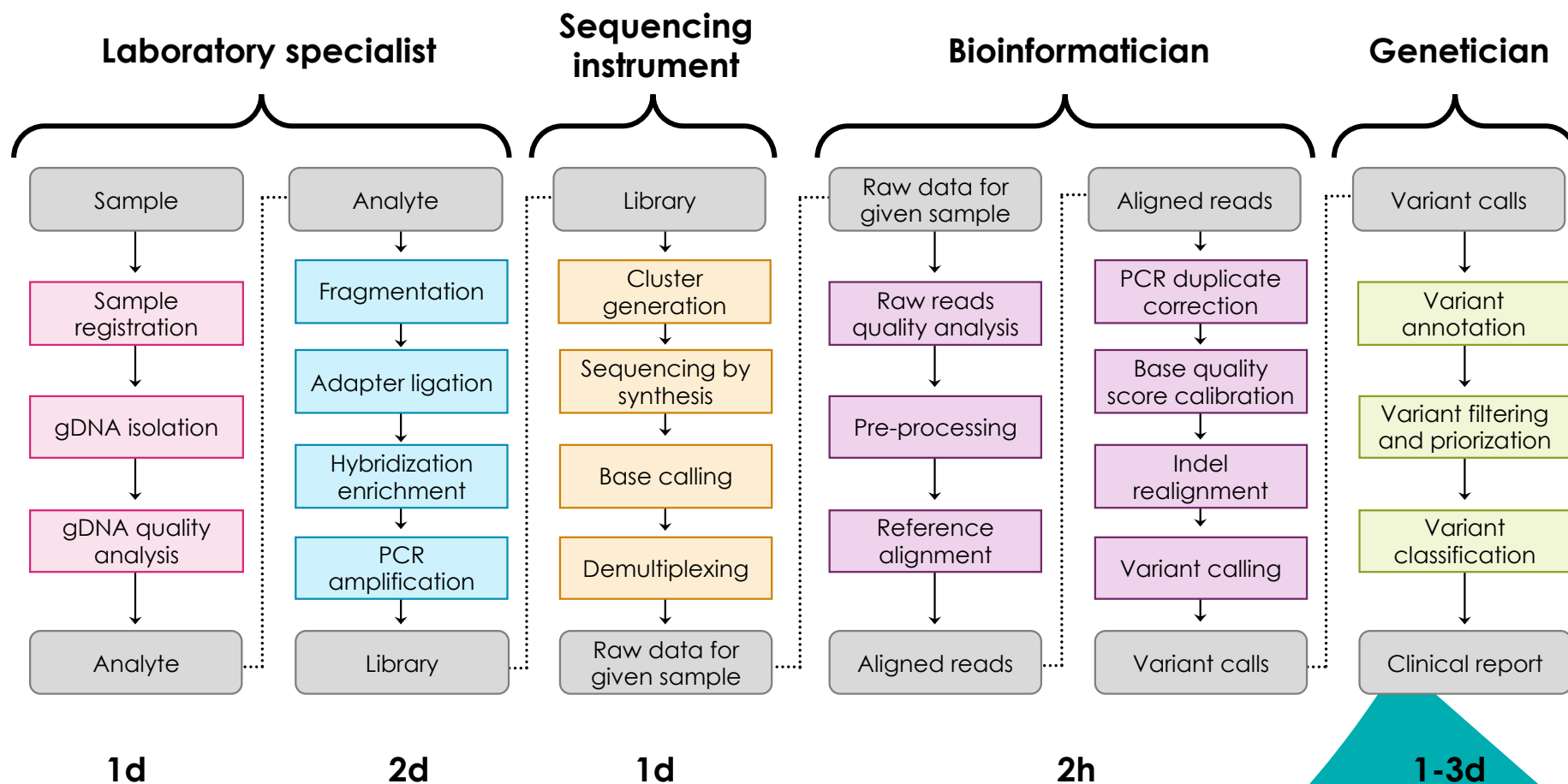
# CLINICAL VS. RESEARCH GENOMICS

- Cohort size:** research cohorts can include thousands of study subjects. Clinical sequencing applied to single individuals (hopefully with controls)
- Sample material:** blood and fresh-frozen samples often used in research project. Clinical sequencing typically applied to blood / FFPE (Formalin Fixing with Paraffin Embedding) samples
- Methods:** Only well-established computational and lab methods should be used in clinical diagnosis. Any inspiring method can be used in research context



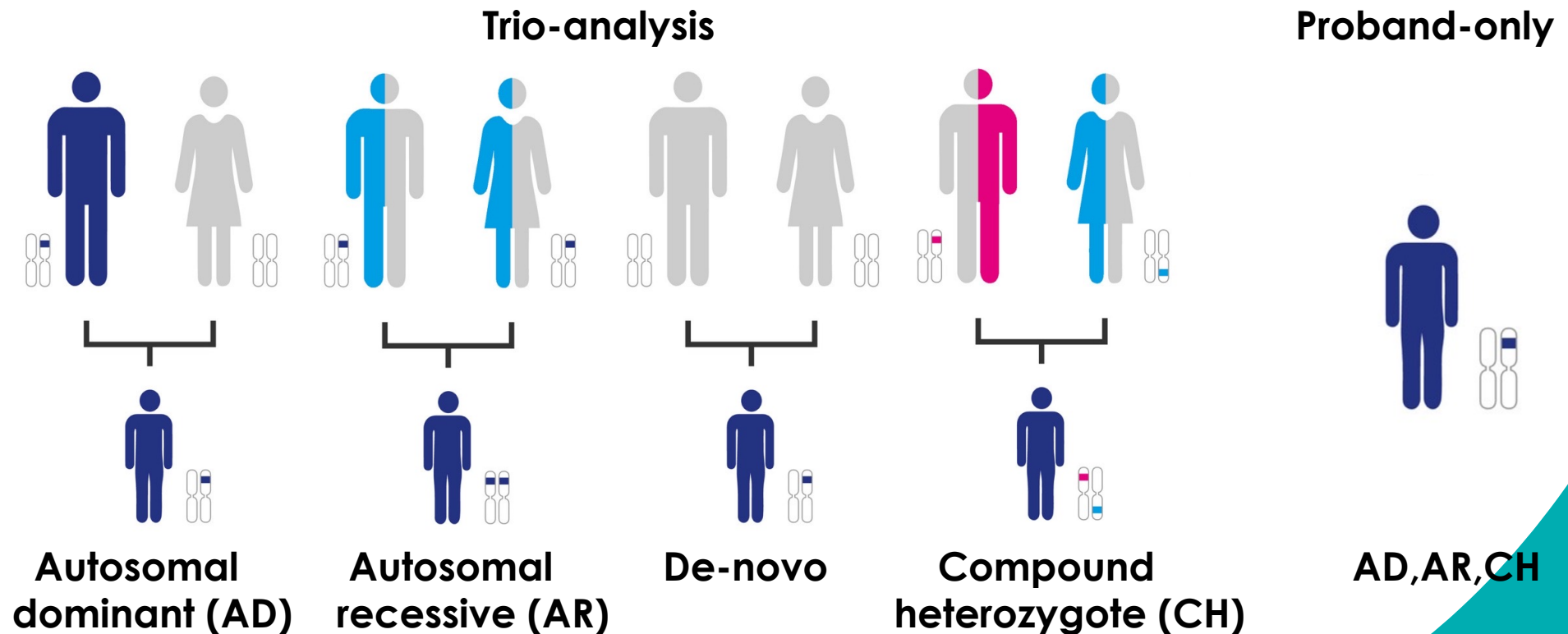
# LABOUR INTENSIVE PROCESS

- Total turnaround time ~10-28 days
- Hands-on time ~5-7 days
- ~50% faster than traditional tests

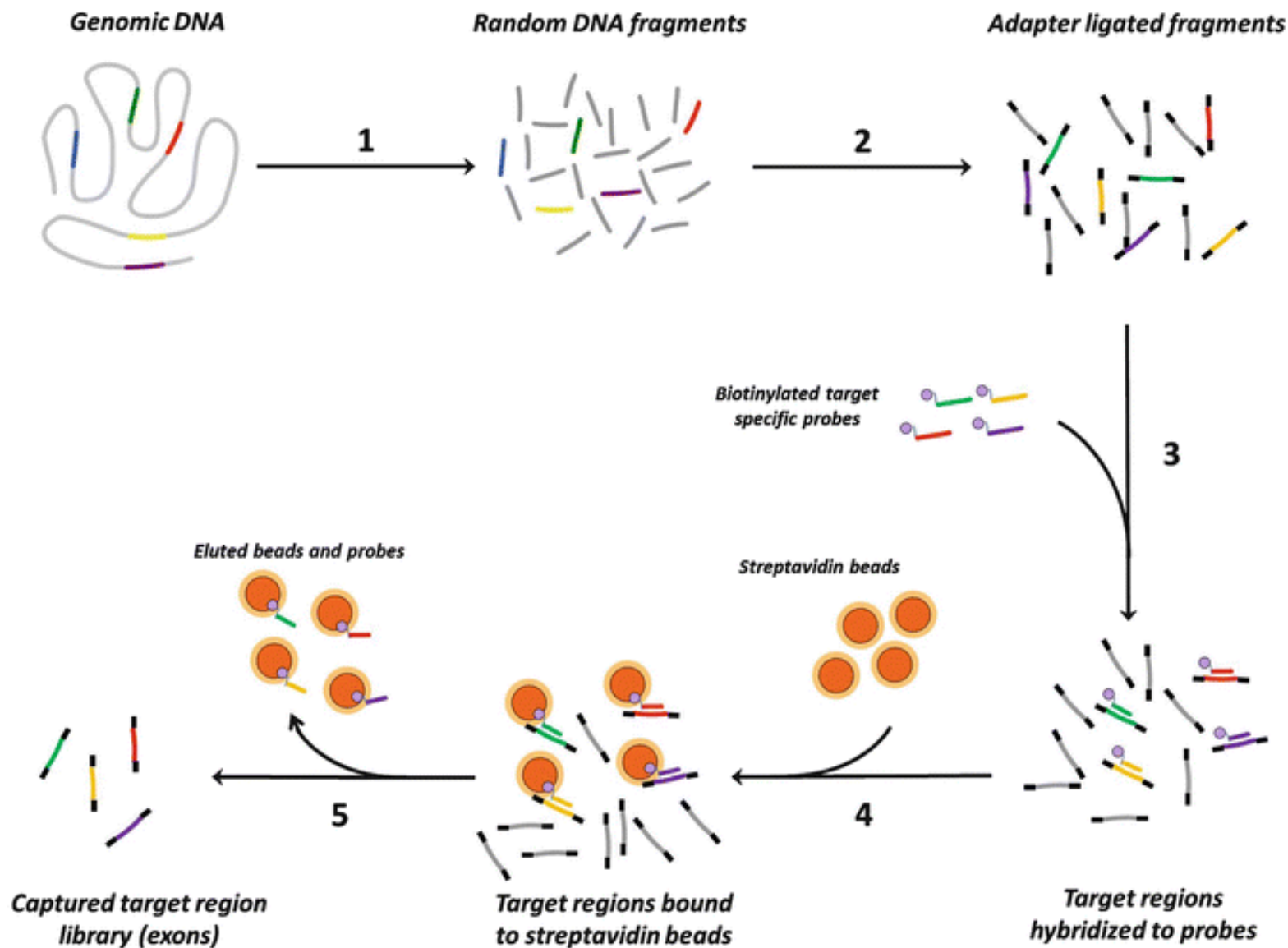


# EXPERIMENTAL DESIGN

- Trio sequencing of the proband and his/her relatives increases diagnostic yield ~10% and streamlines analysis. Enables to filter familial variants and finding de-novo and compound heterozygous events

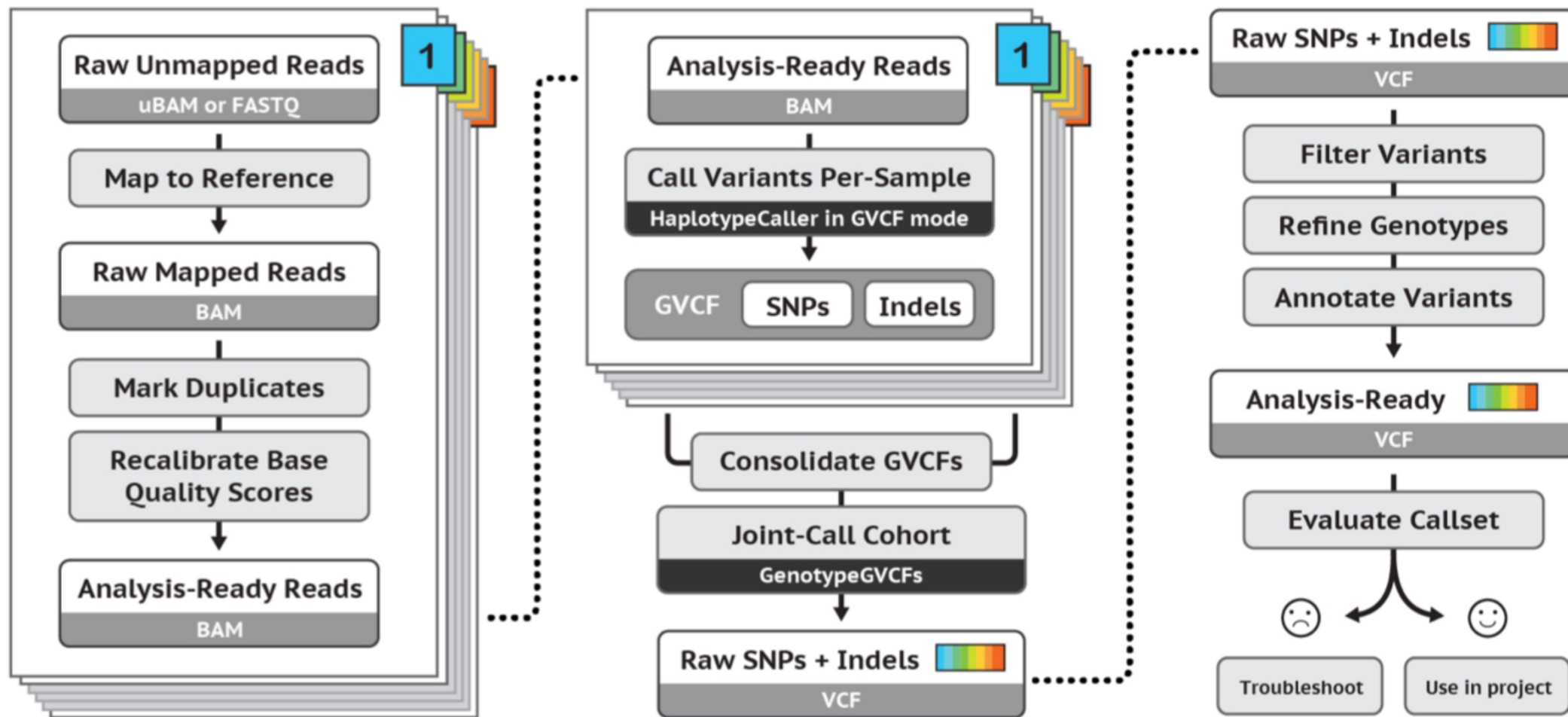


# WES LIBRARY PREPARATION AND SEQUENCING



1. Enzymatic Fragmentation
2. Addition of adapters/barcodes
3. Pre-amplification
4. Pooling of libraries
5. Hybridization
6. Target capture
7. Post-amplification
8. Illumina paired-end sequencing

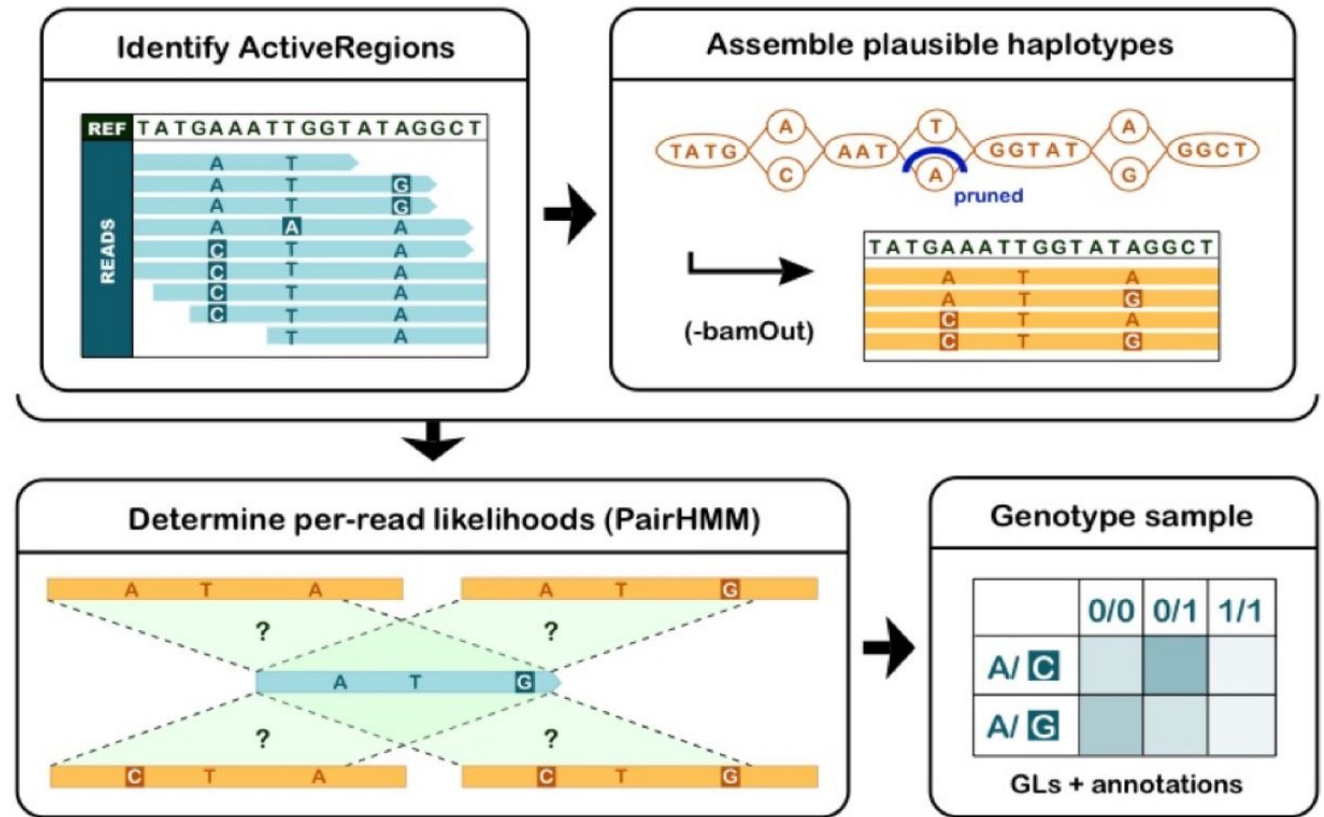
# SMALL VARIANT CALLING PROCESS





# SMALL VARIANT CALLING

1. Identify regions with alignments with mismatch evidence
2. Discover plausible haplotypes by performing de Bruijn-like graph construction
3. Determine the read-support of haplotypes with paired Hidden Markov
4. Define the genotype




# SMALL VARIANT CALLING


- Most small variant callers discover germline variants with exceptional recall (0.996) and precision (0.998). Variant calling performance however varies across genomic regions and is lower on difficult-to-map regions (including segmental duplications etc) and MHC loci

Technology	Genomic region	Participant	Performance metrics			F1 Rank		
			F1	Recall	Precision	All	Diff	MHC
MULTI	all <sup>a</sup>	Sentieon	0.999	0.999	0.999	1	4	1
MULTI	all <sup>a</sup>	Roche Sequencing Solutions	0.999	0.999	0.999	1	1	7
MULTI	all <sup>a</sup>	The Genomics Team in Google Health	0.999	0.999	0.999	1	2	4
MULTI	diff	Roche Sequencing Solutions	0.994	0.992	0.996	1	1	7
MULTI	MHC	Sentieon	0.998	0.998	0.998	1	4	1
ILLUMINA	all	DRAGEN	0.997	0.996	0.998	1	1	5
ILLUMINA	diff	DRAGEN	0.969	0.961	0.978	1	1	5
ILLUMINA	MHC	Seven Bridges Genomics	0.992	0.989	0.996	6	9	1
PACBIO	all	The Genomics Team in Google Health	0.998	0.998	0.998	1	2	4
PACBIO	diff	Sentieon	0.993	0.991	0.994	4	1	1
PACBIO	MHC	Sentieon	0.995	0.993	0.997	4	1	1
ONT	all	The UCSC CGL and Google Health	0.965	0.947	0.984	1	1	2
ONT	diff	The UCSC CGL and Google Health	0.983	0.976	0.988	1	1	2
ONT	MHC	Wang Genomics Lab	0.972	0.964	0.980	3	3	1

# VARIANT FILTERING

- Variant calling results in ~10M of variants. Tertiary analysis aims to short-list these variants into those with relevance for the disease and/or phenotype
    - Filtering of variant calls based on variant caller information (e.g. variants not supported by enough many reads, variant calls supported by low-quality reads, etc)
    - Filtering of variants overly common in population (MAF >5%)
    - Filtering of variants in intronic and/or other non-functional regions
    - Filtering of variants not associated with the disease phenotypes
- 

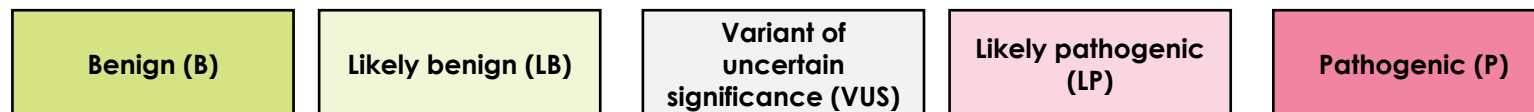
# SMALL VARIANT FILTERING

- All variants **10,000,000**
  - Remove low quality variants **380,000**
  - Remove intronic and intergenic variants **38,000**
  - Remove common variants with  $MAF > 1\%$  **960**
  - Remove synonymous and non-frameshift **700**
  - Return clinically significant variants filtered previously **800**
  - Remove variants not matching disease phenotype **20-40**
- 

# VARIANT CLASSIFICATION

- ACMG/ASCO standards recommend a five-tiered system for indicating variant pathogenicity based on 28 criteria. Classification 18/28 criteria is semi-automated
- Variant assessment requires segregation data on variants / diseases and literature information on variant pathogenicity
- Machine-learning based variant classification and text-mining of gene-disease-variant relationships from medical and scientific literature may ease the process in future
- Classification concordance typically high within laboratory (~78%) but low across different laboratories (~34%)

# ACMG VARIANT CLASSIFICATION



Databases like gnomAD, TopMen

Algorithms like CADD, Revel

In part from software like VEP

Databases like ClinVar, HGMD

	Benign			Pathogenic			
	Strong	Supporting		Supporting	Moderate	Strong	Very Strong
<b>Population Data</b>	MAF is too high for disorder <i>BA1/BS1</i> OR observation in controls inconsistent with disease penetrance <i>BS2</i>				Absent in population databases <i>PM2</i>	Prevalence in affecteds statistically increased over controls <i>PS4</i>	
<b>Computational And Predictive Data</b>		Multiple lines of computational evidence suggest no impact on gene /gene product <i>BP4</i> Missense in gene where only truncating cause disease <i>BP1</i> Silent variant with non predicted splice impact <i>BP7</i>		Multiple lines of computational evidence support a deleterious effect on the gene /gene product <i>PP3</i>	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before <i>PM5</i> Protein length changing variant <i>PM4</i>	Same amino acid change as an established pathogenic variant <i>PS1</i>	Predicted null variant in a gene where LOF is a known mechanism of disease <i>PVS1</i>
<b>Functional Data</b>	Well-established functional studies show no deleterious effect <i>BS3</i>			Missense in gene with low rate of benign missense variants and path. missenses common <i>PP2</i>	Mutational hot spot or well-studied functional domain without benign variation <i>PM1</i>	Well-established functional studies show a deleterious effect <i>PS3</i>	
<b>Segregation Data</b>	Non-segregation with disease <i>BS4</i>			Co-segregation with disease in multiple affected family members <i>PP1</i>	Increased segregation data →		
<b>De novo Data</b>					<i>De novo</i> (without paternity & maternity confirmed) <i>PM6</i>	<i>De novo</i> (paternity & maternity confirmed) <i>PS2</i>	
<b>Allelic Data</b>		Observed in <i>trans</i> with a dominant variant <i>BP2</i> Observed in <i>cis</i> with a pathogenic variant <i>BP2</i>			For recessive disorders, detected in <i>trans</i> with a pathogenic variant <i>PM3</i>		
<b>Other Database</b>		Reputable source w/out shared data = benign <i>BP6</i>		Reputable source = pathogenic <i>PP5</i>			
<b>Other Data</b>		Found in case with an alternate cause <i>BP5</i>		Patient's phenotype or FH highly specific for gene <i>PP4</i>			

Manual assignment



# DEEP LEARNING

- Various deep learning algorithms have in past few years emerged for automated ACMG classification of small variant calls
- Not much information available on algorithms and/or data that were used in model training
- Deep learning methods appear to perform (at least in our laboratory) number-wise well: 77% of 93 test cases solved, the mean rank of the causative variant 4.3, n 55% cases the causative variant ranked in the top 5. However, no productivity improvement observed

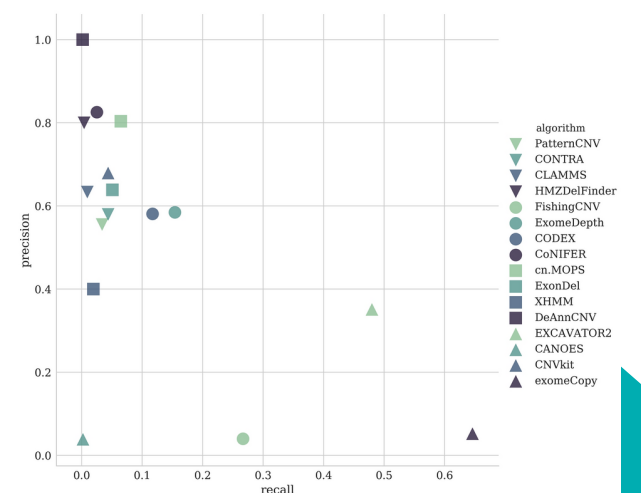
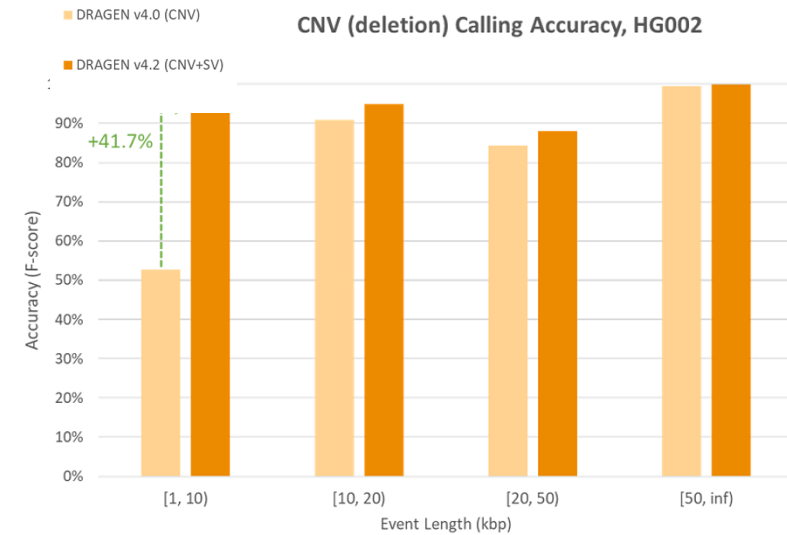


Nostos Genomics



# CNV CALLING

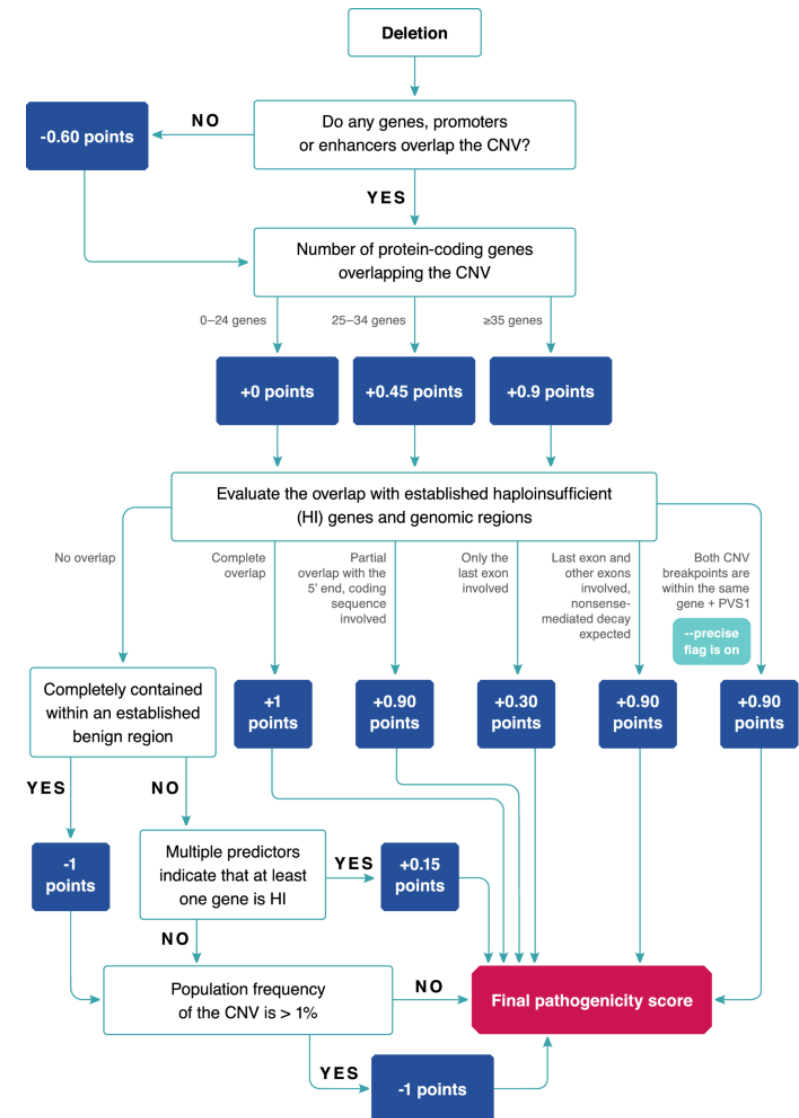
- 10-20% of rare diseases results from small-size (1-3 exons) deletions and duplications
- Read-depth strategy used most commonly due to the absence of reads that span CNV/SV breakpoints
- CNV calling accuracy less than that seen for small variants and results validated with external methods. Longer CNVs detected more reliably
- CNV analysis requires large numbers of sample-type matched control samples analysed using the same platform and assay






# CNV CLASSIFICATION

- ACMG/ASCO standards recommend a five-tiered system for indicating CNV pathogenicity
- Pathogenicity assessment builds upon semi-quantitative point-based scoring metric for CNV classification. Separate metrics exist for deletions and duplications
- Requires segregation data on variants and/or diseases, information on overlap with known benign and haploinsufficiency (i.e. two copies needed for normal function) regions and literature and case information

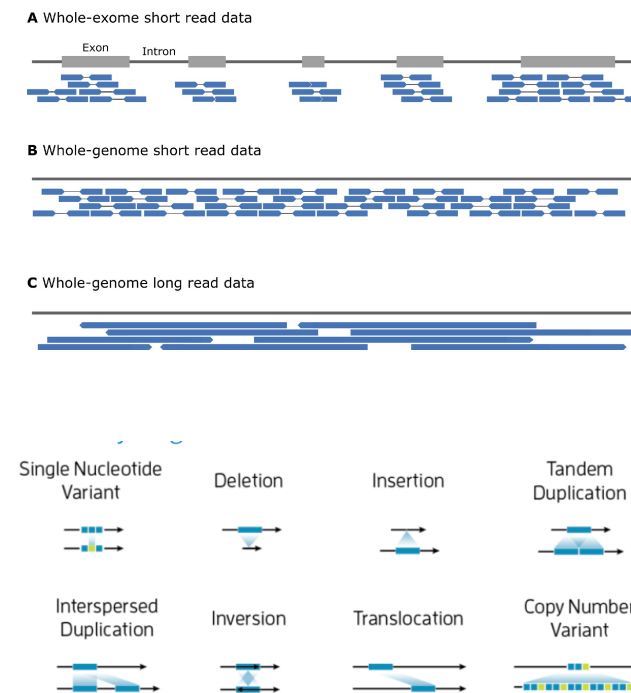


# OVERVIEW

- Basics
  - Clinical genome-wide methods
  - Structural variants
  - Germline variant calling in rare diseases
  - **Future**
- 

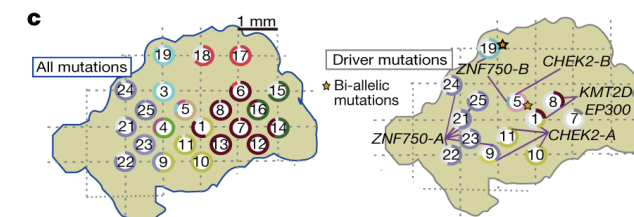
# FUTURE

- Even shorter turn-around-time through the use of GPU-accelerated algorithms allowing to complete WGS analyses in ~60 min. Machine learning will streamline the variant classification process
- Improvements in handling ultra-low DNA inputs and decreasing sequencing costs will make WGS the preferred strategy
- Long-read single-molecule sequencing techniques allow to detect all types of variants accurately and will become the preferred strategy

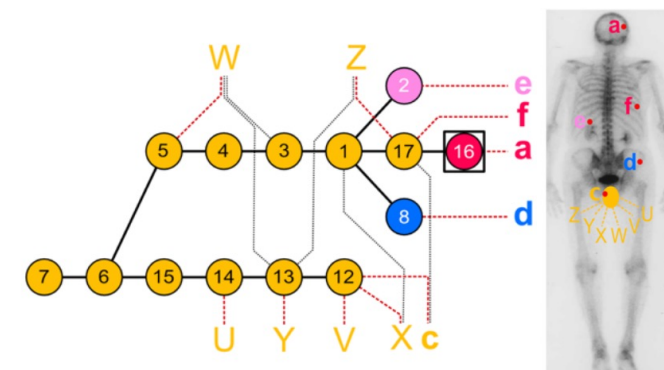


# FUTURE

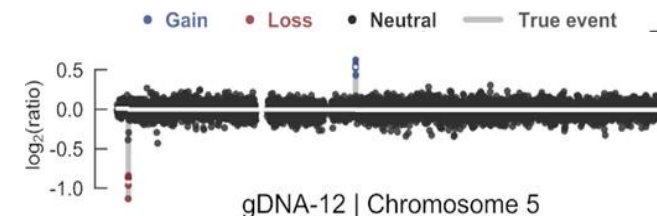
- Multiregional sequencing of tumours from tens of sites will improve cancer diagnosis and treatment and understanding of tumour heterogeneity
- Subclonal reconstruction and inference of cancer evolution to better understand onset of disease and mechanisms that acted then
- Ultra-low coverage (0.1-1X) WGS of cfDNA in CNVs and other variant analysis provides non-invasive means to diagnose and monitor cancer



Yokoyama, Nature, 2020



Woodcock, Nat. comm, 2020



Raman, NAR, 2018

**THANK YOU!**