HG471/BIOS21216: Copy Number Variations and De Novo Mutations

Prof. Xin He

March 1, 2016

Lecture Overview

Structural variation (SV) and copy number variation (CNV).

Detecting SVs and CNVs from microarray and sequencing data.

Role of CNVs in human diseases.

De novo mutations.

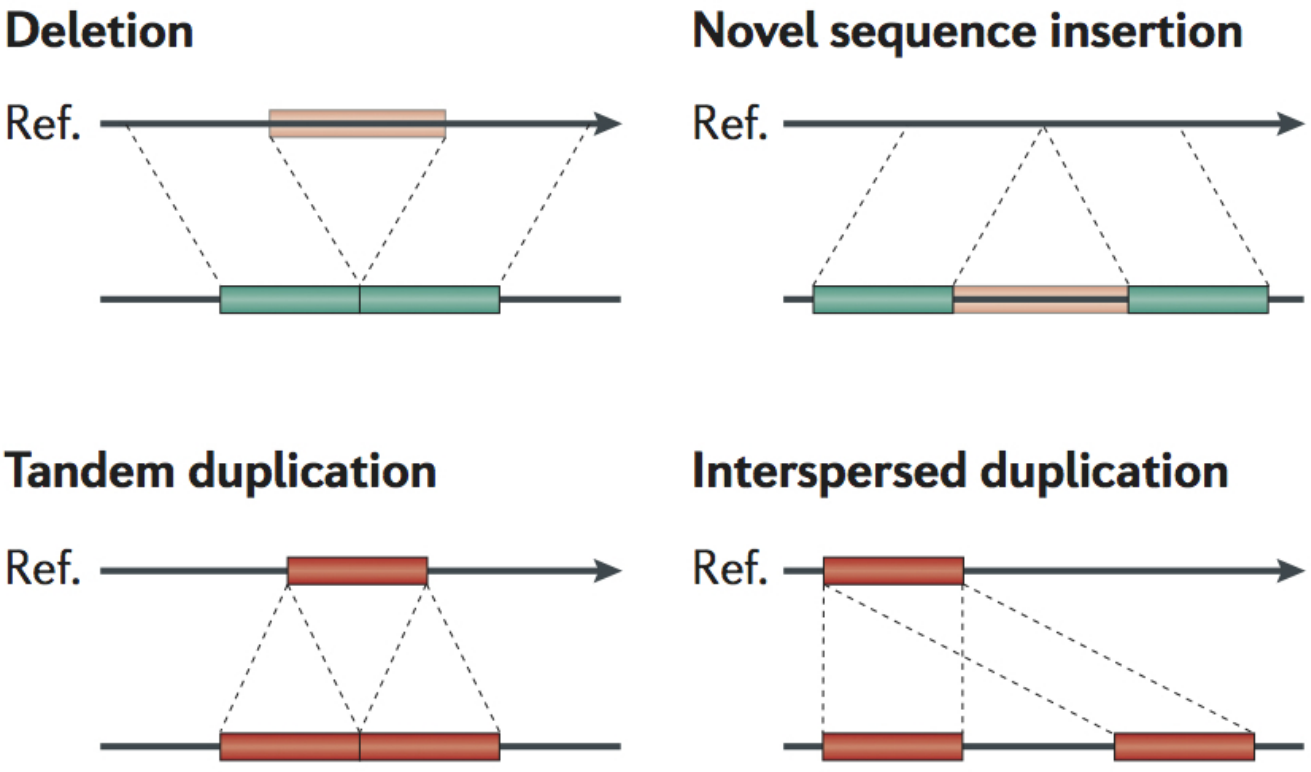
Structural Variations

SV: Variation of DNA regions, greater than > 50 bp in size, and often include deletions, insertions, copy number variations, and so on.

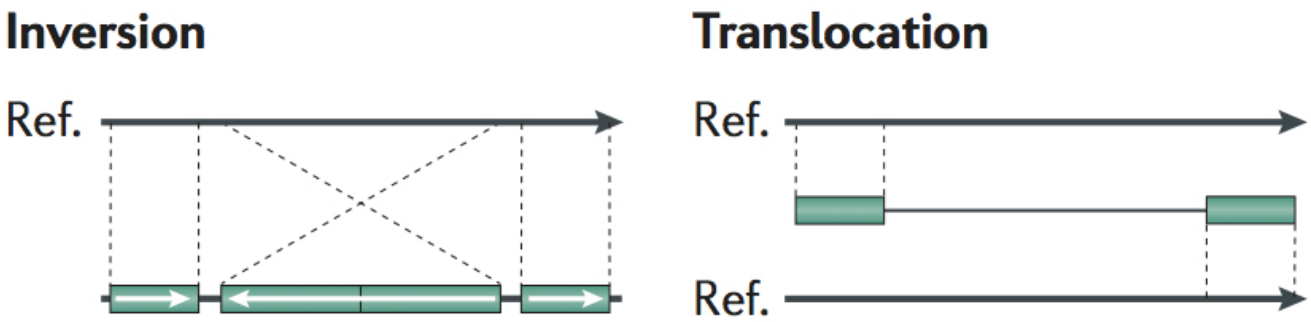
Structural variants are responsible for **more nucleotide differences** between individuals than SNVs: 0.5-1% for structural variants and 0.1% for SNVs.

Structural variants are important for a range of diseases.

Classes of Structural Variation

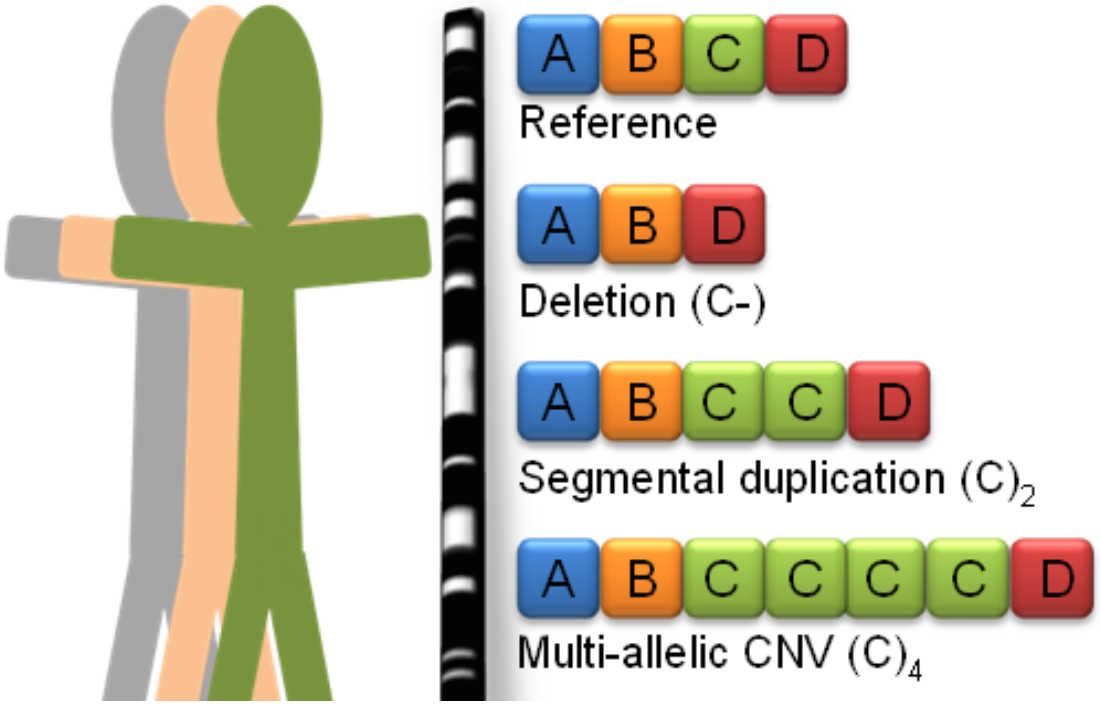


Classes of Structural Variation



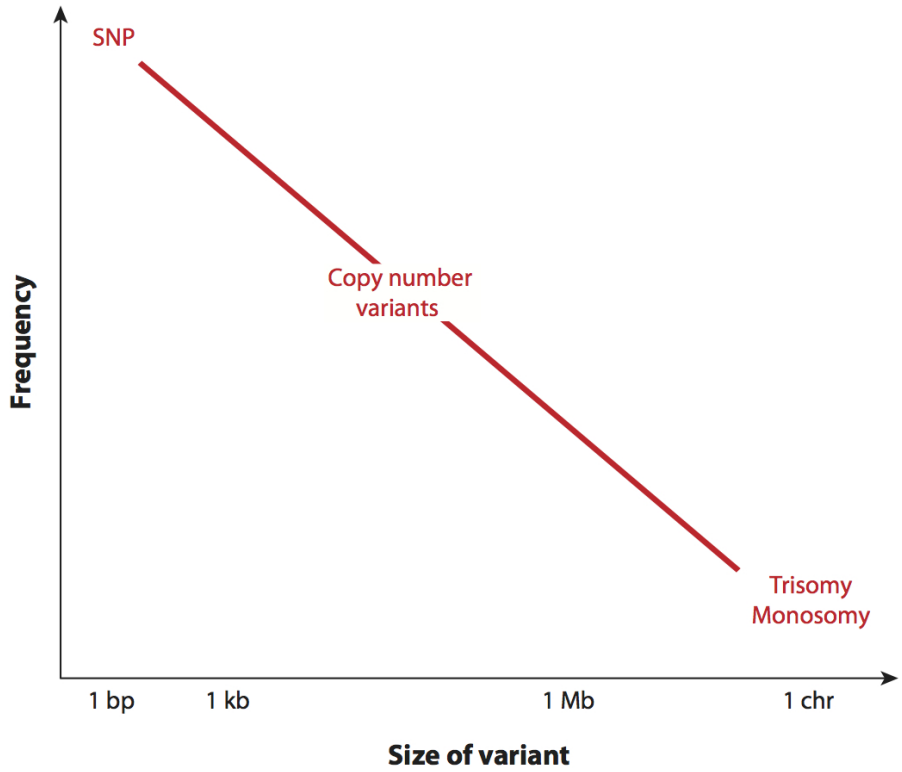
There are more complex SV events: e.g. inversion followed by deletion in one or both ends.

Copy Number Variations



The number of copies in a CNV could range from 0 to 30.

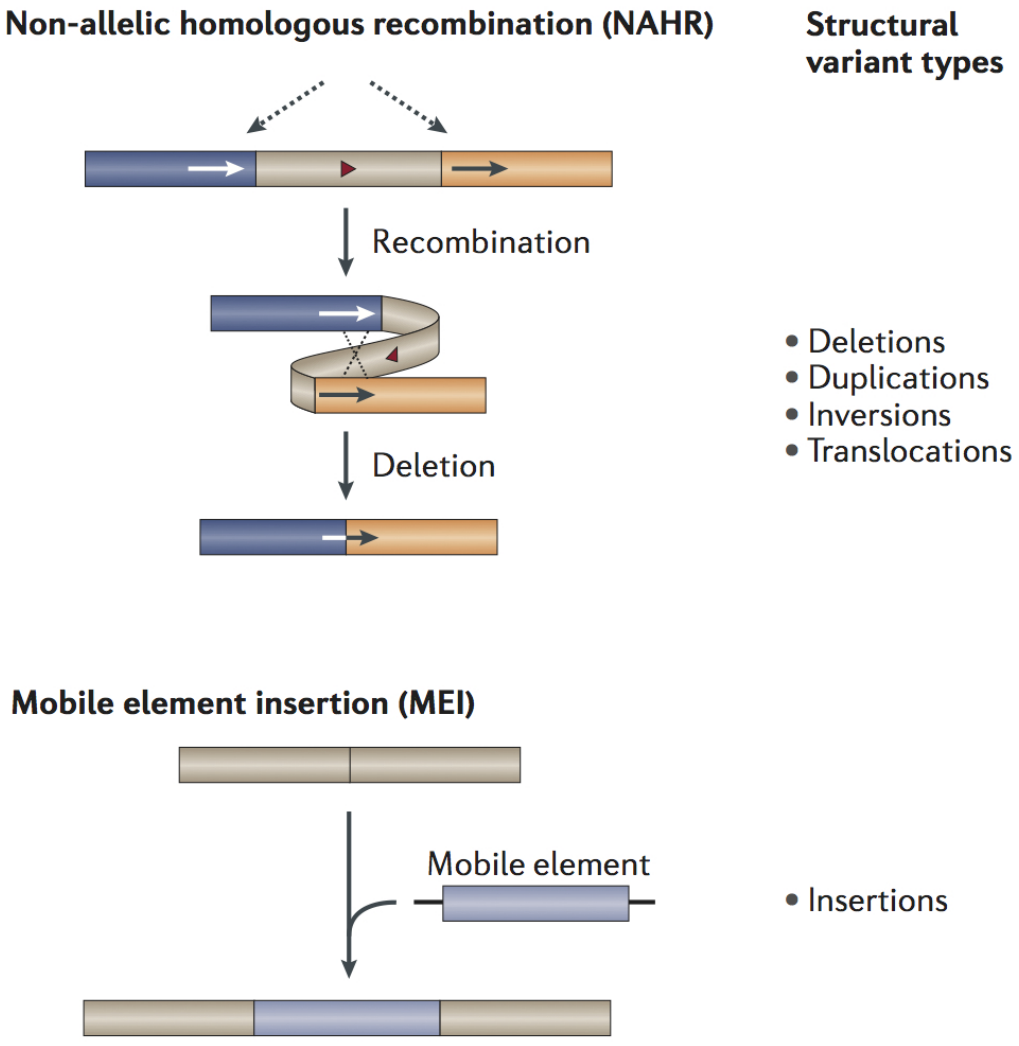
Size and Frequency of CNVs



Large CNVs are rare, in particular, large chromosome aberrations are often associated with major congenital abnormalities.

CNVs $>$ 20kb are studied more extensively because they are easier to detect.

Mechanisms of CNVs



NAHR: involves segmental duplications, large blocks ($>$ 1kb) of 95% or more sequence identity. Most common source of CNVs.

Other non-NAHR events can also contribute: MEIs, replication-based template switching, etc.

Rare CNVs and Copy Number Polymorphisms (CNPs)

Rare CNVs: often large, and many result from *de novo* mutations. They are **individually rare, but collectively common**:

- 65-80% of individuals carry a CNV $>$ 100kb.
- 5-10% of individuals harbor a CNV $>$ 500kb.
- About 1% of individuals carry a large CNV of 1Mb or larger.

Rare CNVs and Copy Number Polymorphisms (CNPs)

Rare CNVs: often large, and many result from *de novo* mutations. They are **individually rare, but collectively common**:

- 65-80% of individuals carry a CNV $>$ 100kb.
- 5-10% of individuals harbor a CNV $>$ 500kb.
- About 1% of individuals carry a large CNV of 1Mb or larger.

CNPs (not the focus of this class):

- Defined as CNV with frequency $>$ 1% in the population.
- Could be bi-allelic, or multi-allelic (0-30 copies). Often associated with segmental duplications.
- Many CNPs were associated with immune diseases.

Detection of SVs and CNVs: an Overview

Arry-based methods: rely on **hybridization** between sample DNA and probes. The signal intensity reflects copy number gain or loss.

- Array comparative genomic hybridization (CGH): 50-75bp long probes, representing millions of DNA segments in the genome.
- SNP arrays: detection of SNPs using allele-specific oligonucleotide (ASO) probes. Copy number events will change the signal pattern.

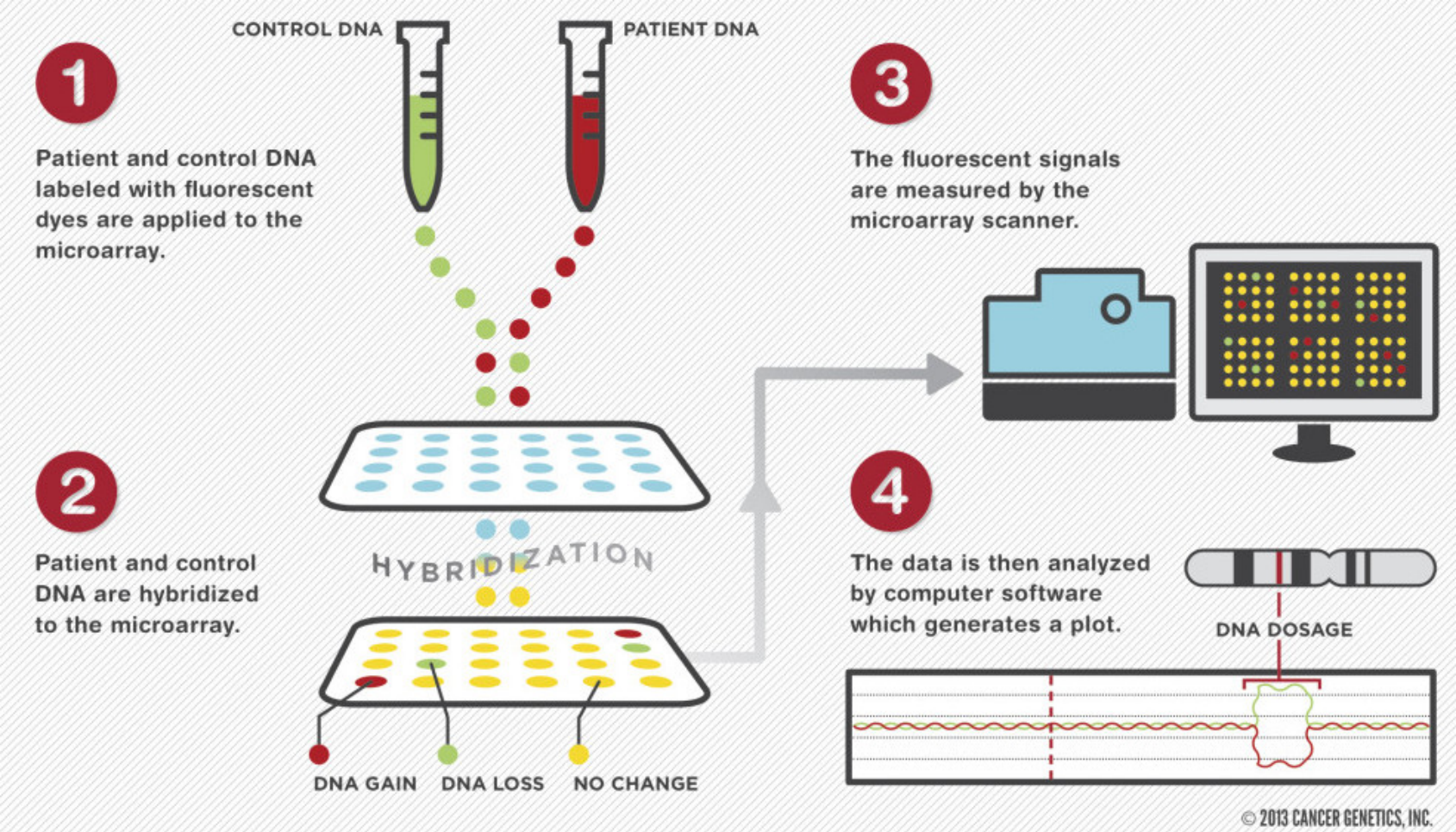
Detection of SVs and CNVs: an Overview

Arry-based methods: rely on **hybridization** between sample DNA and probes. The signal intensity reflects copy number gain or loss.

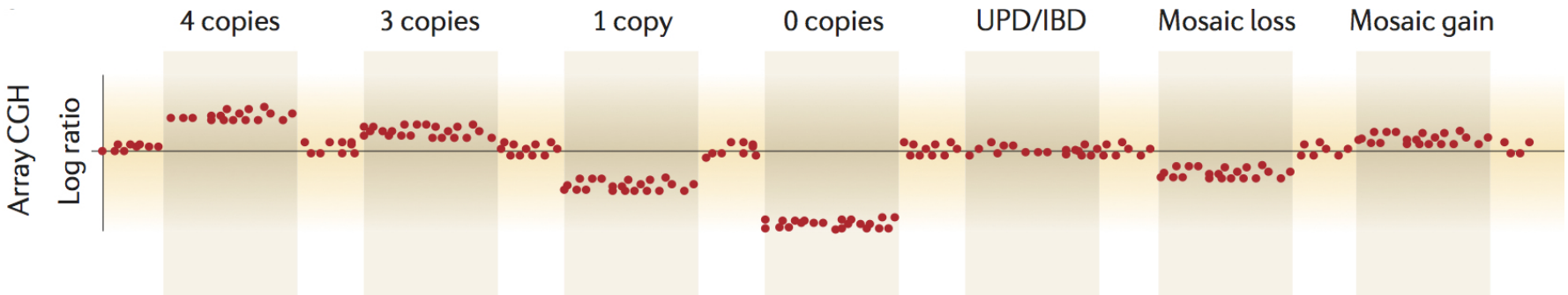
- Array comparative genomic hybridization (CGH): 50-75bp long probes, representing millions of DNA segments in the genome.
- SNP arrays: detection of SNPs using allele-specific oligonucleotide (ASO) probes. Copy number events will change the signal pattern.

Sequencing-based methods: identify various **signatures of SV events**, such as read depth changes and reads containing breakpoints.

The Process of Array CGH



Detection of CNVs from Array CGH Data



Popular aCGH platforms have 1-2M probes, and generally requires 3-10 probes to detect an event. The detection limit is about 20kb or larger.

Cannot detect copy neutral events such as segmental or chromosomal uniparental disomy (UPD).

SNP Arrays Can Be Used to Detect CNVs

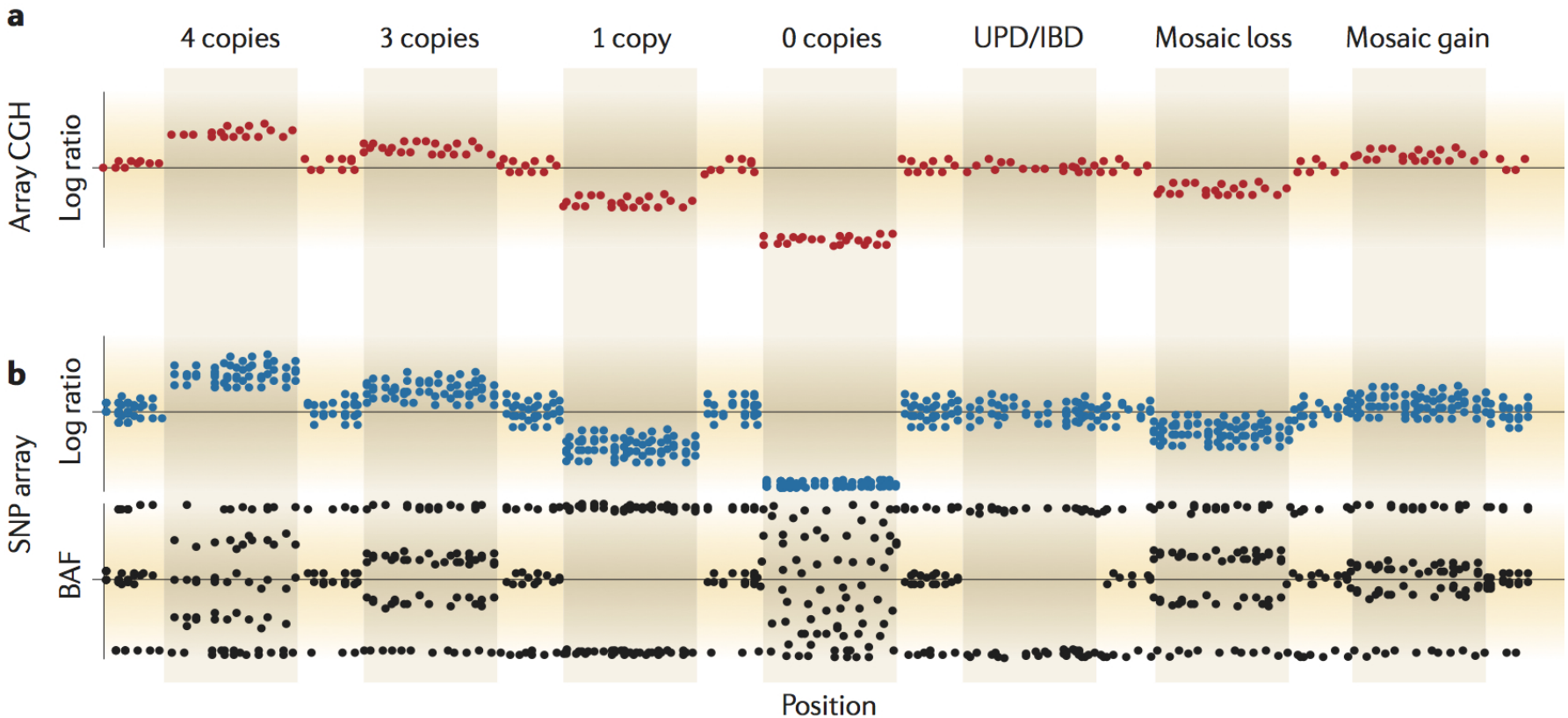
Similar to array CGH, SNP array compares signal intensity between sample and reference or other samples in the data.

SNP array can uses additional metric: B allele frequency (BAF). In normal samples, it is 0 (AA), 0.5 (AB) or 1 (BB). But it can change with copy number events:

- Deletions: 0 (A) or 1 (B).
- Duplications: e.g. 0 (AAA), 1 (BBB), 1/3 (AAB) or 2/3 (ABB).

With BAF, SNP arrays can detect copy neutral events such as UPD: BAF is 0 or 1.

Detection of CNVs from SNP Array Data



Advantages and Limitations of Microarray Methods

Advantage: low cost, especially for SNP arrays. Technologies, including CNV calling algorithms, are mature.

Limitations:

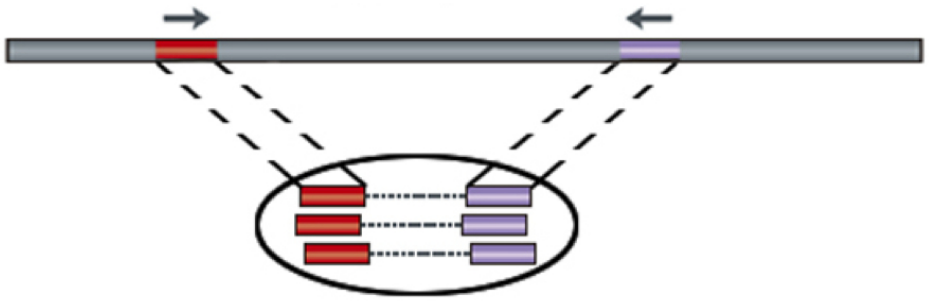
- Deletion bias: lower sensitivity to detect single copy gains. Small CNVs detected are overwhelmingly deletions.
- Unable to precisely define breakpoints.
- Difficulty with smaller CNVs ($<$ 10kb).

Next Generation Sequencing Methods

Compare with microarry-based methods, NGS has many advantages:

- Detecting novel/small CNVs and SVs.
- Precise breakpoint detection.
- Higher sensitivities: can use multple signatures.
- Better estimation of copy numbers.

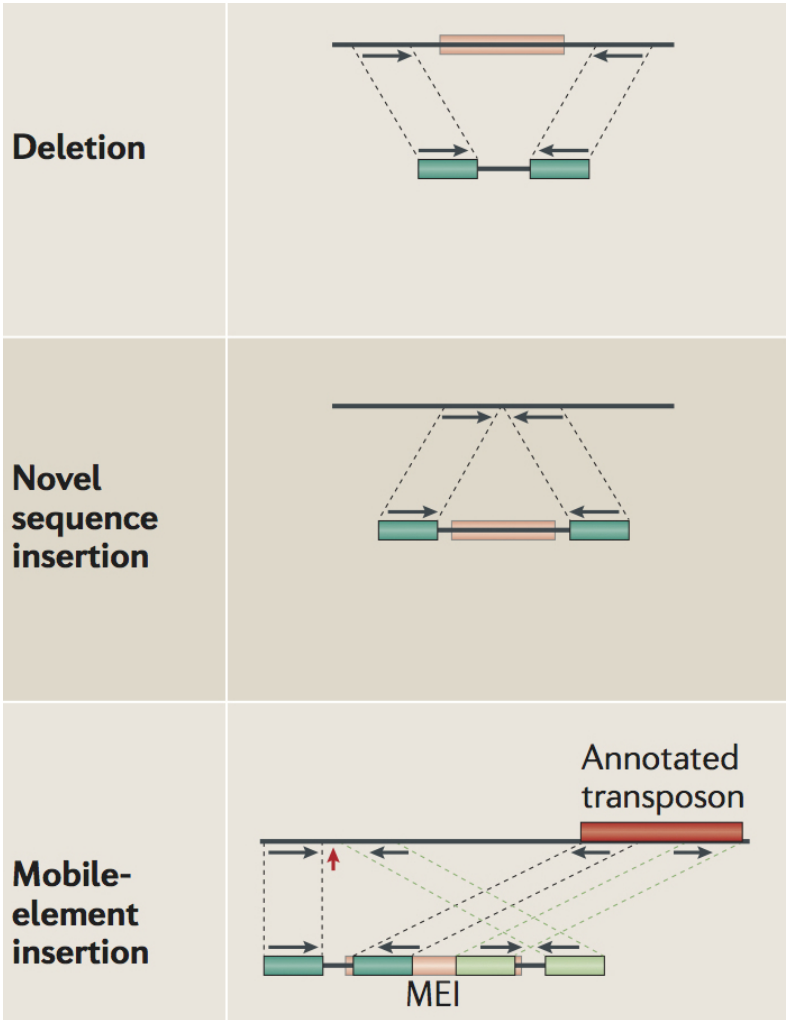
Read Pair Approach

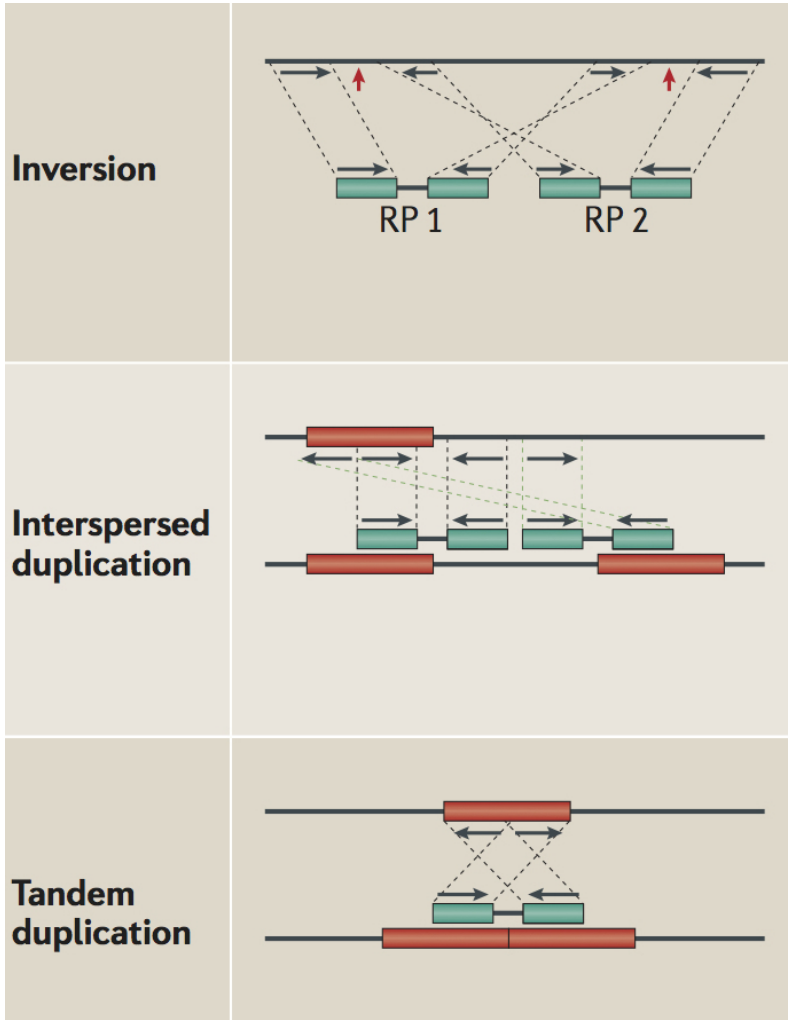


Sequencing libraries contain fragments of different sizes: 200-500 bp for paired-end libraries and 1-10 kb for mate-pair libraries.

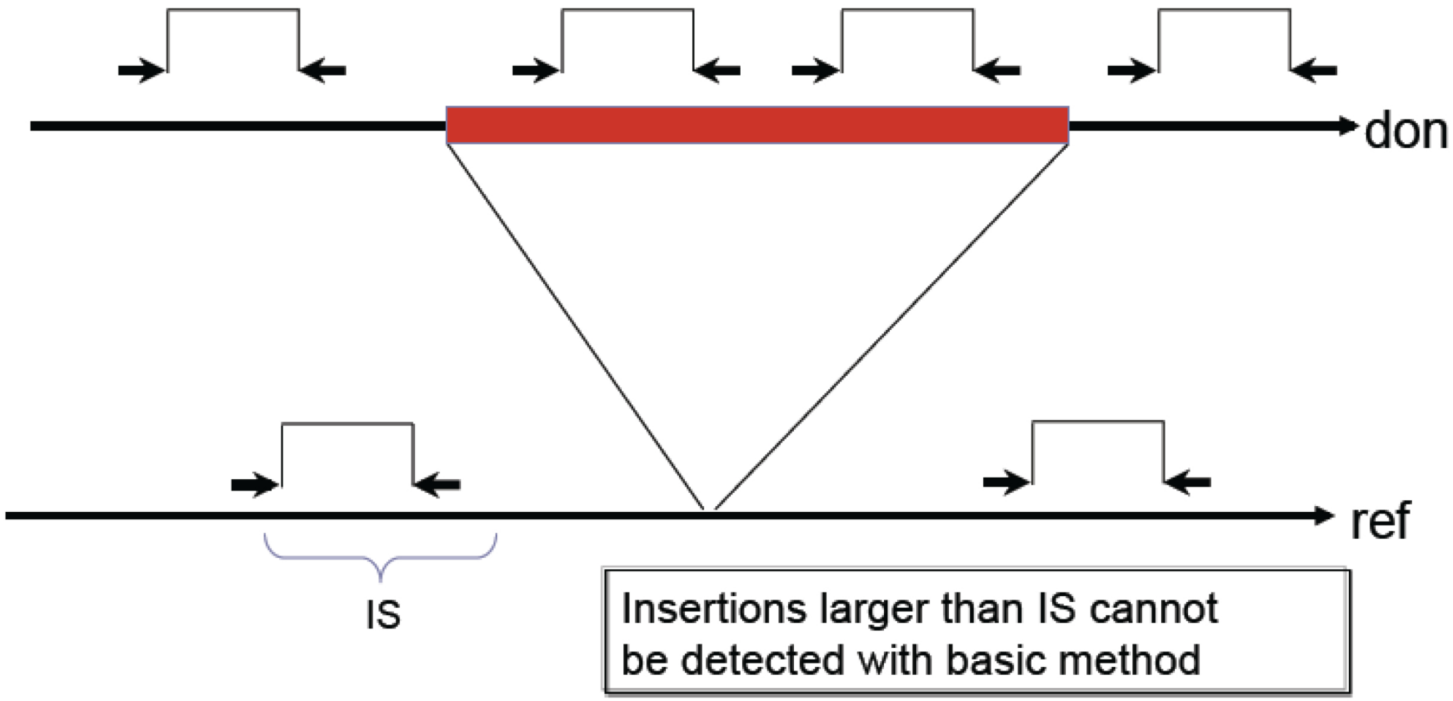
CNVs will increase or decrease the insert size for the mapped read pairs, and/or change the strand orientations, producing *discordant* pairs.

Read Pair Signatures of SVs



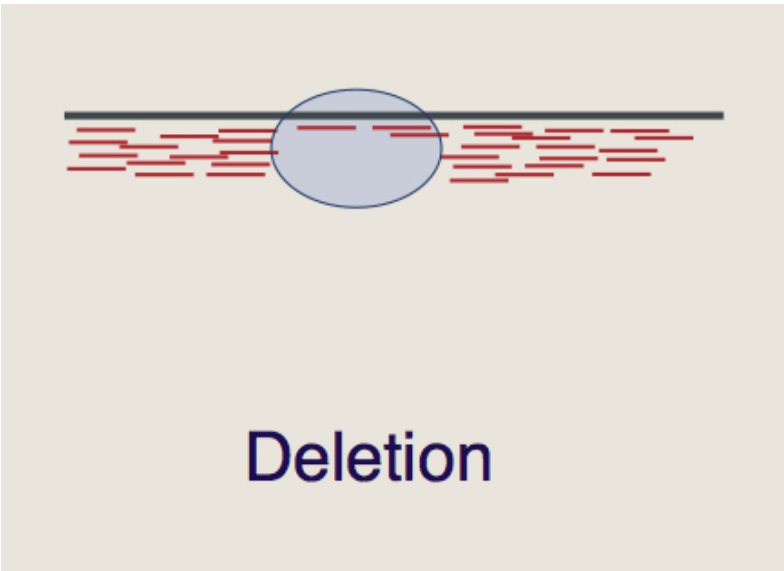


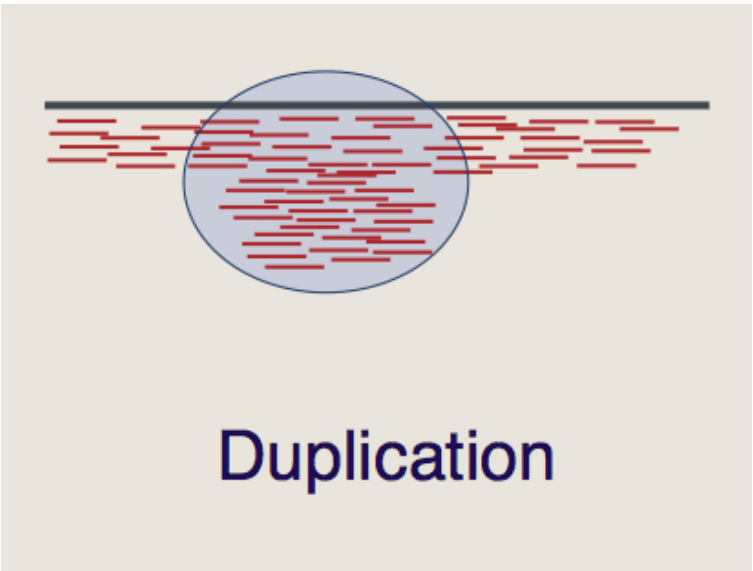
Read Pair Approach May Miss Large Insertions



Read Depth Approach

Copy number changes lead to reduced or increased read depth in the CNV regions. Read depth approach allows estimation of copy numbers, while other methods cannot.





Normalization of Read Depth

Many possible biases may distort the relationship between read depth and copy numbers, arising during exome capture, PCR amplification, sequencing efficiency and read mapping.

As a result, read depth depends on many factors such as GC content, proximity to segmental duplications, read mappability.

Normalization of Read Depth

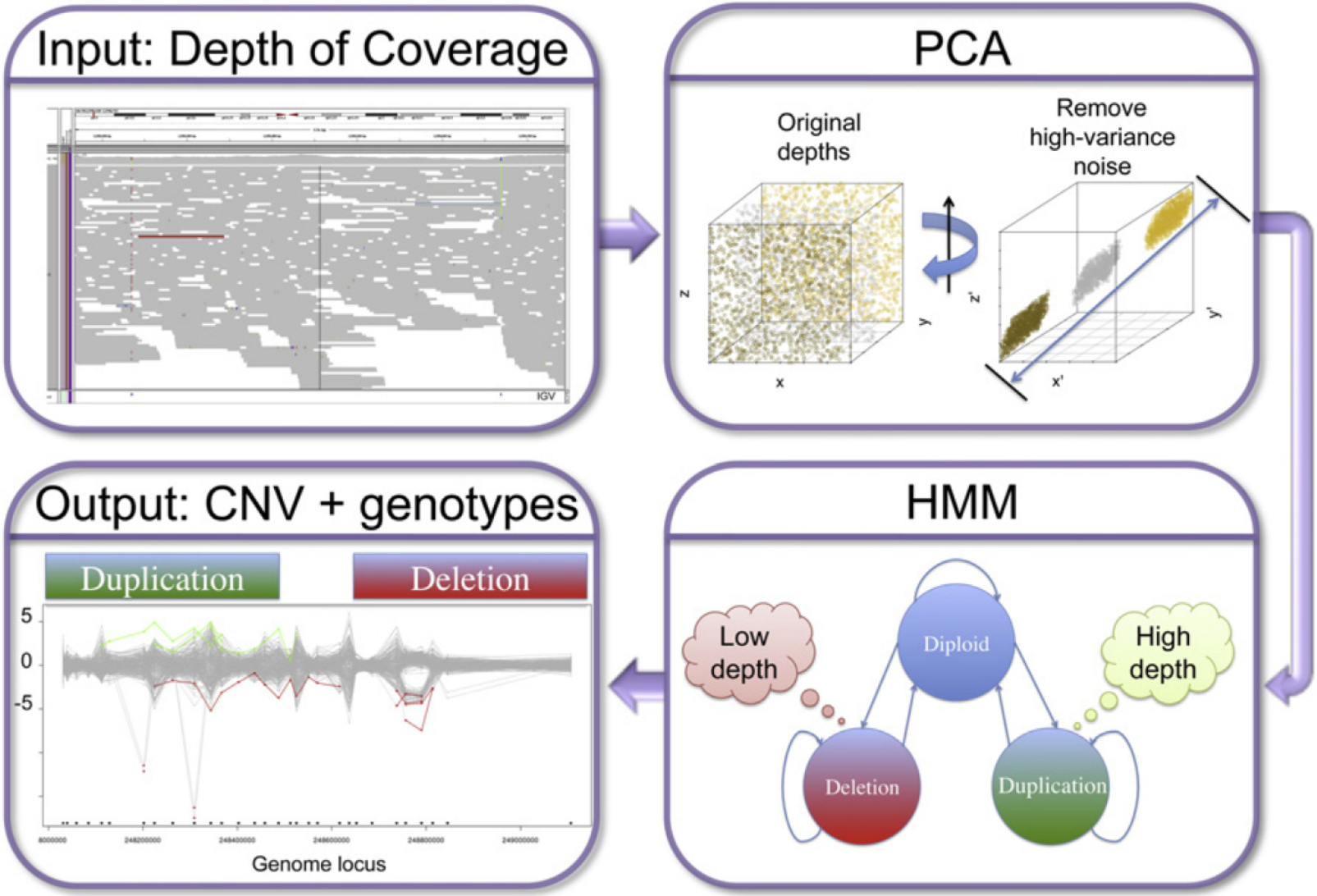
Many possible biases may distort the relationship between read depth and copy numbers, arising during exome capture, PCR amplification, sequencing efficiency and read mapping.

As a result, read depth depends on many factors such as GC content, proximity to segmental duplications, read mappability.

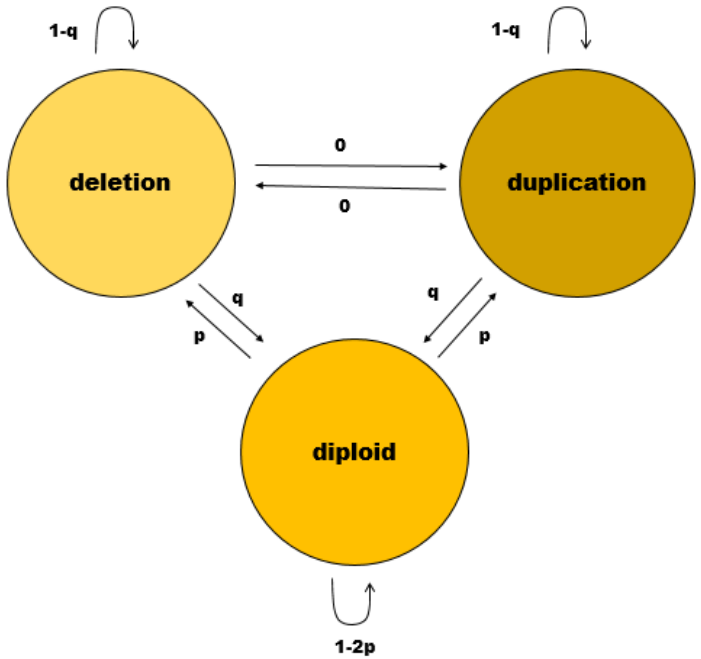
Normalization of read depth is important:

- Control for possible confounders such as GC content and background depth.
- Paired sample: comparison with control (often for cancer).
- PCA approach (XHMM, Fromer et al, AJHG, 2013): PCA on the matrix of exons and read depth.

Read Depth Approach in WES Data (XHMM)



XHMM Model



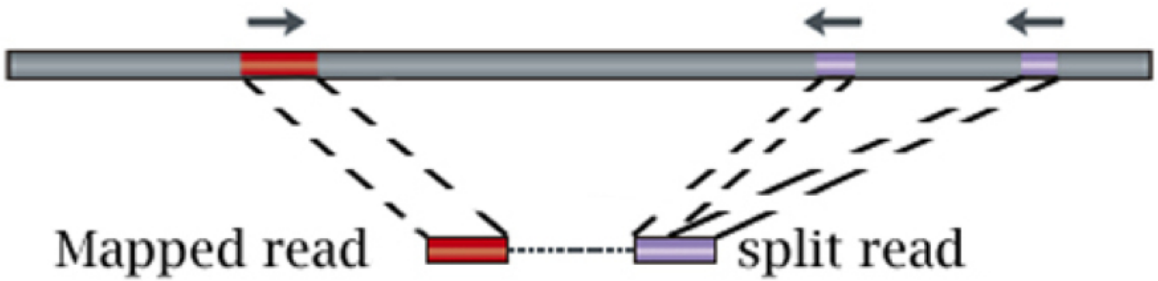
Emission: mean read depth under Diploid, Deletion and Duplication states are $0, -M$ and M respectively.

Transition: p is the exome-wide CNV rate, choose $p = 10^{-8}$. $T = 1/q$ is the average CNV size, choose $T = 6$ (exons).

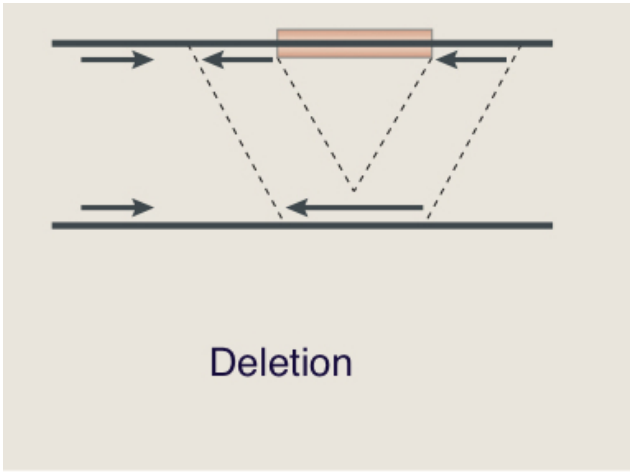
Split Read Approach

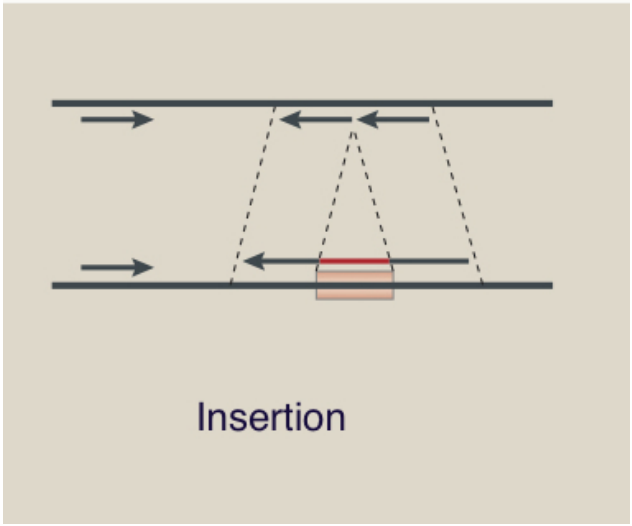
A SV event may create a pair of reads, where one can be mapped, and the other cannot because it contains a breakpoint (called *split read*).

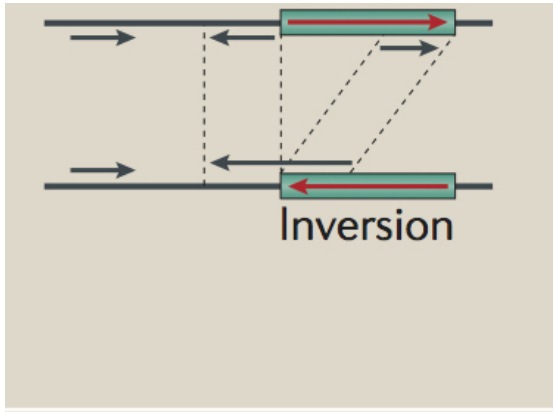
Require one read in the pair to be uniquely mapped. The split read allows precise identification of breakpoints.

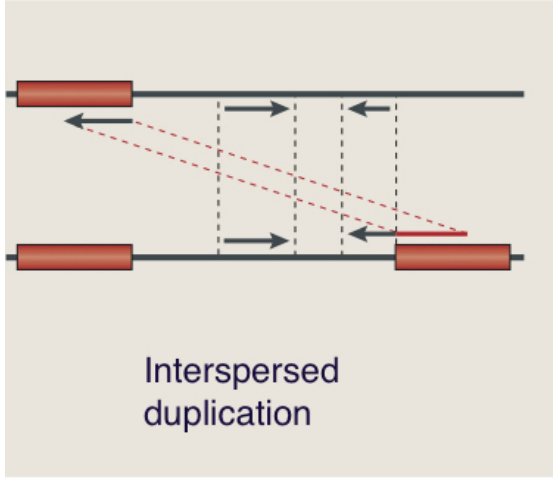


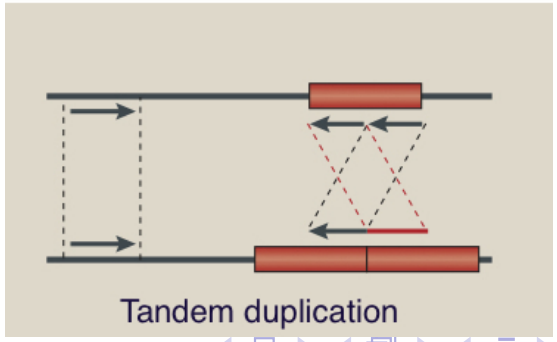
Split Read Signature of SVs





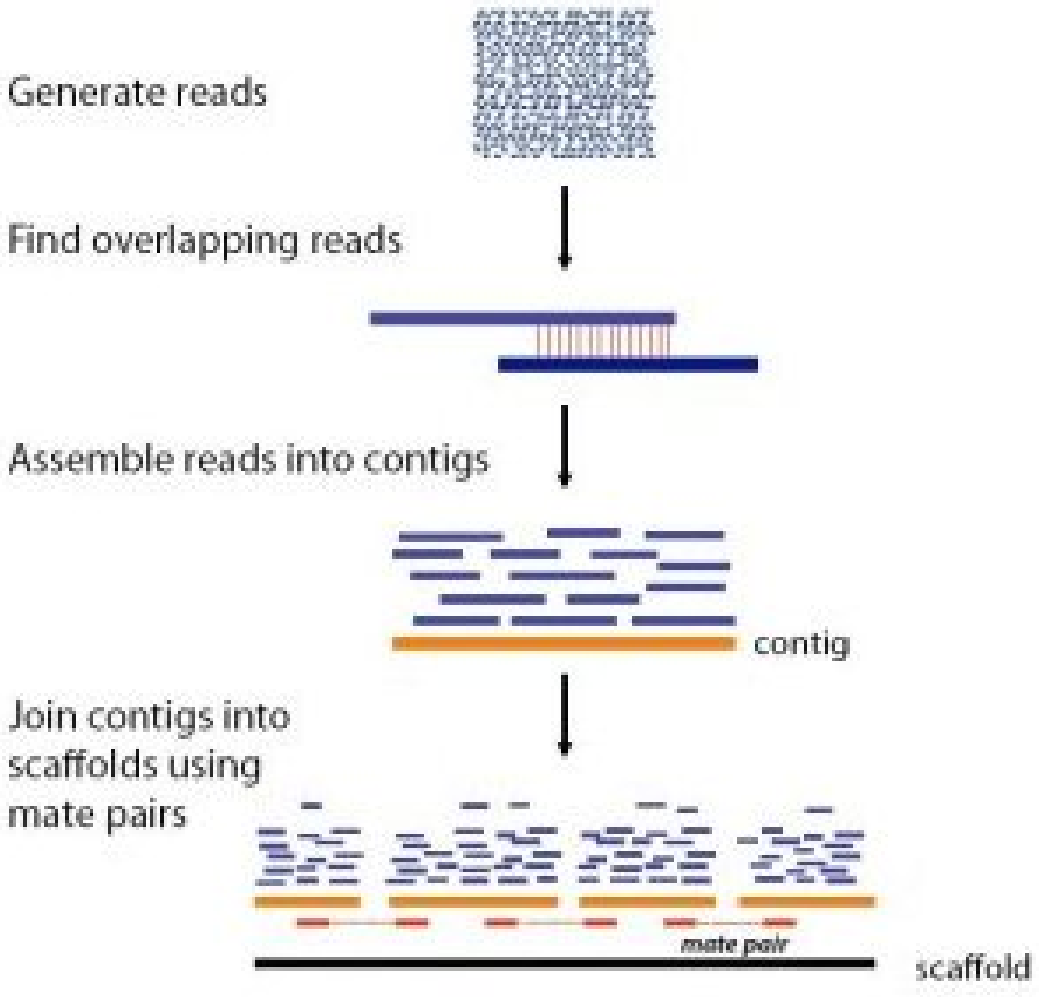






De Novo Assembly Approach

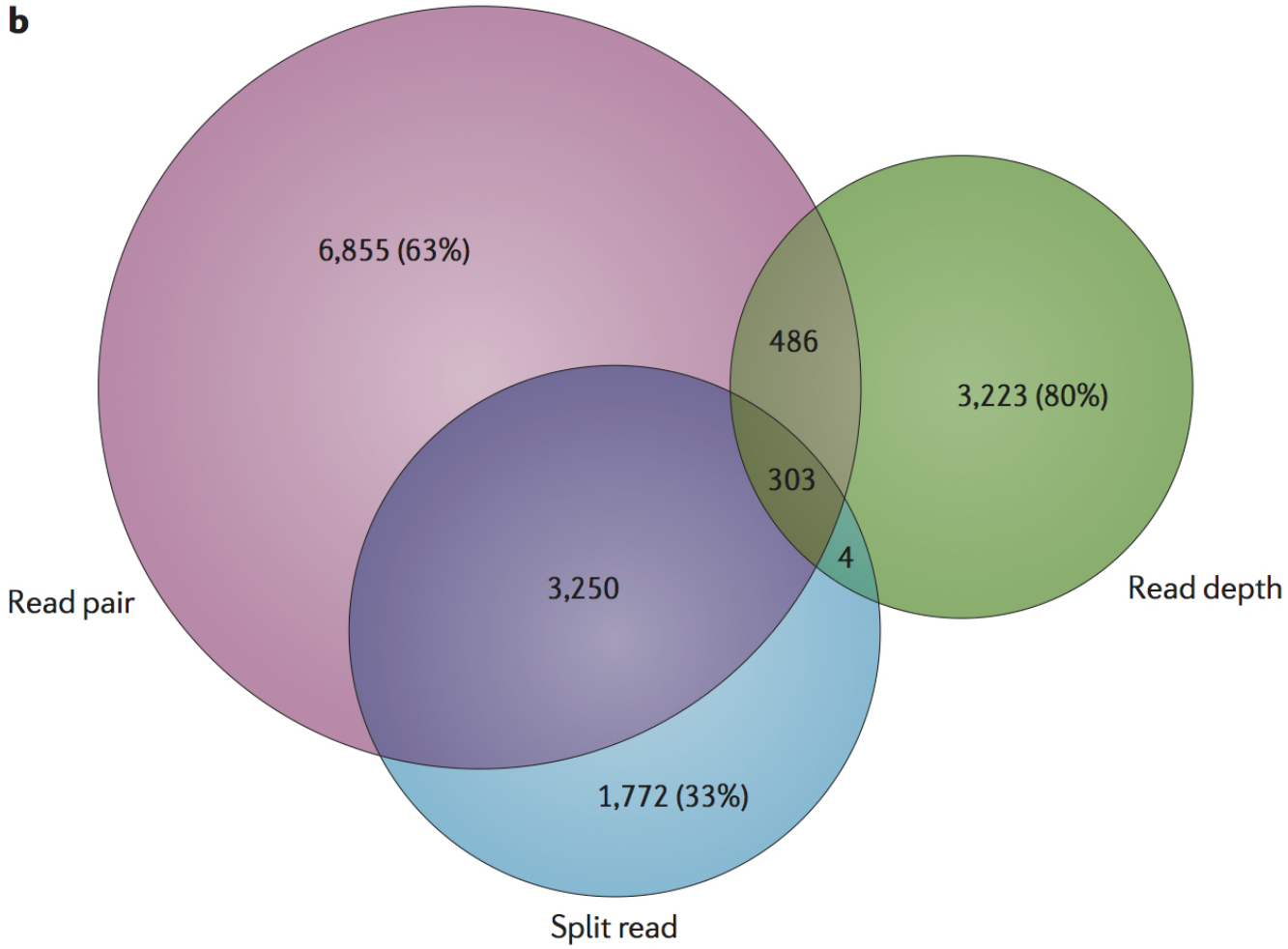
In high coverage regions, directly reconstruct the genomic sequences without using the reference genome.



Comparison of Four Approaches Based on NGS

- Read pair: cant identify all types of SVs, however, it cannot estimate the copy number, and detect large insertions.
- Read depth: estimate copy number, but cannot detect breakpoints, copy neutral events such as inversion or translocations, or small CNVs ($<$ 1kb).
- Split read: precise breakpoints, sensitive to deletions and small insertions, but poor performance in low complexity regions.
- Assembly: able to detect all SVs, but performs poorly on repeat and duplicate regions, and is computationally expensive.

Overlap of Structural Variants Discovered by Different NGS Approaches



CNVs in Human Diseases

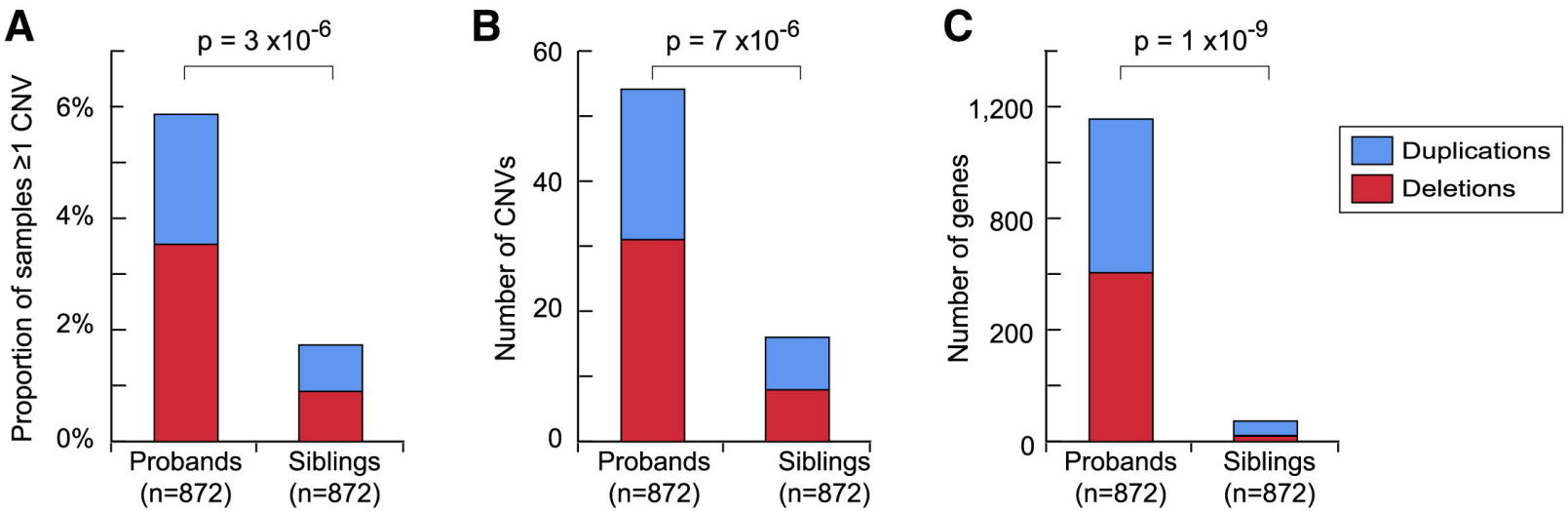
Large, rare CNVs are often found in individuals with developmental delays: 10% have de novo CNVs of size 500kb to 12Mb.

Many pathogenic CNVs are highly penentrant and recurrent (associated with segmental duplications).

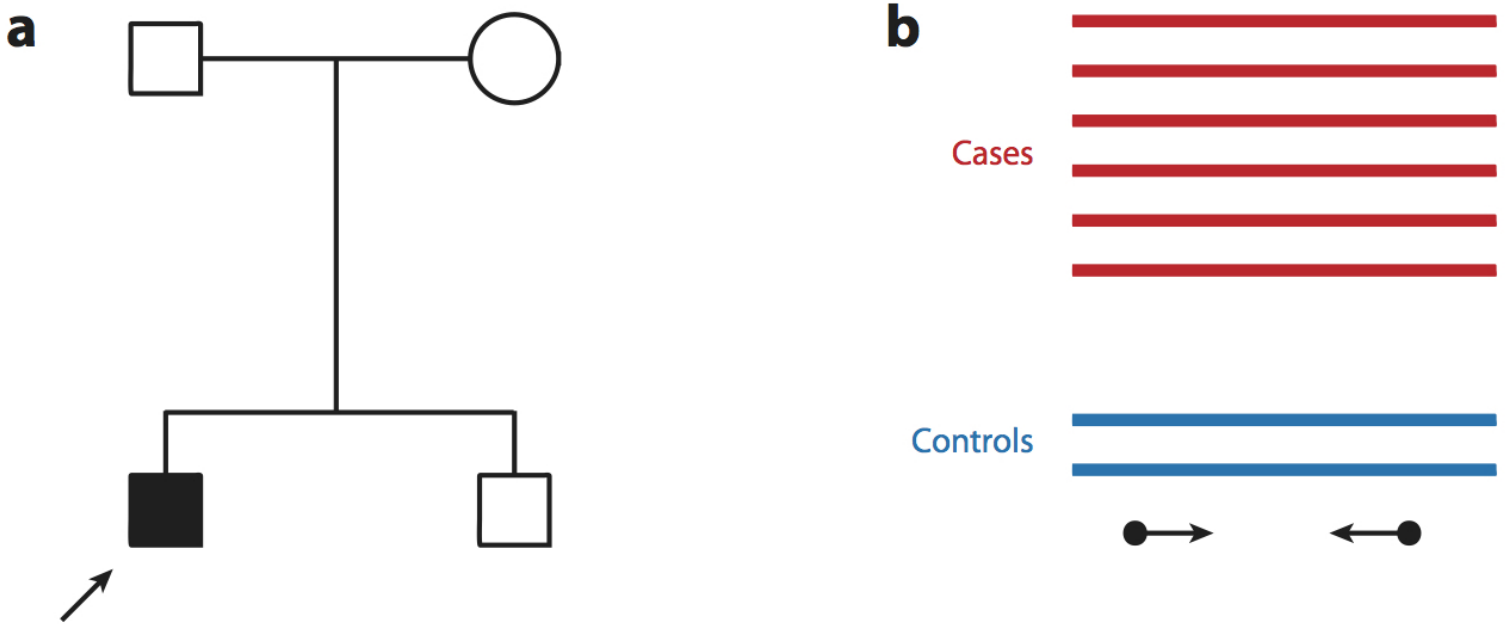
Examples of rare CNVs with variable penetrance and expressivity: 1.5M deletion in 15q13.3 was implicated in developmental delay, autism, schizophrenia and epilepsy.

Enrichment of CNVs in Complex Diseases

De novo CNVs were found to be significantly enriched in autism patients comparing with their unaffected siblings (Sanders et al, Neuron, 2011)



Associating CNVs to Diseases

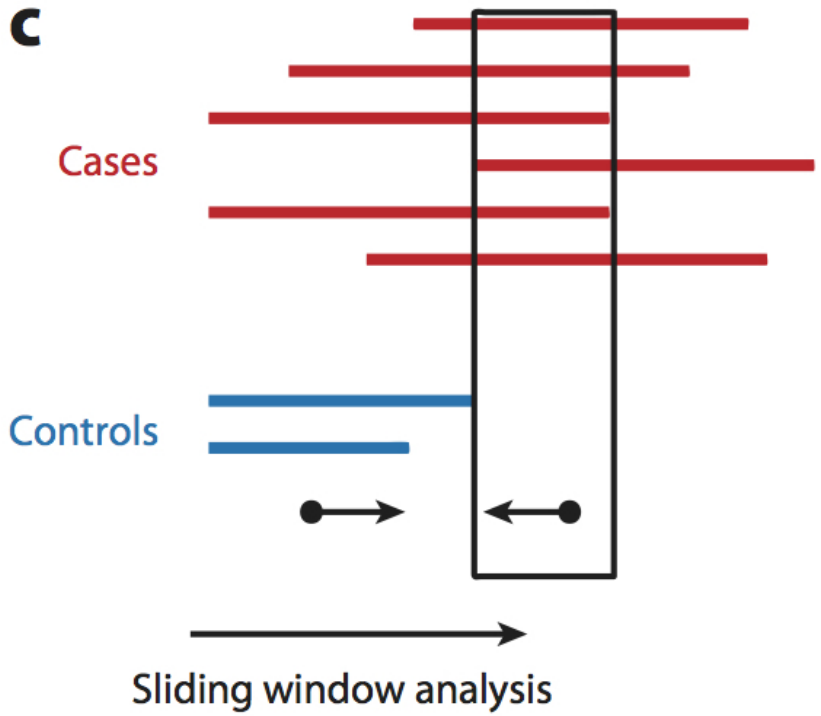


Large, *de novo* and rare ($< 1\%$) CNVs in patients with congenital phenotypes are often classified as pathogenic.

The frequency of a CNV is compared bewteen cases and controls.

Window-based Analysis of CNVs

We can also test enrichment of CNVs in a region (e.g. gene) between cases and controls.



De Novo Mutations

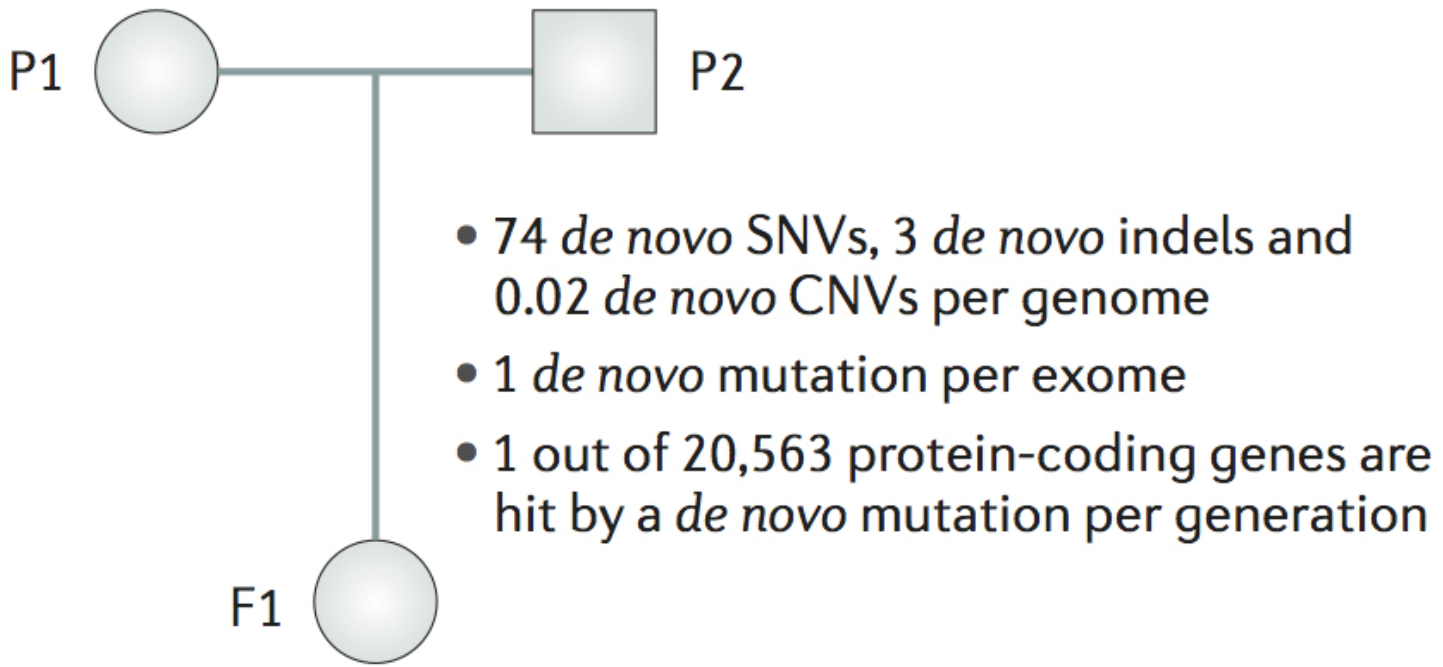
De novo mutations: a genetic alteration that is present for the first time in a family member as a result of a mutation in a germ cell of one of the parents or in the fertilized egg itself.

De novo mutations are subject to much weaker natural selection than standing variations, and can be highly deleterious.

In many early-onset/developmental diseases, de novo mutations play a significant role.

Rates of De Novo Mutations

Whole exome and geome sequencing enables detection of de novo mutations: often use parent-child trios.



De Novo Mutations in Rare Sporadic Genetic Diseases

De novo CNVs: e.g. Down syndrome caused by de novo trisomy of chromosome 21.

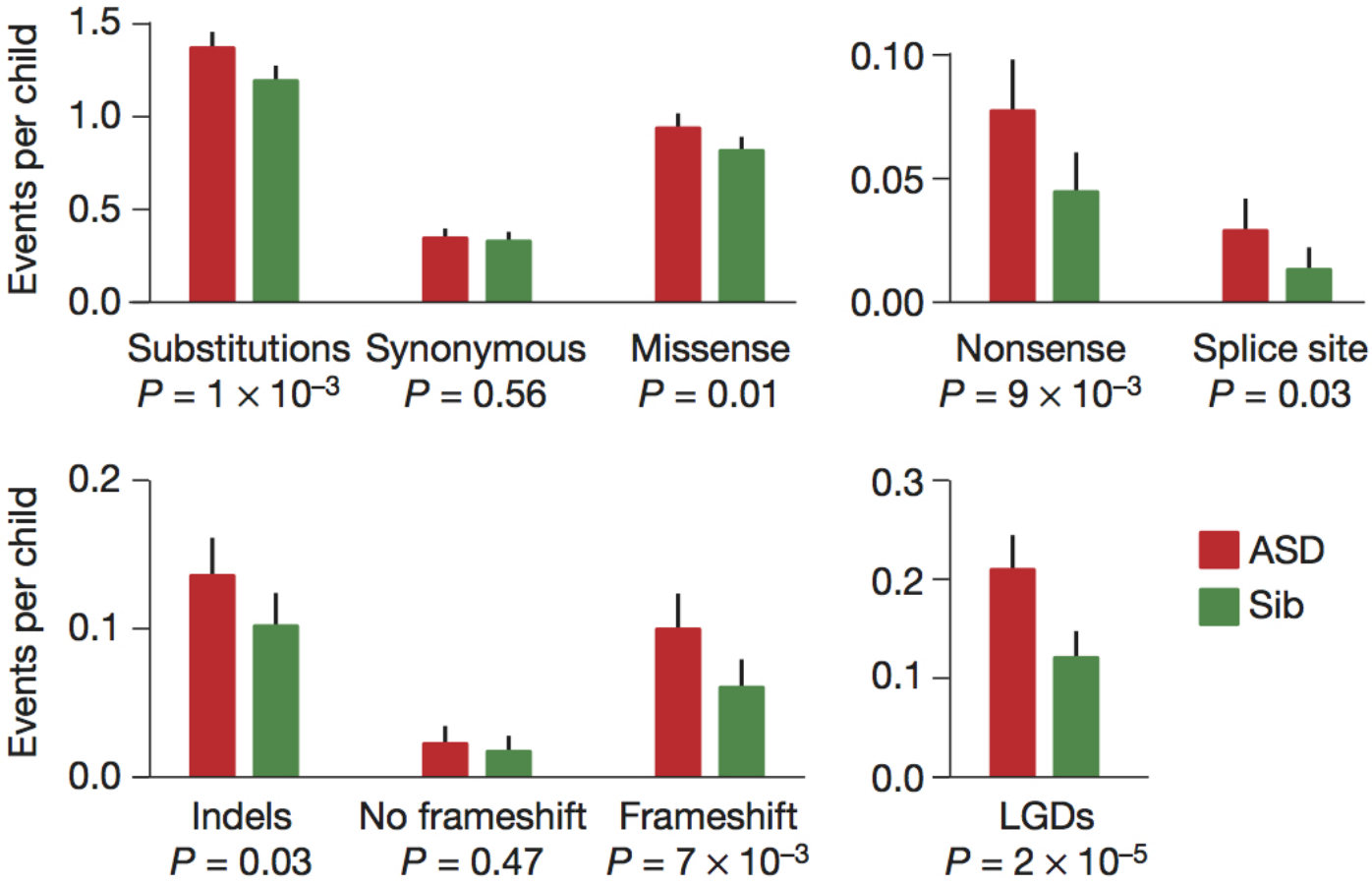
De novo SNVs: WES allows identification of de novo SNVs in a number of Mendelian diseases, e.g. SETBP1 in Schinzel-Giedion syndrome.

De Novo Mutations in Common Diseases

De novo CNVs: large CNVs ($>$ 100kb) are rare in population, about 1 in 50, but large de novo CNVs are found in 10% of patients with schizophrenia, autism and intellectual disability.

De novo SNVs: deleterious SNVs are enriched in patients of autism, intellectual disability, and other neurodevelopmental diseases.

Enrichment of De Novo Mutations in Autism Cases



Comparison of rates of de novo mutations in ASD children and their unaffected siblings (Iossifov et al, Nature, 2014).

Challenge of Interpreting De Novo Mutations

A de novo mutation, almost always, occurs only once in the study samples. How do you interpret a mutation when your sample size is 1?

Challenge of Interpreting De Novo Mutations

A de novo mutation, almost always, occurs only once in the study samples. How do you interpret a mutation when your sample size is 1?

Strategies:

- We can assess how likely a mutation disrupts gene function, based on genetic code, evolutionary conservation and other information. We generally know a lot about genes, even though we know little about individual mutations.
- Genes with unusually high number of de novo mutations in patients are likely risk genes.

Gene Level Analysis of De Novo Mutations

Suppose the mutation rate of a gene is μ, then in N trios, the number of de novo mutations in a non-risk gene follows Poisson distribution with the expected rate $2N\mu$.

Poisson test of the number of de novo mutations in a gene:

- Obtain the mutation rate of the gene. Often we limit the analysis to deleterious mutations, so the rate should be defined on these mutations.
- Let the number of mutations in this gene be x, the p-value of our test:

$$p = \text{Pois}(X \geq x | 2N\mu)$$

Examples of Poisson Test

Table 2 Individually significant genes identified from the analysis of *de novo* mutations in ASD cases

Gene	Mutations	Number of observed loss-of-function mutations	Number of expected loss-of-function mutations	*P* value
DYRK1A	Nonsense, splice site, frameshift	3	0.0072	**6.15×10^{-8}**
SCN2A	Nonsense, nonsense, frameshift	3	0.018	**9.20×10^{-7}**
CHD8	Nonsense, splice site, frameshift	3	0.022	1.76×10^{-6}
KATNAL2	Splice site, splice site	2	0.0049	1.19×10^{-5}
POGZ	Frameshift, frameshift	2	0.013	8.93×10^{-5}
ARID1B	Frameshift, frameshift	2	0.018	1.57×10^{-4}

Shown are genes with multiple *de novo* loss-of-function mutations across 1,078 ASD cases. Loss-of-function mutations include nonsense, frameshift and splice site–disrupting mutations. Number of expected loss-of-function mutations refers to the expected number of *de novo* loss-of-function mutations based on the probability of mutation for the gene as determined by our model. The genome-wide significance threshold is 1×10^{-6}. Significant *P* values are shown in bold.

Samocha et al, Nature Genetics, 2014

Summary

Structural variations: deletions, insertions, duplications, inversions and translocations. CNVs are special form of SVs.

Detection of CNVs by microarrays: array CGH and SNP arrays (use BAF).

Detection of CNVs/SVs by sequencing: read pair, read depth, split reads and de novo assembly.

Testing the role of CNVs in diseases: de novo, association with case/control status, window-based analysis.

De novo mutations: important in neuro-developmental diseases. Gene-level analysis using Poisson test.

Recommended Readings

- Alkan et al (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet.* 12(5):363
- Girirajan et al (2011) Human Copy Number Variation and Complex Genetic Disease. *Annu Rev Genet.* 2011;45:203
- Veltman & Brunner (2012) De novo mutations in human genetic disease, *Nat Rev Genet.* 13(8):565

Optional Readings

- Fromer et al (2012) Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth, *Am J Hum Genet*. 91(4):597
- Samocha et al (2014) A framework for the interpretation of de novo mutation in human disease, *Nat Genet*. 46(9):944