Computational Methods for High-Throughput Omics Data - Genomics

Introduction to NGS Genomics

25.10.2011

Outline

- Sequencing technologies:
    - Sanger, Illumina and other 2nd generation or NGS technologies
    - paired end sequencing, error profiles, quality values
- Applications: RNA-Seq, DNA-Seq, ...
- Computational analysis: Read mapping
    - goal of read mapping, special types of read mapping for special types of reads
    - edit/hamming distance, scores, quality values, mapping qualities
    - problems: sequencing errors, repeats, multi-reads
- Different types of genomic variants:
    - SNPs, small indels, structural variants (SVs), copy number variants (CNVs)
    - functional impact: coding/regulatory region, gene fusions
- Computational analysis: SV detection
    - methods based on read pair, read depth, split-read, de-novo assembly
    - different mapping signatures

Literature & links

Sanger sequencing animation:

http://www.dnalc.org/resources/animations/sangerseq.html

Illumina sequencing video:

http://www.youtube.com/watch?v=77r5p8IBwJk

454 sequencing video:

http://www.youtube.com/watch?v=kYAGFrbGl6E

NGS forum:

www.seqanswers.com

Review papers on wiki page:

- Review of next generation sequencing technologies
- Review of sequence alignment algorithms for NGS
- 2 reviews of structural variant detection methods

Brief history of DNA sequencing

- Sanger sequencing ger
- First complete ge
- Late 1980s: first a systems)
- 1990s: sequencin STs)
- 2001: First draft s ut $1 per base
- Mid 2000s: so-ca emerge, e.g. 454 Life Scien s
- 2008: The1000 G s.org/), cost approaching $1 pe

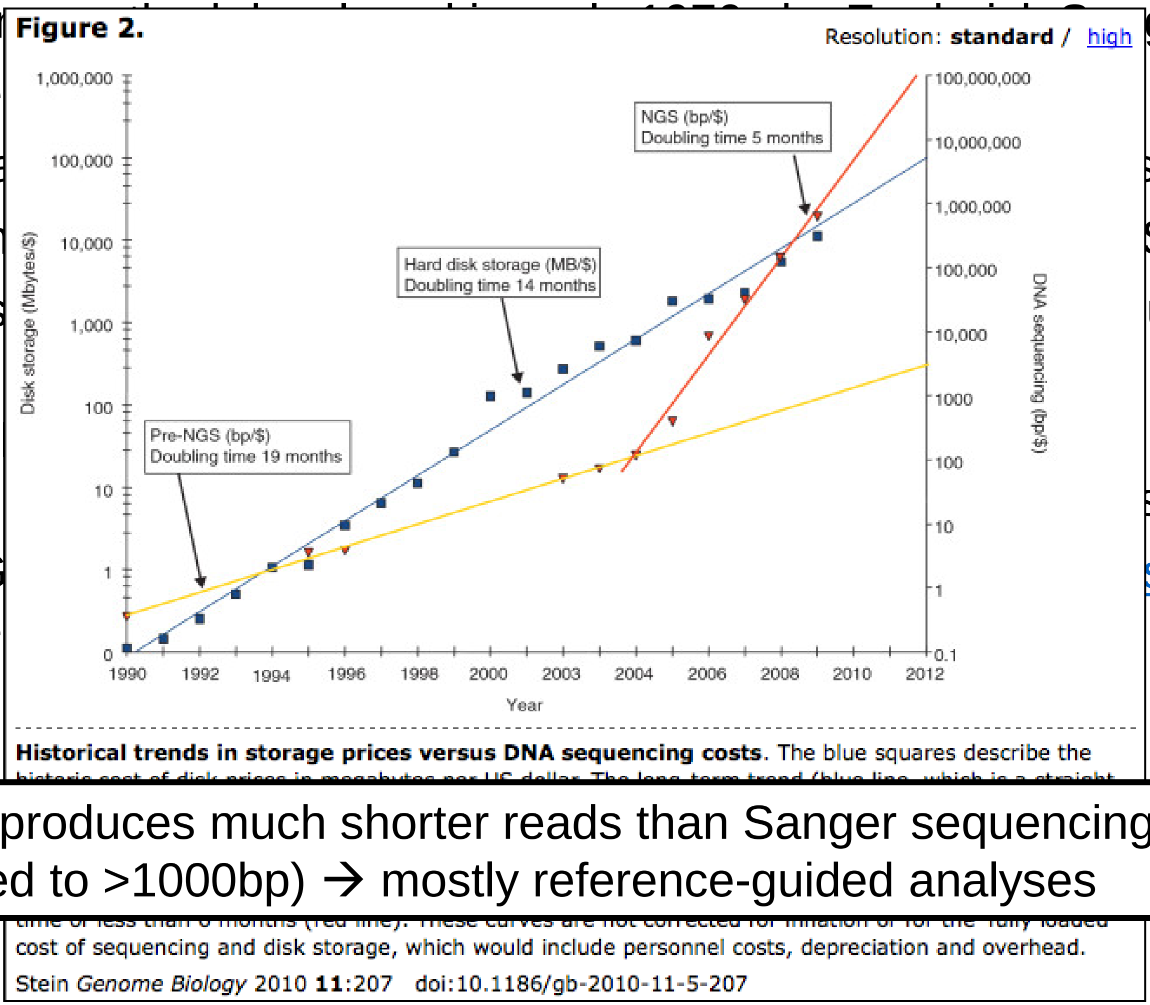


Figure 2. Resolution: **standard** / high

Historical trends in storage prices versus DNA sequencing costs. The blue squares describe the historic cost of disk prices in megabytes per US dollar. The long-term trend (blue line, which is a straight

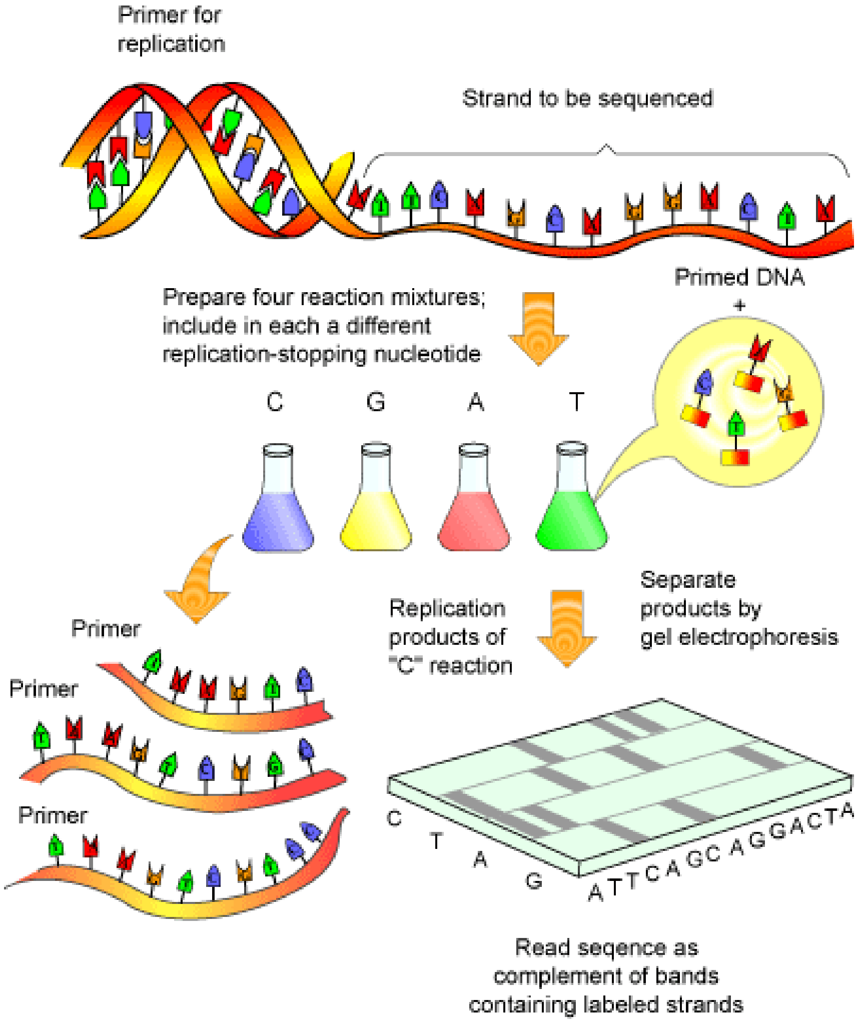
time of less than 6 months (red line). These curves are not corrected for inflation or for the 'fully loaded' cost of sequencing and disk storage, which would include personnel costs, depreciation and overhead.

Stein *Genome Biology* 2010 **11**:207 doi:10.1186/gb-2010-11-5-207

Problem: NGS produces much shorter reads than Sanger sequencing (50-400bp compared to >1000bp) → mostly reference-guided analyses

Sanger Sequencing

- Chain termination through dideoxynucleotide incorporation
- fragments with known last nucleotide are size-separated in a gel
- Produces sequencing reads of length ~1 kb
- Used to produce the first draft sequence of the human genome


Primer for replication
Strand to be sequenced
Prepare four reaction mixtures; include in each a different replication-stopping nucleotide
Primed DNA +
C
G
A
T
Primer
Primer
Primer
Replication products of "C" reaction
Separate products by gel electrophoresis
C
T
A
G
ATTCAGCAGGACTA
Read seqence as complement of bands containing labeled strands

Sanger sequencing animation:

http://www.dnalc.org/resources/animations/sangerseq.html

http://www.scq.ubc.ca/wp-content/uploads/2006/08/sequencing2.gif

The NGS revolution: sequencing-by-synthesis

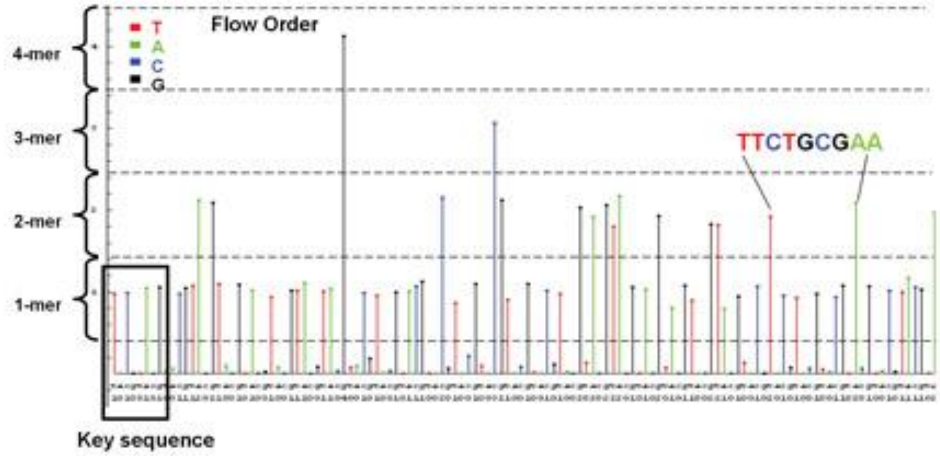
- 454 pyrosequencing is first NGS technology to become comercially available in 2005
- 300x less expensive than Sanger
- Soon followed by: Illumina (Solexa) reversible terminator sequencing (most popular sequencing technology at the moment) and ABI SOLiD two-base encoding

Next-Generation DNA Sequencing Methods, Elaine R. Mardis, Annual Review of Genomics and Human Genetics 2008

454 Pyrosequencing

- First NGS technology to be comercially available
- First instrument in 2005: 300x less expensive than Sanger
- Adds only one type of base at once
- Base incorporation emits light → the higher light intensity, the more bases were incorporated

- Sequencing errors mostly indels, especially in homopolymer runs
- Read length ~300-400bp
- ~500K reads per instrument run

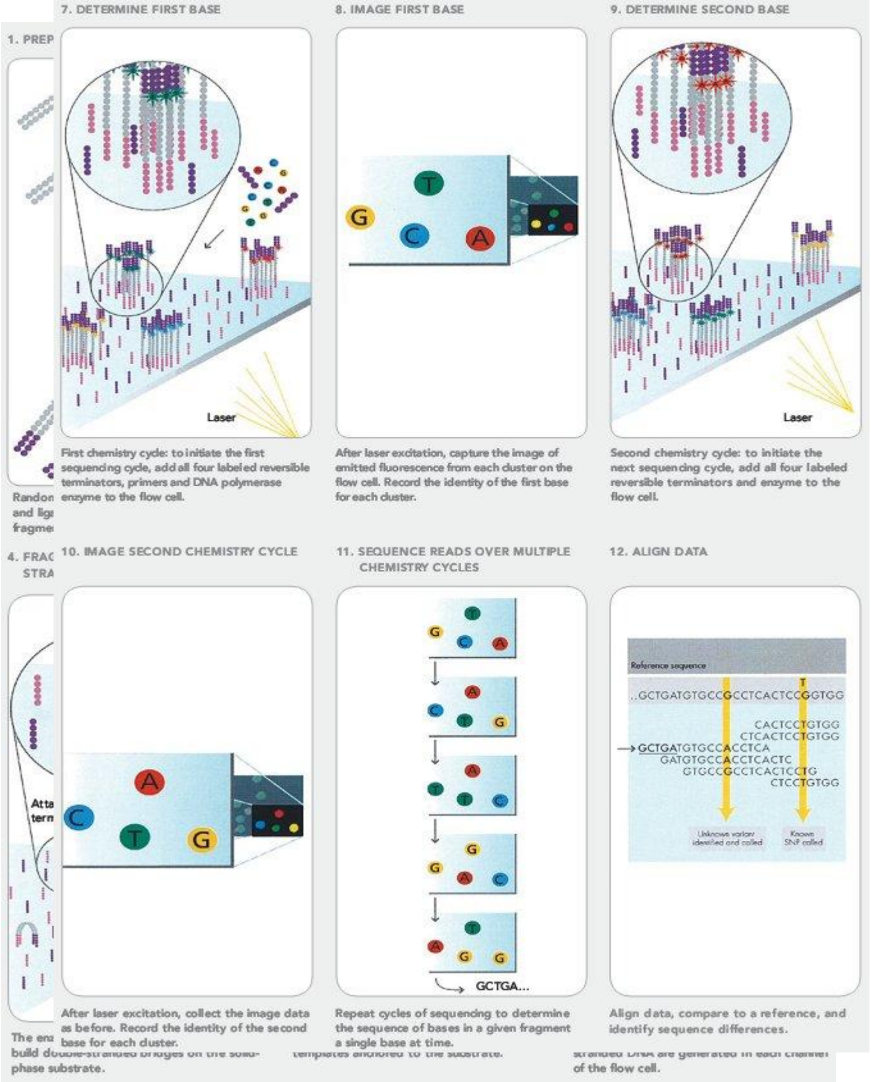


http://www.genengnews.com/sequencing/supp_04.htm

Illumina reversible terminator sequencing

- Another order of magnitude less expensive than 454
- Most popular sequencing technology at the moment
- One base incorporated per cycle, color signal to determine which one

- Sequencing errors mostly subsitutions
- Read length <30bp in the beginning, now >~100bp
- ~ 300 million reads per instrument run


1. PREP
Randon
and lig
fragme
4. FRAC
STRA
Atta
term
Laser
G
T
C
A
Laser
A
C
T
G
GCTGA...
Reference sequence
..GCTGATGTGCCGCCTCACTCCGGTGG
T
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
The en
build double-stranded bridges on the solid-phase substrate.
templates anchored to the substrate.
stranded DNA are generated in each channel of the flow cell.

7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA

Align data, compare to a reference, and identify sequence differences.

Next-Generation DNA Sequencing Methods, Elaine R. Mardis, Annual Review of Genomics and Human Genetics 2008

Sequencing technologies

A bit out-dated:

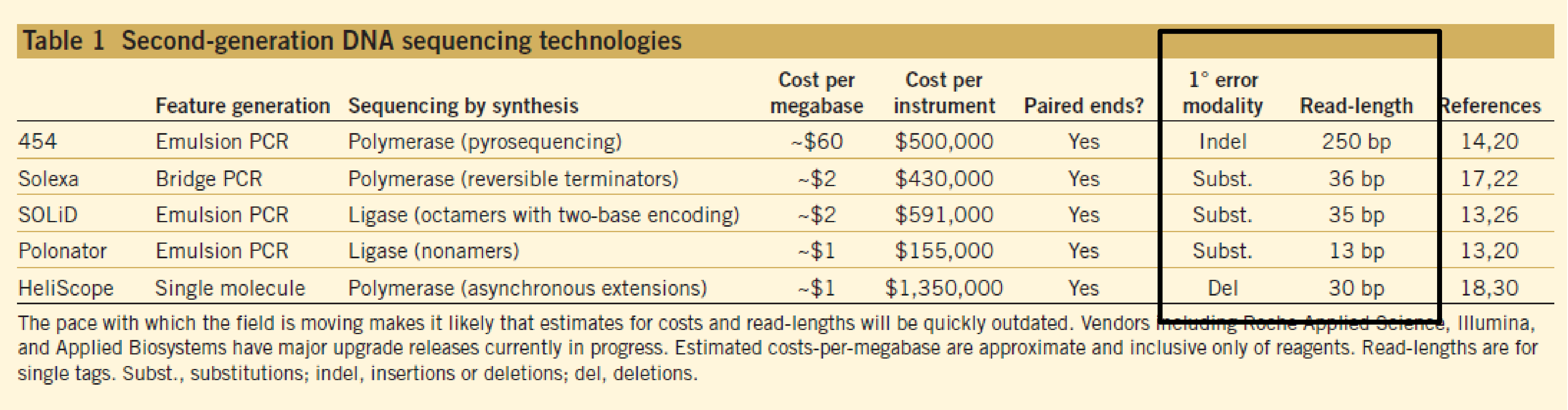


Table 1 Second-generation DNA sequencing technologies

	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length	References
454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	Indel	250 bp	14,20
Solexa	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	Subst.	36 bp	17,22
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	Subst.	35 bp	13,26
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	Subst.	13 bp	13,20
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	Del	30 bp	18,30

The pace with which the field is moving makes it likely that estimates for costs and read-lengths will be quickly outdated. Vendors including Roche Applied Science, Illumina, and Applied Biosystems have major upgrade releases currently in progress. Estimated costs-per-megabase are approximate and inclusive only of reagents. Read-lengths are for single tags. Subst., substitutions; indel, insertions or deletions; del, deletions.

Next generation DNA sequencing, Jay Shendure and Hanlee Ji, Nat. Biotech. 2008

Going from Sanger → sequencing-by-synthesis → single-molecule sequencing

Reads and quality values

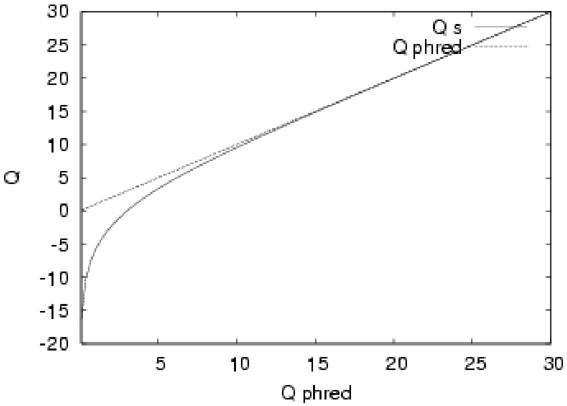
Each base of a read sequence has a *base call quality value*:

- Probability *e* of the base call being wrong (based on signal intensities)
- Log transformed into quality scores:

  $Q_{illumina} = 10 \log10 (e/(1-e))$

  for low values of *e* asymptotically identical to

  $Q_{phred} = -10 \log10 (e)$



http://bioinfo.cgrb.oregonstate.edu/docs/solexa

- Illumina example:

Cycle No. →

Base ↓

A	40	-40	-29	-5	-40	-6
C	-40	40	-40	-5	-40	-3
G	-40	-40	-35	-5	40	-3
T	-40	-40	28	-5	-40	-8

→ Called read seq = A C T N G C

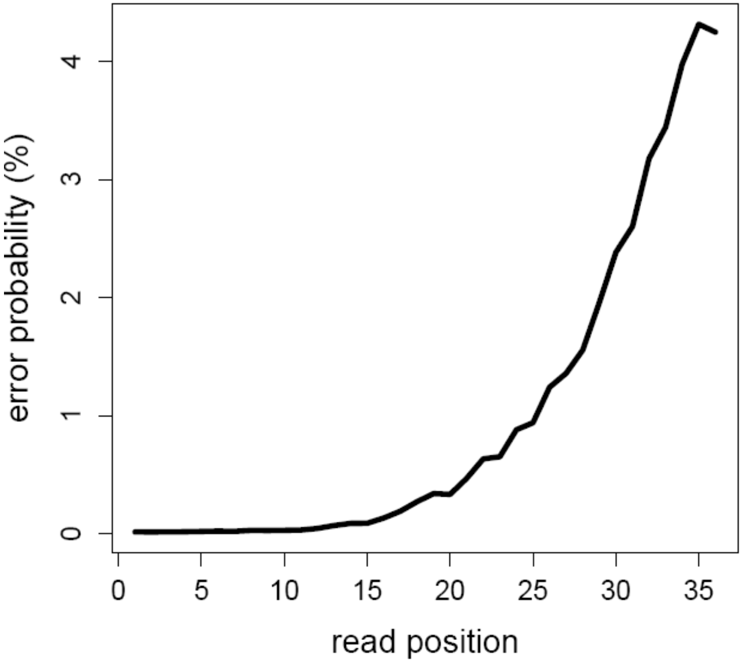
Illumina sequencing errors

Mostly miscalls, rarely under- or overcalls (i.e. mostly substitution, rarely insertion or deletion sequencing errors)

Strong quality trend:

Quality of sequencing decreases with increasing cycle number

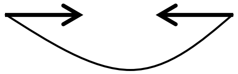
→ reads are sometimes trimmed



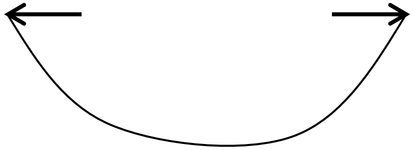
Paired-ends and Mate-Pairs

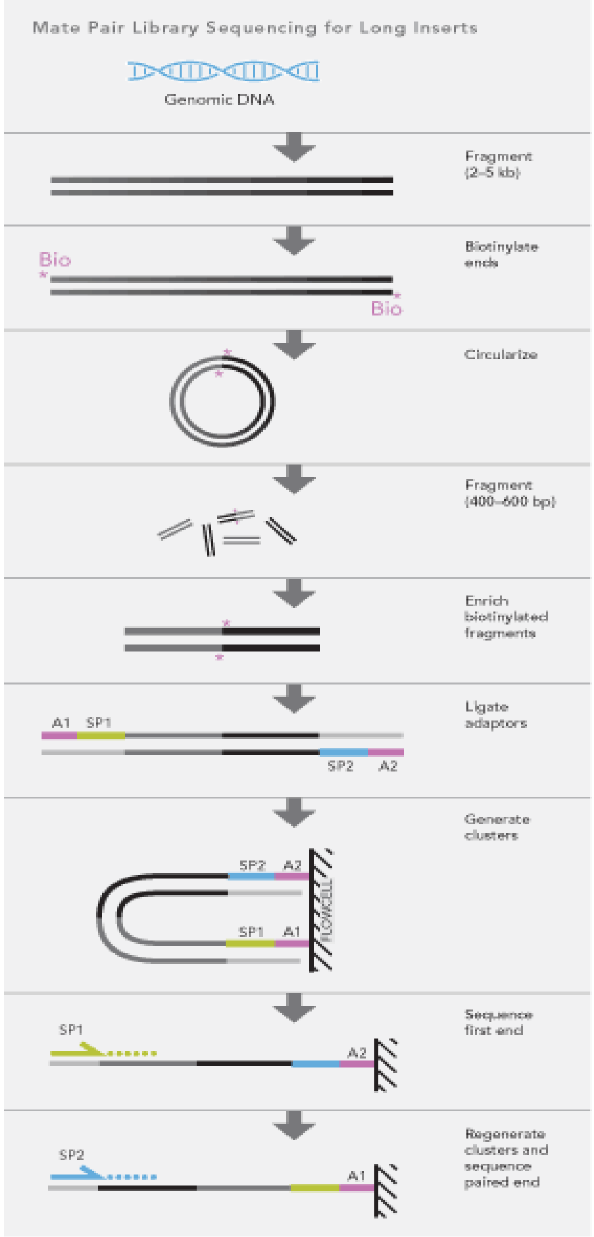
Both ends of a fragment are sequenced, different protocols:

- Paired-ends: ~200-500 insert size



- Mate pairs: ~2-5kb insert size





Illumina Inc.

Applications

- DNA-Seq:
  - Whole-genome resequencing
  - Targeted resequencing: genomic variants in specific regions, e.g. exons

- RNA-Seq:
  - mRNA sequencing
  - Small RNA sequencing

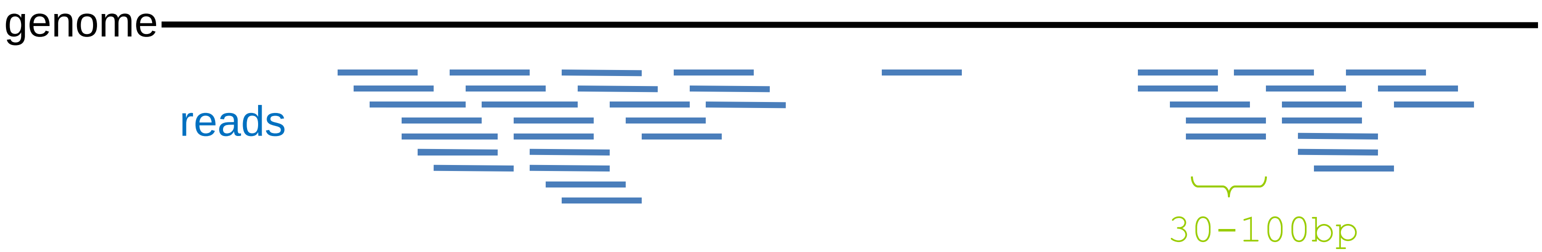
- CHiP-Seq:
  - Sequencing of transcription factor binding sites
  - Histone modification profiling

- Bisulfite sequencing:
  - DNA methylation profiling

(… the list goes on)

Fundamental Step: Read Mapping

Goal: Map each read to the genomic location it originated from



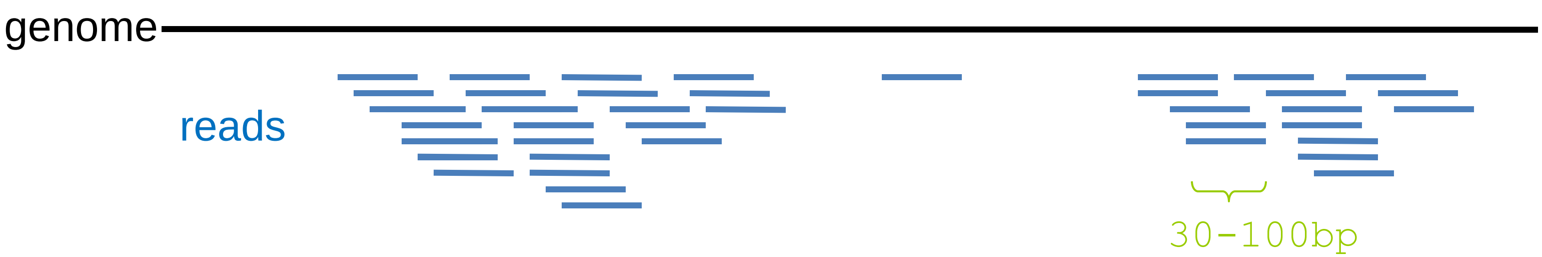
Approximate string matching problem!
And Blast is way too slow...

Difficulties:

- Billions of short reads, long genome
- Sequencing errors + genomic variants → alignment errors
- Repeats, ambiguous regions → multi-reads

Fundamental Step: Read Mapping

Goal: Map each read to the genomic location it originated from

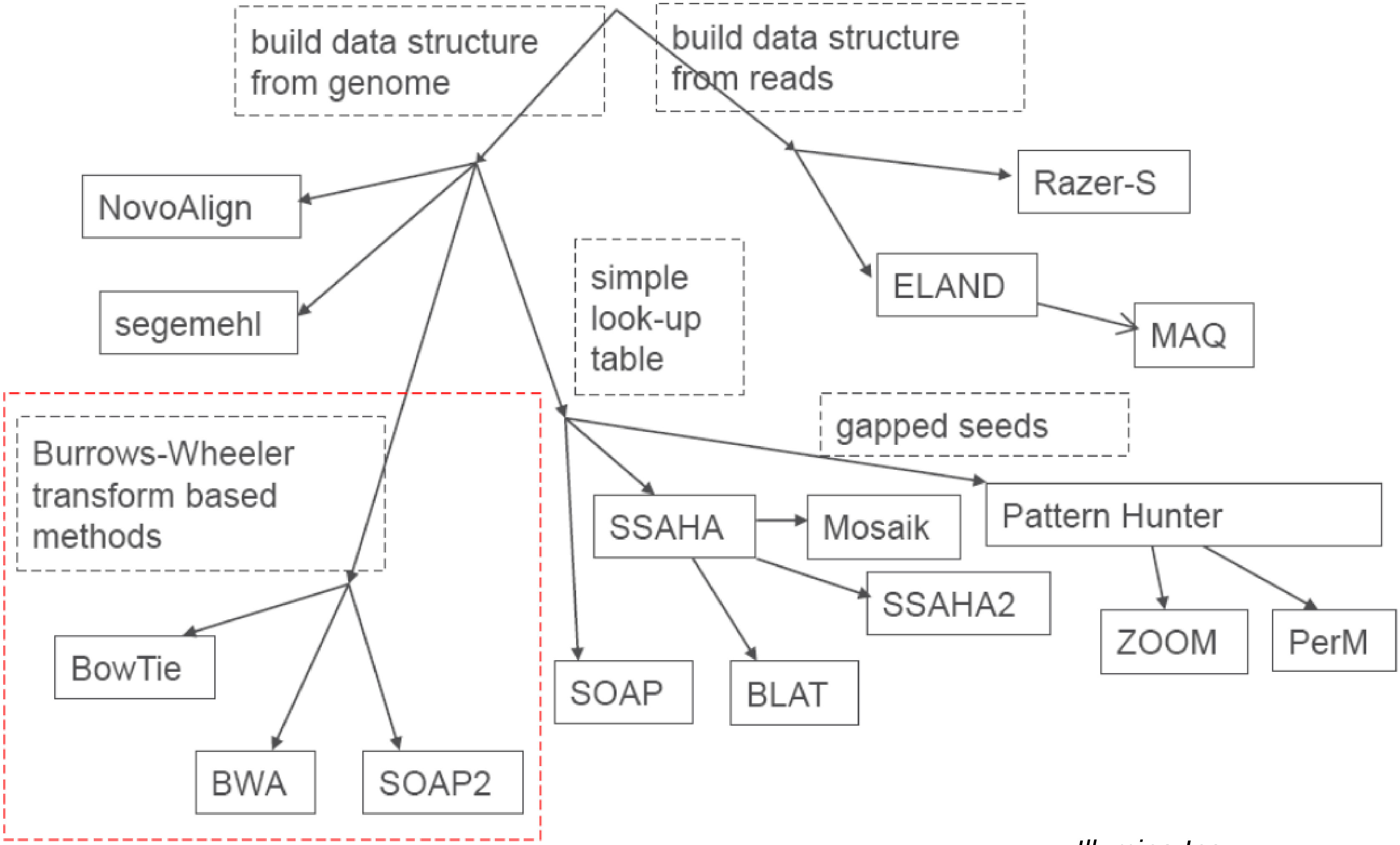


Approximate string matching problem!
And Blast is way too slow...

Edit/Hamming distance read mapping: RazerS paper

Mathematical/computational solutions:
- Hamming/edit distance alignment, allow up to k errors per read
- use base quality values to assign ambiguously mapped reads
- mapping quality: probability of a read being wronlgy mapped (Maq)
- filtering strategies based on index datastructures

Read Mapping Strategies



Illumina Inc.

Read Mapping Strategies: Example

General algorithmic techniques:

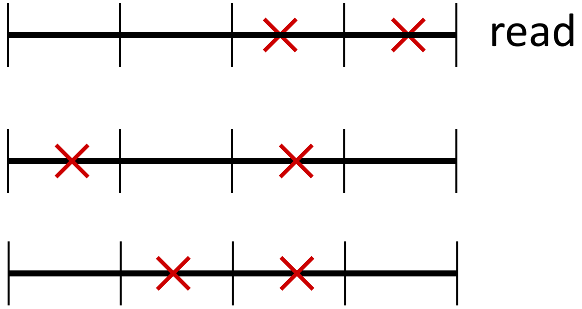
- Filtering (avoid having to check each genomic position for a possible match)
- Indexing and hashing → quick access to exact k-mer matches

Basics of Eland-algorithm (allowing up to 2 errors):

Split each read into 4 segments

2 errors can affect at most 2 segments

→ At least 2 segments match exactly



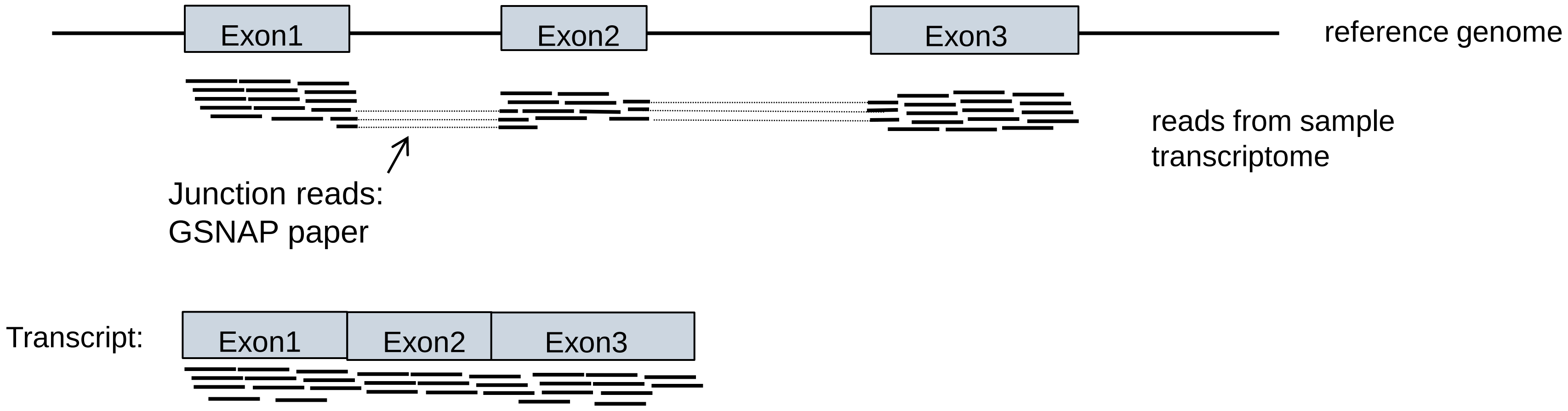
Efficient filtering for short reads, but does not extend well to longer reads

RNA-Seq

→Find out which genes are expressed in a certain tissue at a certain timepoint

→Quantify gene expression

→Identify alternative splicing events, i.e. Transcript isoforms

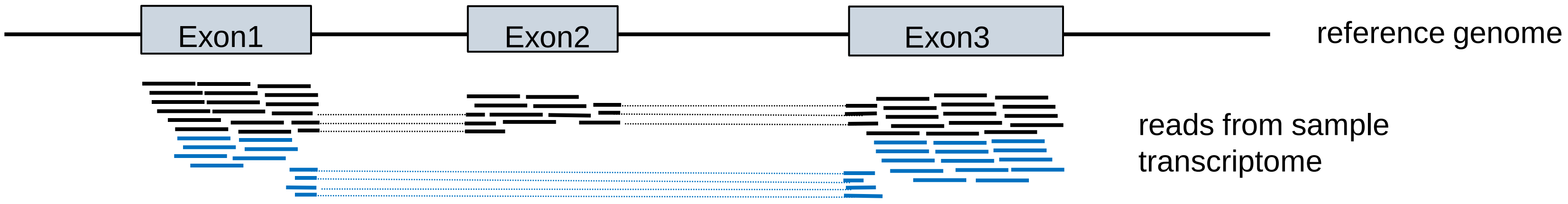


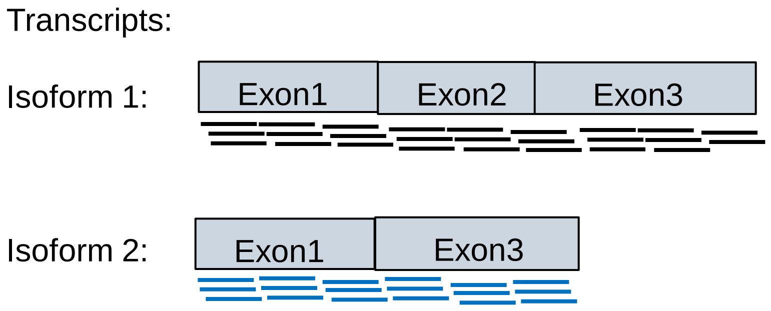
RNA-Seq

→Find out which genes are expressed in a certain tissue at a certain timepoint

→Quantify gene expression

→Identify alternative splicing events, i.e. Transcript isoforms





→ Junction reads and exon expression levels are used to estimate isoform-specific expression levels

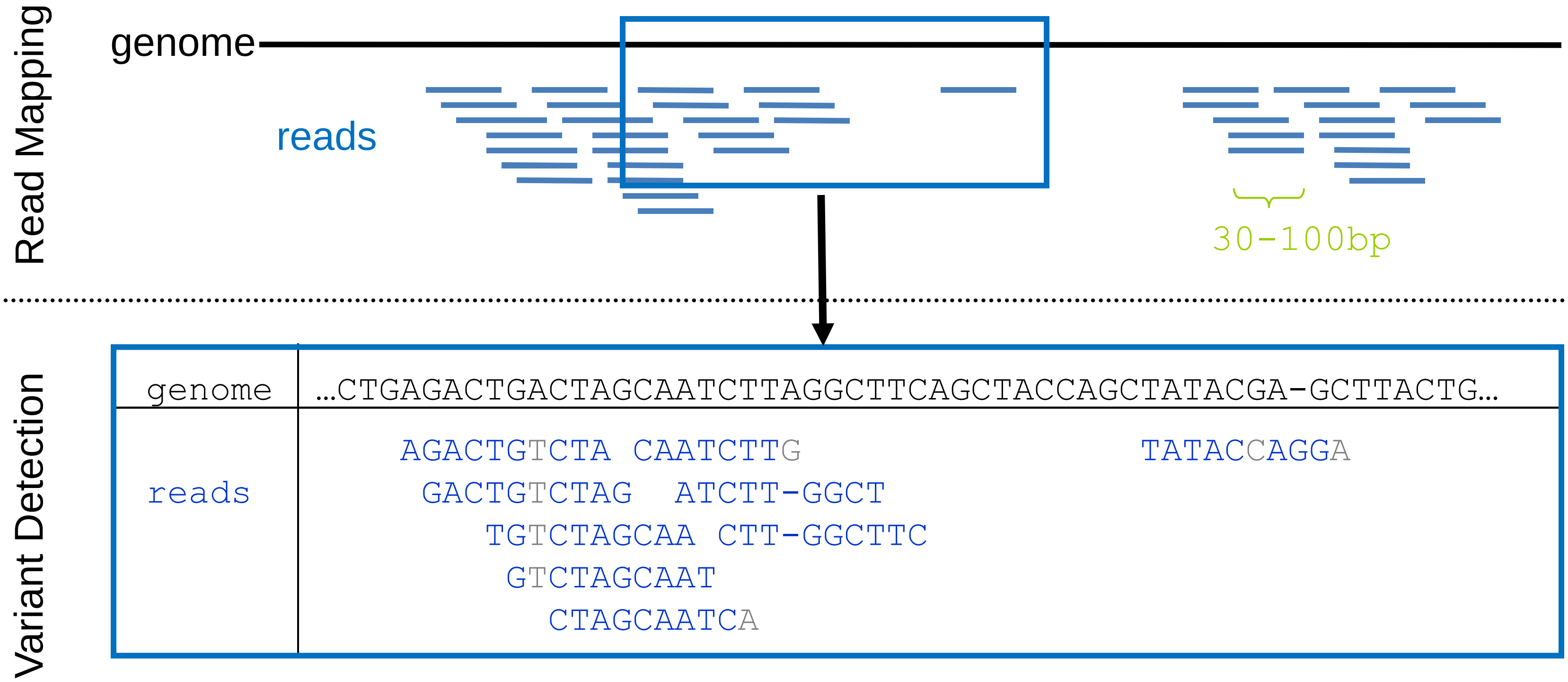
→ Number of mapped reads are used to estimate expression level

DNA-Seq or Genome resequencing

→Reference sequence known

→Identify differences between individuals/cell types, e.g. cancer vs. healthy cells

→SNPs, short indels, large indels, inversions, translocations, copy number variations



```
genome   …CTGAGACTGACTAGCAATCTTAGGCTTCAGCTACCAGCTATACGA-GCTTACTG…
             AGACTGTCTA CAATCTTG                TATACCAGGA
reads         GACTGTCTAG  ATCTT-GGCT
                 TGTCTAGCAA CTT-GGCTTC
                  GTCTAGCAAT
                    CTAGCAATCA
```

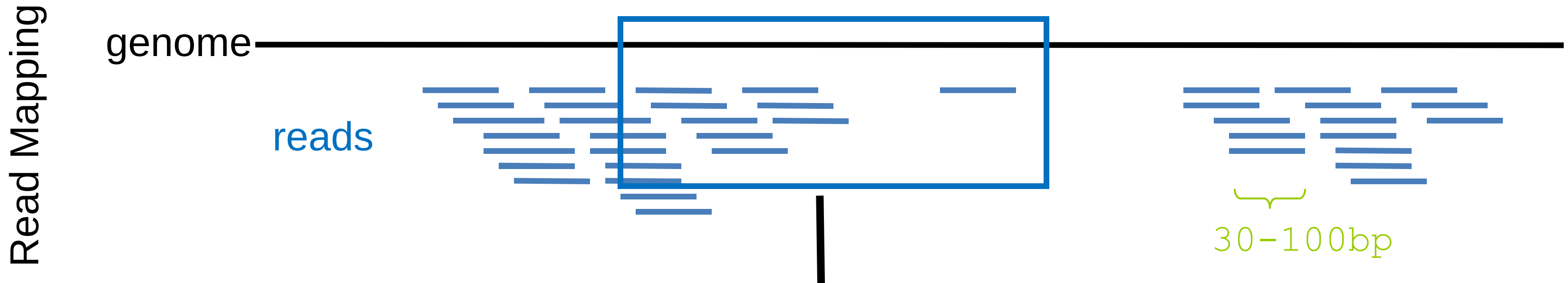
Goal: Detect differences between reference and donor DNA

DNA-Seq or Genome resequencing

→Reference sequence known

→Identify differences between individuals/cell types, e.g. cancer vs. healthy cells

→SNPs, short indels, large indels, inversions, translocations, copy number variations



SRMA paper: realignment

Variant Detection

```
genome   …CTGAGACTGACTAGCAATCTTAGGCTTCAGCTACCAGCTATACGA-GCTTACTG…
             AGACTGTCTA CAATCTT-G               TATACCAGGA
reads         GACTGTCTAG  ATCTT-GGCT
                 TGTCTAGCAA CTT-GGCTTC
                  GTCTAGCAAT
                    CTAGCAATCA
```

Goal: Detect differences between reference and donor DNA

DNA-Seq: Targeted resequencing

Usually of special interest: variants in coding region

→Targeted resequencing often uses exon capture to reduce sequencing cost

→Array with exon specific probes to pull out sequence of interest

Functional impact of small variants:

SNPs in coding region can be non-synonymous or synonymous, i.e. alter the amino acid that is incorporated or not

Indels in coding region can cause a frameshift and thereby non-sense proteins

Splice sites and regulatory elements can be affected, changing protein sequence or dosage

Types of Structural Variation (SV)

Understand role of SVs in

- Disease

- Complex traits

- Evolution

<u>Difficulty:</u>

They tend to reside in repetitive regions

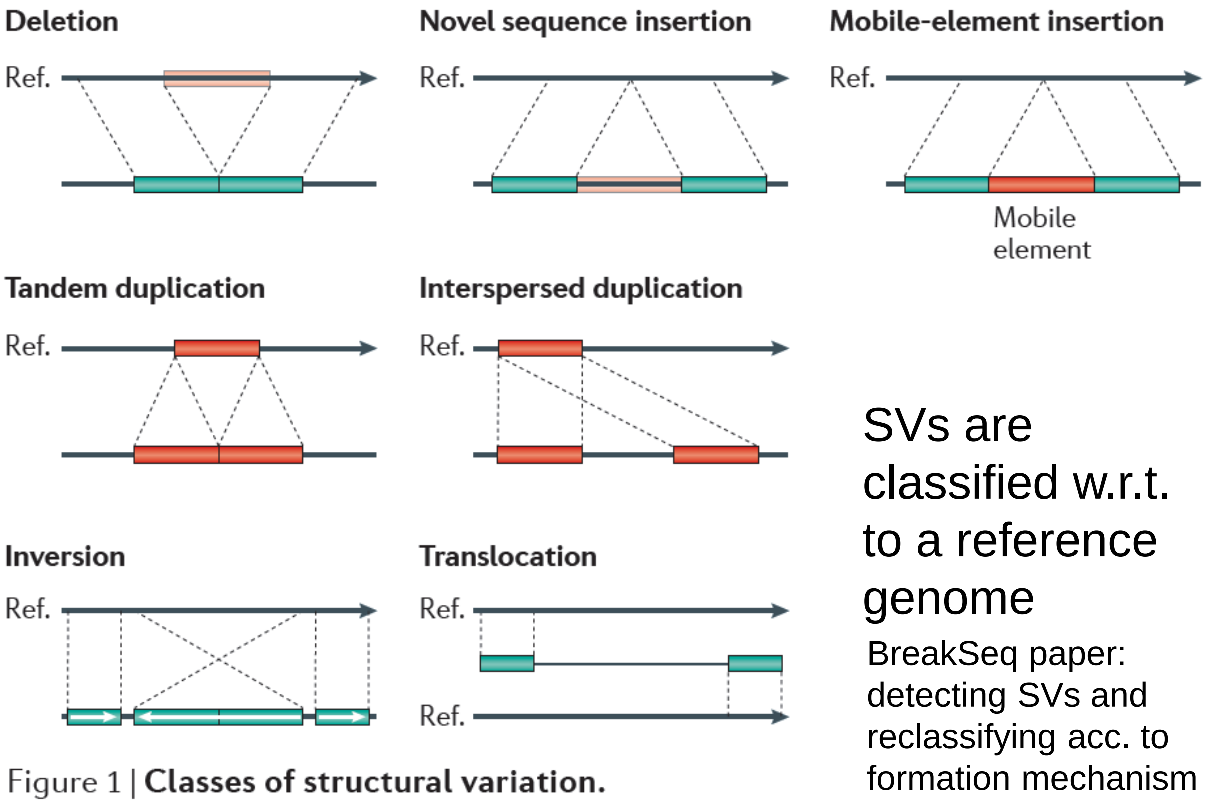


Figure 1 | **Classes of structural variation.**

Alkan *et al*. *„Genome structural variation discovery and genotyping“* Nature Reviews Genetics, 2011

SVs are classified w.r.t. to a reference genome

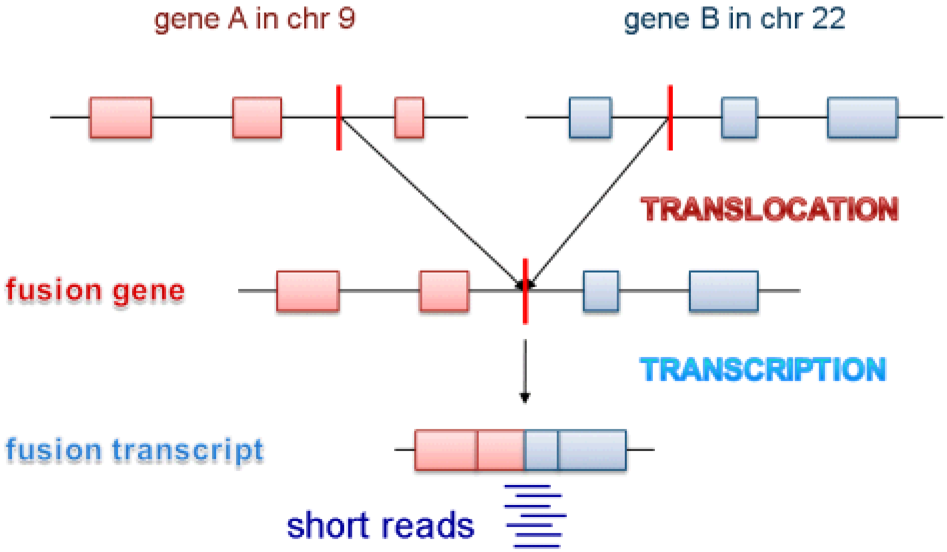
BreakSeq paper: detecting SVs and reclassifying acc. to formation mechanism

Definition of SV varies: > 50bp in some articles, >1kb in others

Special type of SVs: gene fusions

Especially abundant in cancer genome

CimeraScan paper: gene fusions in cancer



ChimerDB, http://ercsb.ewha.ac.kr:8080/FusionGene/

Side note: Other experimental SV discovery approaches

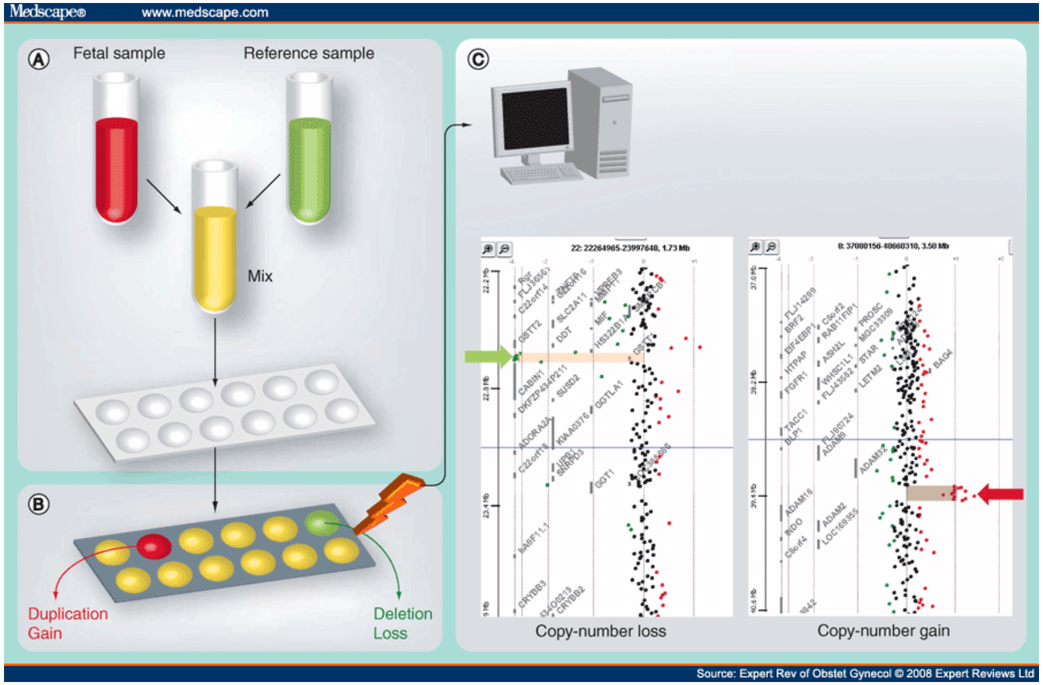
Hybridization-based microarray approaches

- Array CGH (BAC/oligo array)
- SNP microarrays

Sequencing approaches

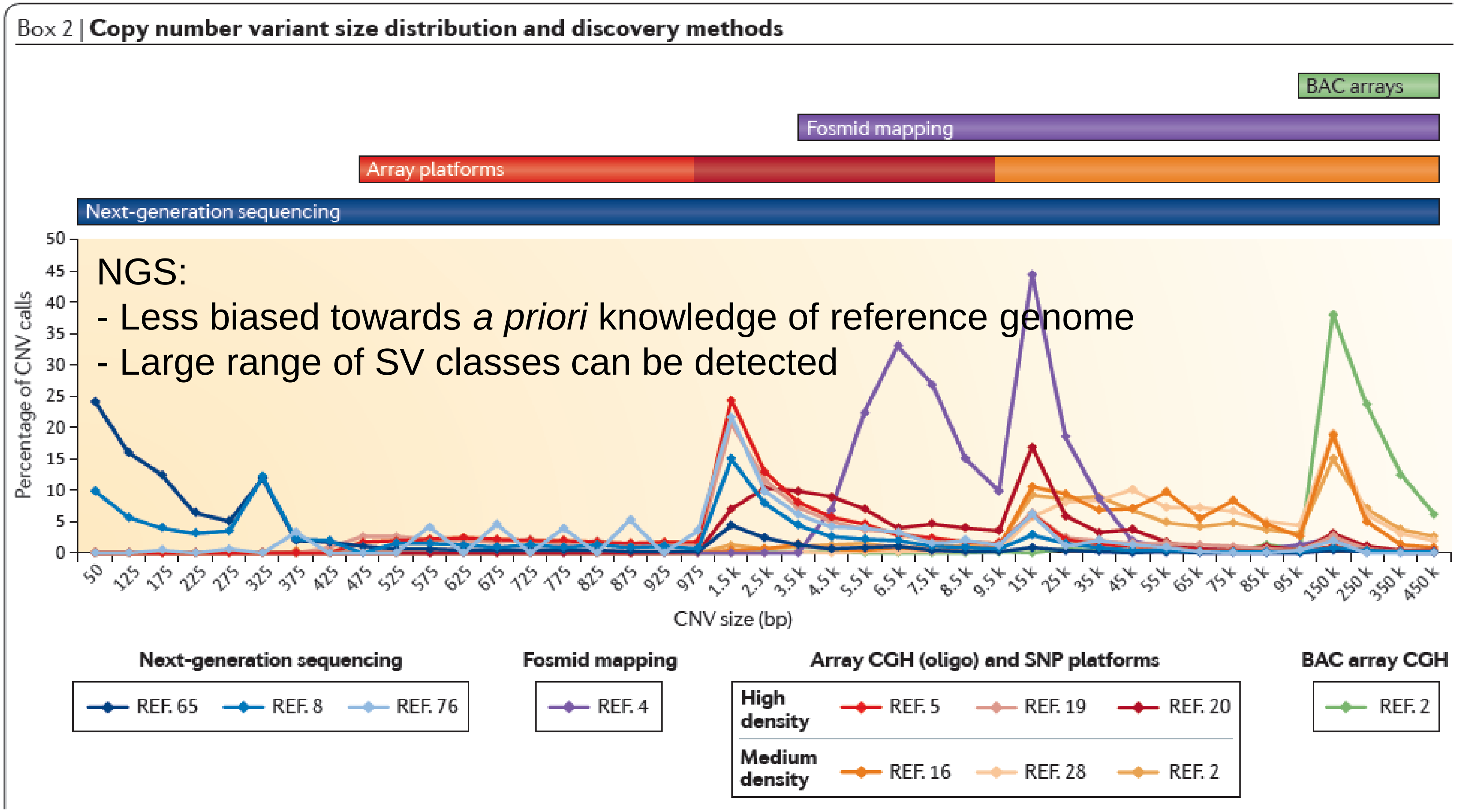
- Fosmid paired-end sequencing
- Next-generation sequencing

Arrays currently still offer higher throughput at lower cost, but this is slowly changing



Medscape education, http://www.medscape.org/viewarticle/585818_2

Experimental SV discovery approaches



NGS:

- Less biased towards *a priori* knowledge of reference genome
- Large range of SV classes can be detected

Alkan *et al*. *„Genome structural variation discovery and genotyping“* Nature Reviews Genetics, 2011

SV sequence signatures (1)

VariationHunter paper,
Lee paper

AGE paper,
SRiC paper

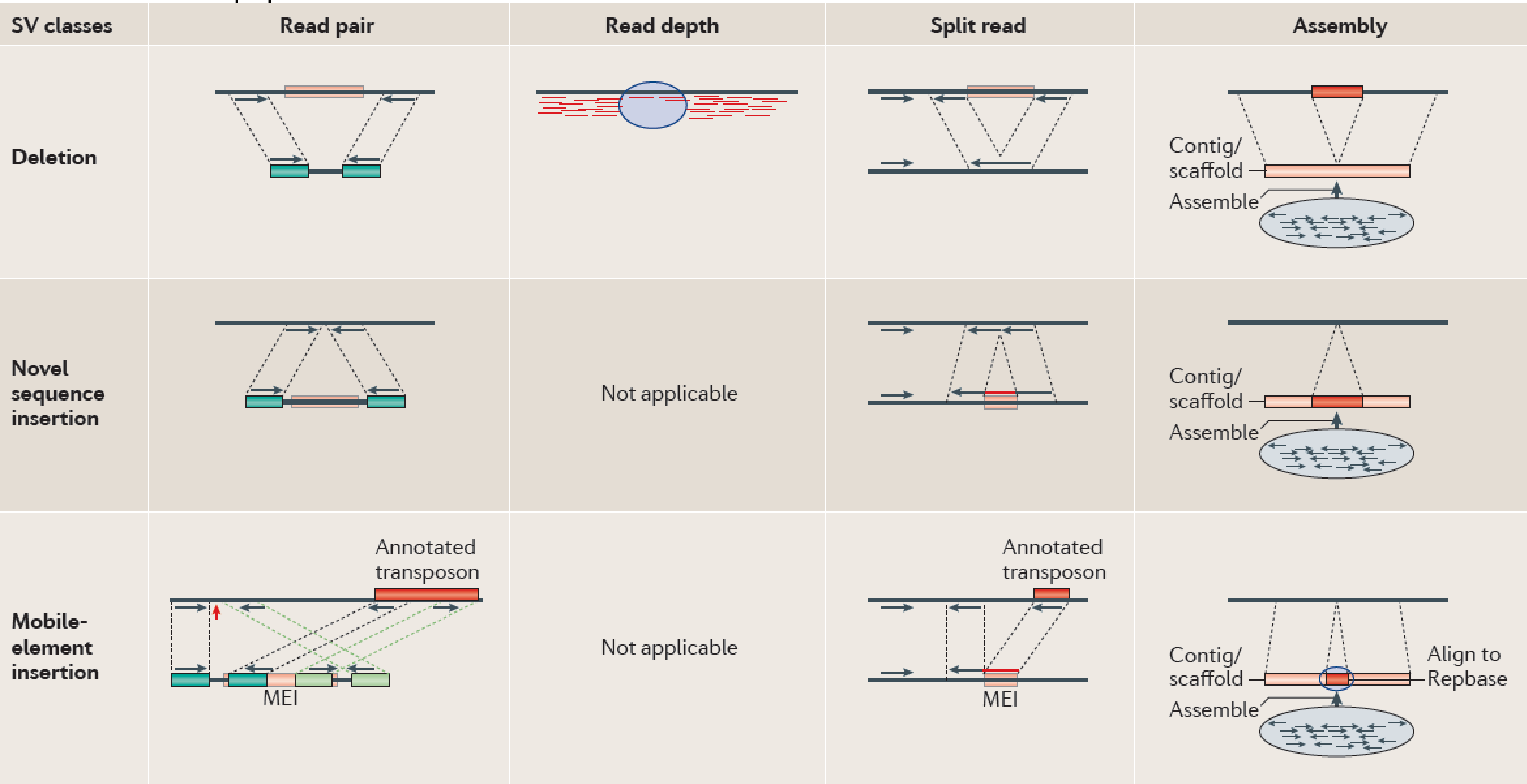


Figure 2 | **Structural variation sequence signatures. (Part 1)**

Alkan *et al*. *„Genome structural variation discovery and genotyping“* Nature Reviews Genetics, 2011

SV sequence signatures (2)

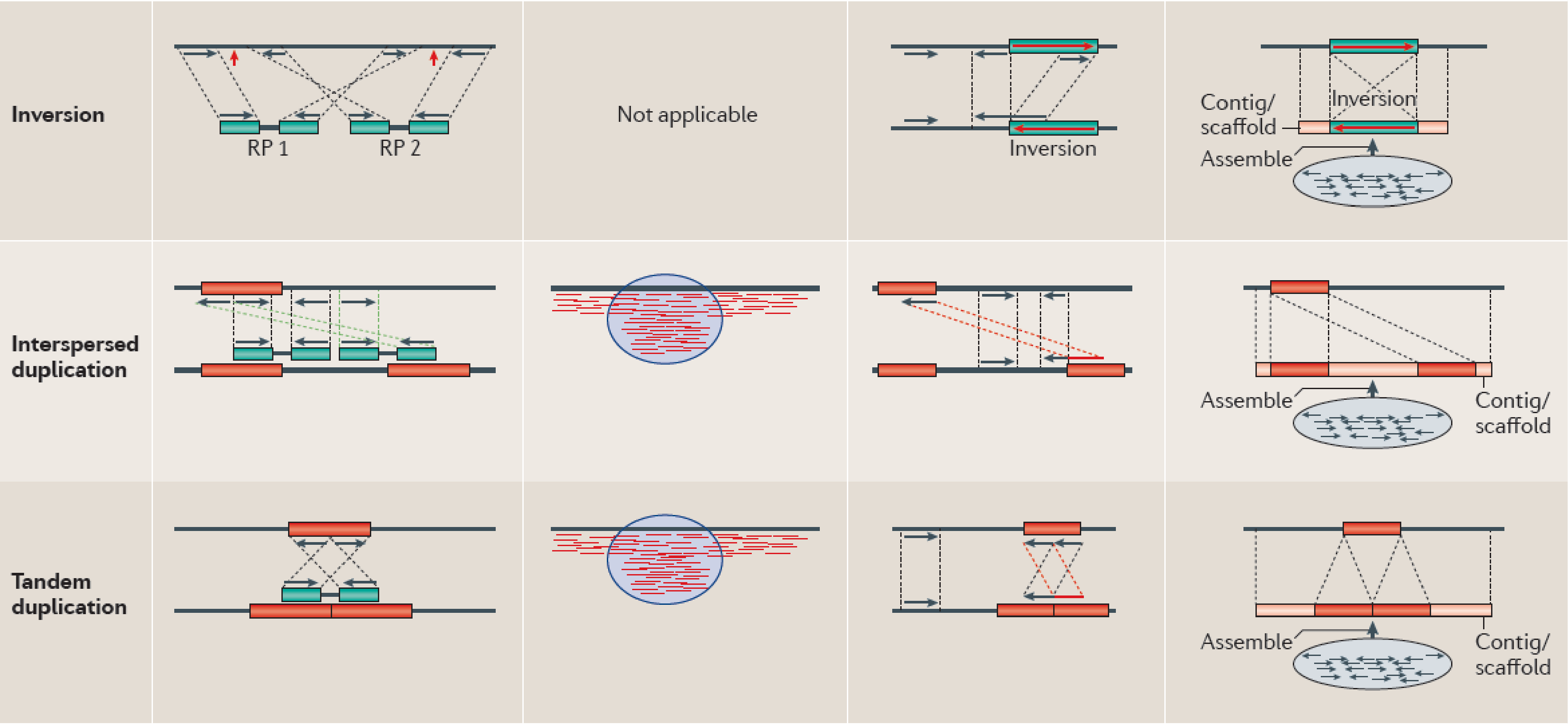
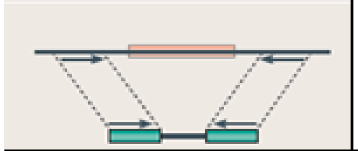
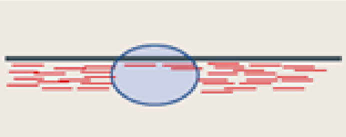
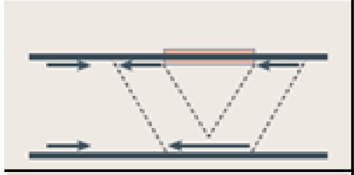
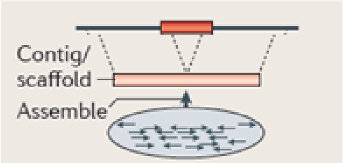


Figure 2 | **Structural variation sequence signatures. (Part 2)**

Alkan *et al*. *„Genome structural variation discovery and genotyping“* Nature Reviews Genetics, 2011

Sequencing-based computational methods

	+	−
Read pair 	Powerful for a wide range of SVs	Not for very small SVs, chimeric reads/fragments lead to wrong mappings
Read depth 	Potential to accurately predict absolute copy numbers of very large SVs	Poor breakpoint resolution, large SVs only, no novel sequence insertions
Split read 	Determines exact size, location and sequence content	Suffers most from hard-to-align repetitive sequence, only detects short insertions
Assembly  	Can theoretically predict all types of SVs	Currently feasible only for local assembly, problematic in repeats

Overlap between methods is low

Indels from 1000 genomes project

Comprehensive methods need to make use of all approaches

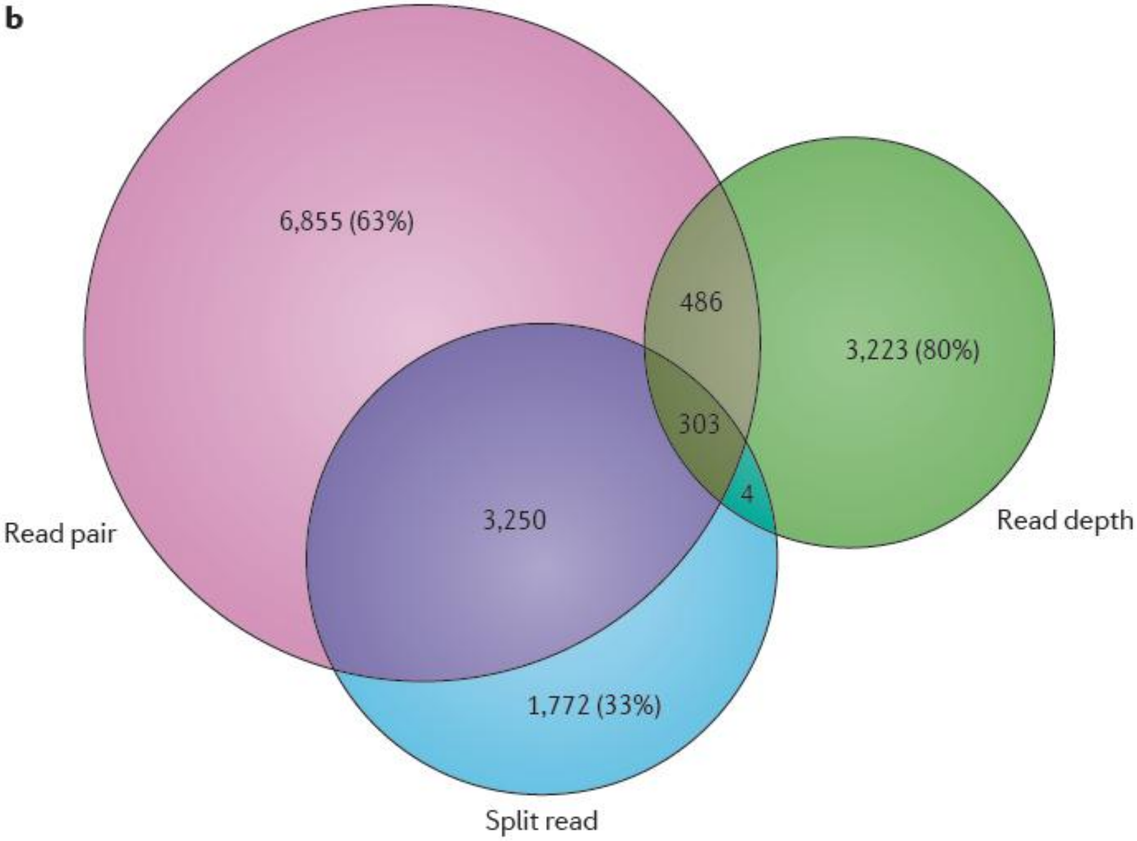


Figure 3 | **Copy number variant discovery biases.**

Alkan *et al*. *„Genome structural variation discovery and genotyping“* Nature Reviews Genetics, 2011

Difficulties/biases

SVs tend to be in repetitive sequence where reads are hard to map

Deletions are „easier“ than insertions → more deletions contained in public databases (e.g. 2:1 ratio in DGV)

With advances in sequencing technology, reads become longer and assembly will become feasible

Conclusions

Different types of sequencing data and applications

→ Different types of read mapping and variant detection algorithms

General difficulties:

- lie in errors and biases of sequencing
- repetitiveness of genome, mapping ambiguities
- computational efficiency