Structure Variant Analysis

Ian Yi-Feng Chang Ph.D
Molecular Medicine Research Center
Chang Gung University
2019/5/3

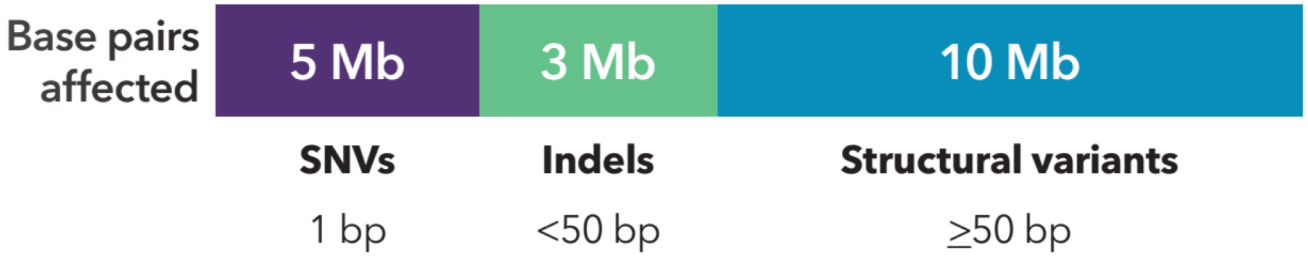
!! Most of the slides were adopted from Bioinformatics.ca (https://bioinformaticsdotca.github.io/CSHL_2019)
High-throughput Biology - From Sequence to Networks 2019 Module 5 !!

Outline

- What are structural variants (SVs)
- How SVs are discovered from NGS data
- The strengths and weaknesses of each SV discovery strategy
- The sequence alignment SV “signals”
- Visually explore read support for SVs

Variation between two human genomes, by number of base pairs impacted

Recent high-quality genome assemblies using PacBio long-read sequencing have revealed that each human genome has approximately 20,000 structural variants, spanning 10 million base pairs, more than twice the number of bases affected by SNVs



SNVs = Single nucleotide variants; indels = insertions and deletions

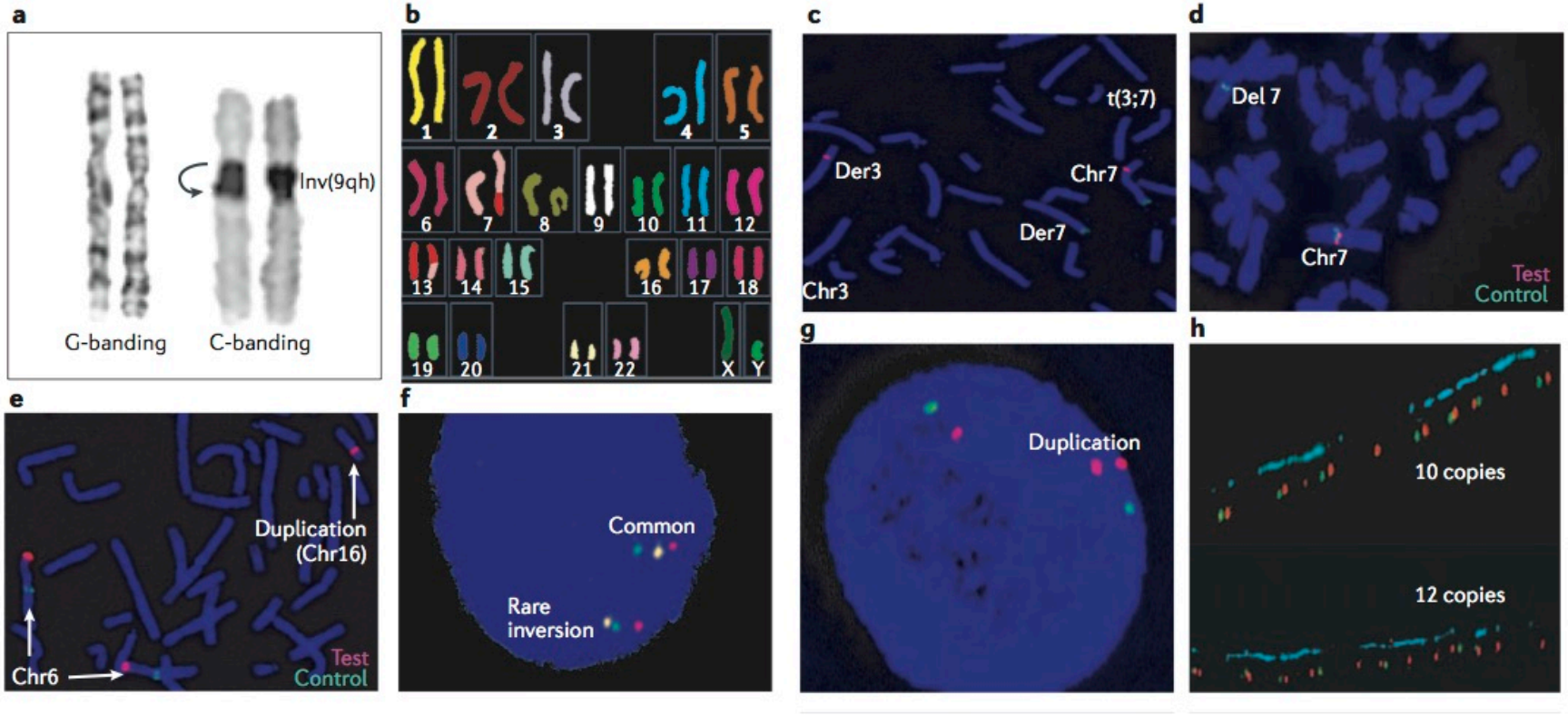
https://www.pacb.com/wp-content/uploads/Whitepaper-Human-Structural-Variation.pdf

What are Structural Variants (SVs) ?

- *Structural Variants (SVs)*: Genomic rearrangements that affect >50bp (or 100bp, or 1Kb) of sequence, including:
  - deletions
  - novel insertions
  - inversions
  - mobile-element transpositions
  - duplications
  - translocations

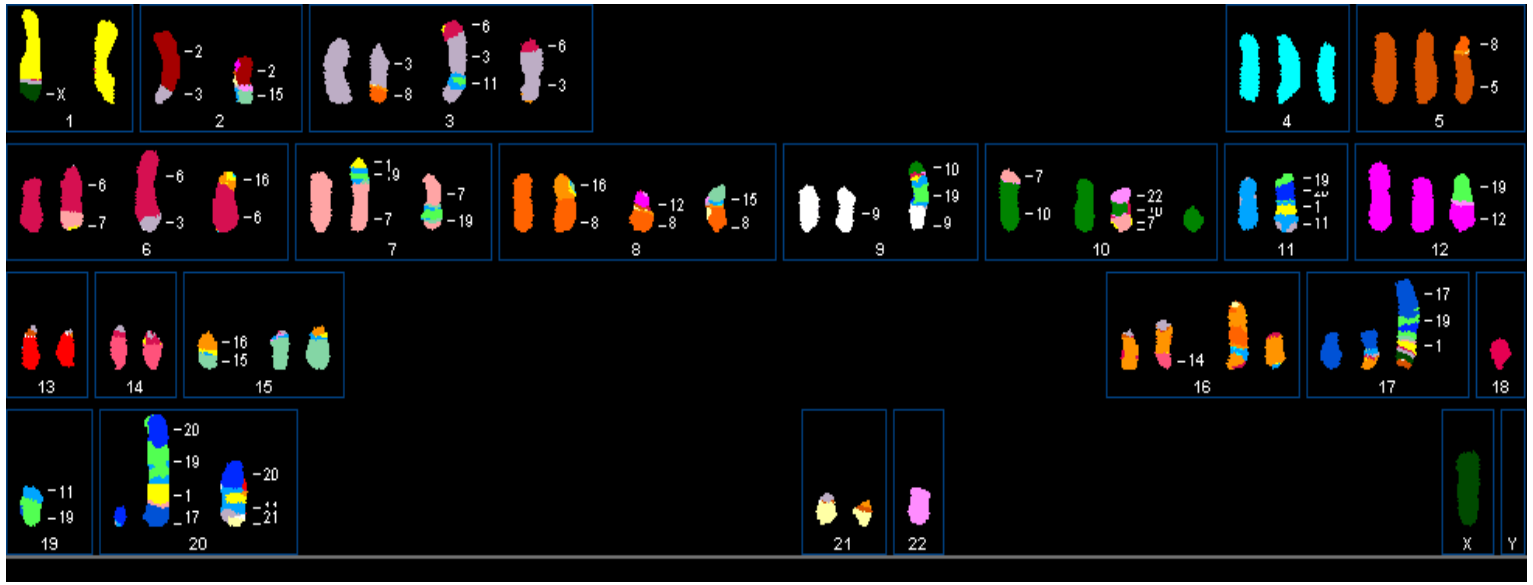
Alkan et al. *Nat Rev Genet* 2011

Detection and confirmation of SVs



Feuk et al. *Nat Rev Genet* 2006

Structural variants in cancer



Can higher resolution maps help identify recurrent aberrations and driver mutations in cancer?

Structural variants of all types are known to cause Mendelian disease and contribute to complex disease

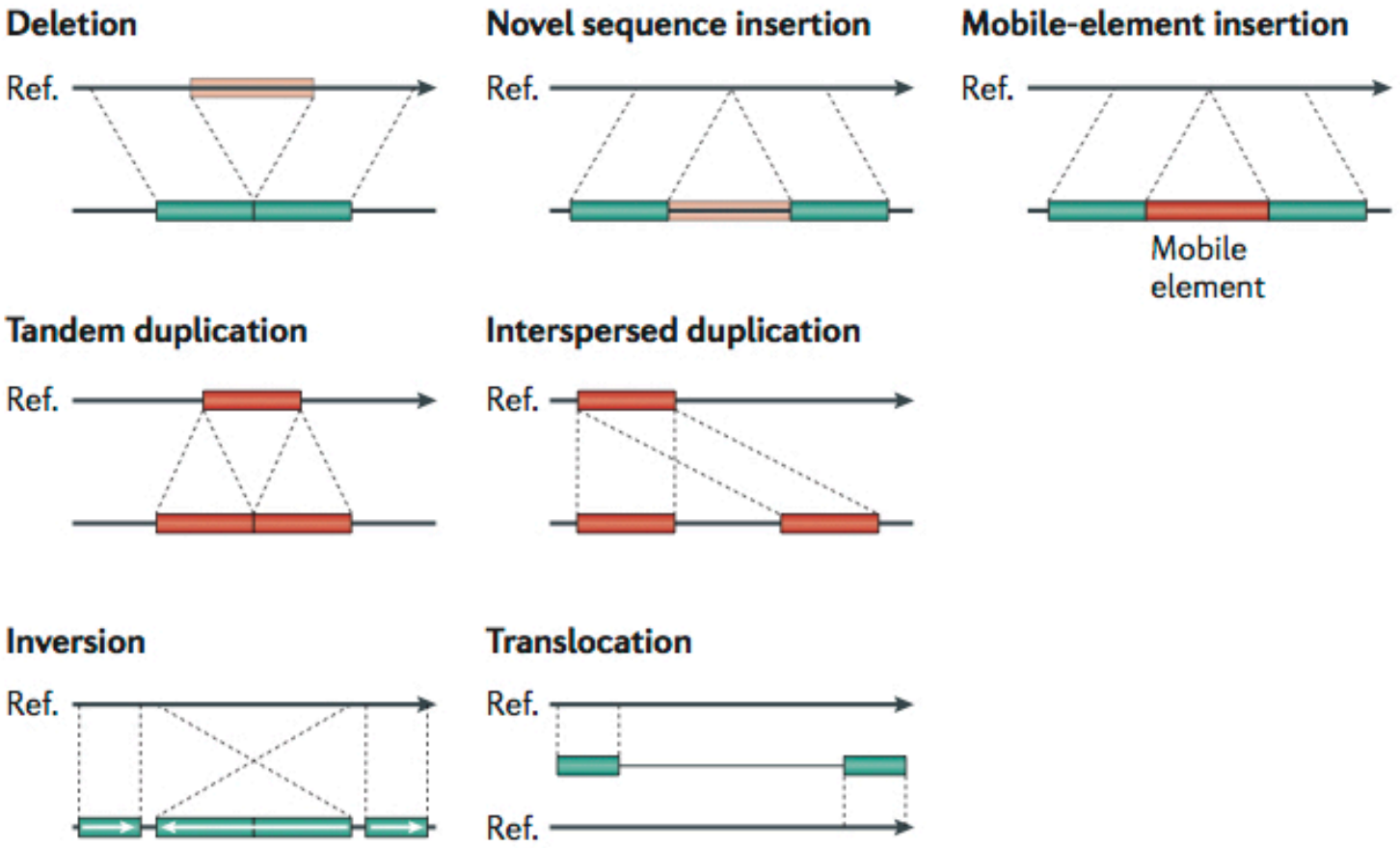
Structural Variant	Disease Examples
INSERTION 	Charcot-Marie Tooth disease, Tay-Sachs disease
DELETION 	Williams syndrome, Duchenne muscular dystrophy, Smith-Magenis syndrome, Carney Complex
INTERSPERSED DUPLICATION 	APP in Alzheimer's disease, Potocki-Lupski syndrome, Prader-Willi syndrome, Angelman syndrome
TRANSLOCATION 	Down syndrome, XX male syndrome (SRY), schizophrenia (chr 11) , Burkitt's Lymphoma
INVERSION 	Hemophilia A, Hunter Syndrome, Emery-Dreifuss muscular dystrophy
TANDEM DUPLICATION 	FMR1 in Fragile-X, Huntington's disease, Spinocerebellar ataxia

https://www.pacb.com/wp-content/uploads/Whitepaper-Human-Structural-Variation.pdf

Classes of SVs

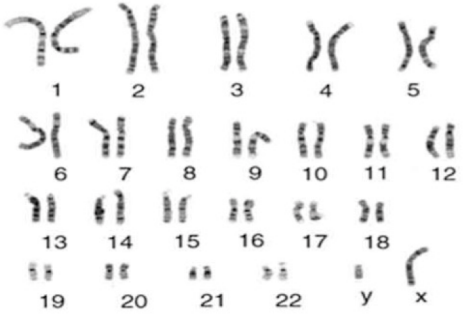
- Copy number variants (CNVs):
  - Deletions
  - Duplications
- Copy neutral rearrangements:
  - Inversions
  - Translocations
- Other structural variants:
  - Novel insertions
  - Mobile-element transpositions

Classes of SVs

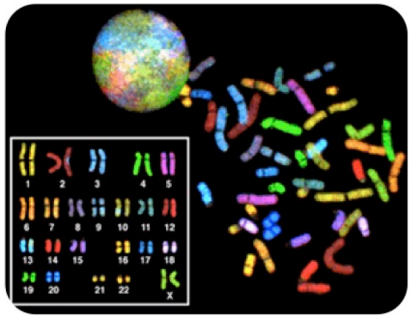


Alkan et al. *Nat Rev Genet* 2011

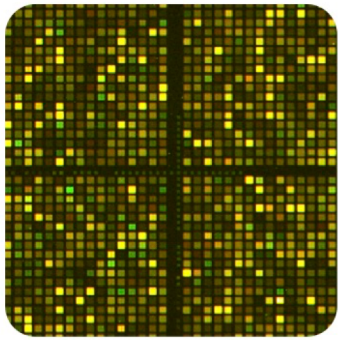
Our understanding is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA

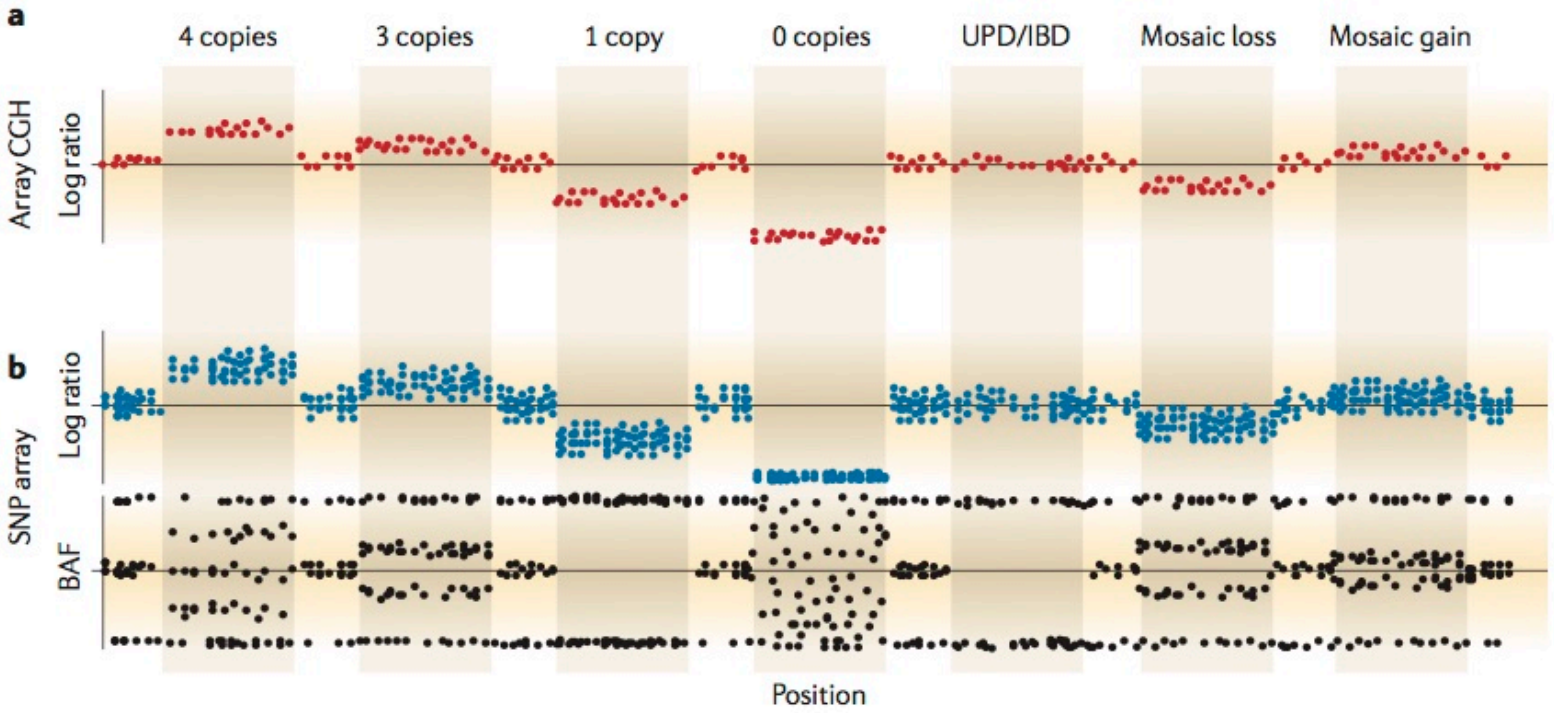


2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

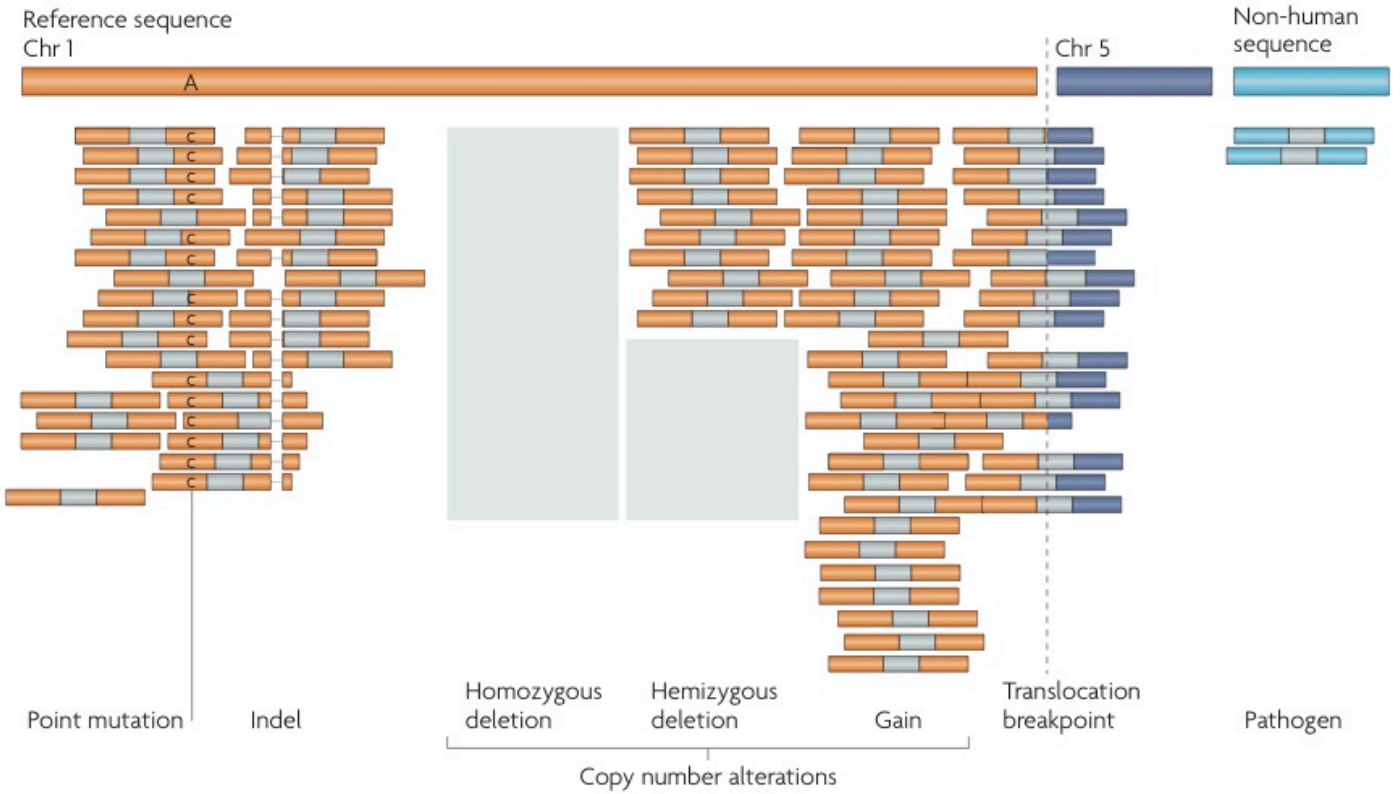


Today
High throughput
DNA sequencing

Array-based detection of CNVs

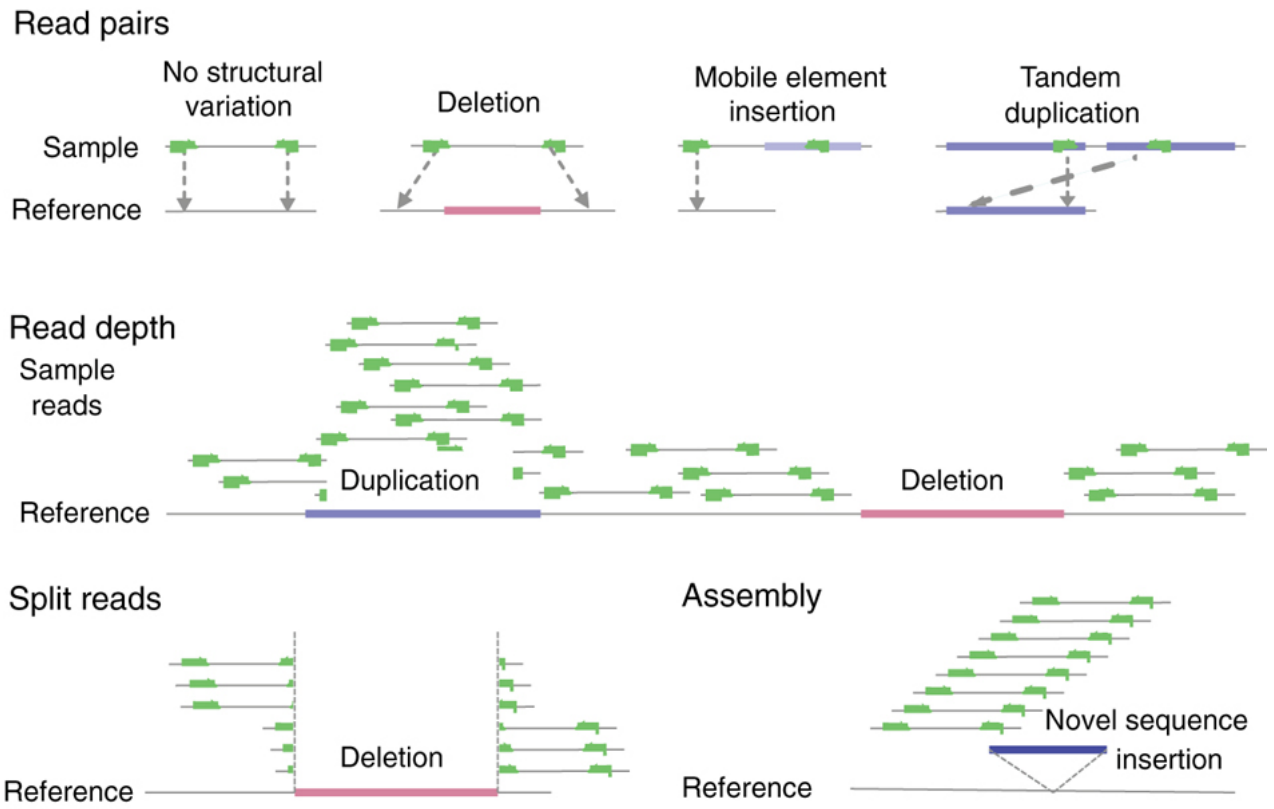


Detecting SVs from NGS data

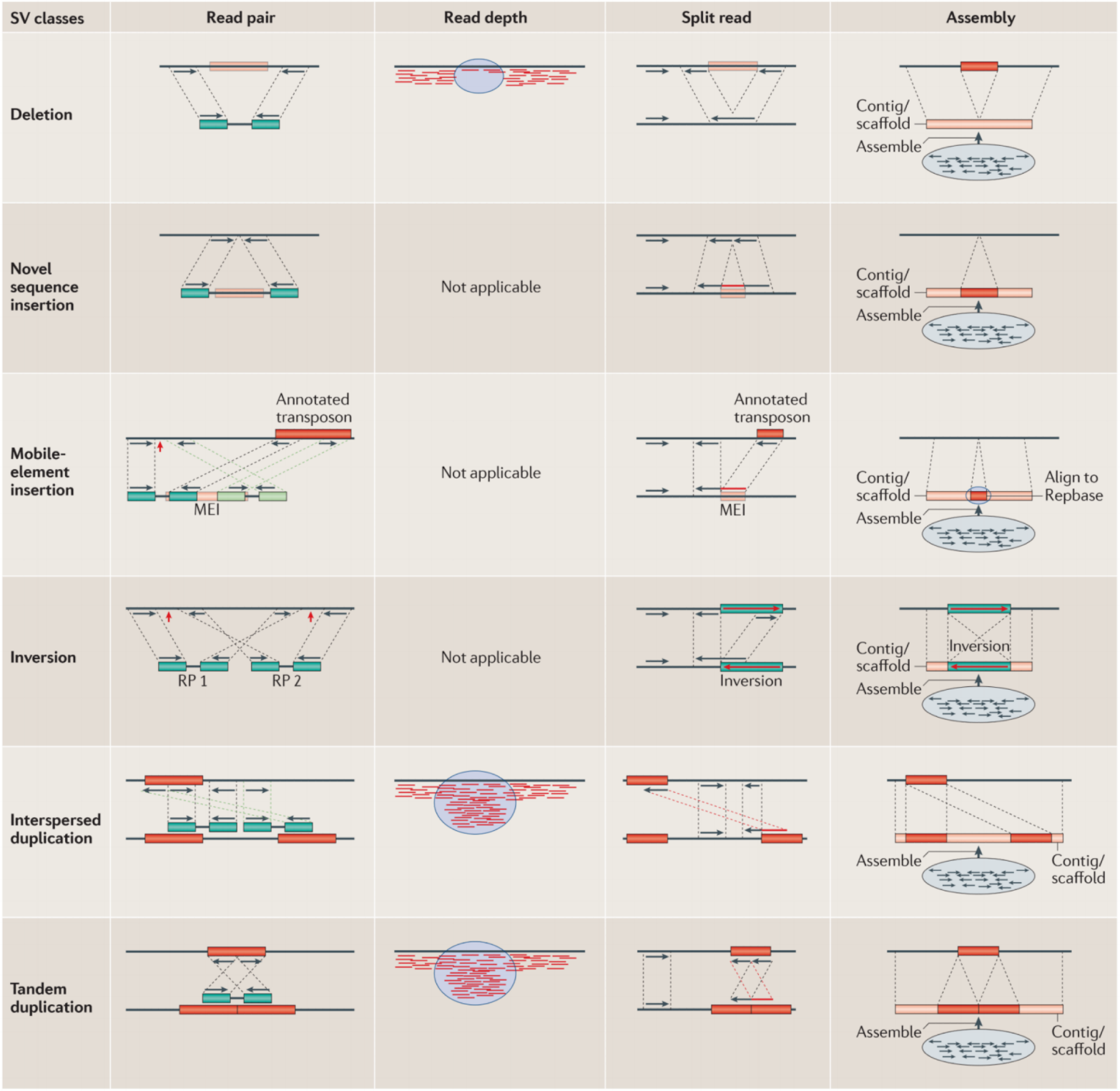


Meyerson et al. *Nat Rev Genet* 2010

Strategies for calling SVs from NGS data



Baker *Nat Methods* 2012



Structural variation sequence signatures

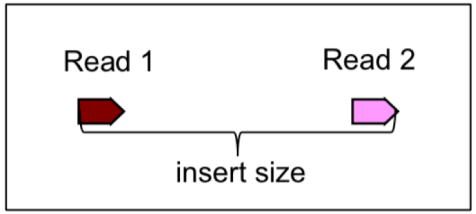
Alkan et al. *Nat Rev Genet* 2011

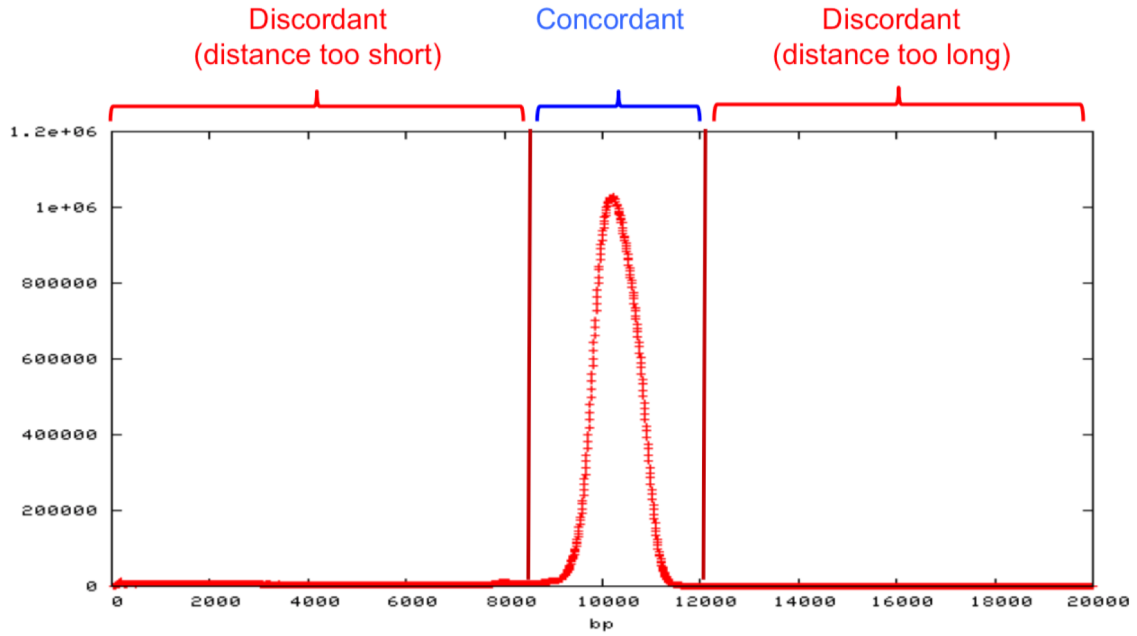
Read pair

Rational

- Identify SV breakpoints by examining the alignments.
- Sequencing libraries are created with fragments of known length.
- Read pairs spanning an SV break point will produce ‘discordant’ alignments with an unexpected alignment distance and/or orientation
- Current fragment sizes limit sensitivity, discordant mapping patterns can be difficult to interpret at complex SVs

Discordant read pairs

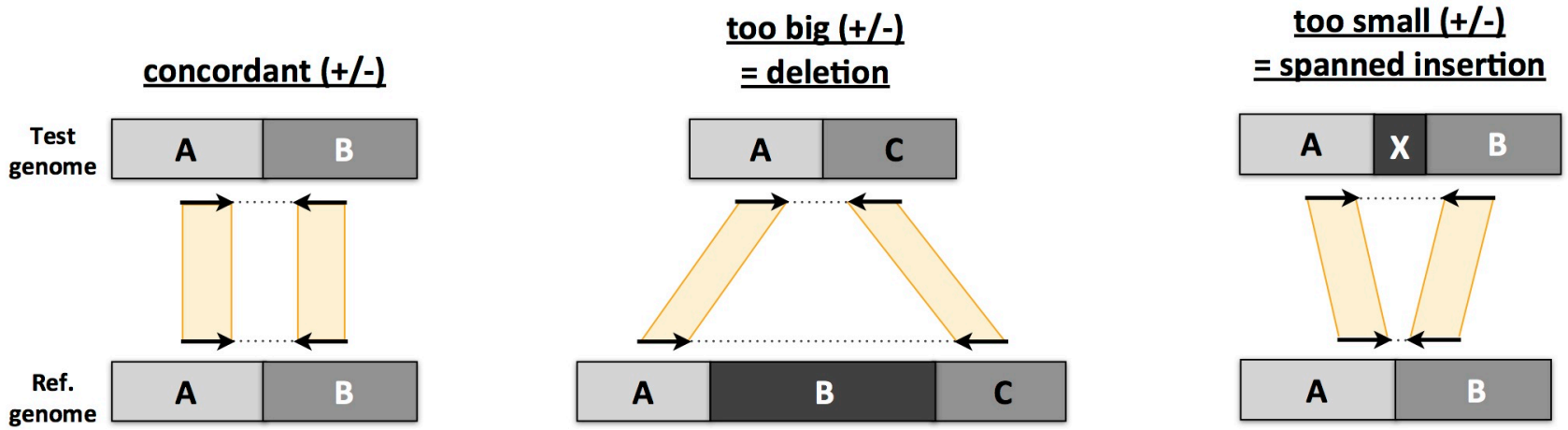




Genomic distance between mapped paired tags

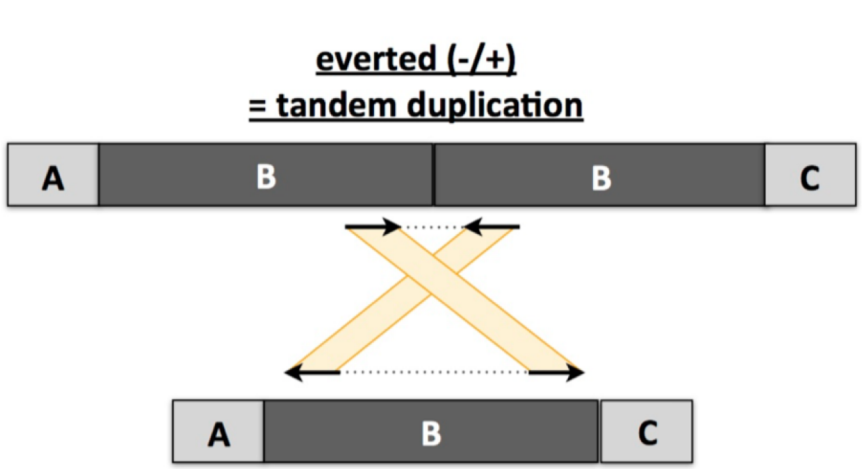
Reads pairs are also Discordant when order or orientation isn't as expected.

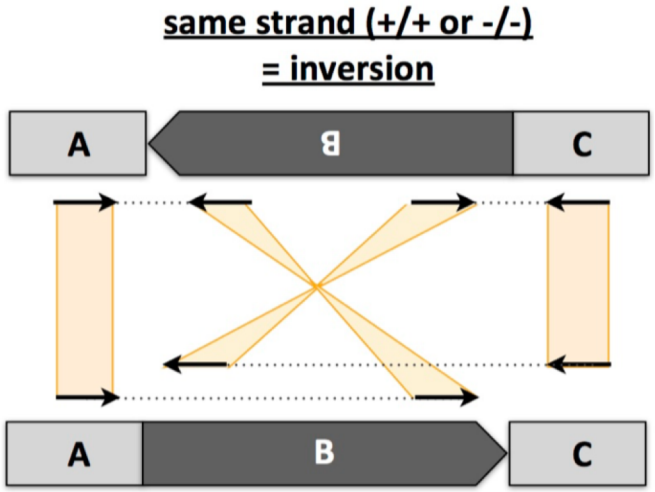
Using discordant reads to detect SVs



From Aaron Quinlan

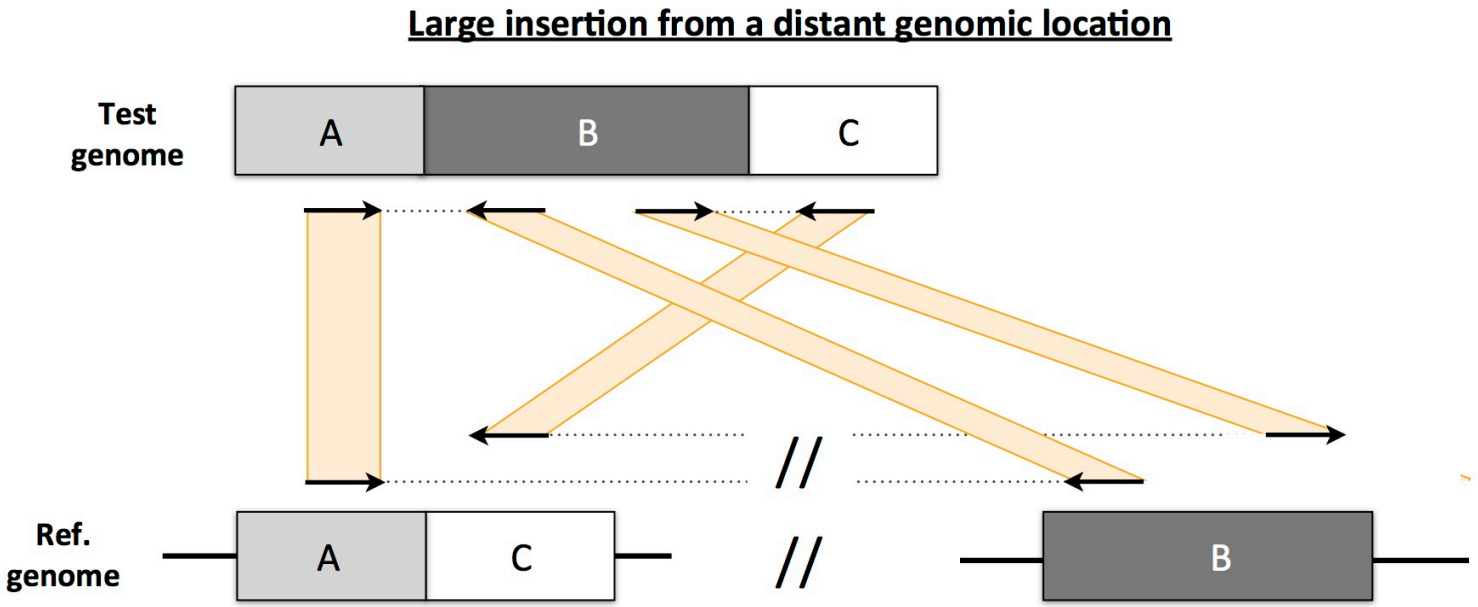
Using discordant reads to detect SVs





From Aaron Quinlan

Using discordant reads to detect SVs



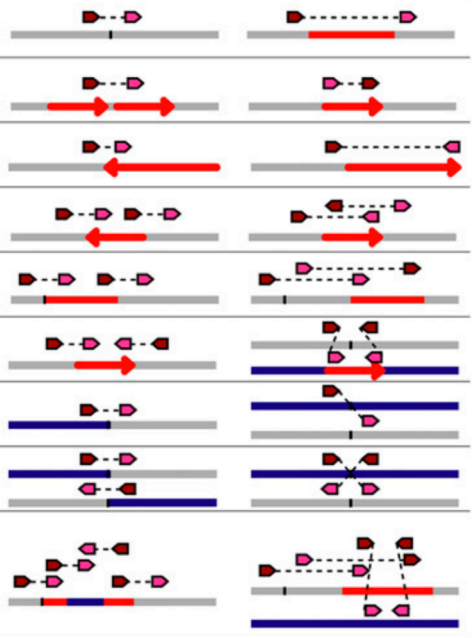
From Aaron Quinlan

Read-pair tools

- BreakDancer
- VariationHunter
- MoDIL
- GASV-PRO
- DELLY
- LUMPY
- GenomeSTRiP

Read-pairs have difficulty to detect balanced translocation

	Normal		Breast tumor					Breast cancer cell line			Gastric tumor				Gastric cancer cell line	Other cancer cell lines		Interpretation	Mapping to reference
	European	African	Breast tumor 1	Breast tumor 2	Breast tumor 5	Breast tumor 13	Breast tumor 14	MCF7	SKBR3	T47D	Gastric tumor 17	Gastric tumor 26	Gastric tumor 28	Gastric tumor 38	TMK1	HCT116	K562		
Deletion	460	189	226	280	131	682	159	352	606	223	640	341	1027	375	571	614	500		
Tandem duplication	28	25	13	40	22	39	46	203	78	32	71	62	29	28	103	74	140		
Unpaired inversion	43	38	31	32	24	85	70	83	135	37	180	68	38	40	183	49	86		
Inversion	58	64	26	38	38	38	56	60	56	40	80	48	62	54	76	58	56		
Intra-chr. insertion	20	16	6	10	8	22	16	14*	14	6	19*	15*	16	15*	27*	24	16		
Inter-chr. insertion	6	0	0	6	2	6	7*	8	6	0	14	4	6	4	11*	12	13*		
Isolated translocation	18	14	11	13	12	34	22	59	45	24	77	28	14	14	48	13	36		
Balanced translocation	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0		
Complex intra-chr.	18	2	0	2	0	33	40	146	158	6	32	6	31	12	150	26	73		
Complex inter-chr.	15	8	0	5	5	18	18	122	47	6	13	14	15	8	82	13	19		
Physical coverage	130	72	17	35	24	61	57	101	68	45	141	40	106	80	233	77	92		



Hillmer et al. Genome Research 2011

Read-pair summary

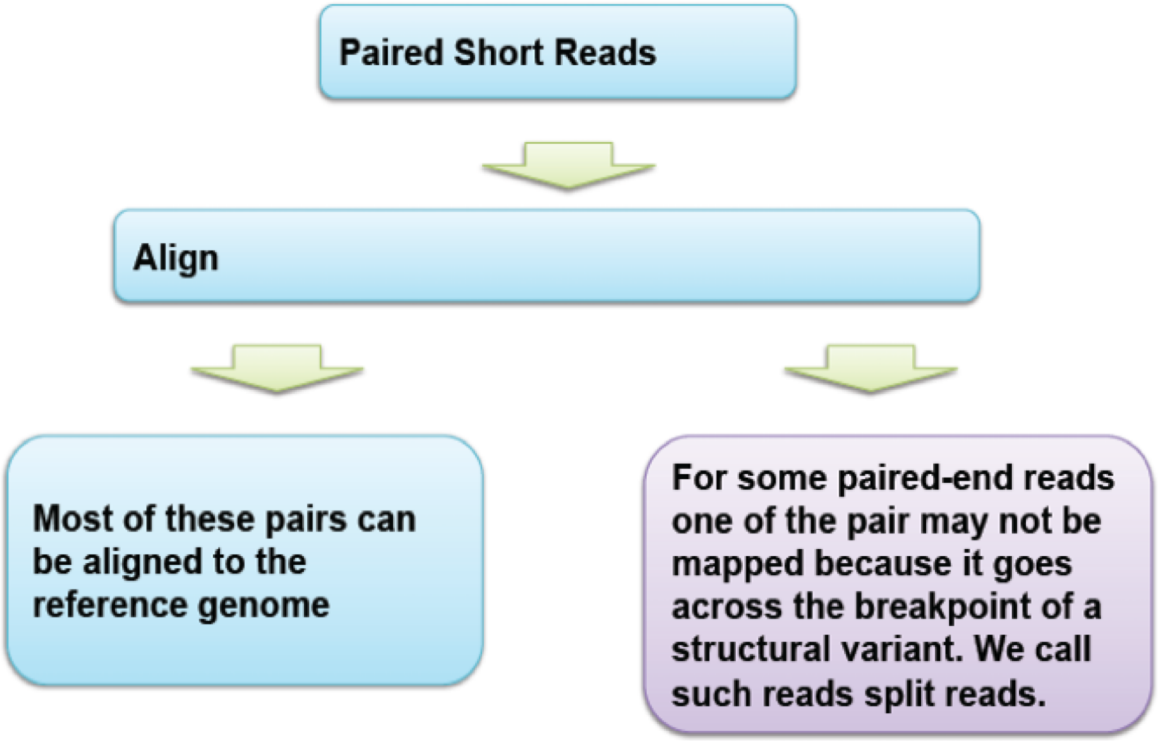
- Weaknesses
  - Difficult to interpret read-pairs in repetitive regions
  - Difficult to fully characterize highly rearranged regions
  - High rate of false positives
- Strengths:
  - Most classes of variation can, in principle, be detected

Split read

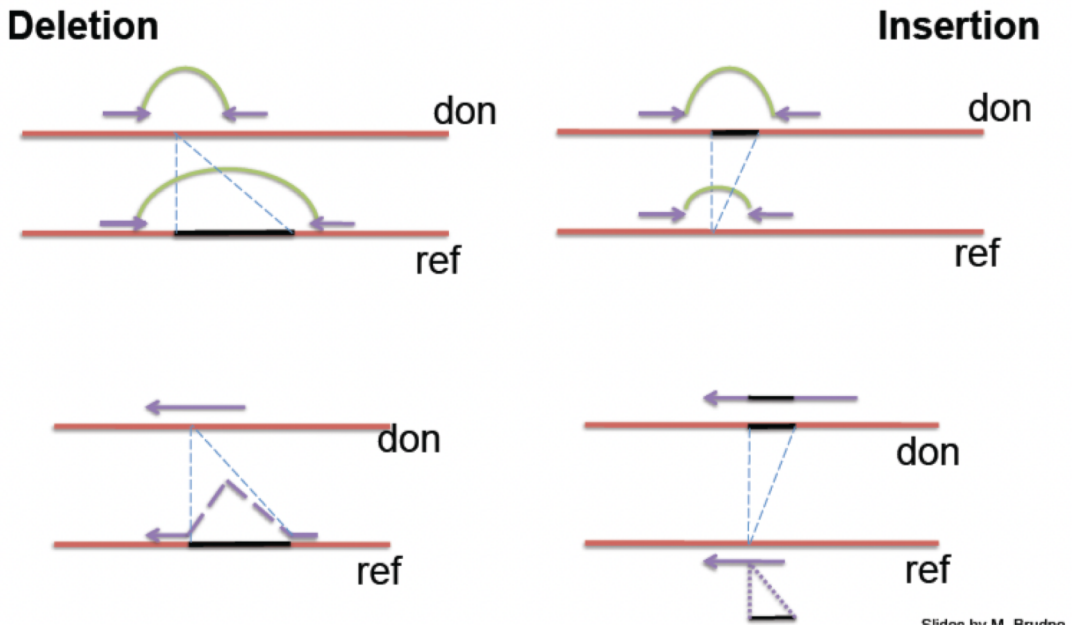
Rational

- SR identifies sequences that actually contain a breakpoint
- The alignments for such sequences are ‘split’ because DNA segments flanking the breakpoint align to disjoint locations in the reference genome
- SR inherently maps breakpoints to single base
- Long-read (>500 bp) SR is a particularly powerful approach for studying complex SV because multiple breakpoints can potentially be captured by a single read, aiding in variant locus reconstruction

Split reads mapping



Insertion and deletion signatures



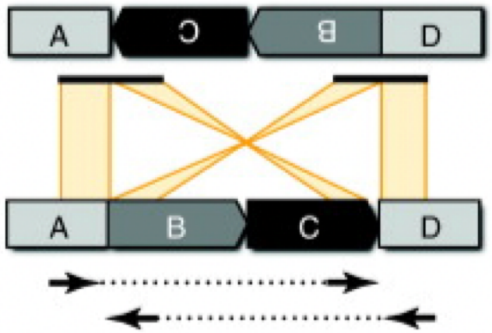
RP signature

SRM signature

Inversion signature

don

ref

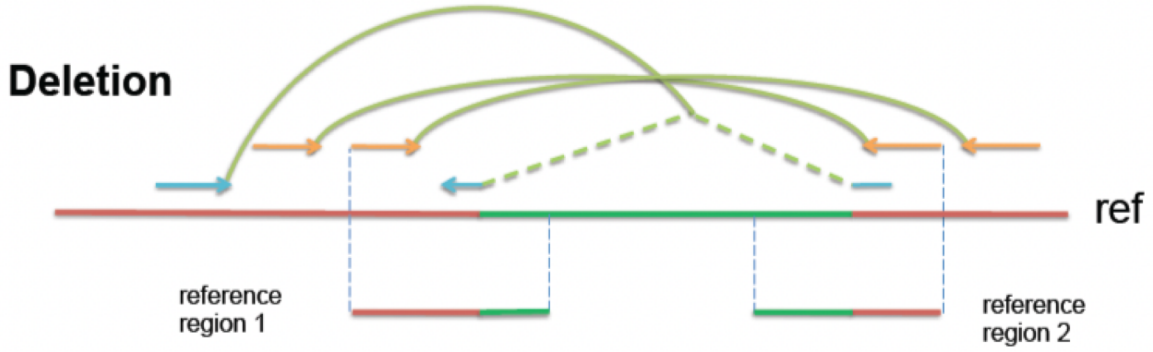


SRM signature

PEM signature

Importance of the paired-reads usage

- searching the whole genome for split mappings
  - gives a lot of false mappings
  - too slow
- can exactly estimate breakpoint and indel sizes
- can detect very small deletions



Split read tools

- Pindel
- DELLY
- LUMPY
- PRISM
- Mobster

Split reads summary

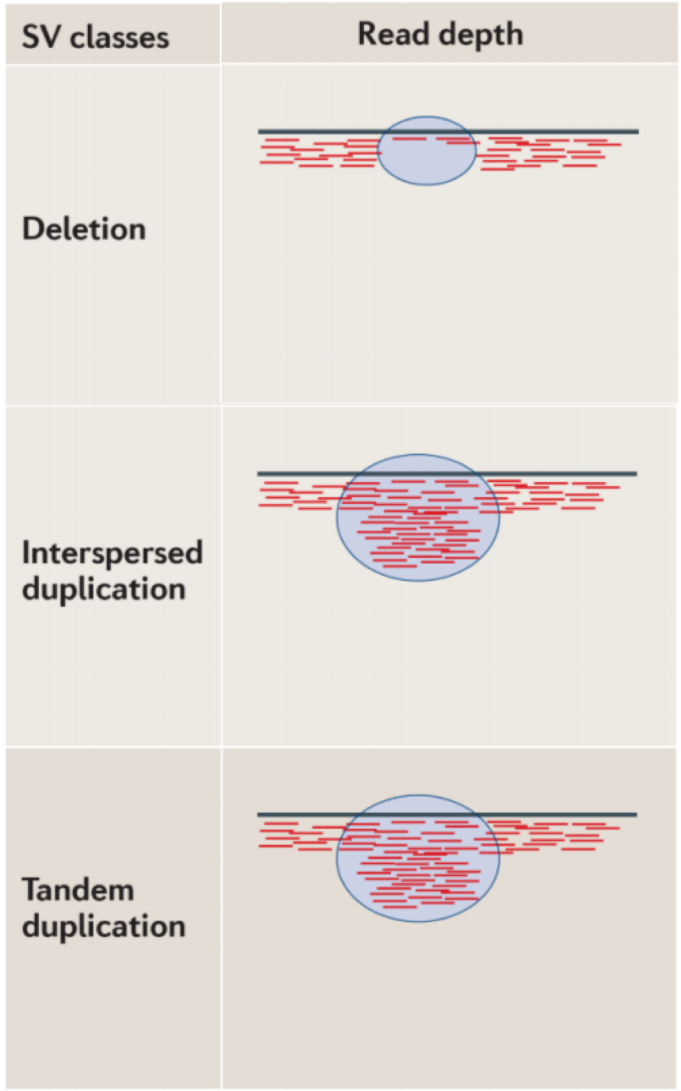
- Weaknesses
  - Requires sufficient coverage
  - Can have false positives especially in repetitive regions
- Strengths:
  - Can be added to read-pairs methods Base-pair resolution of breakpoints
  - Can detect very short events

Read depth

Rational

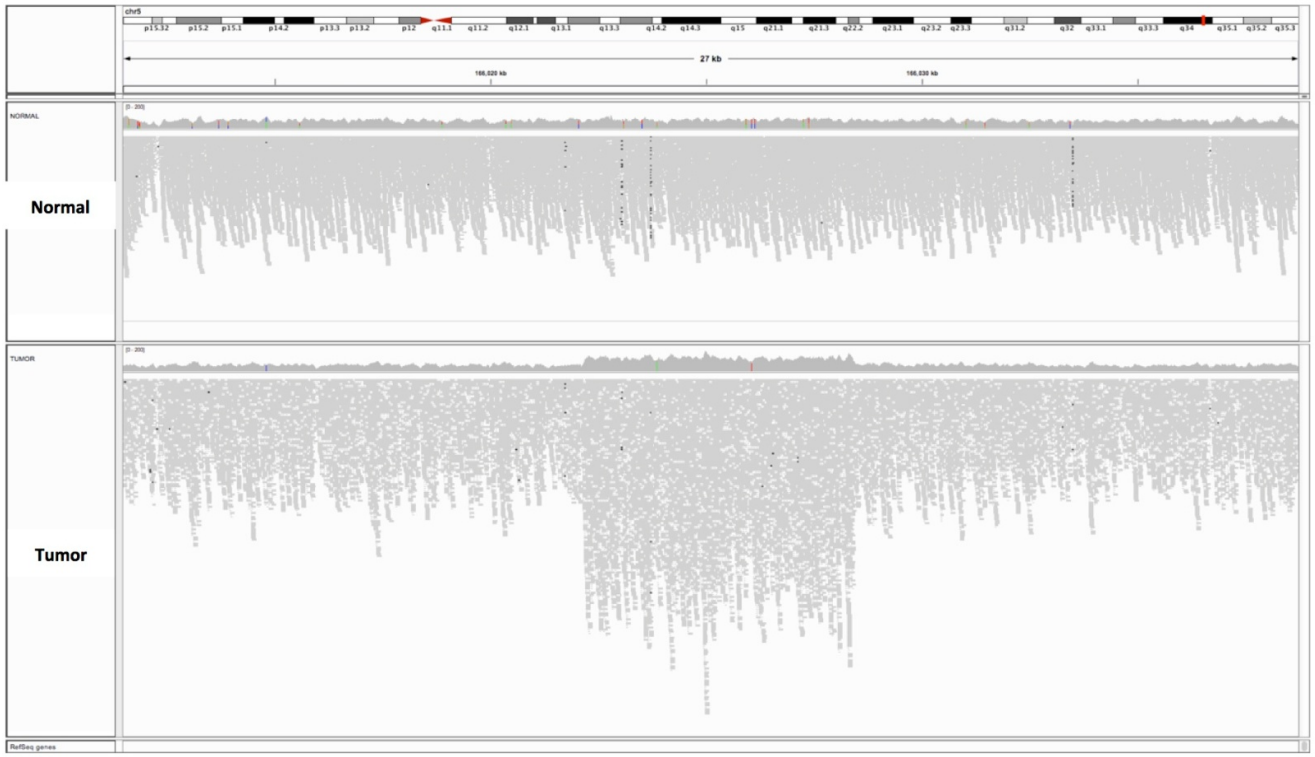
- Assuming an homogenous distribution of read (WGS): CNVs are evidence of significant increases or decreases in the depth of aligned sequence
- Divide the genome in bins of equal size
- Estimate the DOC in each bin
- Looking for cluster of consecutive bins that shows significant excess or loss of coverage

Read-depth signature



Alkan et al. *Nat Rev Genet* 2011

Read-depth



Basic approach:

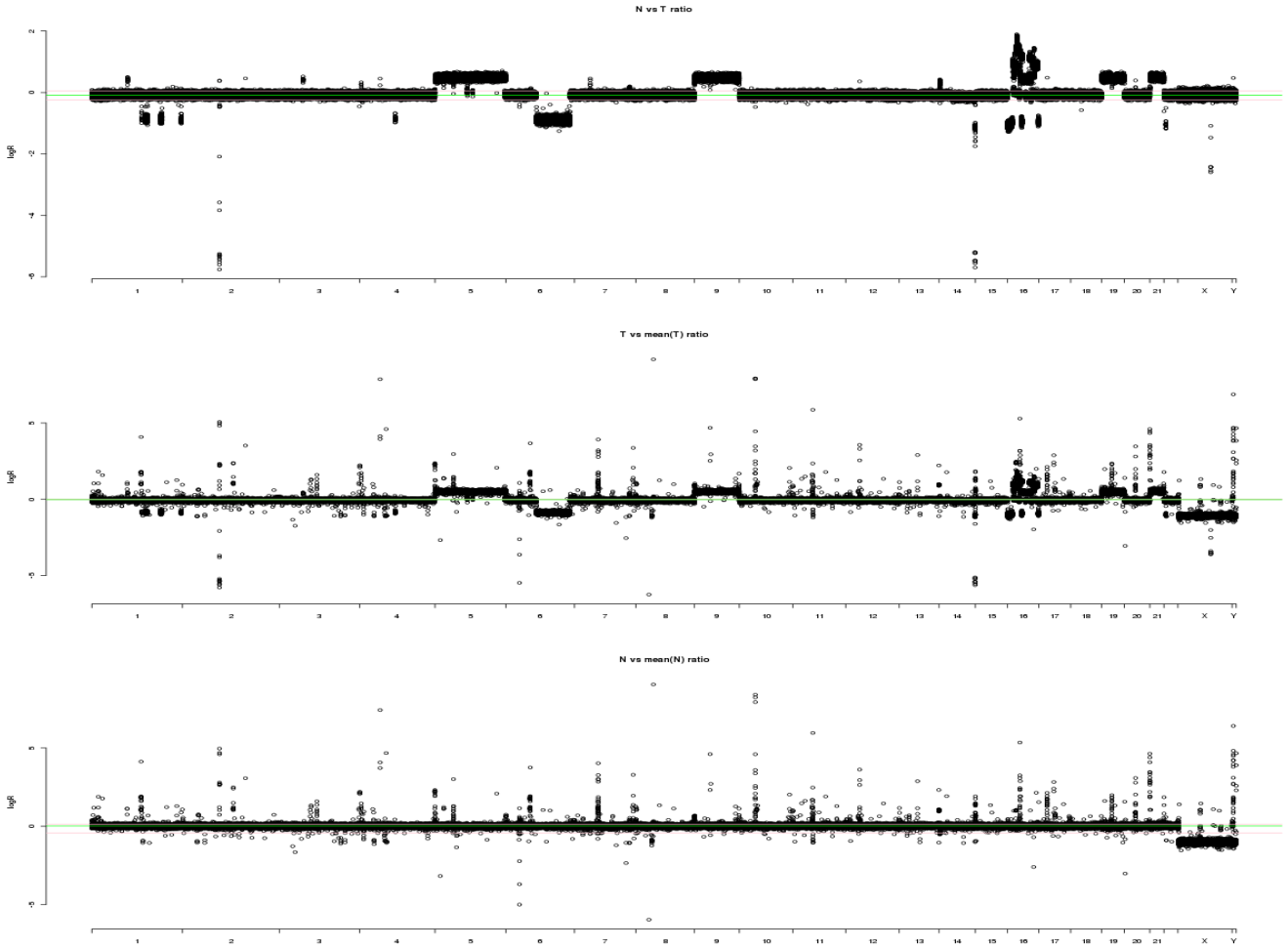
1) Count reads in sliding windows (e.g., 5kb) across the genome.

2) Normalize for GC bias.

3) Use segmentation to define CNAs (similar to array-CGH data).

From Aaron Quinlan

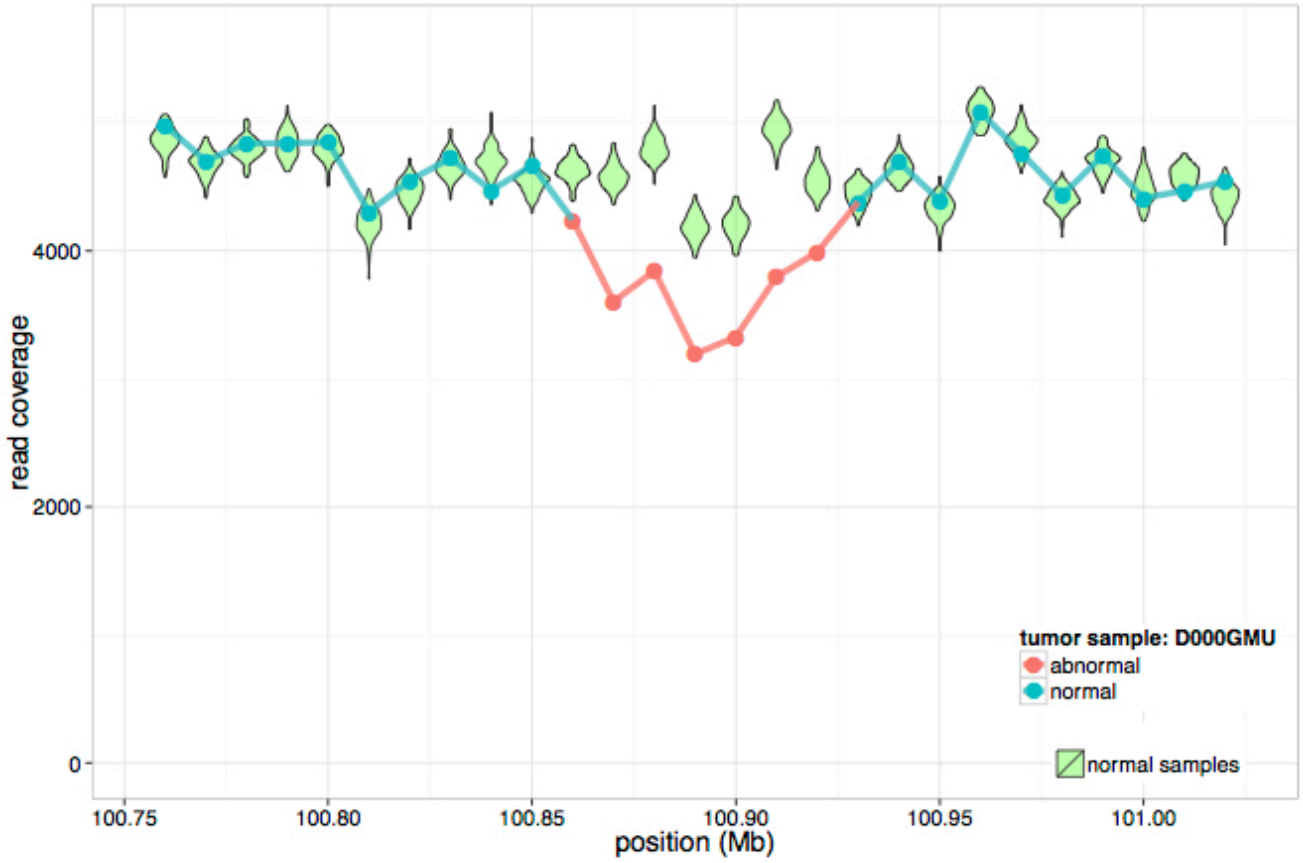
Read-depth genomic visualization



Normalization issues

- Noisy or reduced signal in repeat-rich regions, centromeres, and telomeres.
- Unpredictable segmentation ➔ reduced sensitivity/specificity.
- Filtering problematic regions reduced the genome range tested.

Population based SV detection : PopSV



Monlong et al. *BioRxiv* 034165

Read depth tools

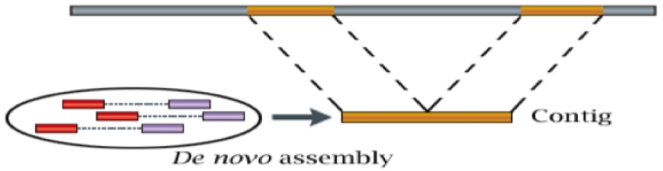
- ReadDepth
- RDXplorer
- cnvSeq
- CNVer
- CopySeq
- GenomeSTRiP
- CNVnator
- SCoNEs
- PopSV
- CNVKits
- CopywriteR

Read depth summary

- Weaknesses
  - Relatively low resolution (normally ~5-10Kb)
  - Breakpoint are ambiguous
  - Cannot detect balanced rearrangements (e.g.,
  - inversions), or transposon insertions
- Strengths:
  - Fast and simple
  - Easy to interpret
  - Determines DNA copy number (unlike most other methods)
  - Provides useful information even with low coverage, albeit at low resolution

Assembly

Rational



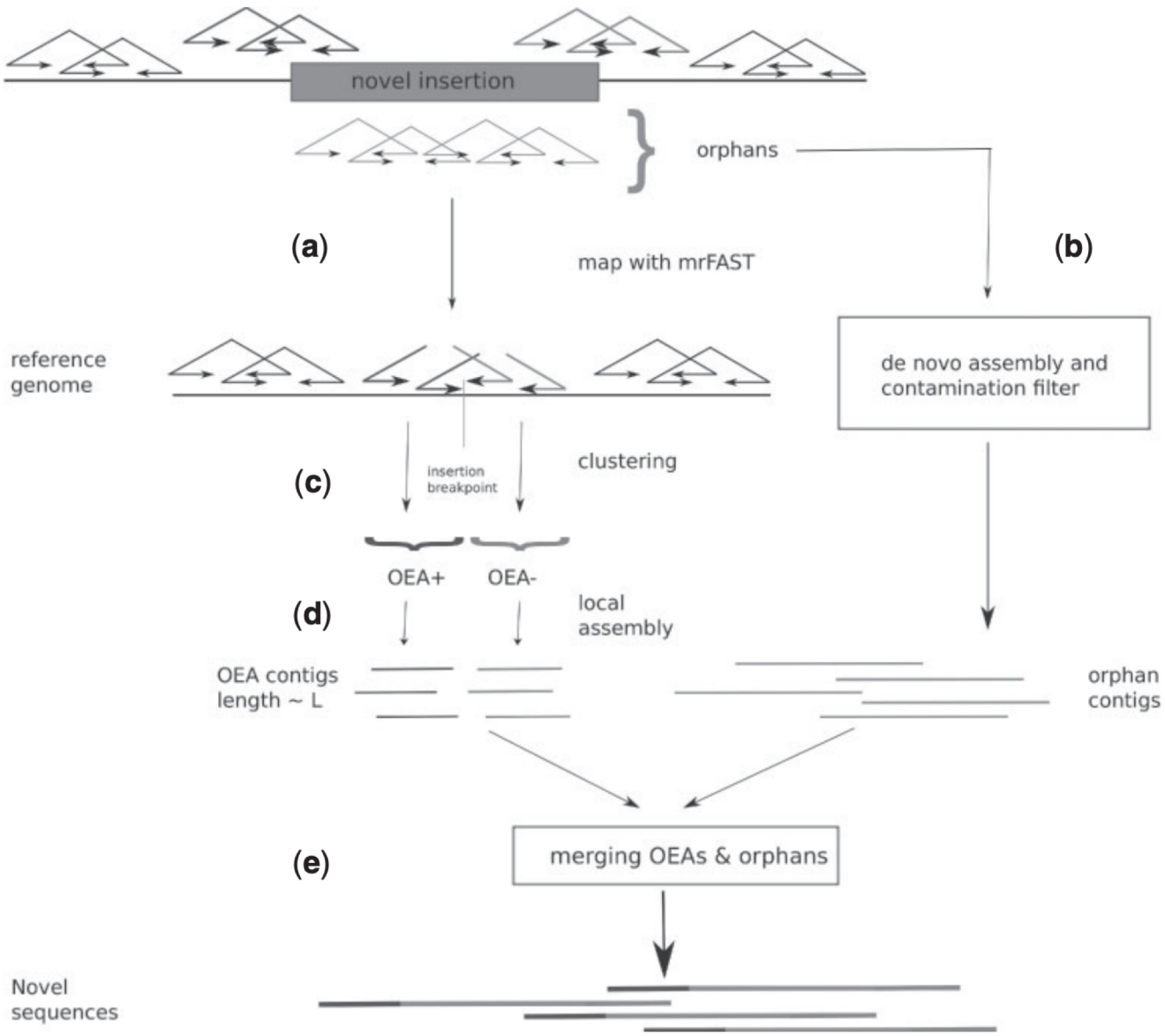
- *De novo* sequence assembly can be used to reconstruct a variant locus.
- Assembled sequences are aligned to the reference genome, SV are discerned similarly than in SR methods.
- Assembly typically generates substantially larger '**contigs**' that are more suitable to characterizing complex SV.
- Two different approaches:
  - Whole genome assembly
  - Local assembly

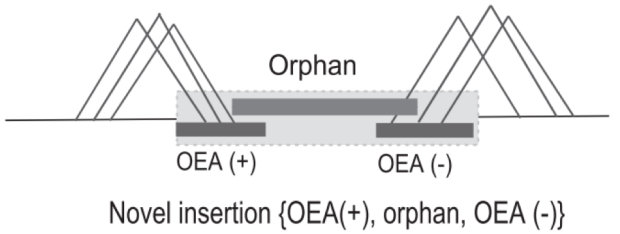
Whole genome de novo assembly

- De novo sequence assembly
- Comparison with ref sequence
  - BLAT : pre-align all assembled scaffold
  - Group scaffold located on the same chromosome
  - Align scaffold set to specific chromosome using an
  - insertion tolerant algorithm
  - Report hits that contribute most to the colinearity between a scaffold and a chromosome if confirmed by at least several different alignments
- These hits potentially contain de novo insertion and deletion events

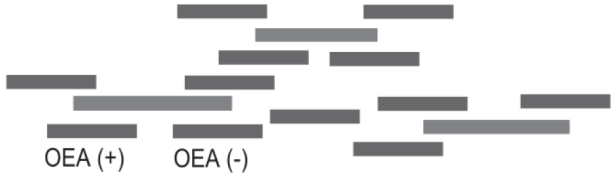
Local *de novo* assembly

- *De novo* assembly of orphan reads and one-end anchored (OEA) reads
- Mark contaminated contigs (i.e, Epstein-Barr virus)
- Blast on the reference genome
- Use OEA mapped ends to validate blast hits

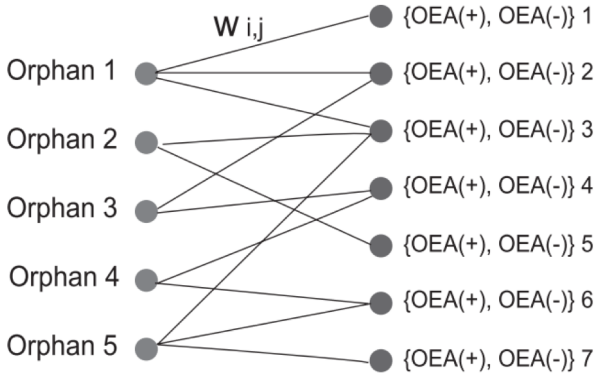




Novel insertion {OEA(+), orphan, OEA (-)}



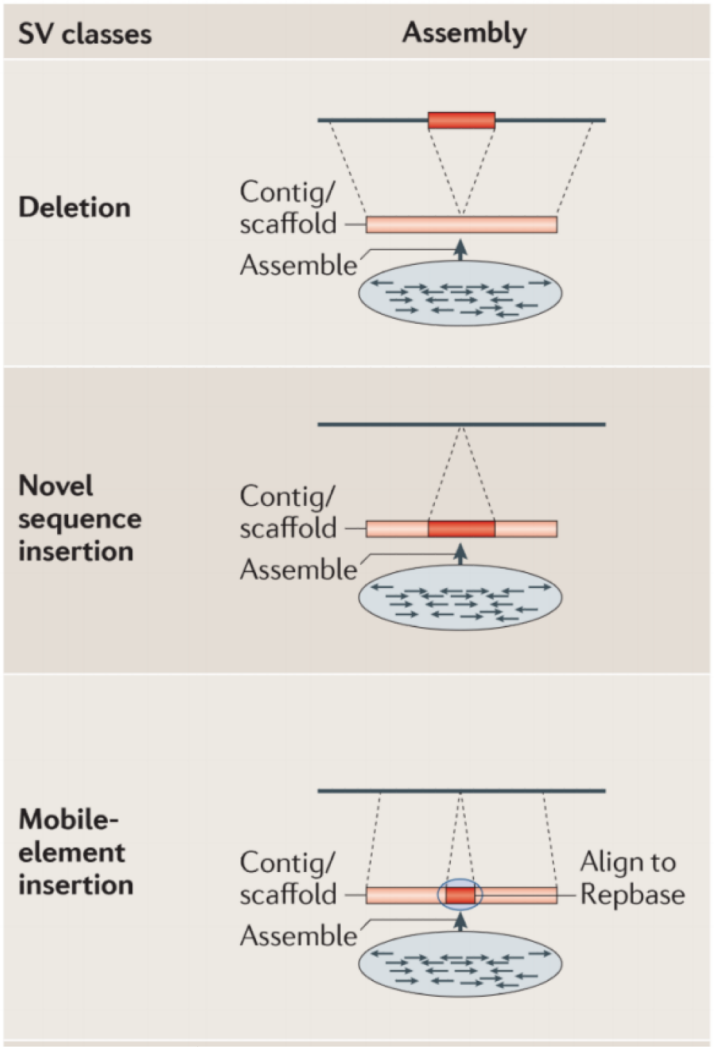
Overlaps between {OEA(+), OEA(-)} and orphan contigs

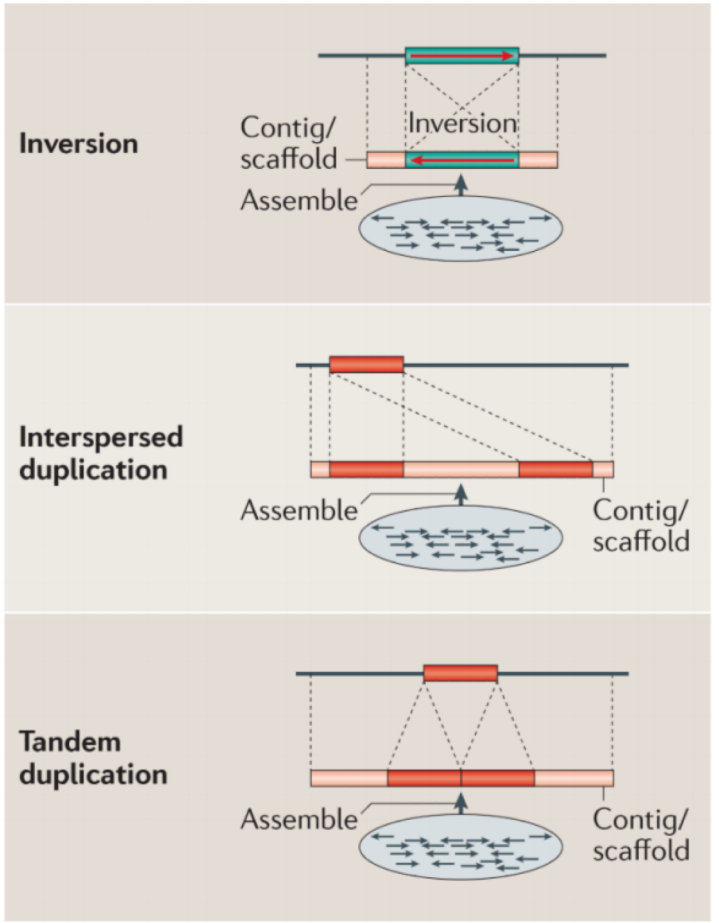


Maximum Weighted Matching

Hajirasouliha, et al., Bioinformatics 2010: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC286586

Assembly signature





Alkan et al. *Nat Rev Genet* 2011

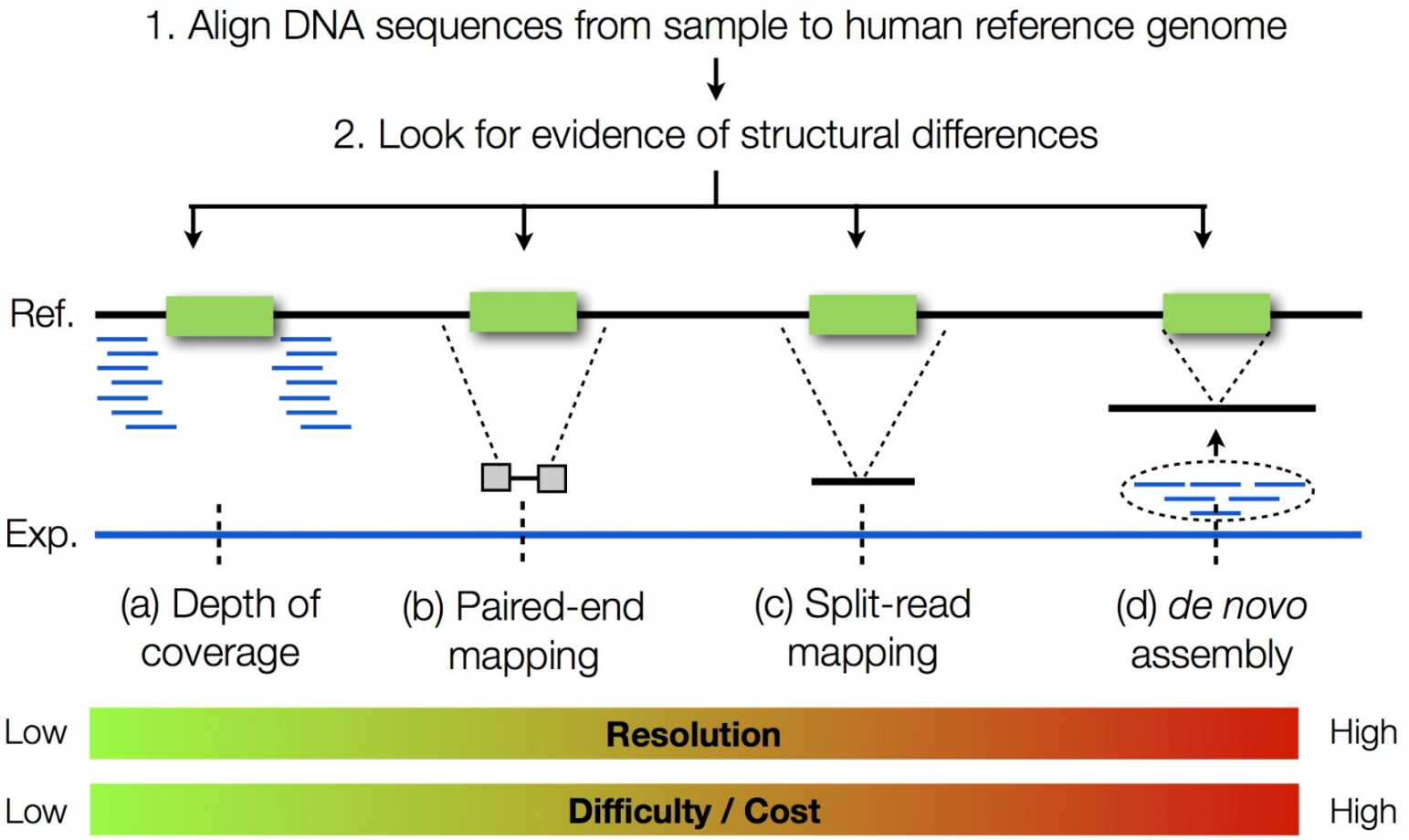
De novo assembly tools for SVs

- Cortex
- SGA
- DISCOVAR
- ABySS
- Ray
- SvABA

Assembly for SVs summary

- Weaknesses
  - Computationally very intensive
  - Hard to resolve repetitive and complex regions
- Strengths:
  - Base-pair resolution of breakpoints
  - All classes of variation can, in principle, be detected

Summary of strategies for calling SVs



From Aaron Quinlan

SV detection summary

- Four different methods exist with their own strength and weakiness.
- More recent tools are now combining several approach to gain in accuracy
  - Delly
  - Lumpy
- Major challenges are
  - Complex events
  - The breakpoint
  - The validation

SV detection benchmarking

GIAB Gold standard

- NA24385 (HG002): Based on union of >1 mil sequence-resolved calls >=20bp using 30+ callers from 5 short, long, linked read dataset from GIAB community – includes optical mapping
  - Tier 1: >=49b isolated, sequence-resolved SVs (edit distance <20% and from 2+ technologies and 4+ callsets across the trio)
  - Tier 2: Regions with at least one likely SV but likely complex or unable to determine consensus sequence change

Dataset	Total SVs	Number of SVs	Size distributions
tier1	12745	5464 DEL, 7281 INS	DEL: -50 bp to -114 kb INS: 50 bp to 125kb
tier2	33279	17664 DEL, 15615 INS	DEL: -11 bp to -997kb INS: 4 bp to 47kb

SV: NA24385 – HG002

- Evaluate 12745 tier 1 DELs from 5 datasets across 3 technologies

Dataset	Precision	Recall	F1	RAW SV size	TP SV size
Pacbio_pbsv_2.0.1*	0.9521	0.9555	0.9538	-20bp to -98.1kb	-50 bp to -69.1kb
Pacbio_HySA	0.8702	0.2279	0.3612	-11bp to -23.3kb	-39 bp to -23.3kb
Illumina_metasv	0.8320	0.2069	0.3314	-30bp to -4.271Mb	-100 bp to -114.3kb
10x_52x_comb	0.9662	0.5917	0.7340	-41bp to -90.25Mb	-50 bp to -114.2 kb
10x_84x_comb	0.9764	0.5522	0.7055	-41bp to -90.25Mb	-50 bp to -114.2kb

DNASeq pipeline: Germline HG002

- Evaluate 12745 tier 1 DELs from 5 datasets across 3 technologies

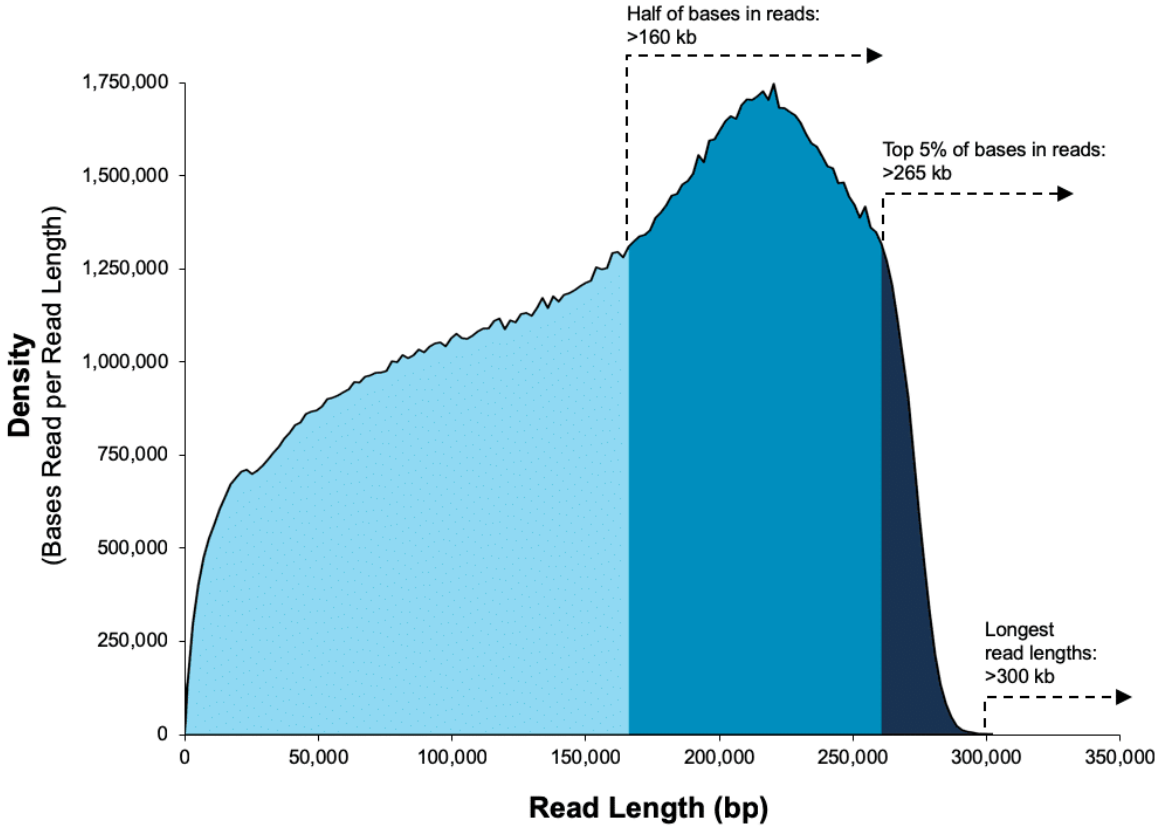
caller	version	genome	precision	recall	F1-score
manta	1.4.0	b37	0.8572	0.7357	0.7918
wham+svtyper	1.8.0+0.7.0	b37	0.8823	0.5153	0.6506
lumpy	0.2.13+0.7.0	b37	0.8892	0.5306	0.6646
delly	0.7.9	b37	0.8508	0.1547	0.2617
delly	0.8.1	b37	0.7019	0.3321	0.4509
cnvkit	0.9.3	b37	0.8333	0.0095	0.0188
ensemble_metasv_1caller_raw	0.5.4_noassembly	b37	0.8248	0.7426	0.7815
ensemble_metasv_1caller_pass	0.5.4_noassembly	b37	0.9150	0.6529	0.7620

PacBio **the Sequel II System**

Single Molecule High-Fidelity (HiFi) Sequencing

Half the data in reads: >160 kb

Data per SMRT Cell: Up to 320 Gb





https://www.pacb.com/smrt-science/smrt-sequencing/read-lengths/

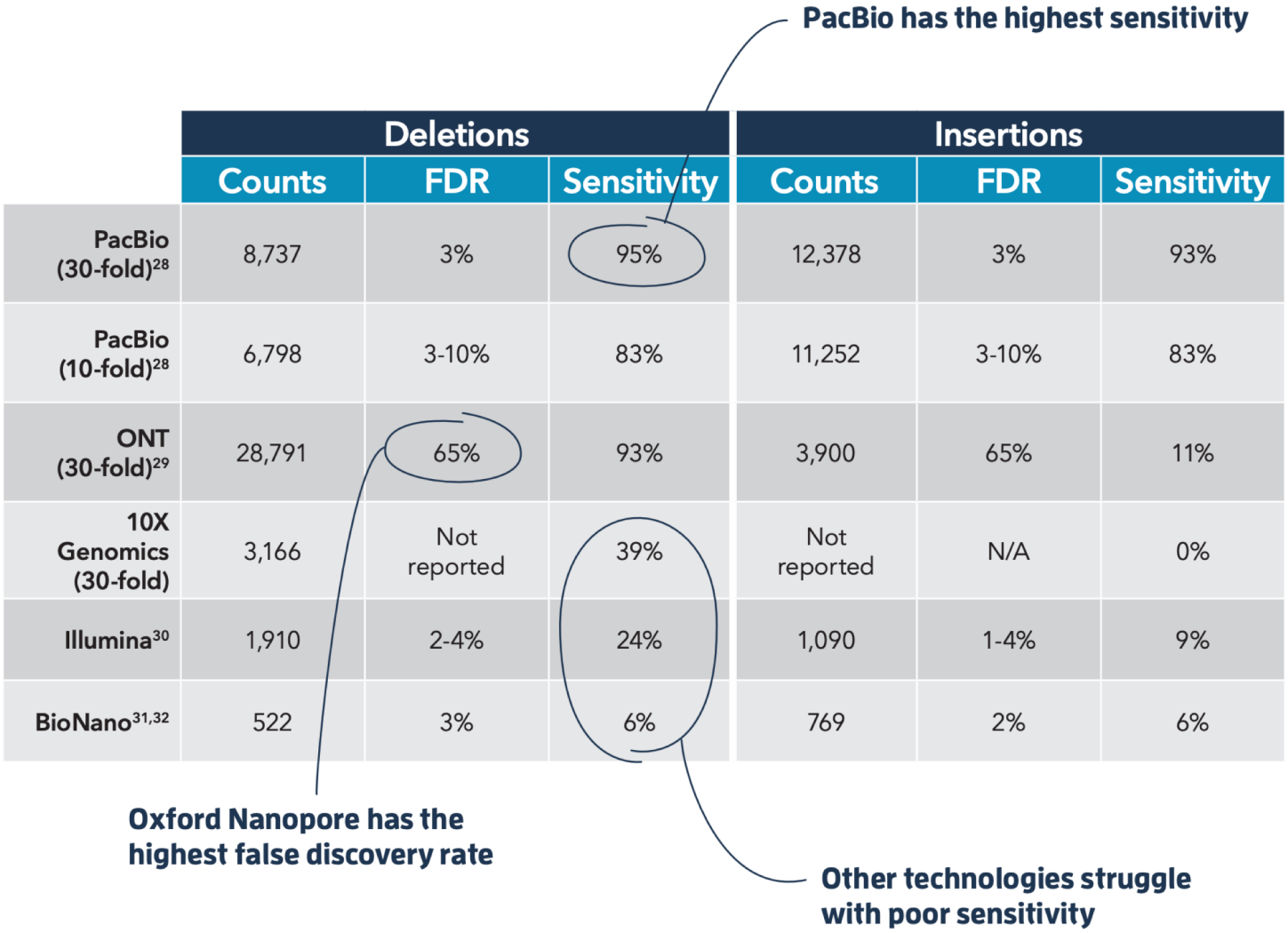
Using long read sequencing technology to detect structural variants

Structural Variant	Disease Examples	PacBio Advantage
INSERTION 	Charcot-Marie Tooth disease, Tay-Sachs disease	- Base pair resolution of breakpoints - Complete inserted sequence
DELETION 	Williams syndrome, Duchenne muscular dystrophy, Smith-Magenis syndrome, Carney Complex	- Base pair resolution of break points - High sensitivity even in repeats
INTERSPERSED DUPLICATION 	APP in Alzheimer's disease, Potocki-Lupski syndrome, Prader-Willi syndrome, Angelman syndrome	- Precise copy number - Base pair resolution of the duplicated sequence - Genomic context of additional copies
TRANSLOCATION 	Down syndrome, XX male syndrome (SRY), schizophrenia (chr 11) , Burkitt's Lymphoma	- Detection of balanced events - Complete sequence information - Unambiguous resolution of genomic context
INVERSION 	Hemophilia A, Hunter Syndrome, Emery-Dreifuss muscular dystrophy	- Detection of balanced events - Continuous sequence information - Base pair resolution of break points
TANDEM DUPLICATION 	FMR1 in Fragile-X, Huntington's disease, Spinocerebellar ataxia	- Complete repeat sequence, including interruptions - Quantitation of repeat expansions

Table 1. Structural variants of all types are known to cause Mendelian disease and contribute to complex disease. *All of these variants can be most robustly detected by PacBio SMRT Sequencing.*

https://www.pacb.com/wp-content/uploads/Whitepaper-Human-Structural-Variation.pdf

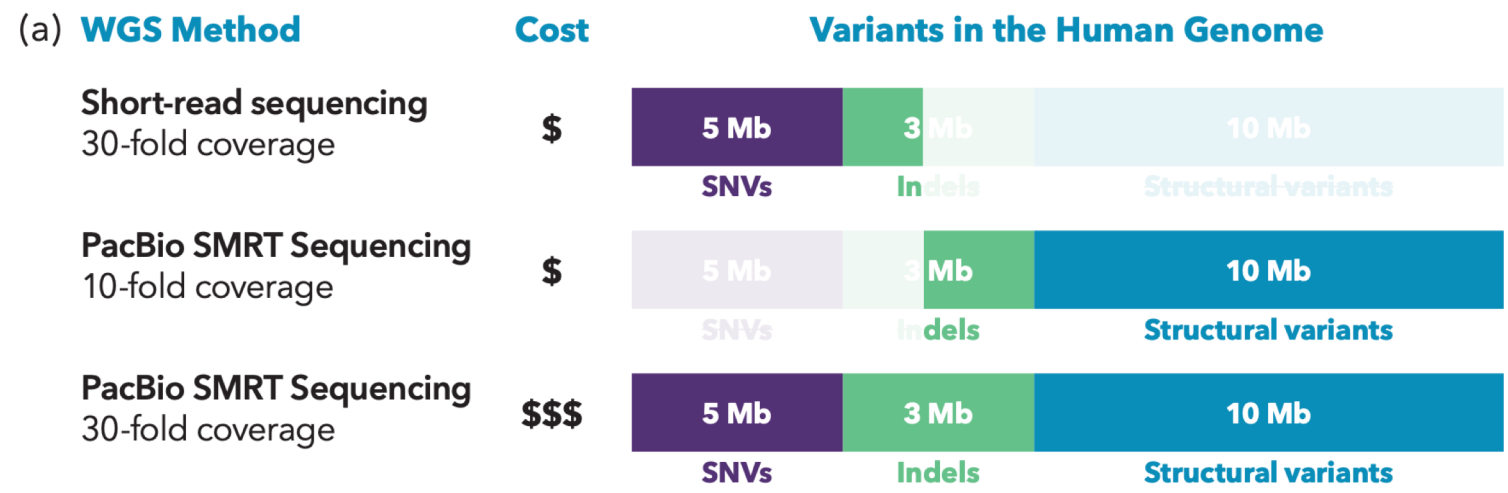
Sensitivity and Reliability of Structural Variant Calling Platforms



	Deletions			Insertions		
	Counts	FDR	Sensitivity	Counts	FDR	Sensitivity
PacBio (30-fold)[28]	8,737	3%	95%	12,378	3%	93%
PacBio (10-fold)[28]	6,798	3-10%	83%	11,252	3-10%	83%
ONT (30-fold)[29]	28,791	65%	93%	3,900	65%	11%
10X Genomics (30-fold)	3,166	Not reported	39%	Not reported	N/A	0%
Illumina[30]	1,910	2-4%	24%	1,090	1-4%	9%
BioNano[31,32]	522	3%	6%	769	2%	6%

Table 2. *PacBio has the highest sensitivity and specificity for structural variants greater than 50 bp, even at low coverage. Other technologies struggle with poor sensitivity and/or high false positive rates. PacBio: NGM-LR (GRCh37/hg19) & PBHoney; Illumina: 1000genomes.org; 10X: http://www.slideshare.net/GenomeInABottle/sept2016-sv-10x; ONT: https://github.com/nanopore-wgs-consortium/NA12878, 12/5/2016 release*

https://www.pacb.com/wp-content/uploads/Whitepaper-Human-Structural-Variation.pdf



(b)

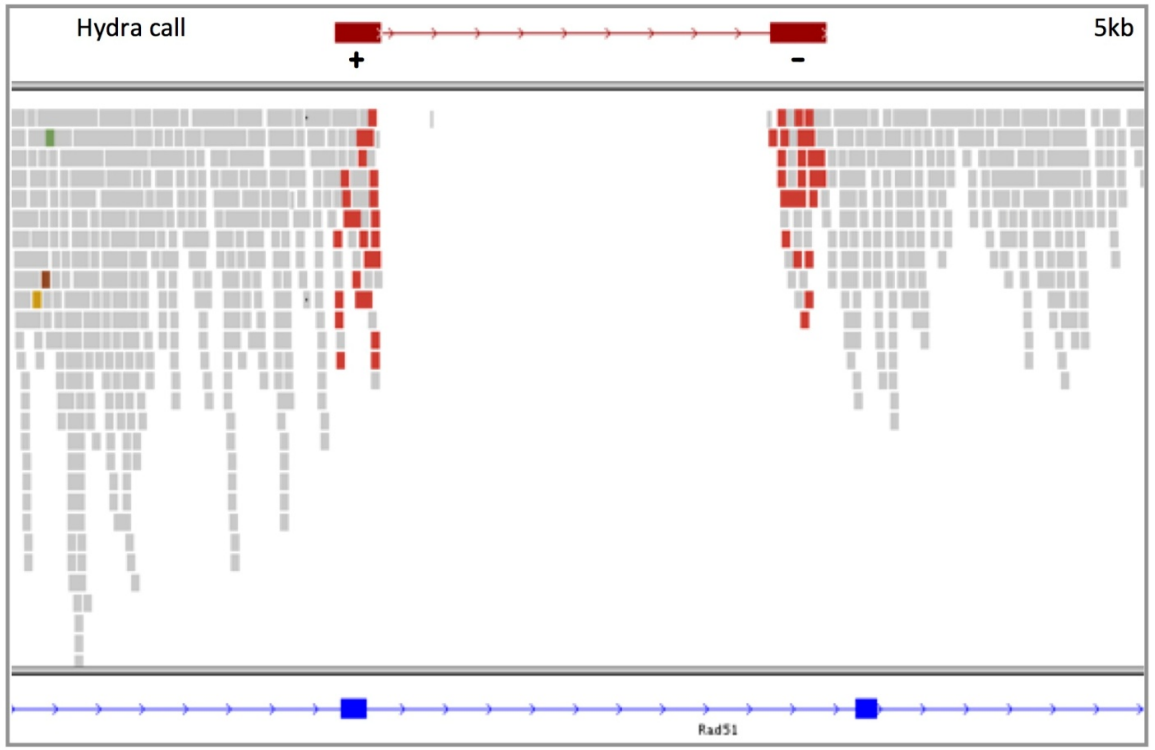
Application	Study Design
Disease Research • Find causative mutations in undiagnosed cases • Find novel disease genes in disease cohorts	≥10 samples at 10-fold PacBio coverage
Population Genetics • Identify polymorphic structural variants present at ≥1% frequency in the population	≥250 samples at 5-fold PacBio coverage
Clinical Genetics • Find causative mutations in patients with genetic disease	1 sample at 30-fold PacBio coverage

Figure 5. *Low-coverage SMRT Sequencing for structural variant discovery. (a) Low-coverage PacBio SMRT Sequencing is a cost-effective complement to short-read sequencing to obtain the most comprehensive view of human genetic variation. (b) The optimal balance of cohort size and sequencing coverage depends on the specific goals of a study. For many studies, low-coverage sequencing is the most effective approach.*

https://www.pacb.com/wp-content/uploads/Whitepaper-Human-Structural-Variation.pdf

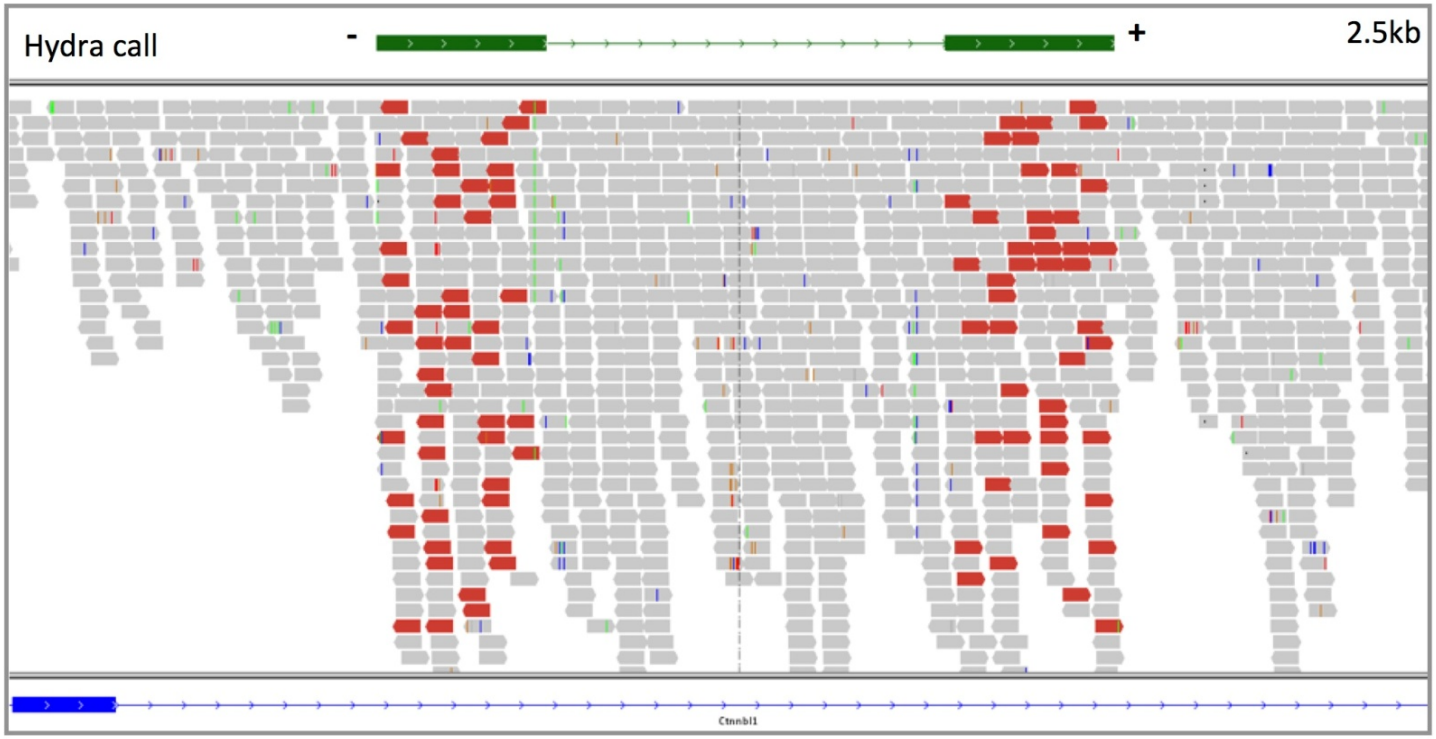
Visualization

Visual validation: a deletion

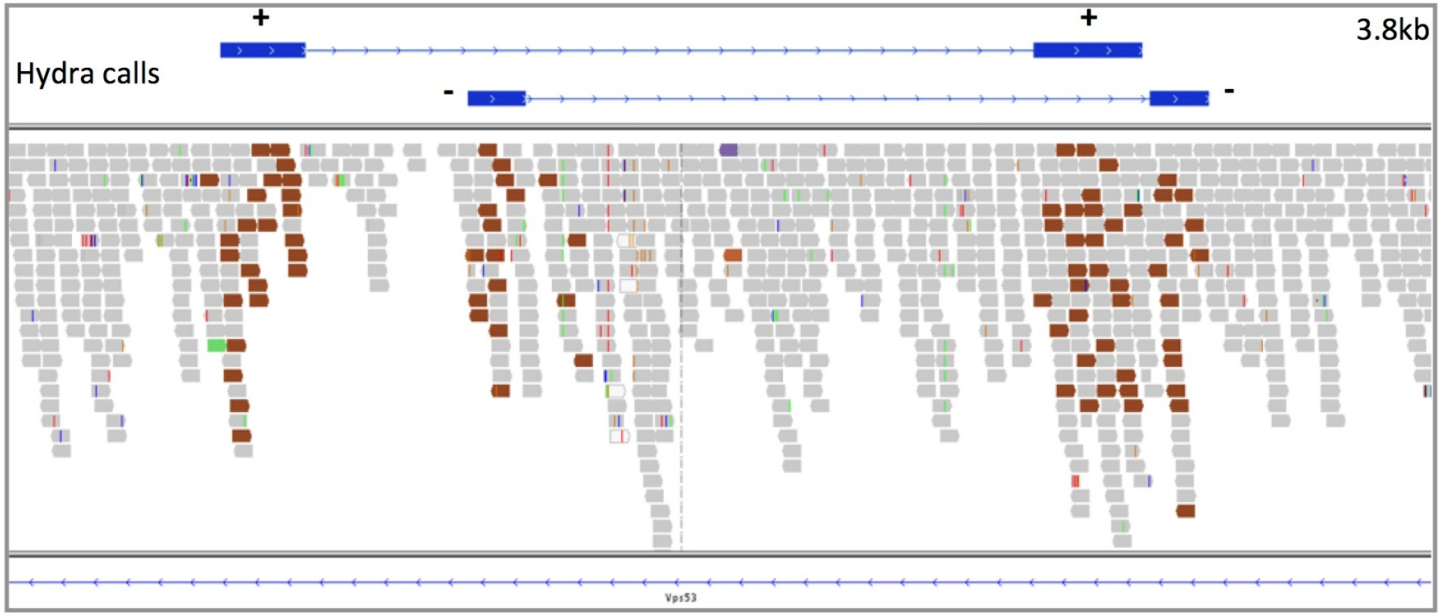


From Aaron Quinlan

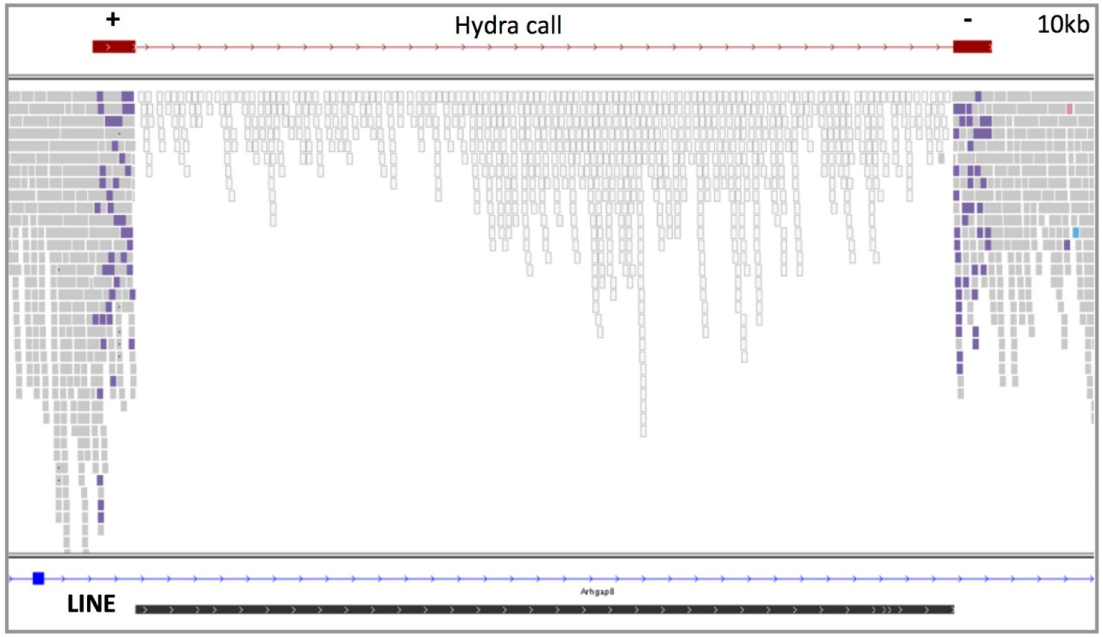
Visual validation: a tandem duplication



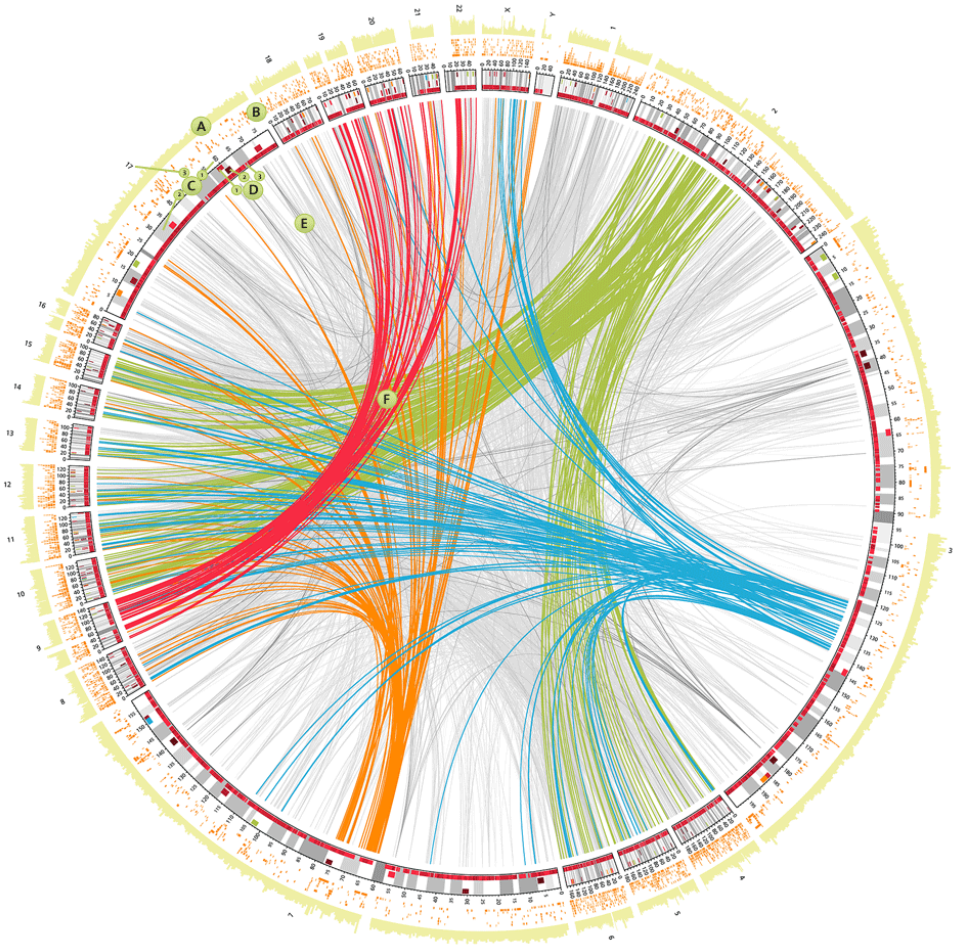
Visual validation: an inversion



Visual validation: an insertion



SVs summary view : Circos plots



circos.ca

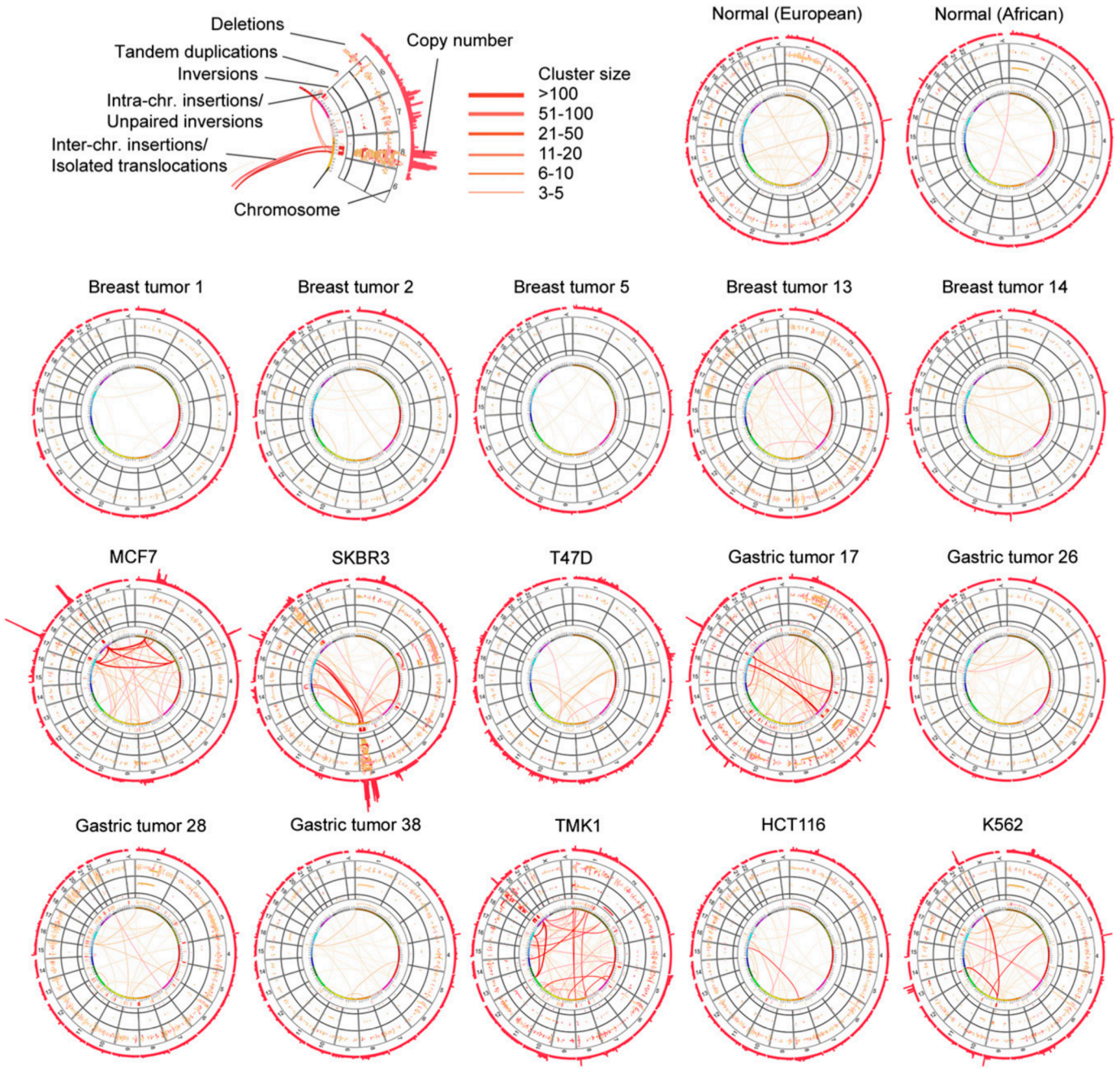


Figure 2. Karyo-genomic maps of 15 cancer and two normal human genomes. Genomes are arranged in a circular manner with SV categories arranged in concentric layers as indicated on the *top, left*. Circular plots have been generated using Circos (Krzywinski et al. 2009).

Hillmer et al. Genome Research 2011

Lab

CNVKit

Build a reference from normal samples and infer tumor copy ratios

