Lecture 8. Structural Variation Analysis

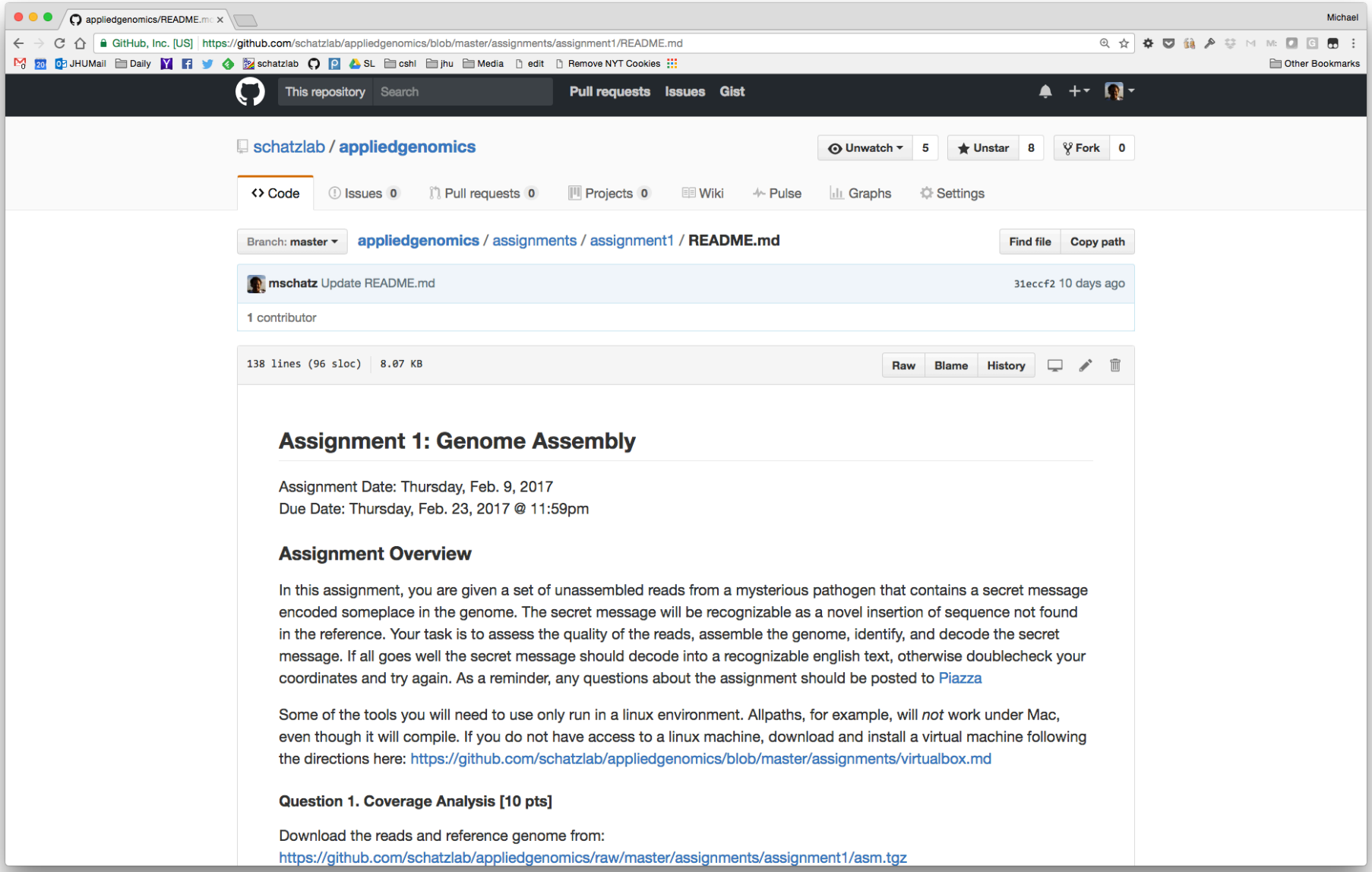
Michael Schatz

Feb 23, 2017

JHU 600.649: Applied Comparative Genomics

Assignment 1: Due Thursday @ 11:59pm

Email PDF to: jhuappliedgenomics@gmail.com



appliedgenomics/README.md

GitHub, Inc. [US] https://github.com/schatzlab/appliedgenomics/blob/master/assignments/assignment1/README.md

This repository Search Pull requests Issues Gist

schatzlab / appliedgenomics

Unwatch 5 Unstar 8 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Pulse Graphs Settings

Branch: master appliedgenomics / assignments / assignment1 / README.md Find file Copy path

mschatz Update README.md 31eccf2 10 days ago

1 contributor

138 lines (96 sloc) 8.07 KB Raw Blame History

Assignment 1: Genome Assembly

Assignment Date: Thursday, Feb. 9, 2017
Due Date: Thursday, Feb. 23, 2017 @ 11:59pm

Assignment Overview

In this assignment, you are given a set of unassembled reads from a mysterious pathogen that contains a secret message encoded someplace in the genome. The secret message will be recognizable as a novel insertion of sequence not found in the reference. Your task is to assess the quality of the reads, assemble the genome, identify, and decode the secret message. If all goes well the secret message should decode into a recognizable english text, otherwise doublecheck your coordinates and try again. As a reminder, any questions about the assignment should be posted to Piazza

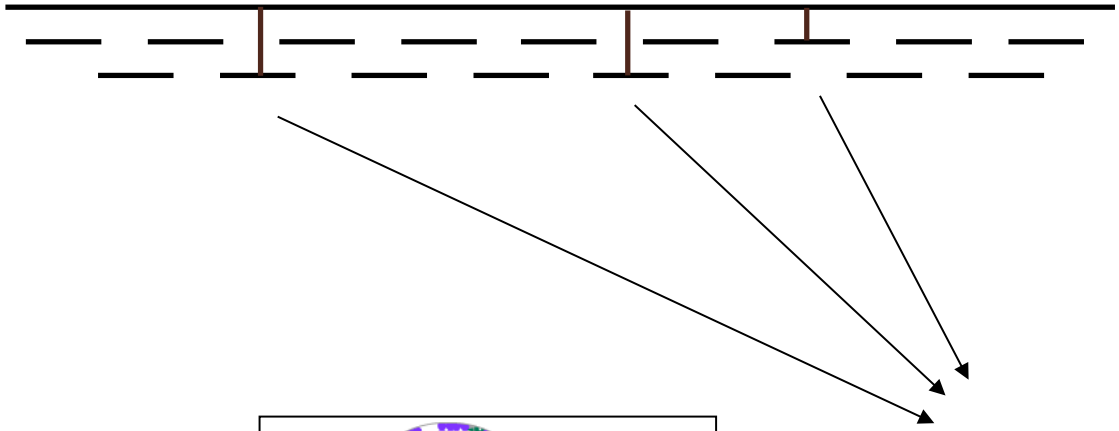
Some of the tools you will need to use only run in a linux environment. Allpaths, for example, will *not* work under Mac, even though it will compile. If you do not have access to a linux machine, download and install a virtual machine following the directions here: https://github.com/schatzlab/appliedgenomics/blob/master/assignments/virtualbox.md

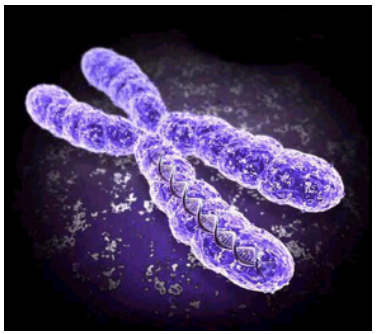
Question 1. Coverage Analysis [10 pts]

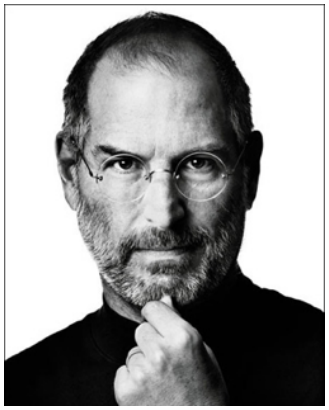
Download the reads and reference genome from:
https://github.com/schatzlab/appliedgenomics/raw/master/assignments/assignment1/asm.tgz

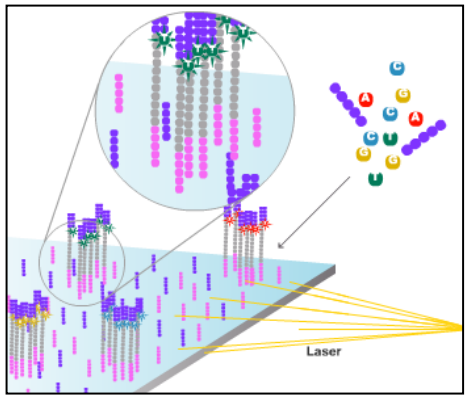
Personal Genomics

How does your genome compare to the reference?







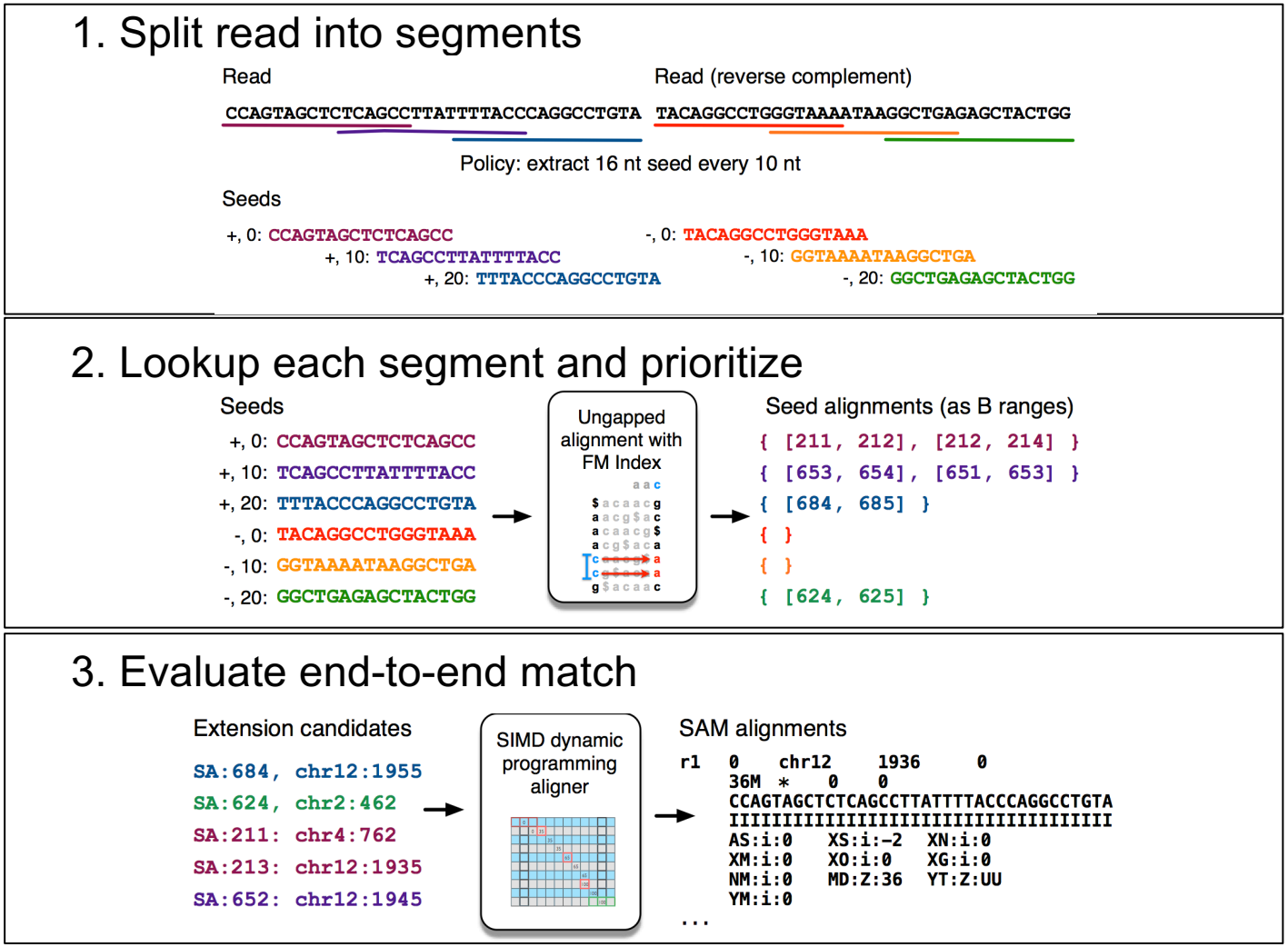


Heart Disease

Cancer

Creates magical technology

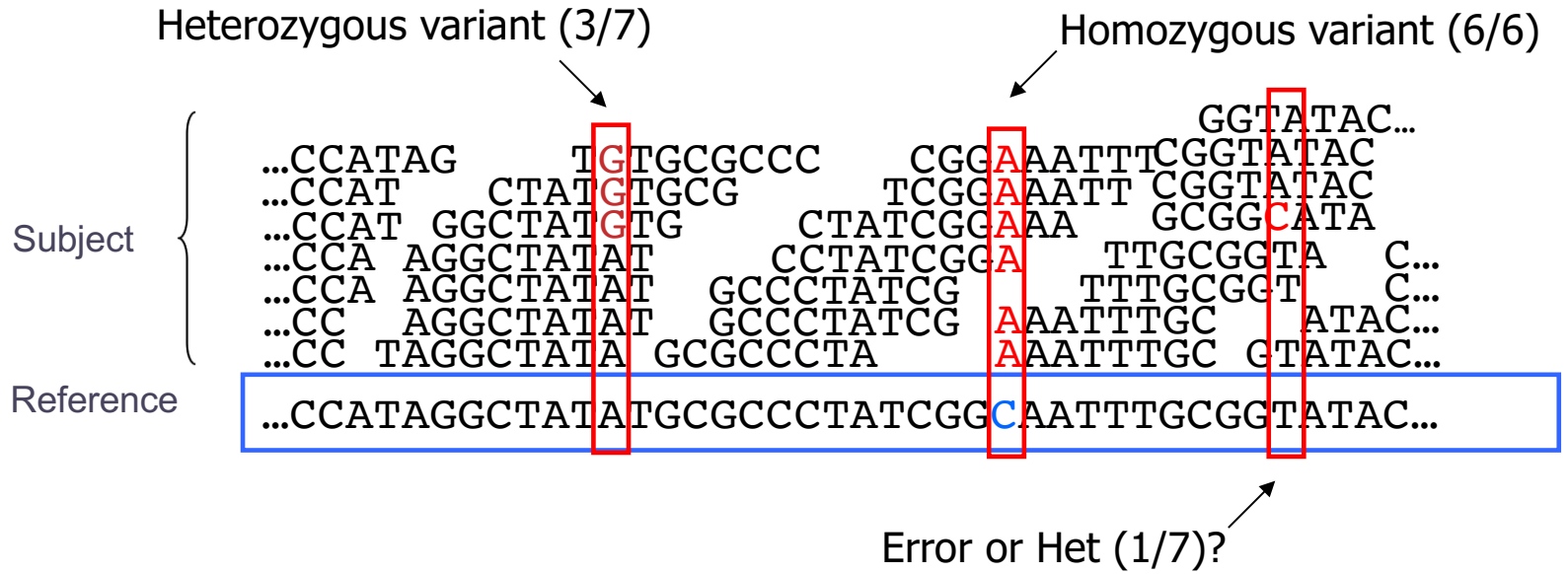
Read Mapping Overview



Fast gapped-read alignment with Bowtie 2

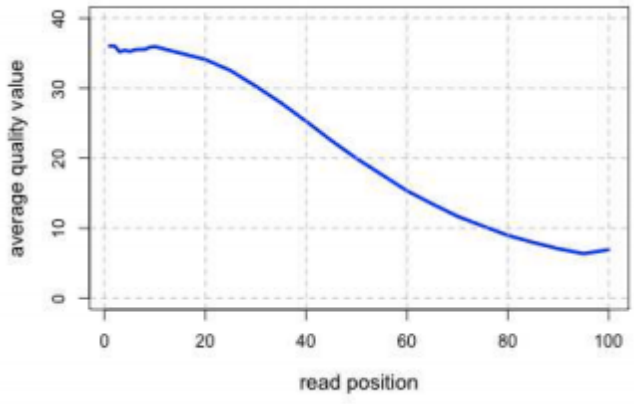
Langmead & Salzberg (2012) Nature Methods. doi:10.1038/nmeth.1923

Genotyping Theory



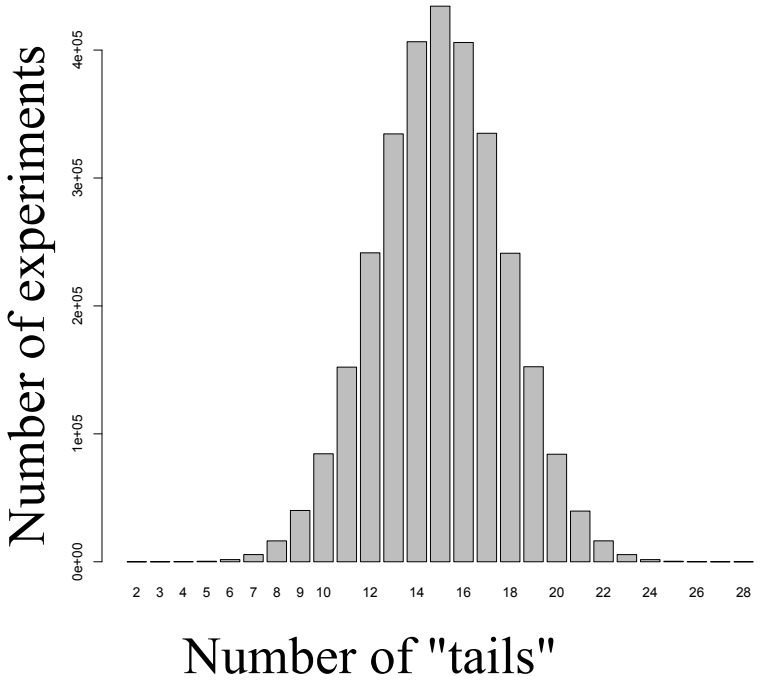
- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!

- Sequencing instruments make mistakes
  - Quality of read decreases over the read length



- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome



This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

P(3/30 het) <?> P(3/30 err)

So, with 30 tosses (reads), we are much more likely to see an even mix o at a e

Number of experiments

a "30X" verage) nded: it wer to ygous mere ors

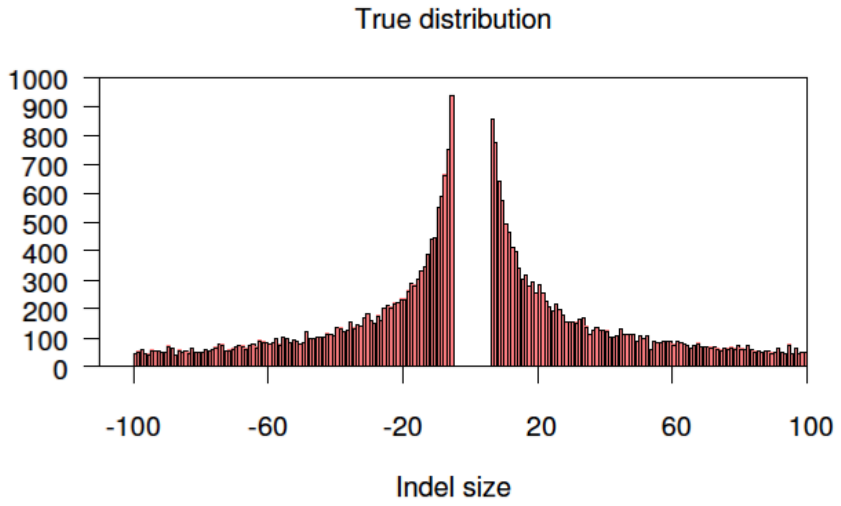
/30 err)

Variation Detection Complexity

SNPs + Short Indels

High precision and sensitivity

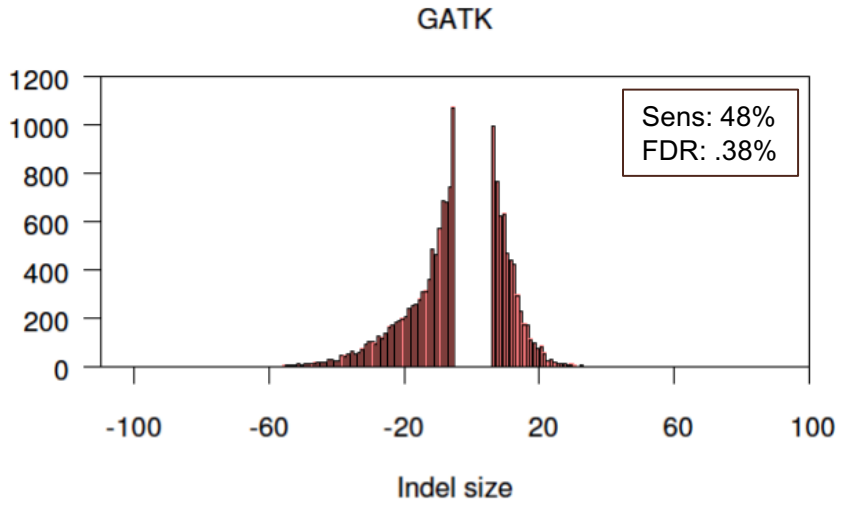
```
..TTTAGAATAG-CGAGTGC...
      ||||||| ||||
     AGAATAGGCGAG
```



"Long" Indels (>5bp)

Reduced precision and sensitivity

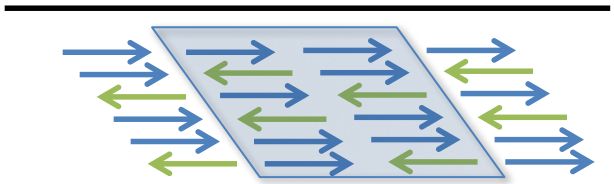
```
..TTTAG--------AGTGC...
   |||||        |||||
  TTTAGAATAGGC |||||
        ATAGGCGAGTGC
```



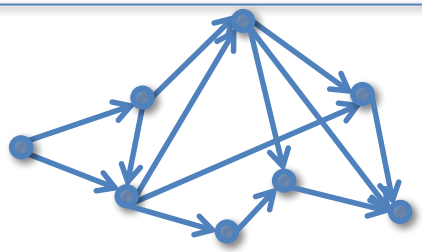
Analysis confounded by sequencing errors, localized repeats, allele biases, and mismapped reads

Scalpel Algorithm

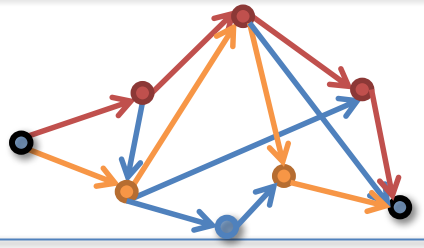
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



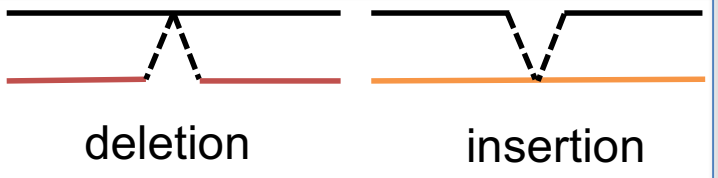
Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads



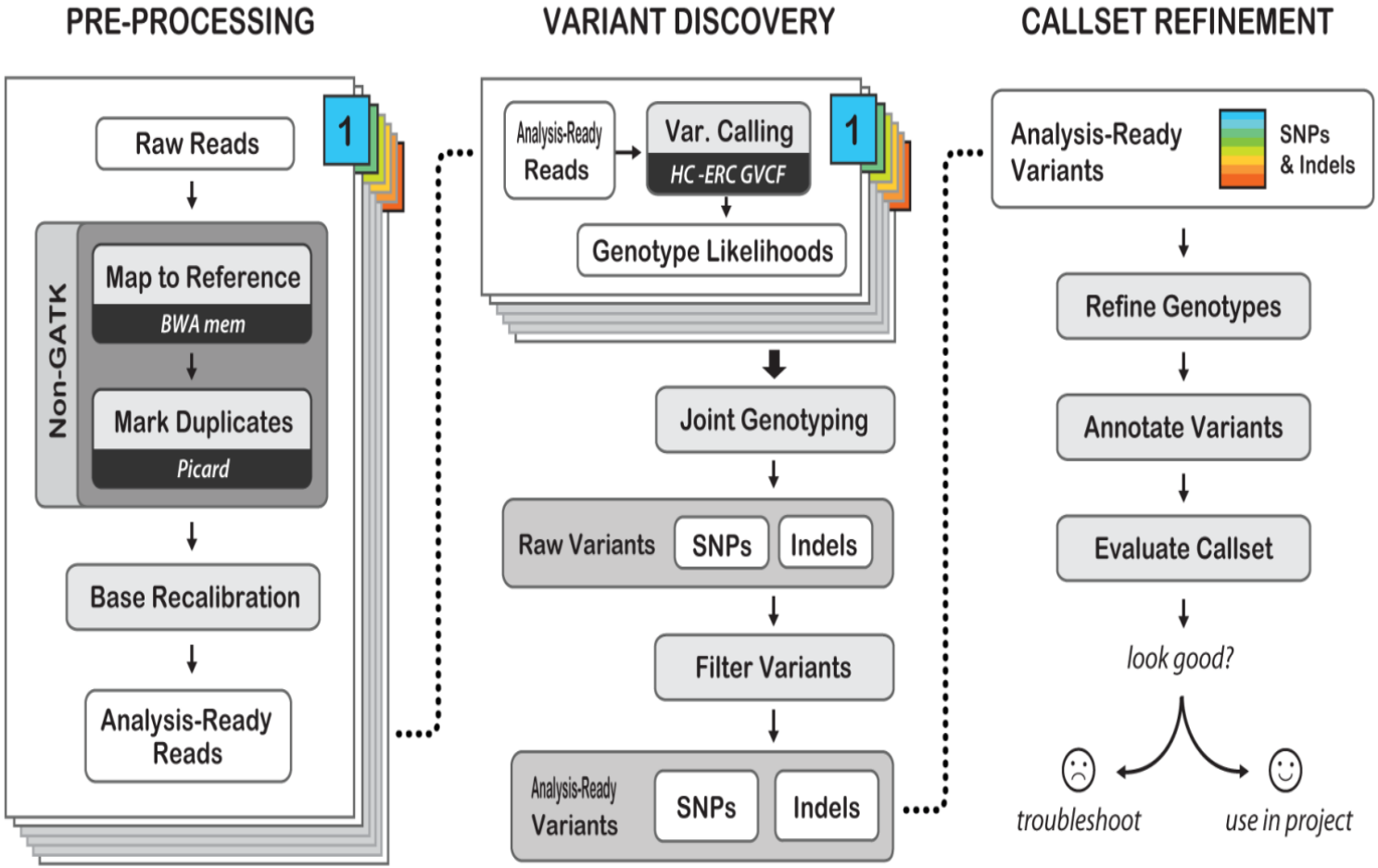
Find end-to-end haplotype paths spanning the region



Align assembled sequences to reference to detect mutations



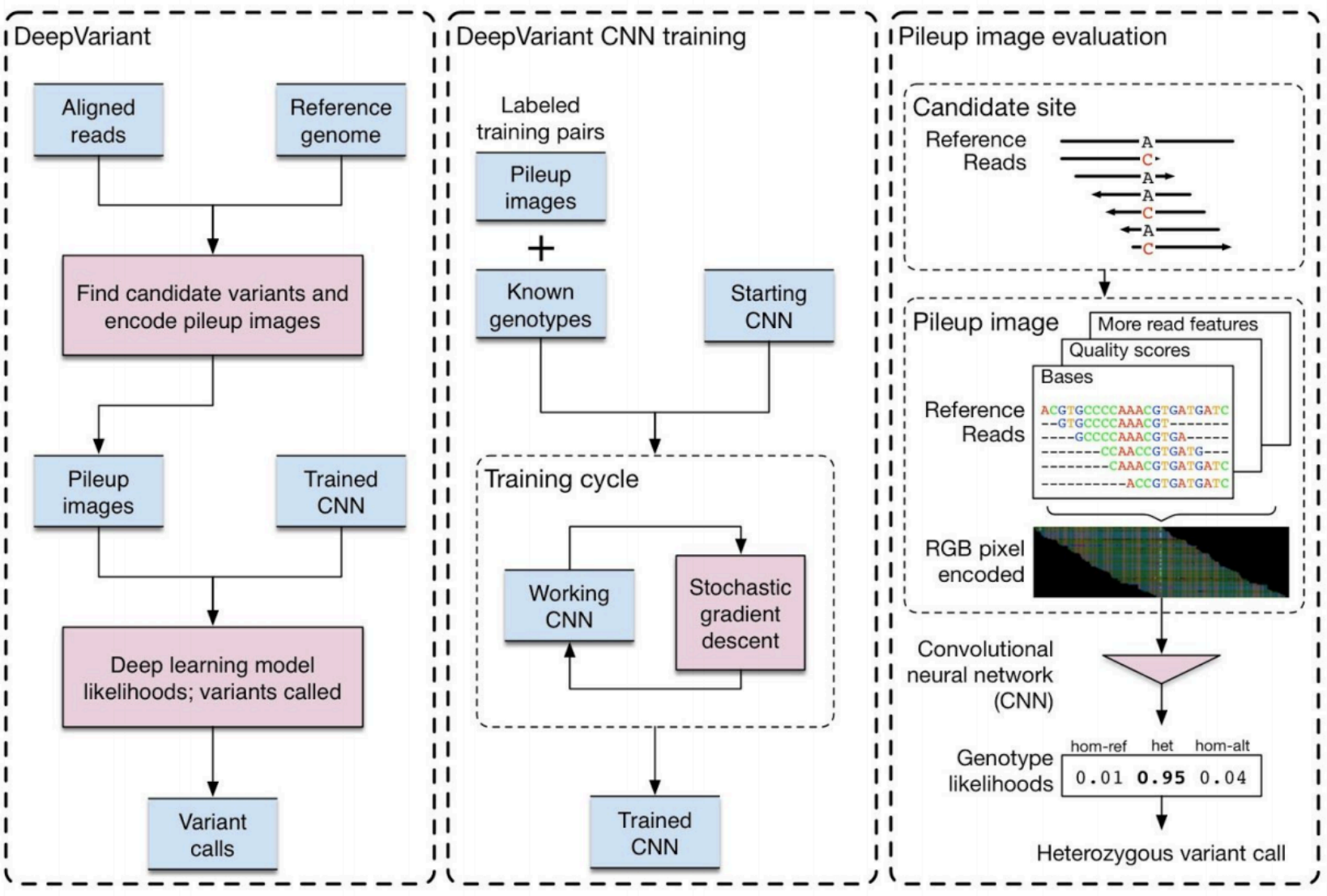
GATK workflow



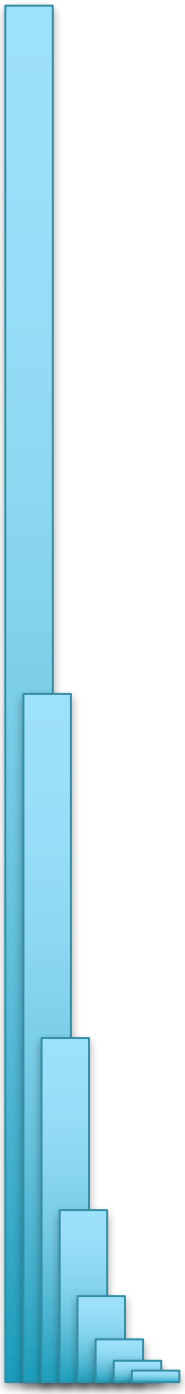
Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

https://software.broadinstitute.org/gatk/img/BP_workflow_3.6.png

Deep Variant

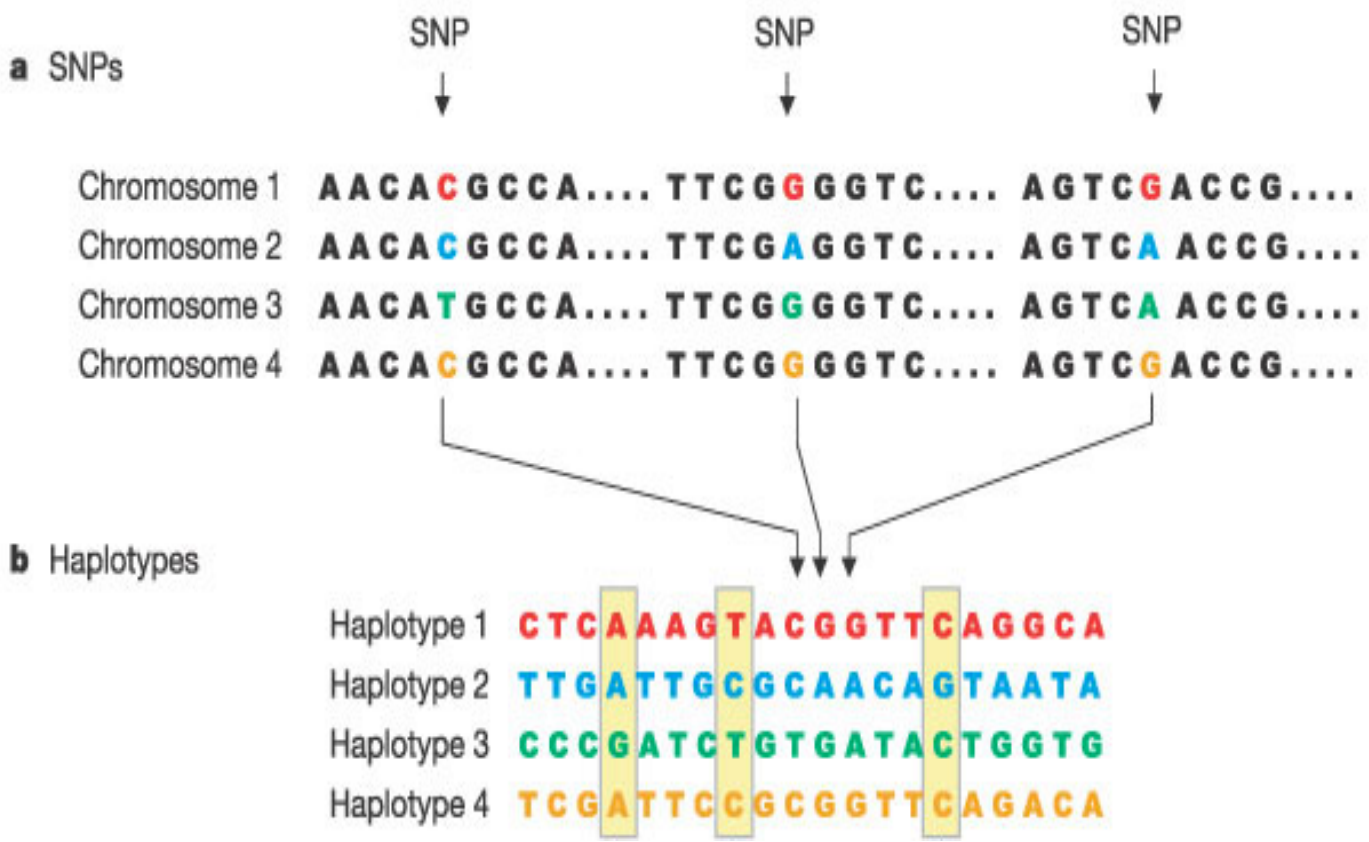


Creating a universal SNP and small indel variant caller with deep neural networks
Poplin et al. (2016) bioRxiv. doi: https://doi.org/10.1101/092890



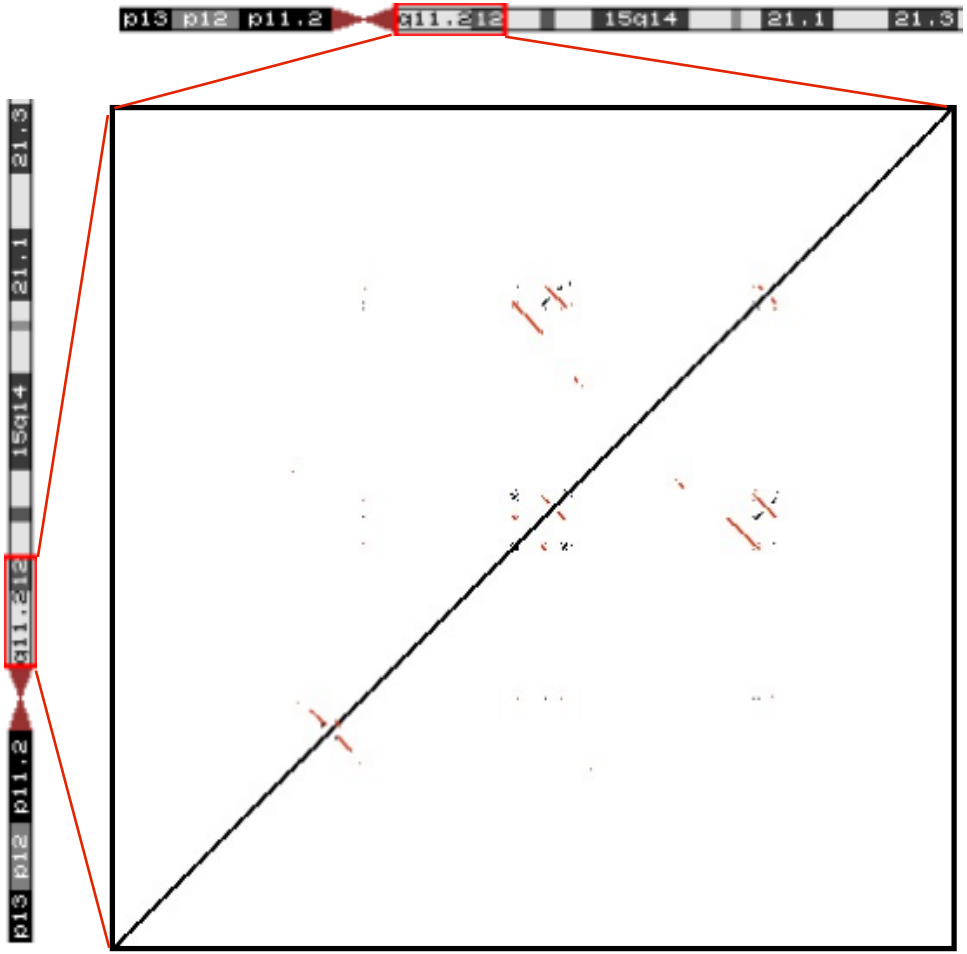
Part II: Structural Variations

Early 2000s dogma: SNPs account for most human genetic variation



https://hapmap.ncbi.nlm.nih.gov

Segmental duplications (a.k.a. Low copy repeats)

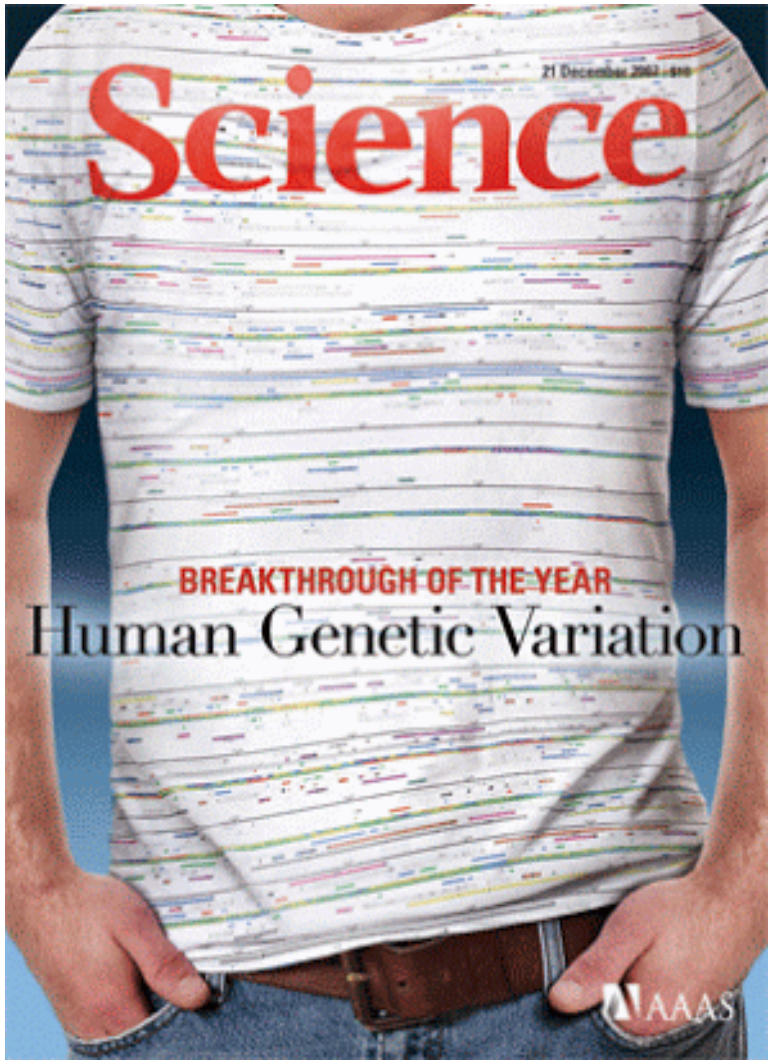


Self Dotplot:
10 megabases of Chr 15
(dot = 1 kb exact match)

~5% of the human genome is duplicated!

Bailey et al, 2002

2007 Science magazine breakthrough of the year? Why?



Discovery of abundant copy-number variation

Science, **July 2004**

Large-Scale Copy Number Polymorphism in the Human Genome

Jonathan Sebat,[1] B. Lakshmi,[1] Jennifer Troge,[1] Joan Alexander,[1] Janet Young,[2] Pär Lundin,[3] Susanne Månér,[3] Hillary Massa,[2] Megan Walker,[2] Maoyen Chi,[1] Nicholas Navin,[1] Robert Lucito,[1] John Healy,[1] James Hicks,[1] Kenny Ye,[4] Andrew Reiner,[1] T. Conrad Gilliam,[5] Barbara Trask,[2] Nick Patterson,[6] Anders Zetterberg,[3] Michael Wigler[1]*

76 CNVs in 20 individuals
70 genes

Nature Genetics, **Aug. 2004**

Detection of large-scale variation in the human genome

A John Iafrate[1,2], Lars Feuk[3], Miguel N Rivera[1,2], Marc L Listewnik[1], Patricia K Donahoe[2,4], Ying Qi[3], Stephen W Scherer[3,5] & Charles Lee[1,2,5]

255 CNVs in 55 individuals
127 genes

- 331 CNVs, only 11 in common
- Half observed in only 1 individual
- Impact "plenty" of genes
- Correlated with segmental duplications in the reference genome

Why is structural variation relevant / important?

- They are common and affect a large fraction of the genome
  - In total, SVs impact more base pairs than all single-nucleotide differences.

- They are a major driver of genome evolution
  - Speciation can be driven by rapid changes in genome architecture
  - Genome instability and aneuploidy: hallmarks of solid tumor genomes

Why is structural variation relevant / important?

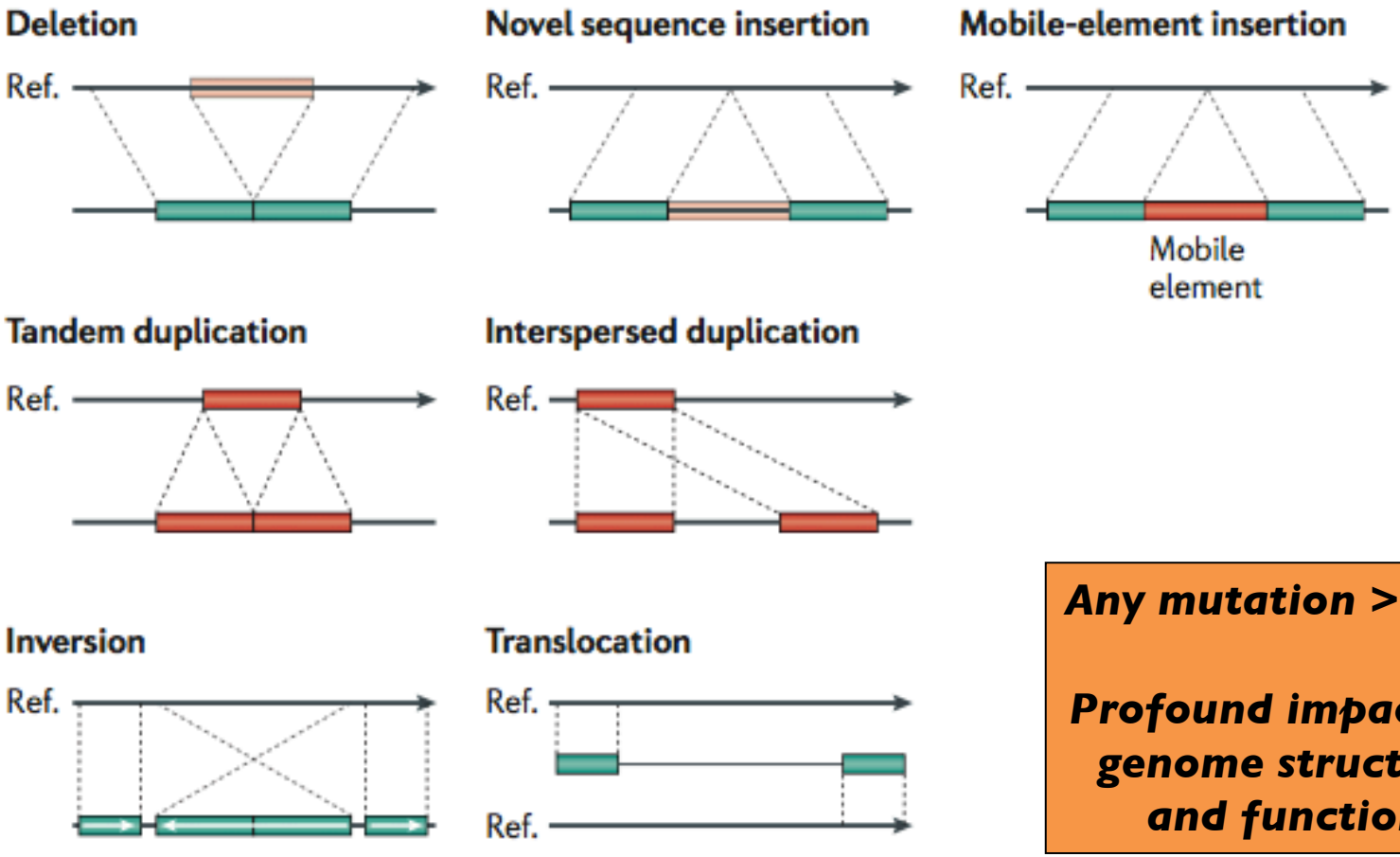
- Genetic basis of traits
  - Gene dosage effects.
  - Neuropsychiatric disease (e.g., autism, schizophrenia)
  - Spontaneous SVs implicated in so-called “genomic” and developmental disorders
  - Somatic genome instability; age-dependent disease

SV and human disease phenotypes

Mendelian (X-linked)			
Hemophilia A	306700	*F8*	inv/del
Hunter syndrome	309900	*IDS*	del/inv
Ichthyosis	308100	*STS*	del
Mental retardation	300706	*HUWE1*	dup
Pelizaeus-Merzbacher disease	312080	*PLP1*	del/dup/tri
Progressive neurological symptoms (MR+SZ)	300260	*MECP2*	dup
Red-green color blindness	303800	opsin genes	del
Complex traits			
Alzheimer disease	104300	*APP*	dup
Autism	612200	3q24	inherited homozygous del
	611913	16p11.2	del/dup
Crohn disease	266600	*HBD-2*	copy number loss
	612278	*IRGM*	del
HIV susceptibility	609423	*CCL3L1*	copy number loss
Mental retardation	612001	15q13.3	del
	610443	17q21.31	del
	300534	Xp11.22	dup
Pancreatitis	167800	*PRSS1*	tri
Parkinson disease	168600	*SNCA*	dup/tri
Psoriasis	177900	*DEFB*	copy number gain
Schizophrenia	612474	1q21.1	del
	181500	15q11.2	del
	612001	15q13.3	del
Systemic lupus erythematosus	152700	*FCGR3B*	copy number loss
	120810	*C4*	copy number loss

Zhang et al, 2009

Structural Variations



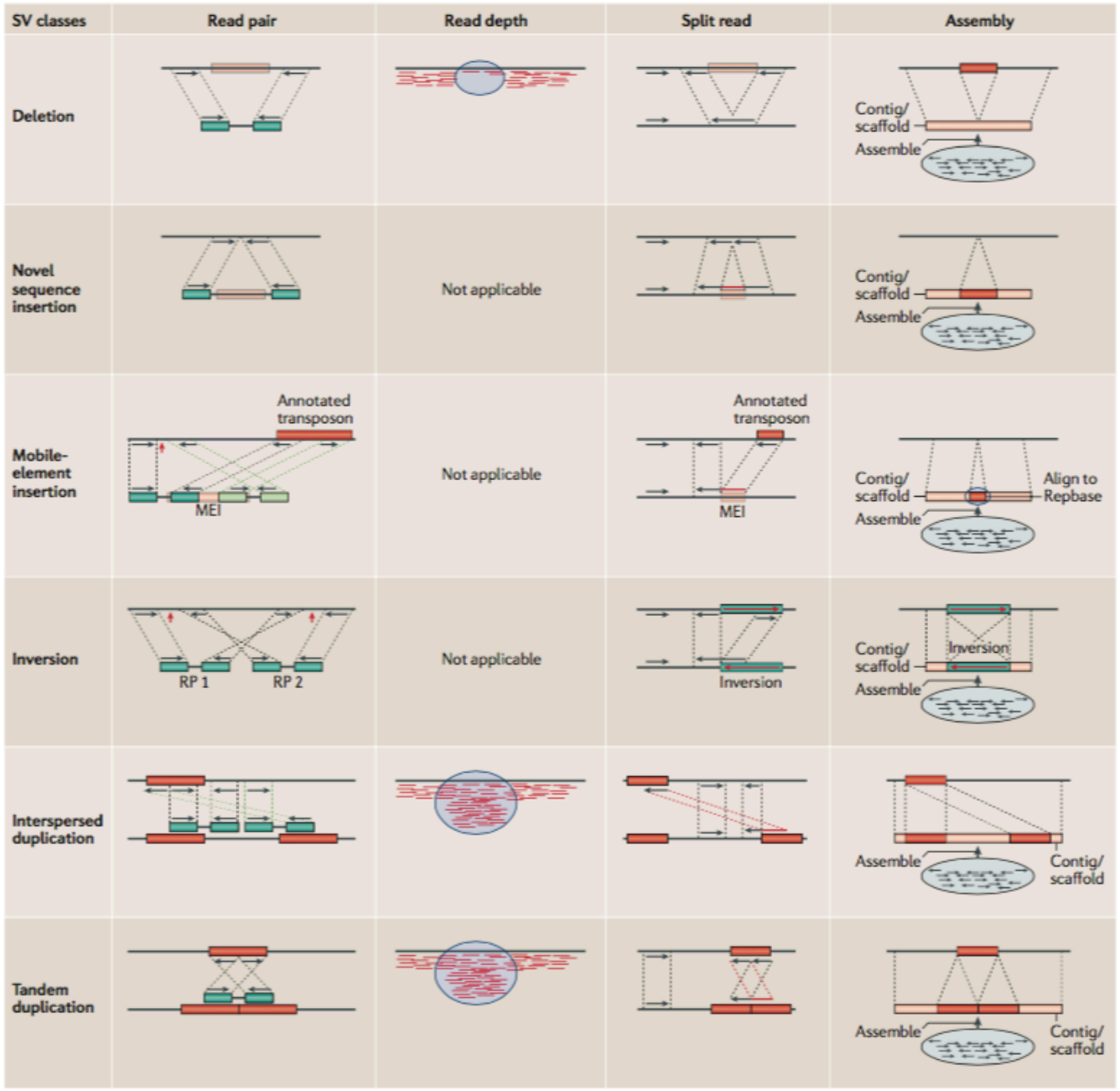
Any mutation >50bp

Profound impact on genome structure and function

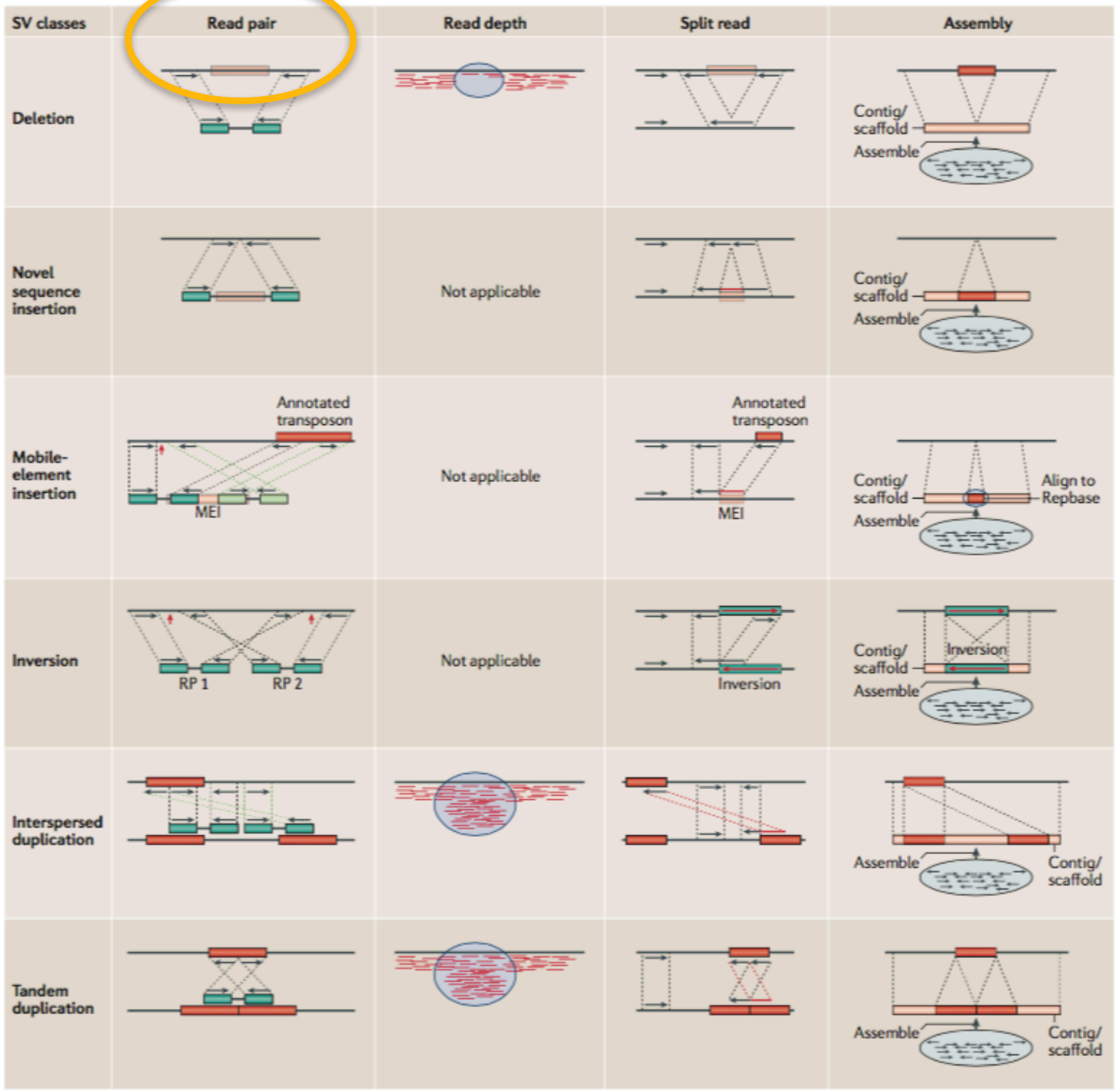
Genome structural variation discovery and genotyping

Alkan, C, Coe, BP, Eichler, EE (2011) *Nature Reviews Genetics*. May;12(5):363-76. doi: 10.1038/nrg2958.

Structural Variation Sequence Signatures



Structural Variation Sequence Signatures



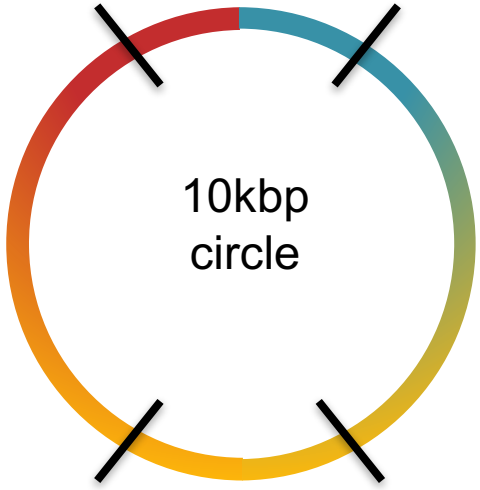
Paired-end and Mate-pairs

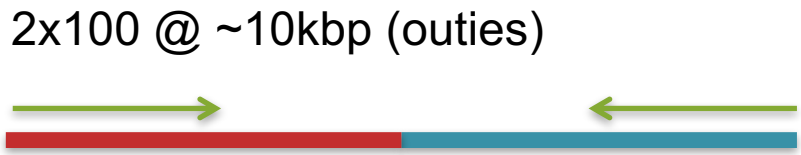
Paired-end sequencing

- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



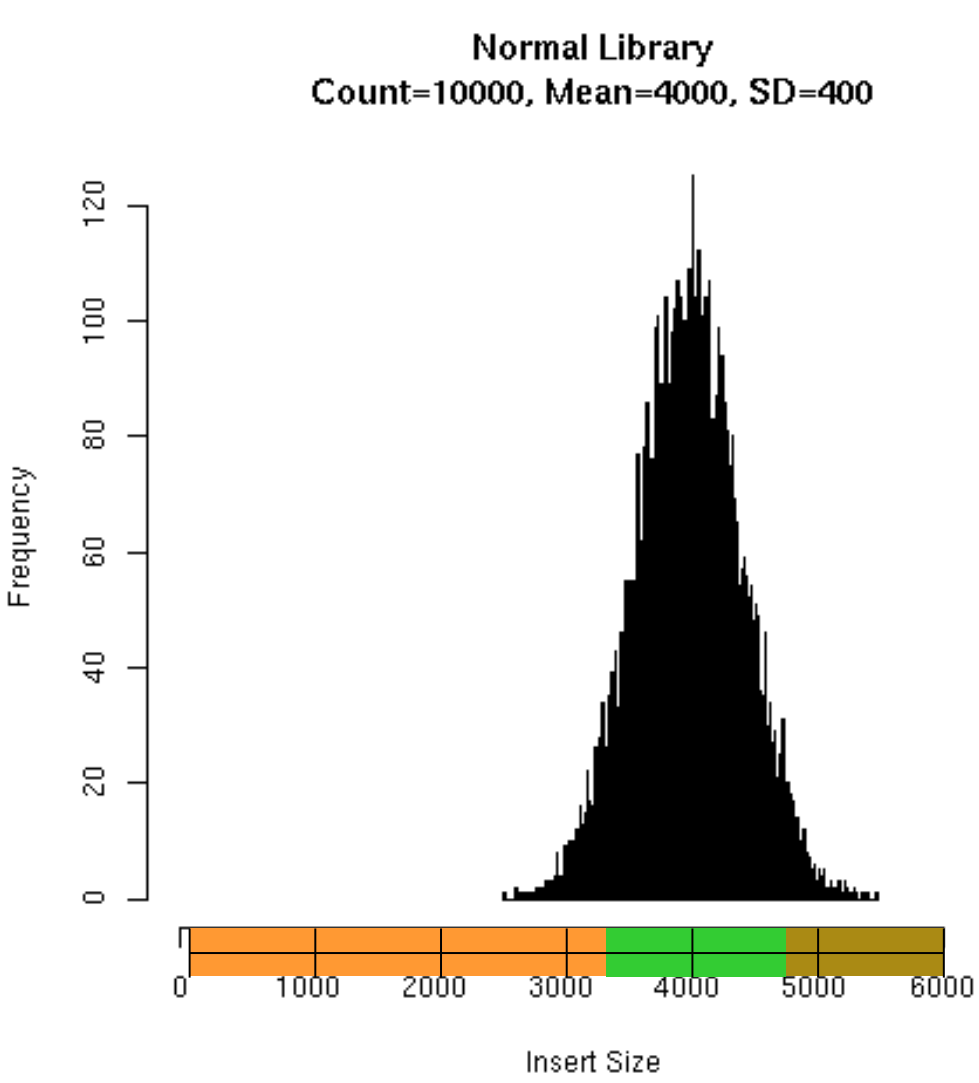


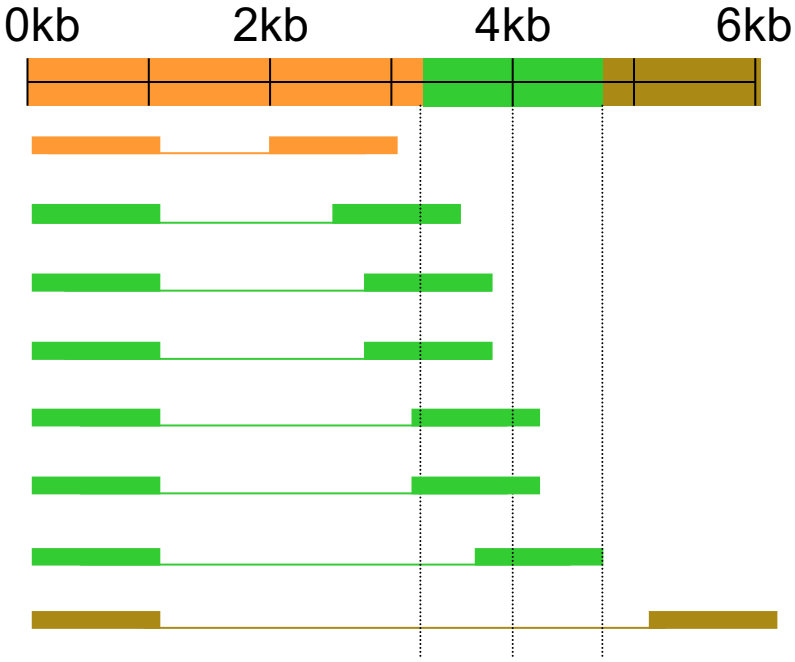
C/E Statistic

- The presence of individual compressed or expanded mates is rare but expected.
- Do the inserts spanning a given position differ from the rest of the library?
  - Flag large differences as potential structural variation / misassemblies
  - Even if each individual mate is "happy"
- Compute the statistic at all positions
  - (Local Mean – Global Mean) / Scaling Factor
- Introduced by Jim Yorke's group at UMD

Assembly reconciliation
Zimin, Smith, Sutton, Yorke (2008) Bioinformatics. doi 10.1093/bioinformatics/btm542

Sampling the Genome





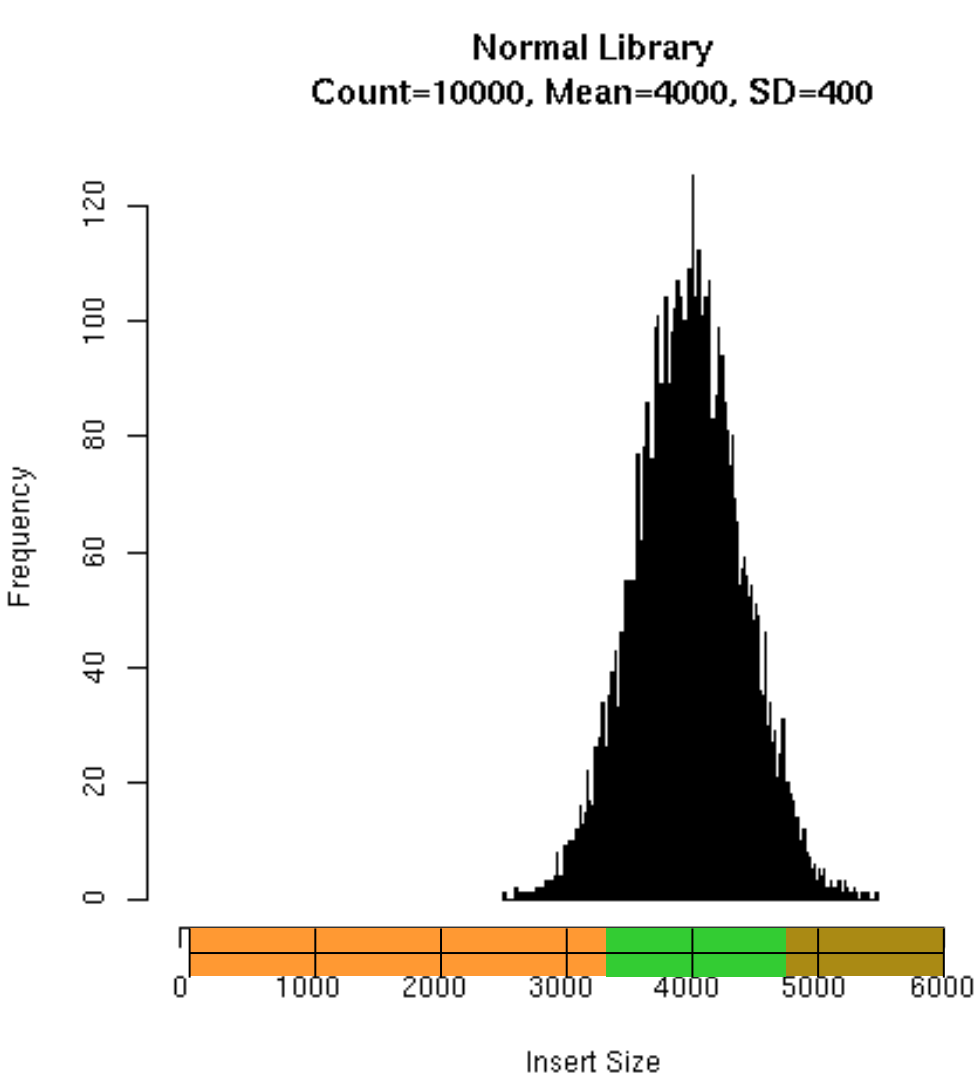
8 inserts: 3kb-6kb

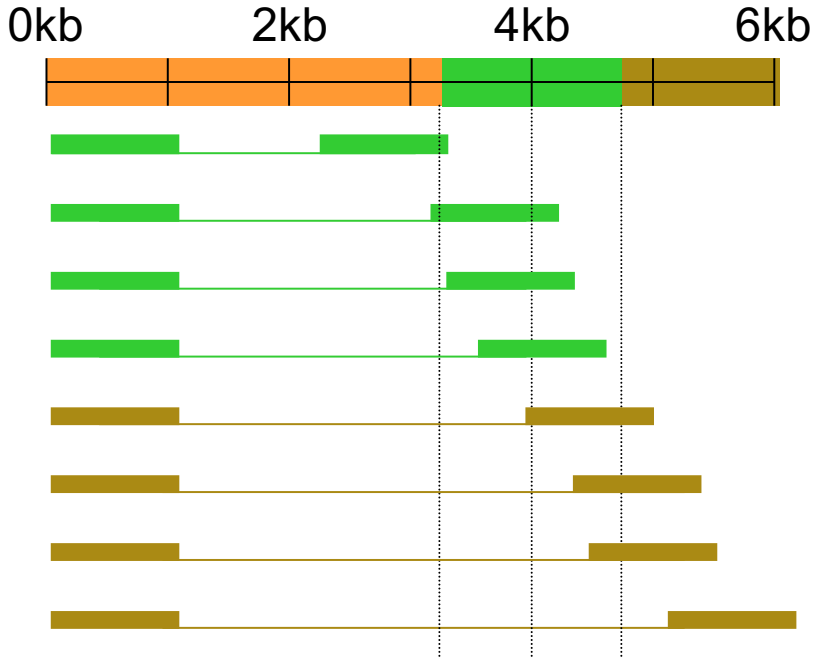
Local Mean: 4048

C/E Stat: $\frac{(4048-4000)}{(400 / \sqrt{8})}$ = +0.33

Near 0 indicates overall happiness

C/E-Statistic: Expansion





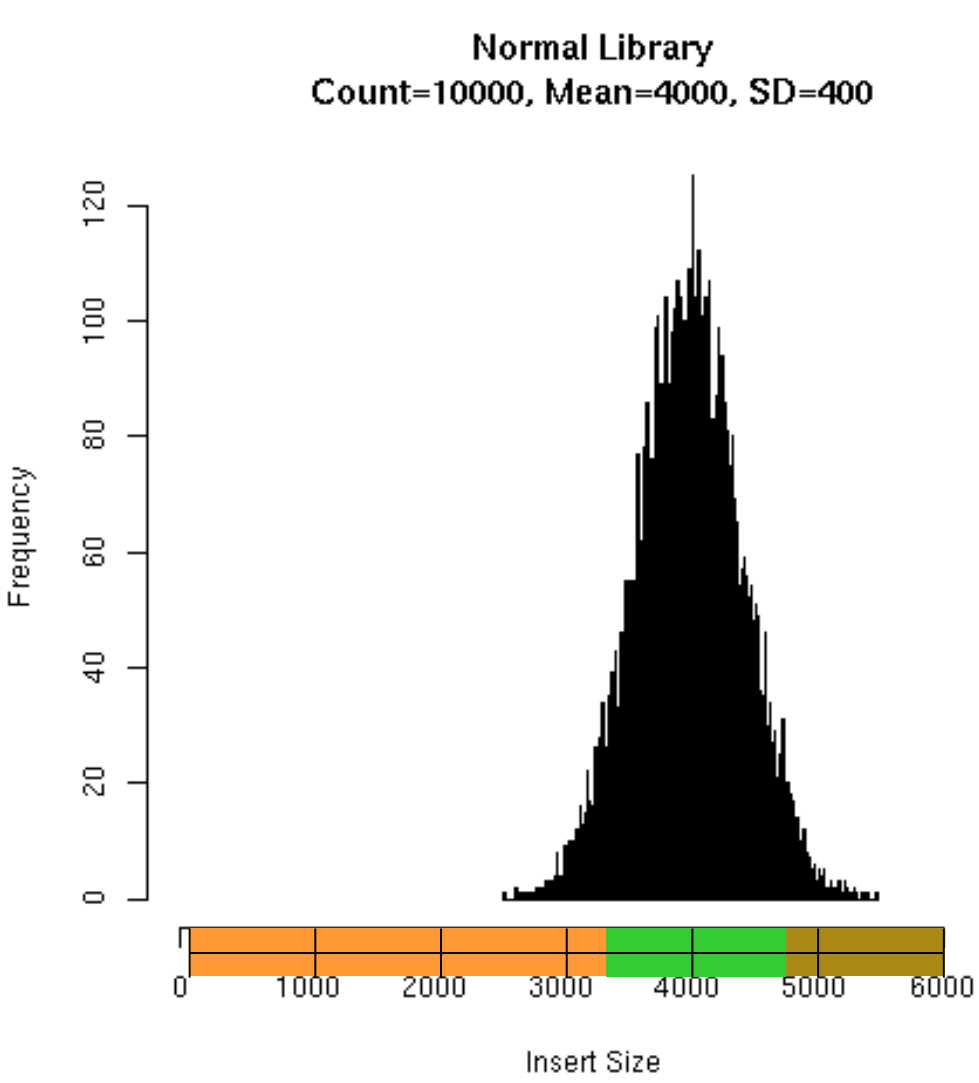
8 inserts: 3.2kb-6kb

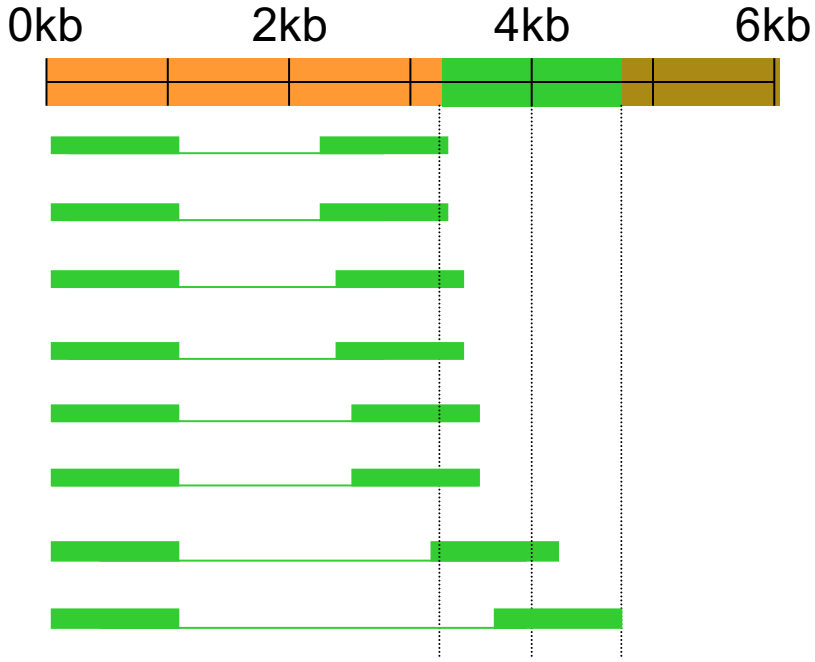
Local Mean: 4461

C/E Stat: $\frac{(4461-4000)}{(400 / \sqrt{8})}$ = +3.26

C/E Stat ≥ 3.0 indicates Expansion

C/E-Statistic: Compression





8 inserts: 3.2 kb-4.8kb

Local Mean: 3488

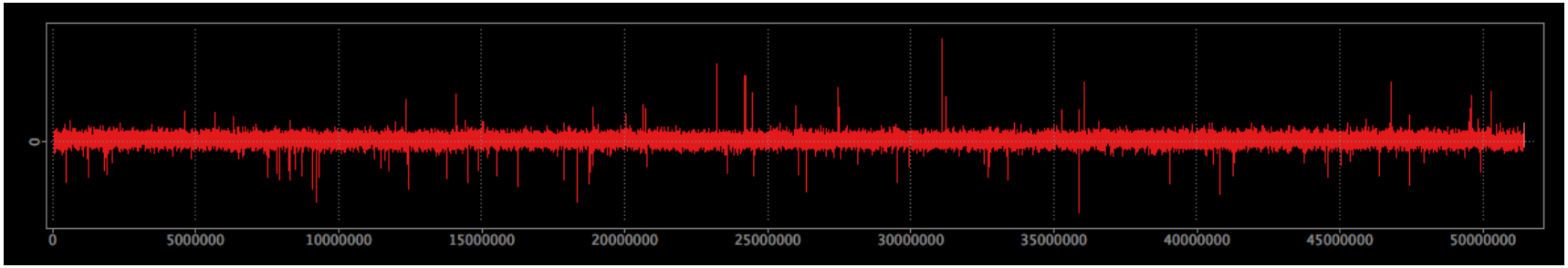
C/E Stat: $\frac{(3488-4000)}{(400 / \sqrt{8})} = -3.62$

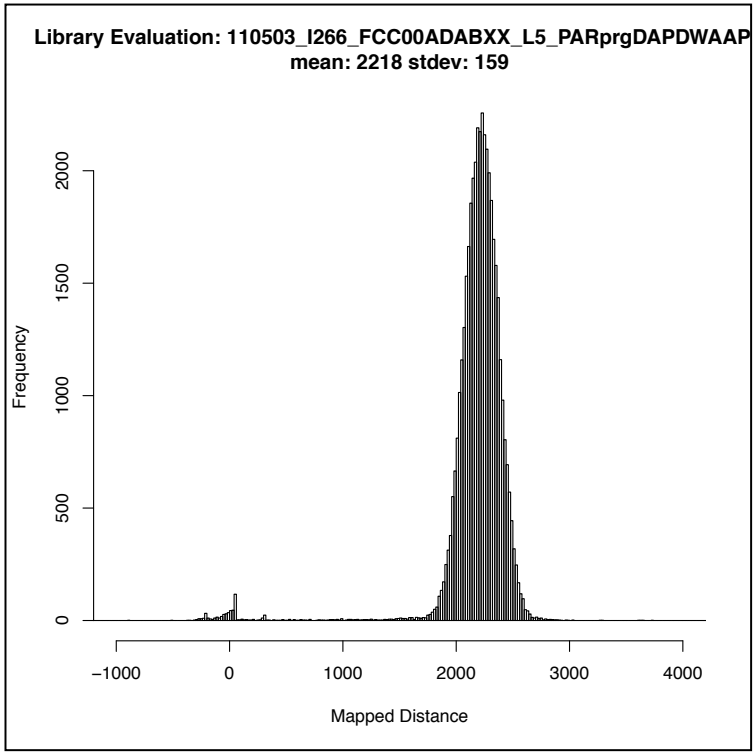
C/E Stat ≤ -3.0 indicates Compression

Parrot Metassembly

CE statistic (projected) across 51.1 Mbp scaffold

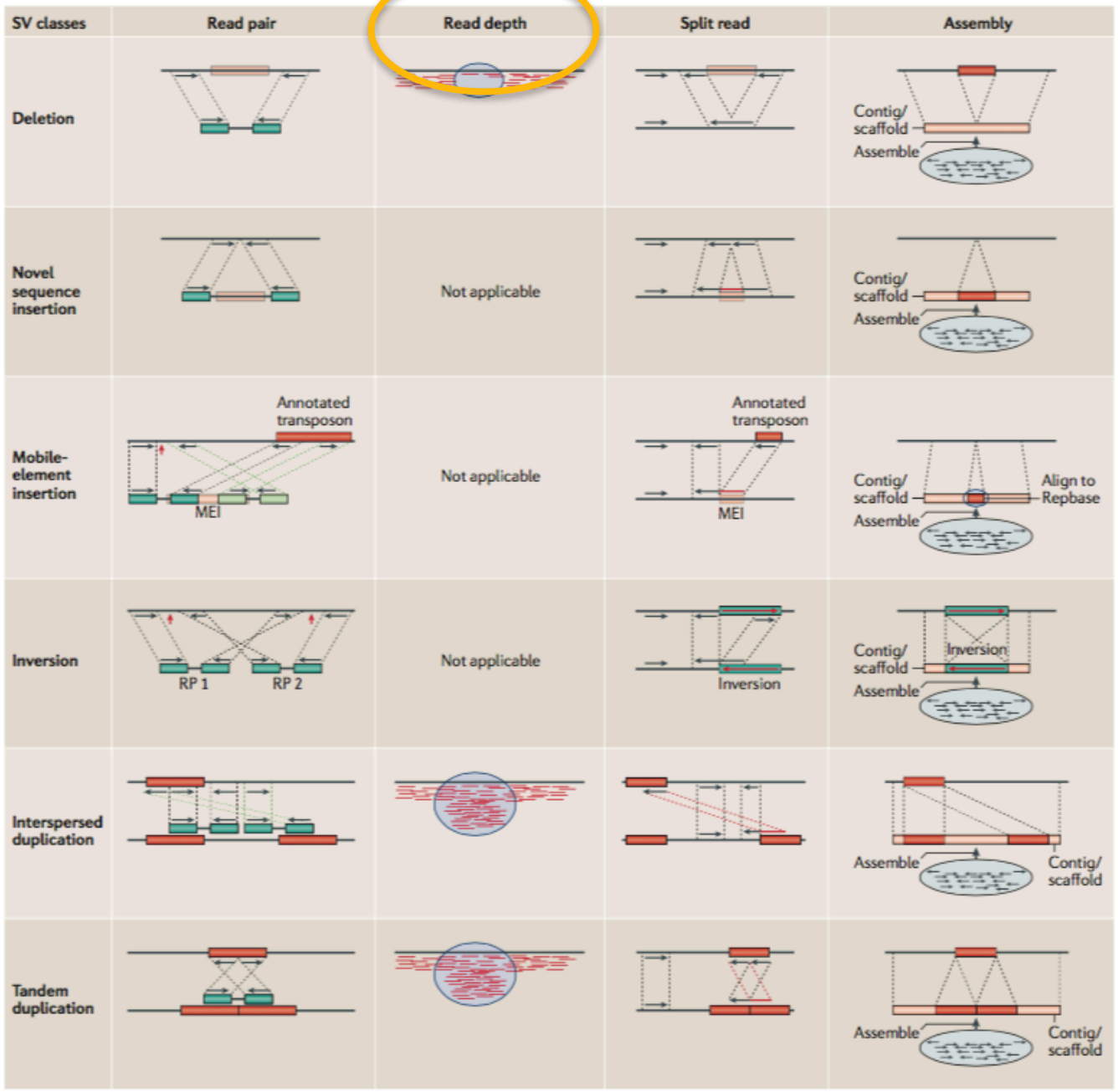
6C



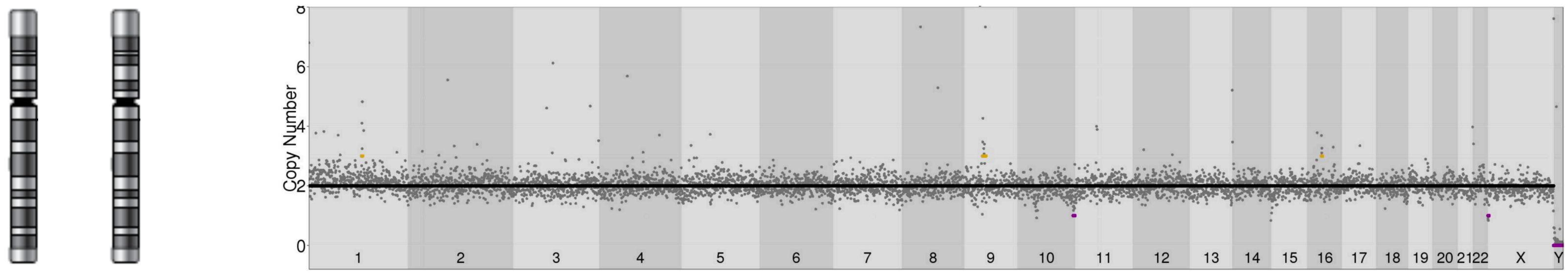


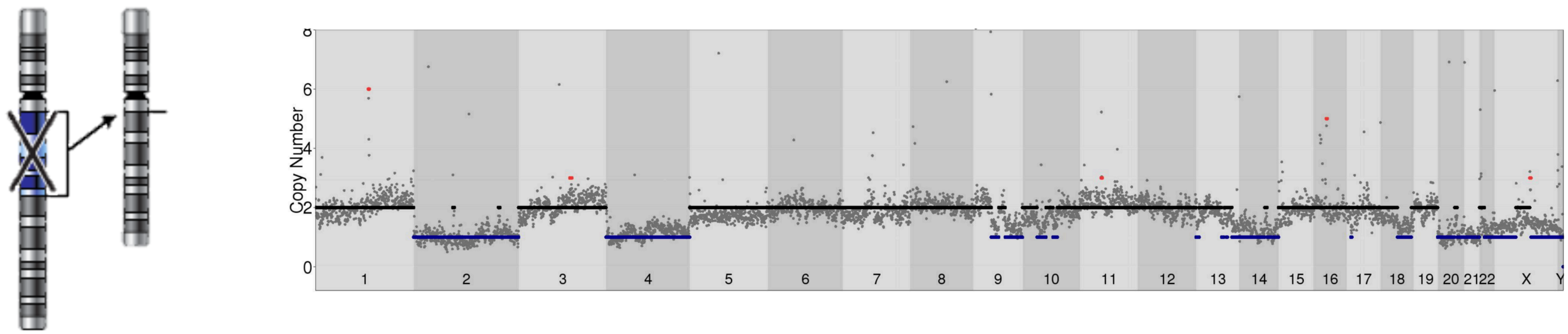
- Re-map 2kbp mates to each draft assembly, compute CE statistic at every position
- Extreme CE values are likely to be mis-assemblies
  - Can also look at coverage, mis-oriented mates, and other forensics features
  - Approximately 1.4 major events per Mbp

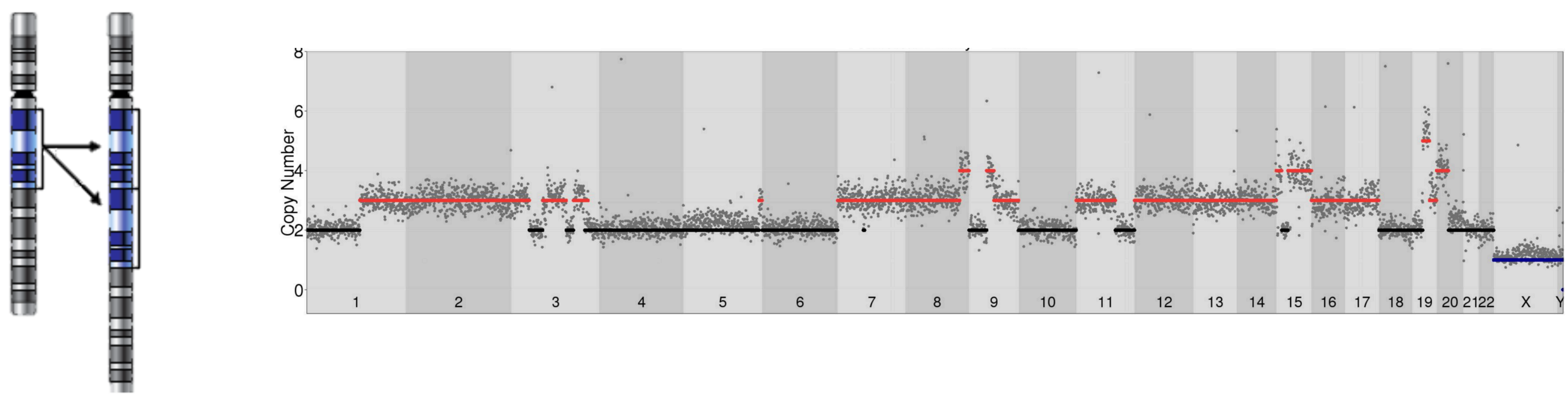
Structural Variation Sequence Signatures



What are Copy Number Variations?

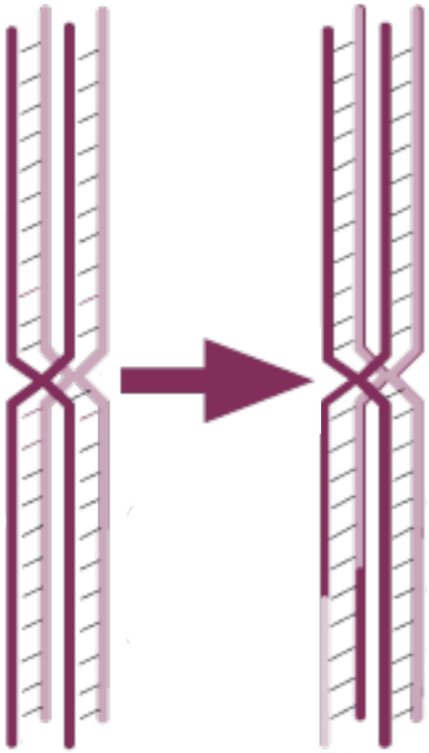






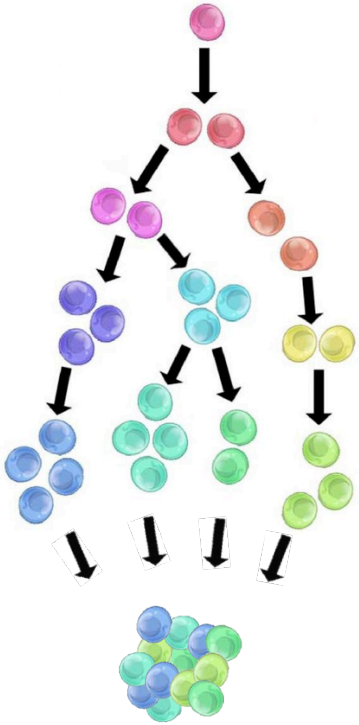
Single Cell CNV analysis

Germ Cells



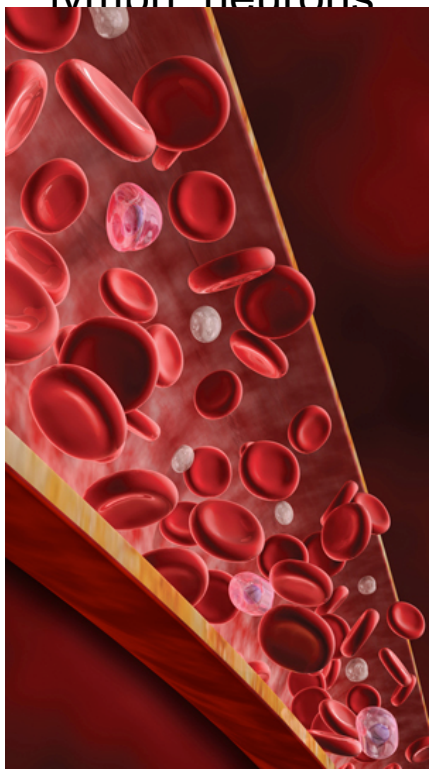
Recombination & crossover events

Heterogeneous Tumors



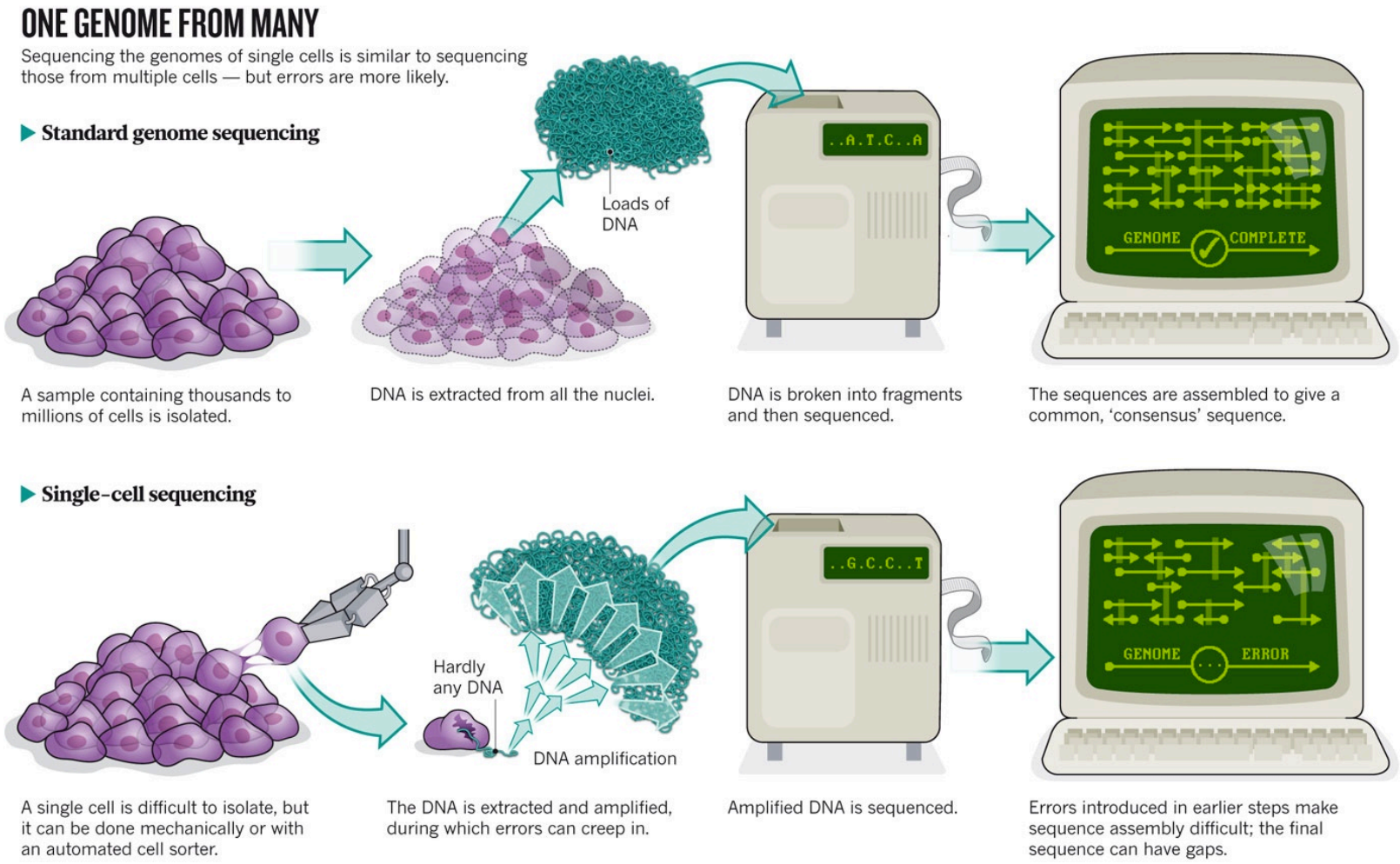
Clonal expansion

Heterogeneous Tissues: blood, lymph, neurons

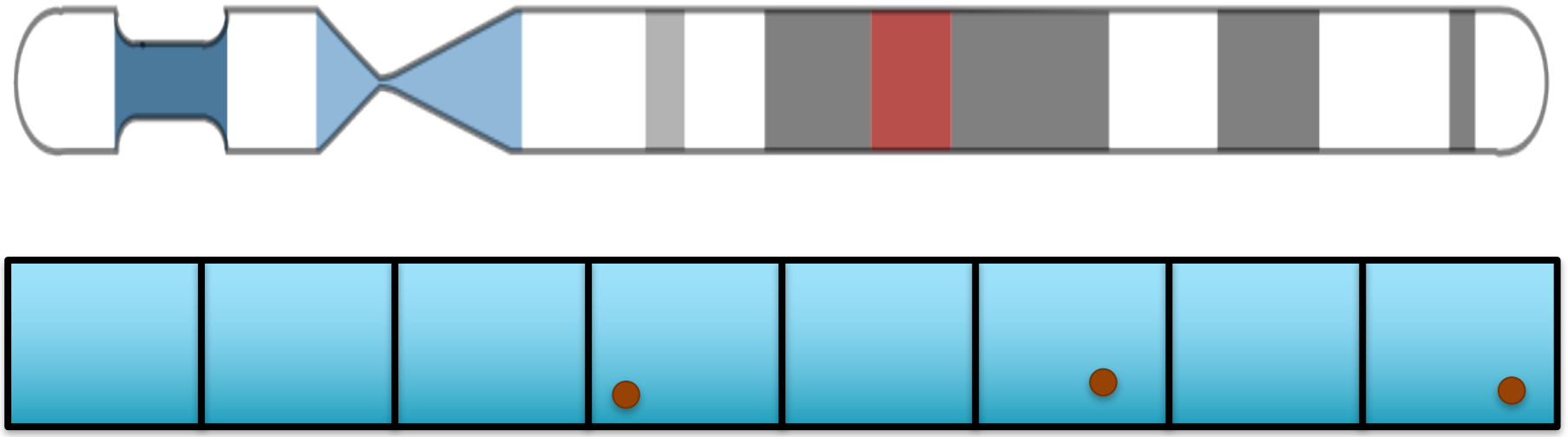


Isolating unique cell types

Whole Genome Amplification

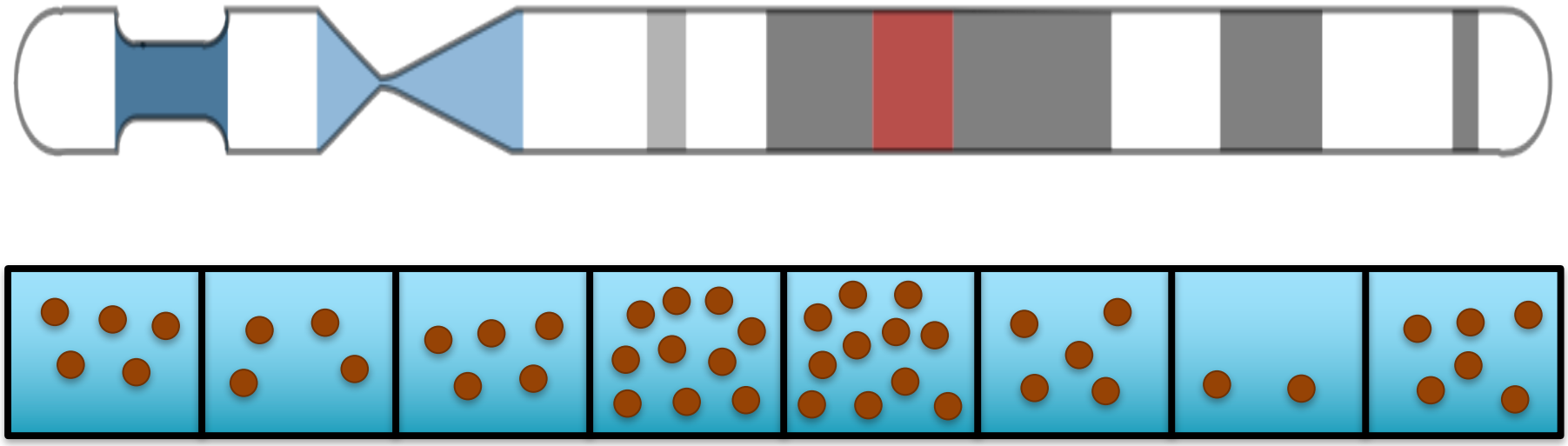


1) Binning



- Single Cell CNV analysis
  - Coverage is too sparse to identify point mutations
  - Divide the genome into “bins” with ~50 – 100 reads / bin
  - Map the reads and count reads per bin

1) Binning



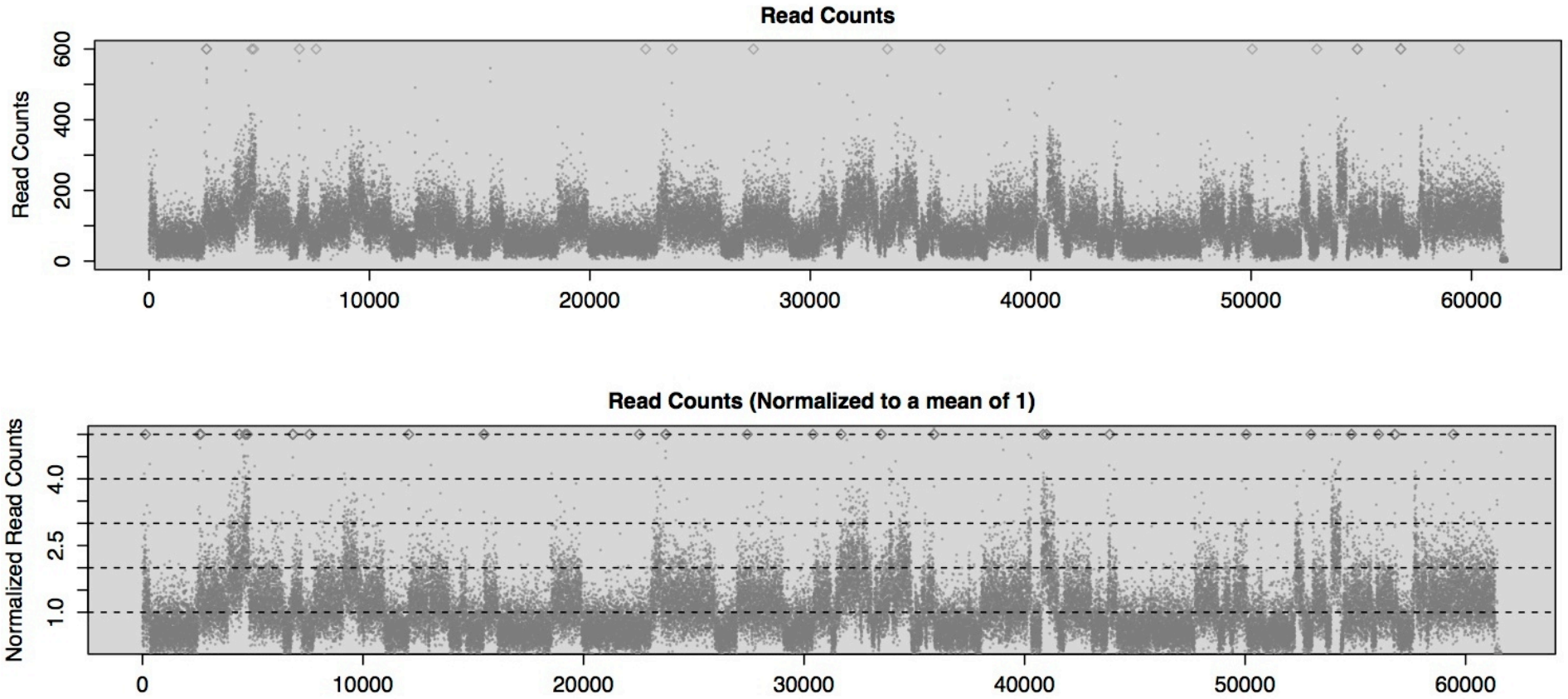
- Single Cell CNV analysis
  - Coverage is too sparse to identify point mutations
  - Divide the genome into “bins” with ~50 – 100 reads / bin
  - Map the reads and count reads per bin

1) Binning

5	4	5	10	11	5	2	5

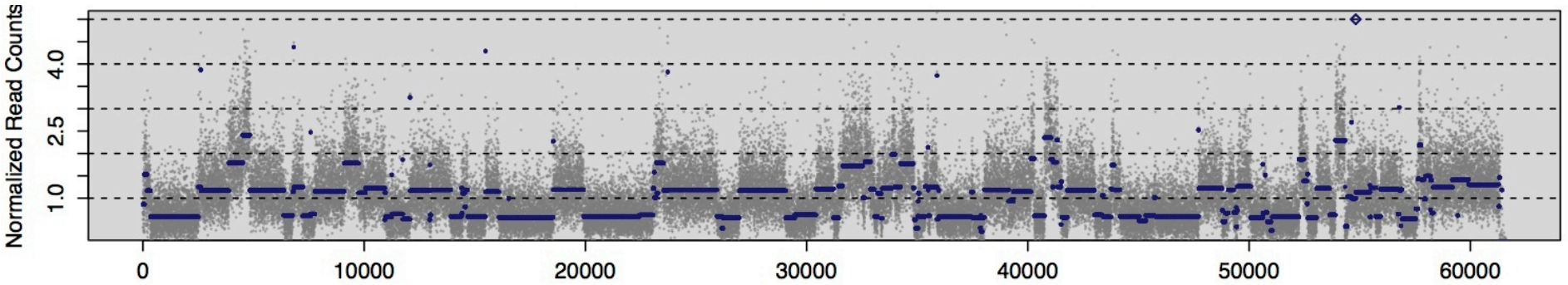
- Single Cell CNV analysis
  - Coverage is too sparse to identify point mutations
  - Divide the genome into “bins” with ~50 – 100 reads / bin
  - Map the reads and count reads per bin

2) Normalization

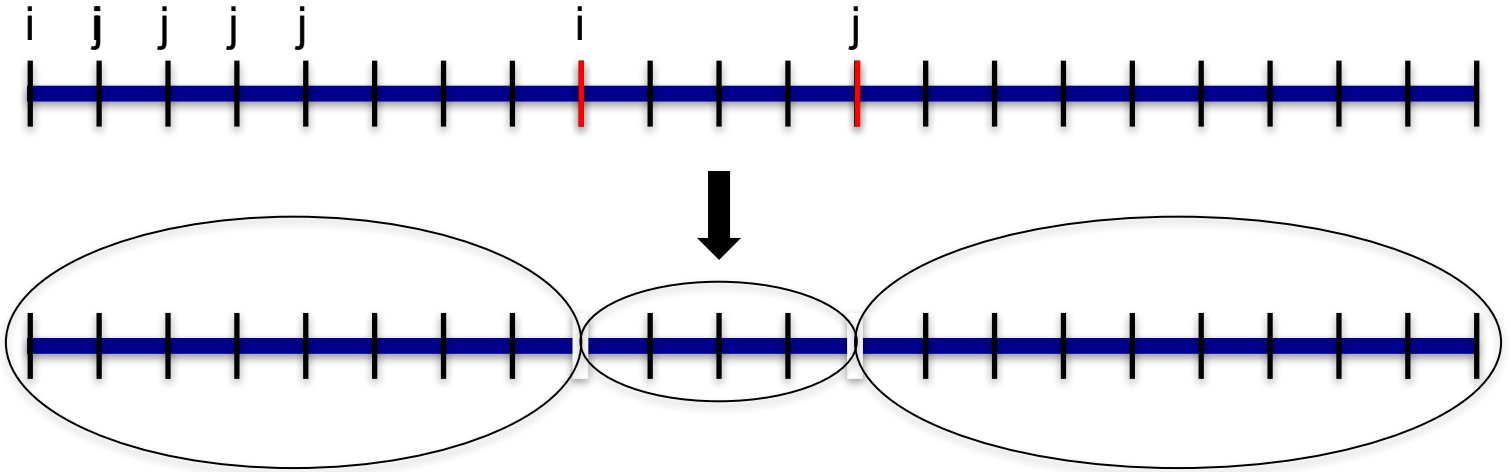


With real data, also correct for mappability, GC content, amplification biases

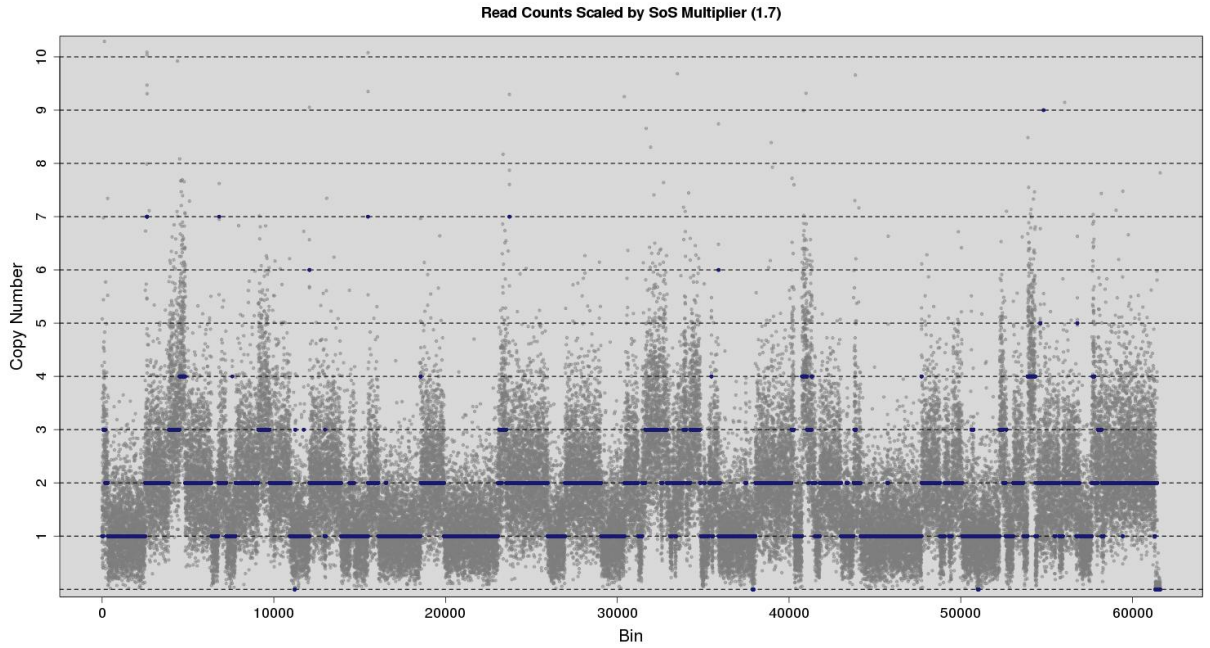
3) Segmentation



Circular Binary Segmentation (CBS)

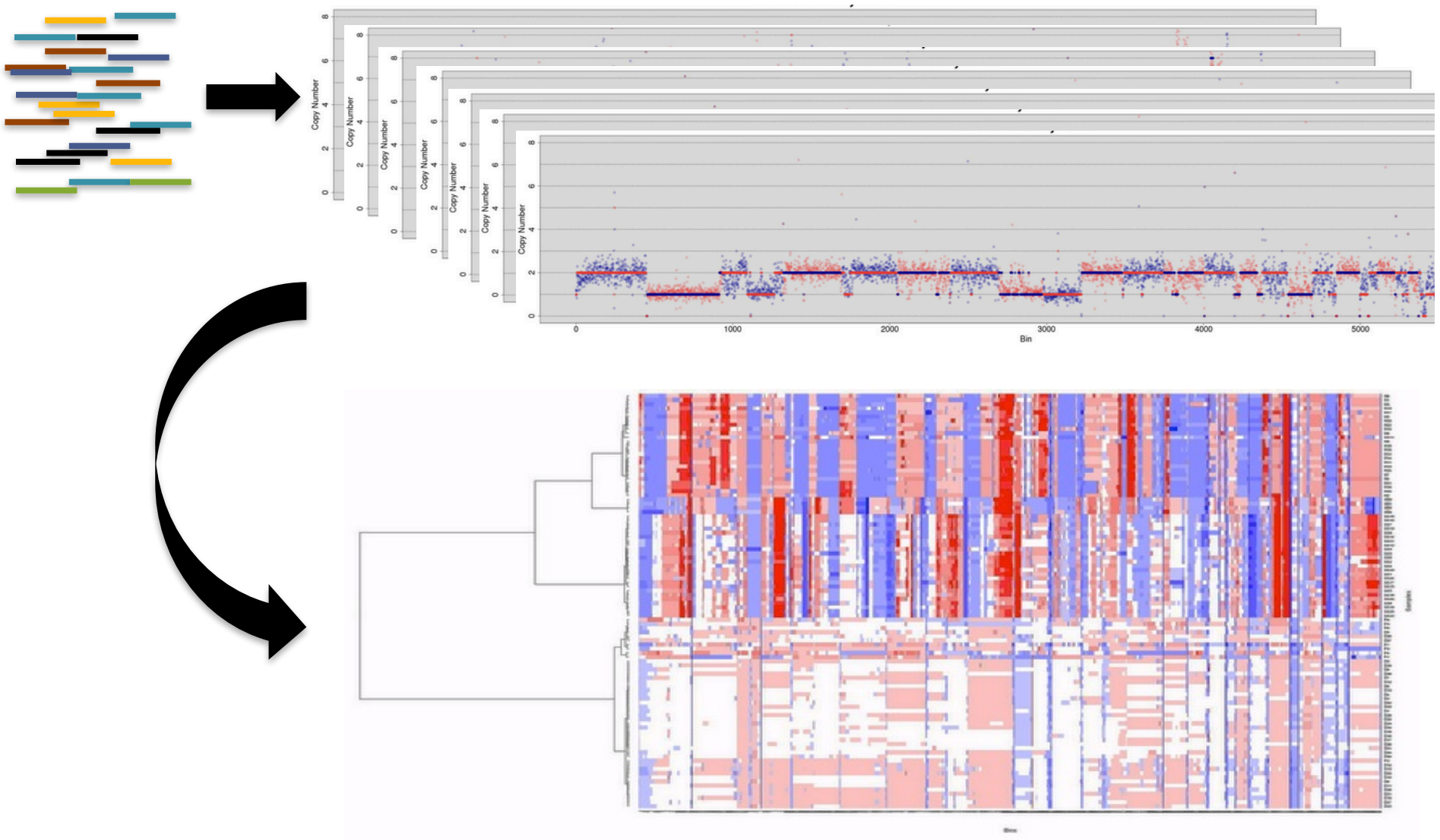


4) Estimating Copy Number



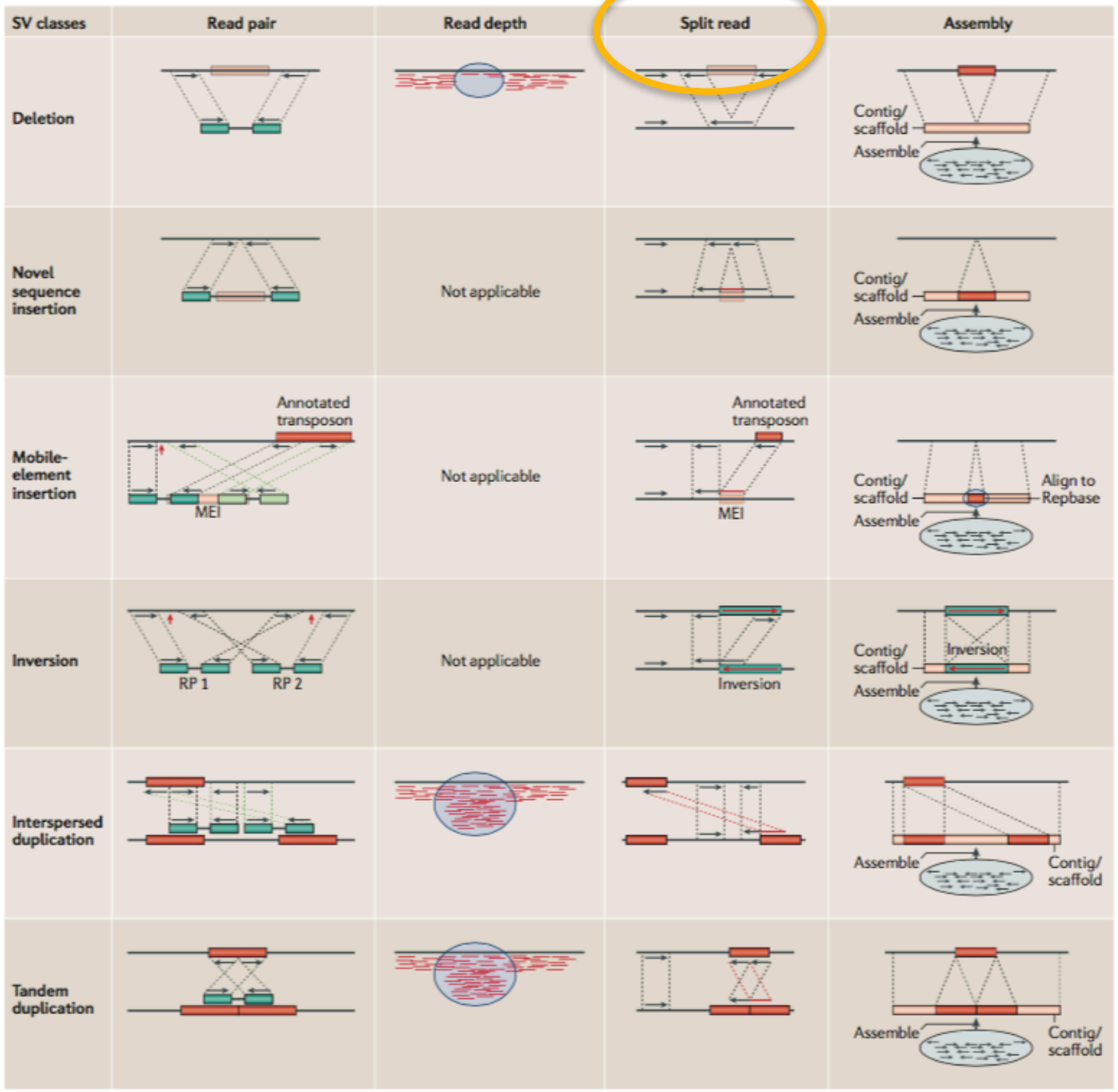
$$CN = argmin \left\{ \sum_{i,j} (\hat{Y}_{i,j} - Y_{i,j}) \right\}$$

5) Cells to Populations



Interactive analysis and assessment of single-cell copy-number variations.
Garvin et al (2015) Nature Methods doi:10.1038/nmeth.3578

Structural Variation Sequence Signatures



A probabilistic framework for SV discovery

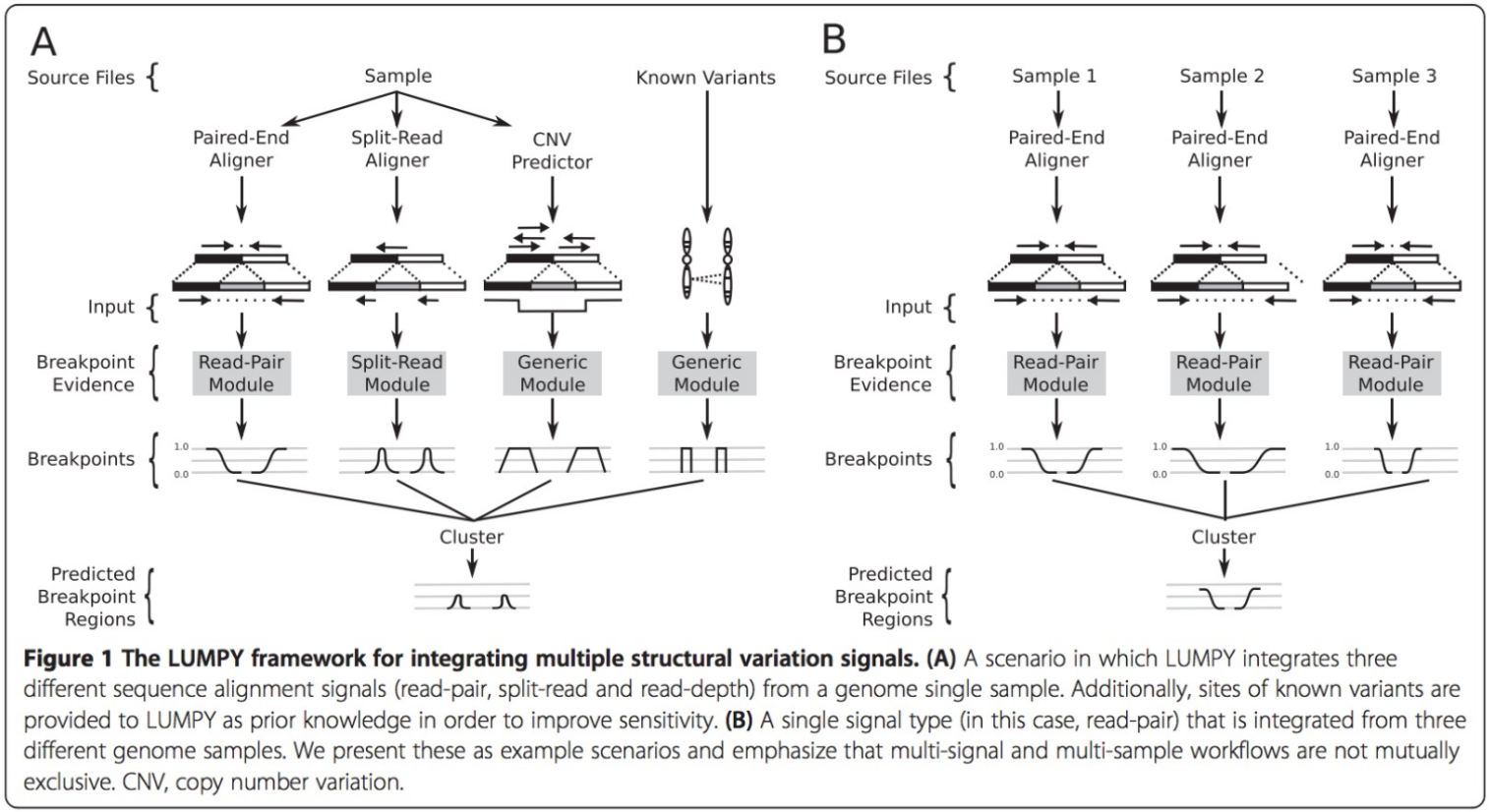
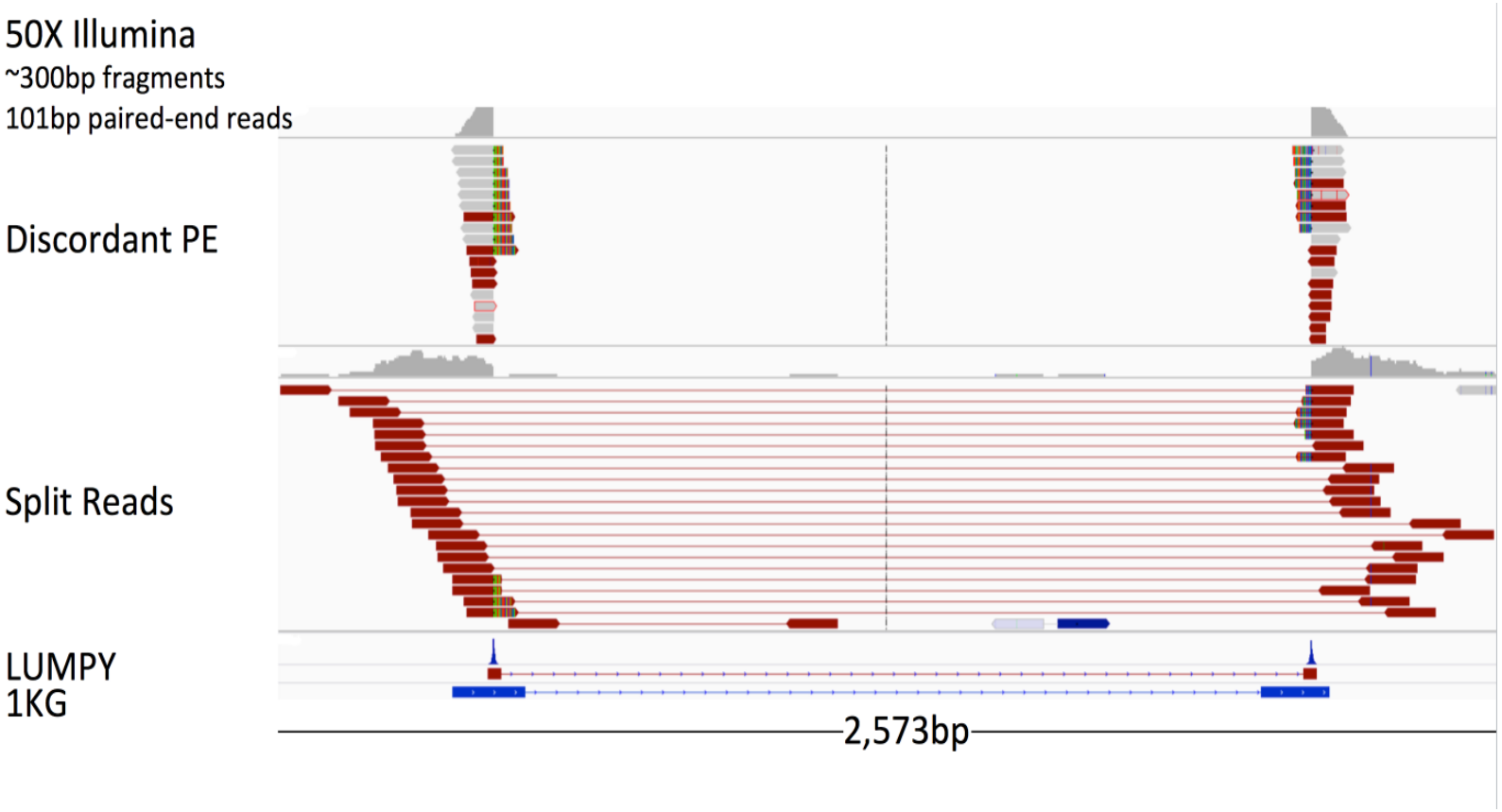


Figure 1 The LUMPY framework for integrating multiple structural variation signals. (A) A scenario in which LUMPY integrates three different sequence alignment signals (read-pair, split-read and read-depth) from a genome single sample. Additionally, sites of known variants are provided to LUMPY as prior knowledge in order to improve sensitivity. **(B)** A single signal type (in this case, read-pair) that is integrated from three different genome samples. We present these as example scenarios and emphasize that multi-signal and multi-sample workflows are not mutually exclusive. CNV, copy number variation.

Lumpy integrates paired-end mapping, split-read mapping, and depth of coverage for better SV discovery accuracy

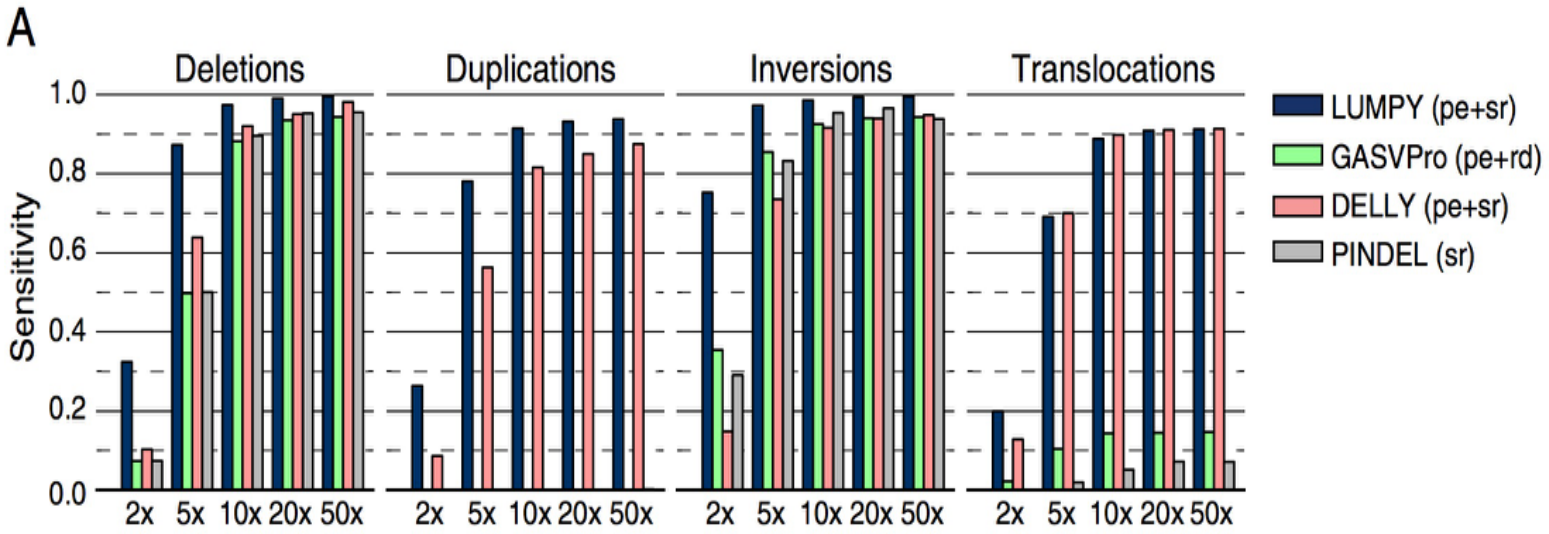
Ryan Layer

A probabilistic framework for SV discovery



Layer et al, 2014

A probabilistic framework for SV discovery



Sequencing depth

Layer et al, 2014

The dirty secrets of SV discovery

Secret #1: Often many false positives

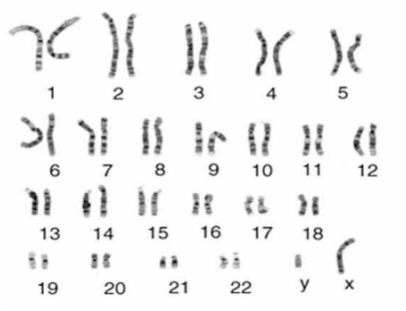
- Short reads + heuristic alignment + rep. genome = **systematic alignment artifacts (false calls)**
- Chimeras and duplicate molecules
- Ref. genome errors (e.g., gaps, mis-assemblies)
- **ALL** SV mapping studies use strict filters for above

Secret #2: The false negative rate is also typically high

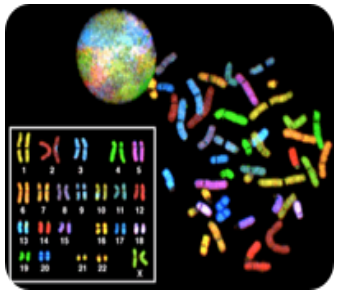
- Most current datasets have low to moderate ***physical*** coverage due to small insert size (~10-20X)
- Breakpoints are enriched in repetitive genomic regions that pose problems for sensitive read alignment
- FILTERING!
- The false negative rate is usually hard to measure, but is thought to be extremely high for most paired-end mapping studies (>30%)
- When searching for spontaneous mutations in a family or a tumor/normal comparison, a false negative call in one sample can be a false positive somatic or de novo call in another.

PacBio Single Molecule Real Time Sequencing (SMRT-sequencing)

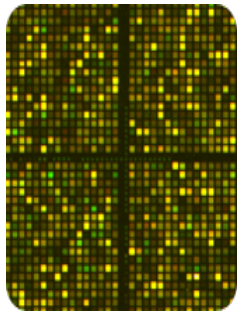
Our understanding of structural variation is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



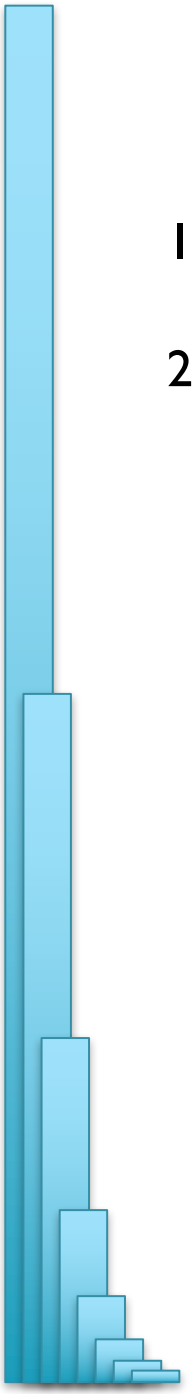
2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

Today
High throughput
DNA sequencing

Next Steps

1. Finish Assignment 1

2. Check out the course webpage




Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics

Questions?