CS681: Advanced Topics in Computational Biology

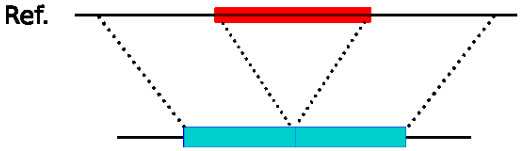
Week 6 Lecture 1

Can Alkan
EA224
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

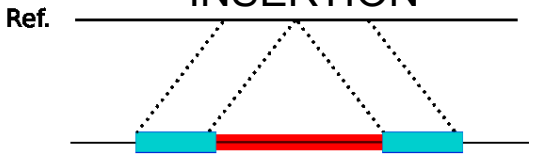
Structural Variation Classes

DELETION

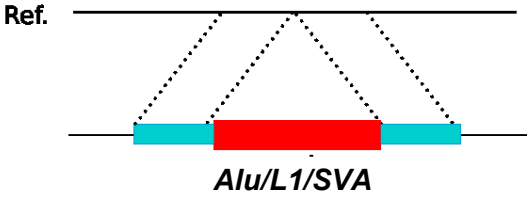


Autism, mental retardation, Crohn's

NOVEL SEQUENCE INSERTION

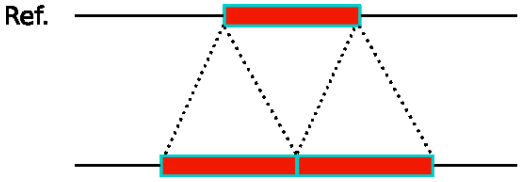


MOBILE ELEMENT INSERTION

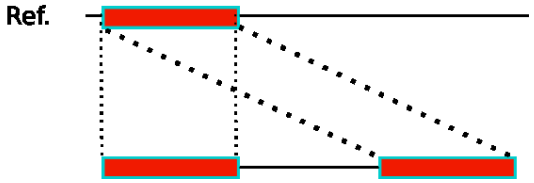


Haemophilia

TANDEM DUPLICATION



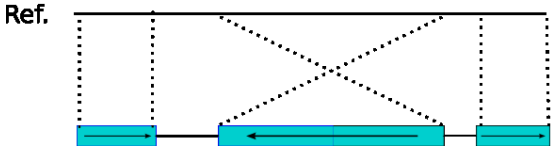
INTERSPERSED DUPLICATION



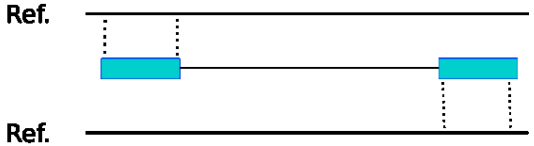
Schizophrenia, psoriasis

CNV: Copy number variants

INVERSION



TRANSLOCATION



Chronic myelogenous leukemia

Balanced rearrangements

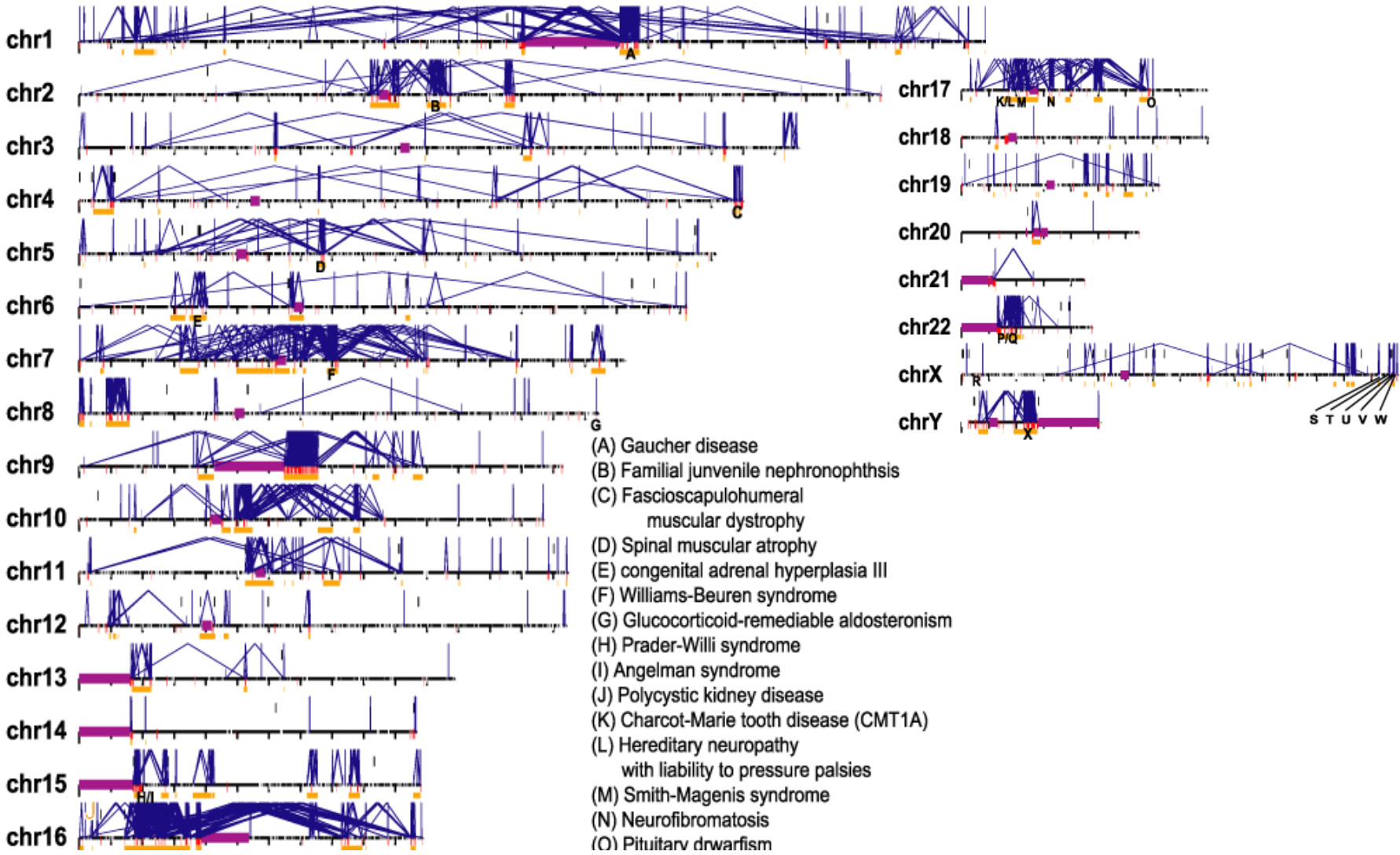
Structural variation discovery with NGS data

- SVs: genomic alterations > 50 bp.
  - Databases:
    - dbVar: http://www.ncbi.nlm.nih.gov/dbvar/
    - DGV: http://projects.tcag.ca/variation/
- Input: sequence data and reference genome
- Output: set of SVs and their genotypes (homozygous/heterozygous)
- Often there are errors, filtering required
- SV detection methods can be based on statistical analysis or combinatorial optimization
- Tools: VariationHunter, BreakDancer, MoDIL, CommonLAW, Genome STRiP, Spanner, HYDRA, etc.

Challenges

- Most SVs are embedded within or around segmental duplications or long repeats
  - If you use unique mapping, you will lose sensitivity
  - Ambiguous mapping of reads will increase false positives
  - Reference genome is incomplete; missing portions are duplications which cause more problems in accurate detection
- Many SVs are complex; many rearrangements at the same site
- CNV discovery is heavily studied but still not perfect; detection of balanced rearrangements are still problematic

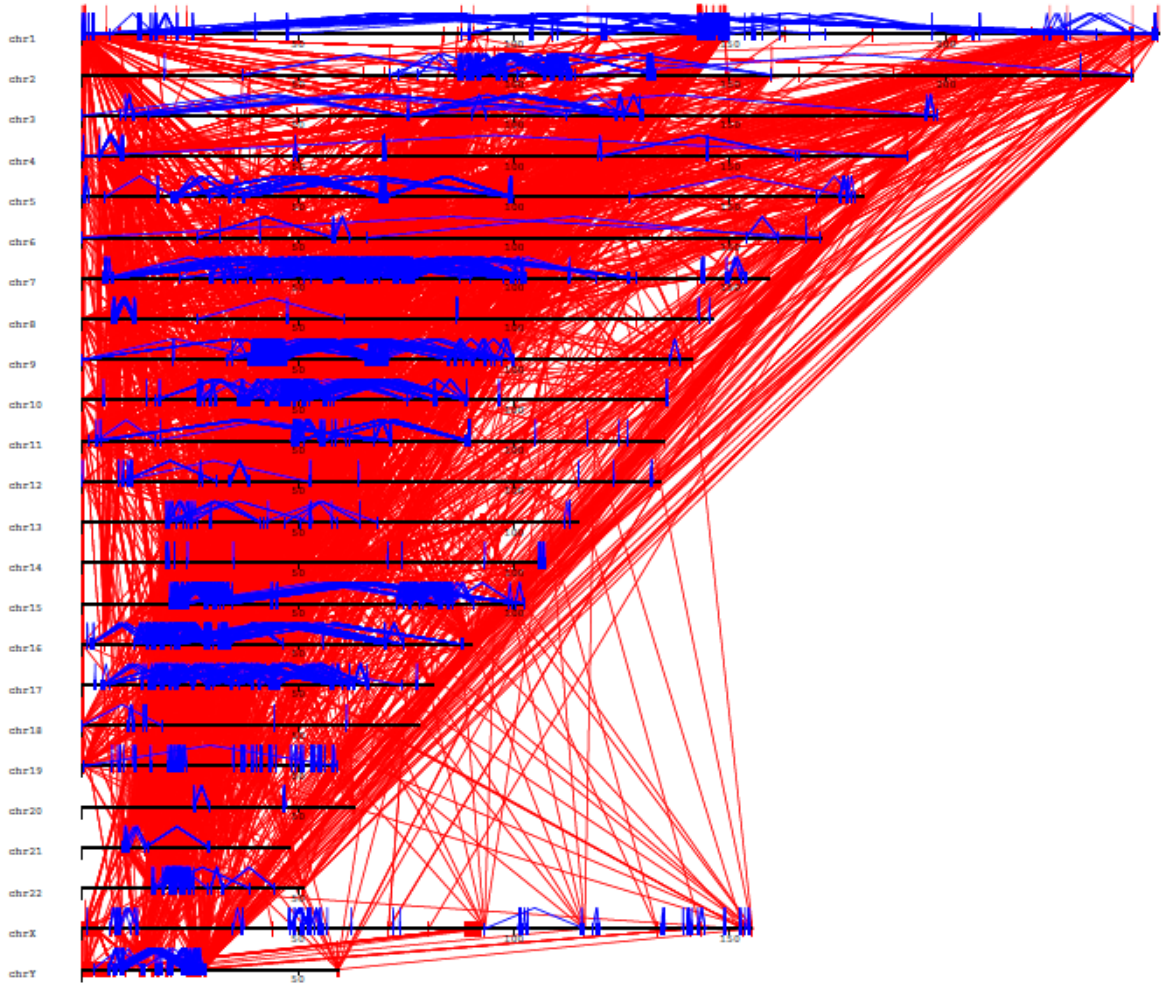
Duplications and CNV hotspots



Human genome

Bailey et al., Science, 2002

Duplications: inter & intra



- 51,599 pairs of SDs
  - 18,559 pairs intrachromosomal
  - 32,740 pairs interchromosomal
- Non-redundant corresponds to 166 Mb (~5% of genome)

Human genome

Bailey et al., Science, 2002

Genome-wide SV Discovery Approaches

Hybridization-based

- Iafrate et al., 2004, Sebat et al., 2004
- SNP microarrays: McCarroll *et al*., 2008, Cooper *et al.*, 2008, Itsara *et al.*, 2009
- Array CGH: Redon *et al.* 2006, Conrad *et al.,* 2010, Park *et al.,* 2010, WTCCC, 2010

Single molecule analysis

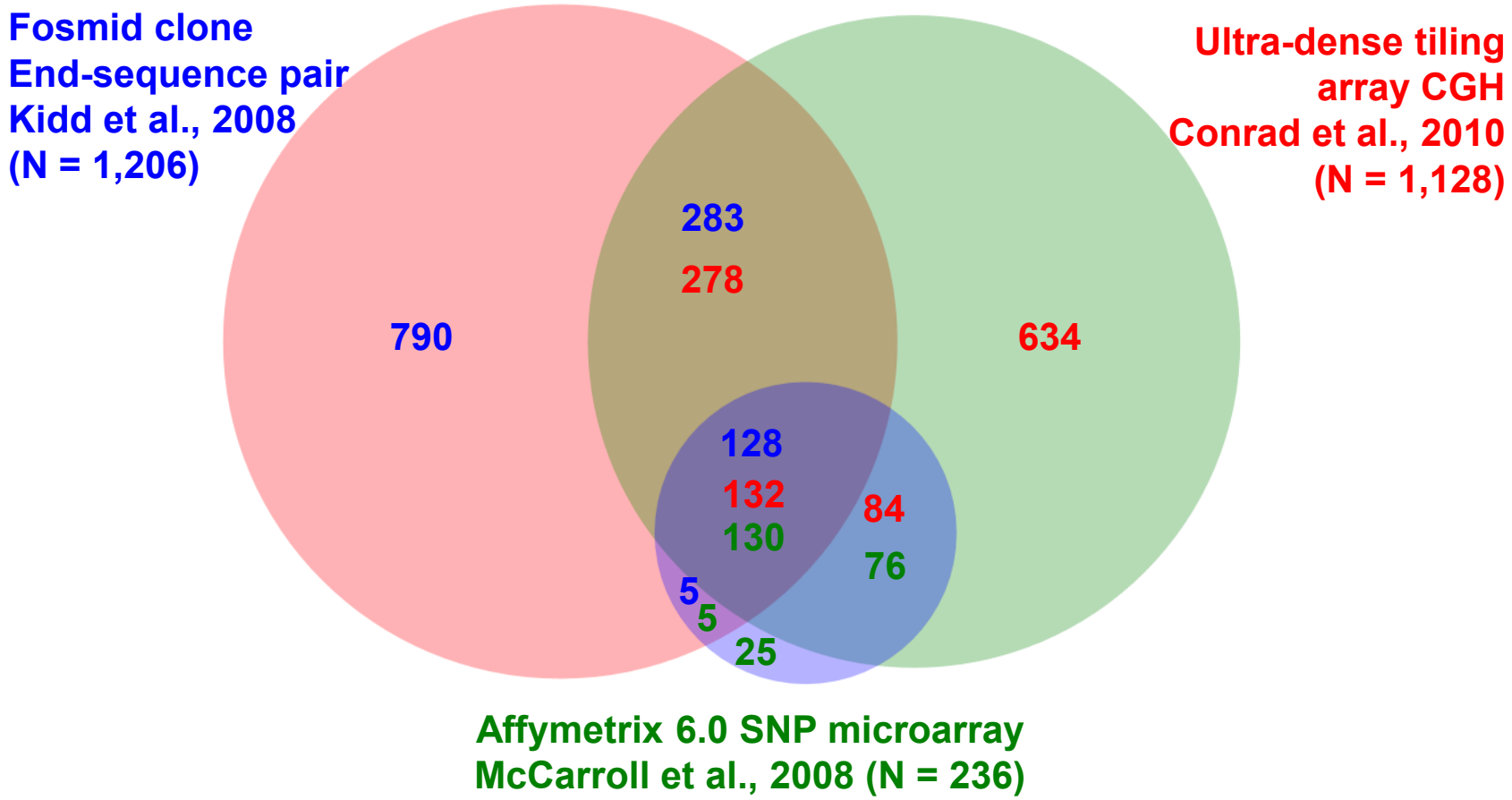
- **Optical mapping:** Teague et al., 2010

Sequencing-based

- Read-depth: Bailey et al, 2002
- Fosmid ESP: Tuzun *et al.* 2005, Kidd *et al.* 2008
- Sanger sequencing: Mills *et al.*, 2006
- Next-gen sequencing: Korbel *et al.* 2007, Yoon *et al.*, 2009, Alkan et al., 2009, Hormozdiari *et al.* 2009, Chen *et al.* 2009,
  - 1000 Genomes Project

Detection diversity

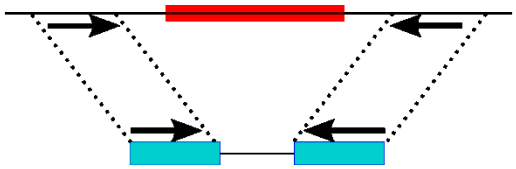
Gains & Losses > 5 Kbp in the same 5 individuals



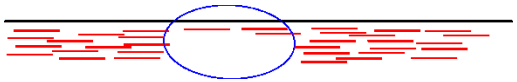
Kidd et al. Cell, 2010

Sequence signatures of structural variation

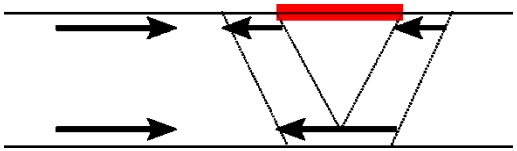
- Read pair analysis
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size



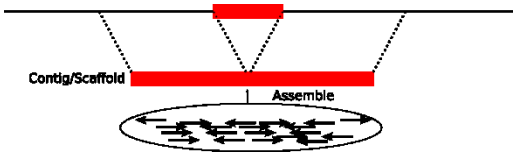
- Read depth analysis
  - Deletions and duplications only
  - Relatively poor breakpoint resolution



- Split read analysis
  - Small novel insertions/deletions, and mobile element insertions
  - 1bp breakpoint resolution



- Local and *de novo* assembly
  - SV in unique segments
  - 1bp breakpoint resolution



SV by sequencing: first algorithms

Read Depth

799

Recent Segmental Duplications in the Human Genome

Jeffrey A. Bailey,[1] Zhiping Gu,[2] Royden A. Clark,[1] Knut Reinert,[2] Rhea V. Samonte,[1] Stuart Schwartz,[1] Mark D. Adams,[2] Eugene W. Myers,[2] Peter W. Li,[2] Evan E. Eichler[1]*

Science, 2002

Read Pair

662

nature genetics

Fine-scale structural variation of the human genome

Eray Tuzun[1,5], Andrew J Sharp[1,5], Jeffrey A Bailey[2,5], Rajinder Kaul[3], V Anne Morrison[1], Lisa M Pertz[2], Eric Haugen[3], Hillary Hayden[3], Donna Albertson[4], Daniel Pinkel[4], Maynard V Olson[3] & Evan E Eichler[1]

Nature Genetics, 2005

Split read

196

An initial map of insertion and deletion (INDEL) variation in the human genome

Ryan E. Mills,[1,2] Christopher T. Luttig,[1] Christine E. Larkins,[3] Adam Beauchamp,[4] Circe Tsui,[1,2] W. Stephen Pittard,[2,5] and Scott E. Devine[1,2,3,4,6]

Genome Research, 2006

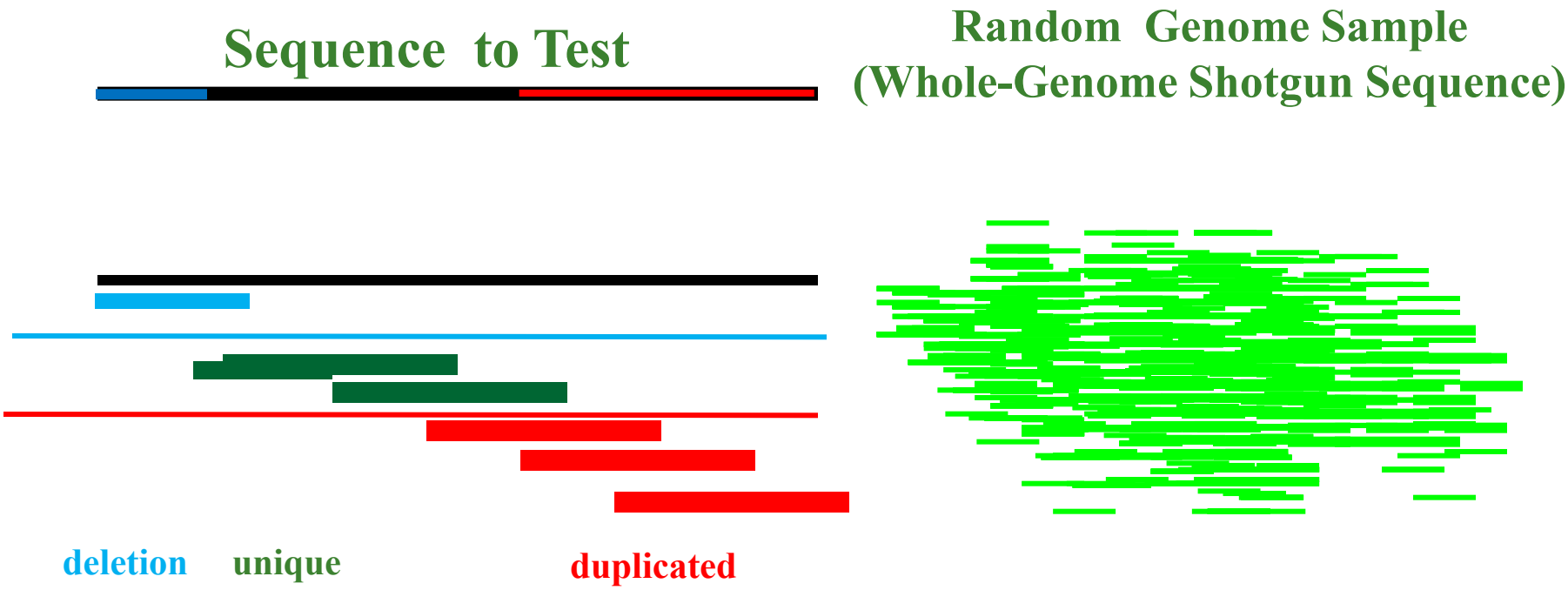
All these first algorithms used Sanger sequence, but laid out the basic principles for NGS analysis

Read depth based algorithms

- Assume random (Poisson) distribution in read depth
- Multiple mapping:
  - WSSD (whole genome shotgun sequence detection)
- Unique mapping:
  - Low resolution: Campbell et al. Nat Genet 2008, Chiang et al. Nat Meth, 2009 (SegSeq)
  - High(er) resolution: CNVnator, EWT (RDXplorer)

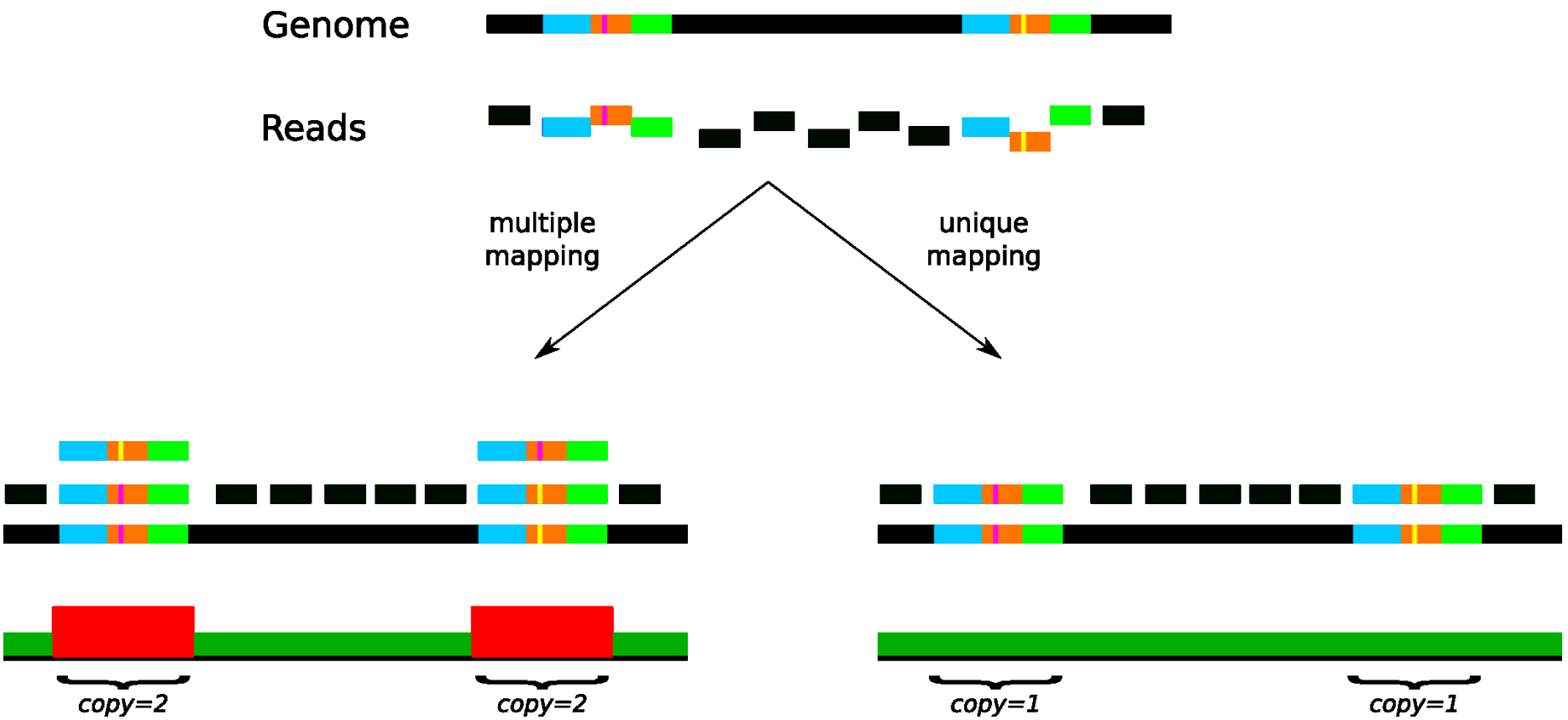
Read depth analysis: WSSD

- Uses database of random reads to confirm duplicated nature of the sequence
  - increased # of copies => increased number of reads
  - decreased # of copies => decreased number of reads
- Compute depth-of-coverage in 5kb windows (sliding by 1kb); select regions with increased depth as **duplications,** regions with reduced depth as **deletions** (WSSD method)



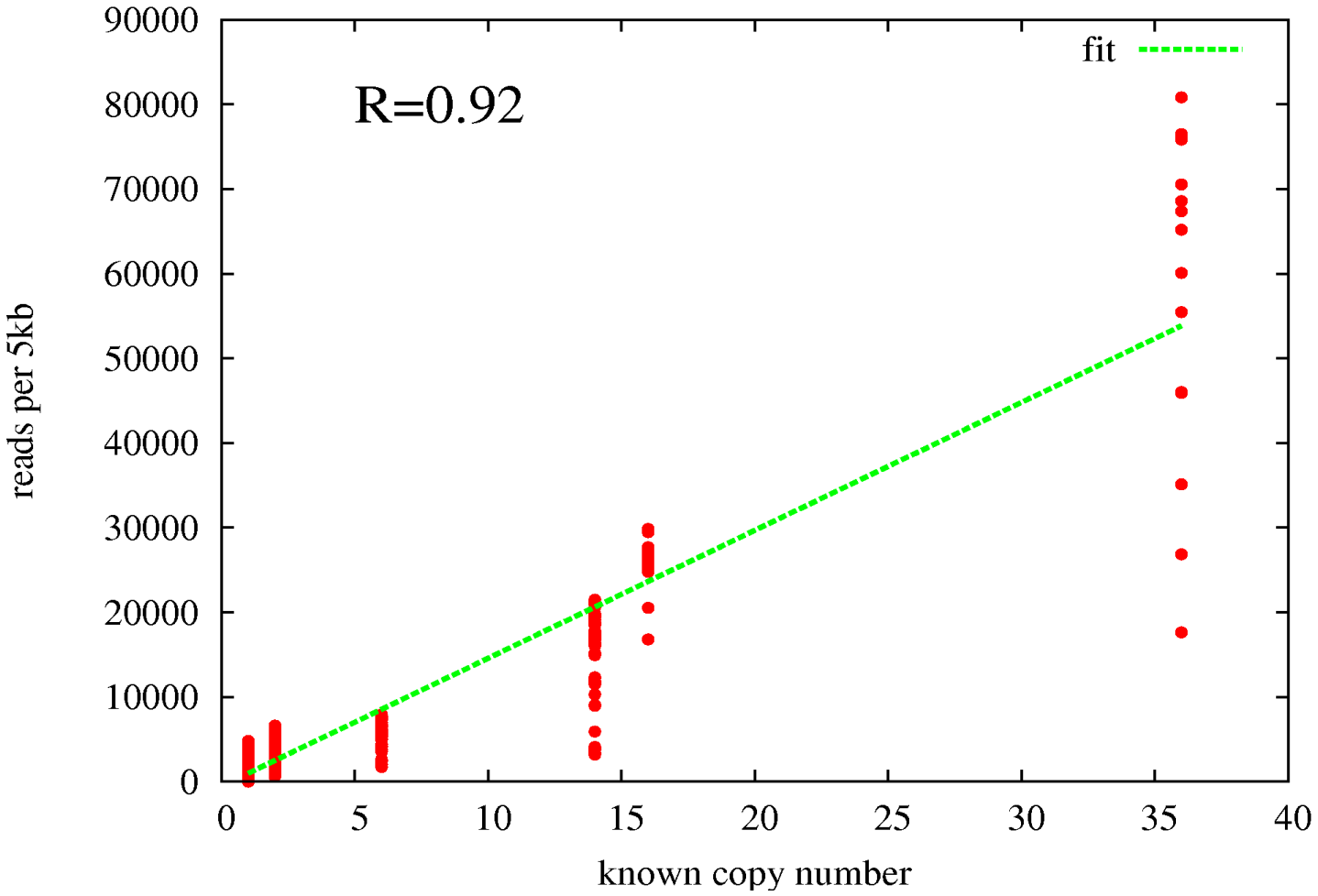
Bailey et al., Science, 2002

Multiple vs. unique mapping



Modified from Chiang & McCarroll, Nat Biotech, 2009

Read depth - Copy number correlation



Alkan et al., Nature Genetics, 2009

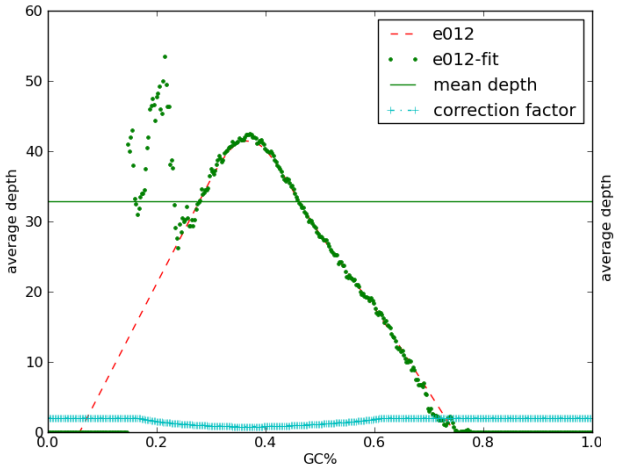
WSSD: next-gen

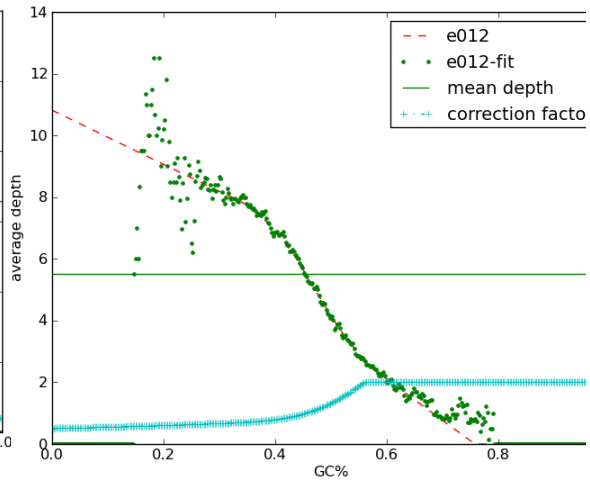
- NGS specific problems
  - Short reads: MegaBLAST is replaced by mrFAST / mrsFAST
  - Common repeats: all repeats need to be masked
  - GC % bias needs to be fixed
- Improvement
  - Absolute copy number detection in 1 kb non-overlapping windows
  - Genotyping highly identical paralogs

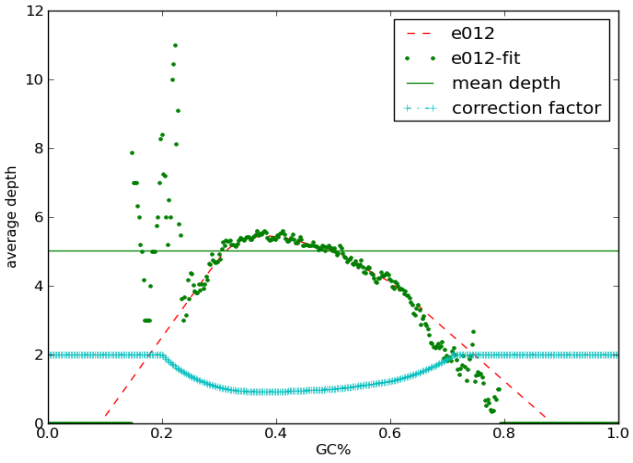
Alkan et al., Nat Genet, 2009

Read depth distribution

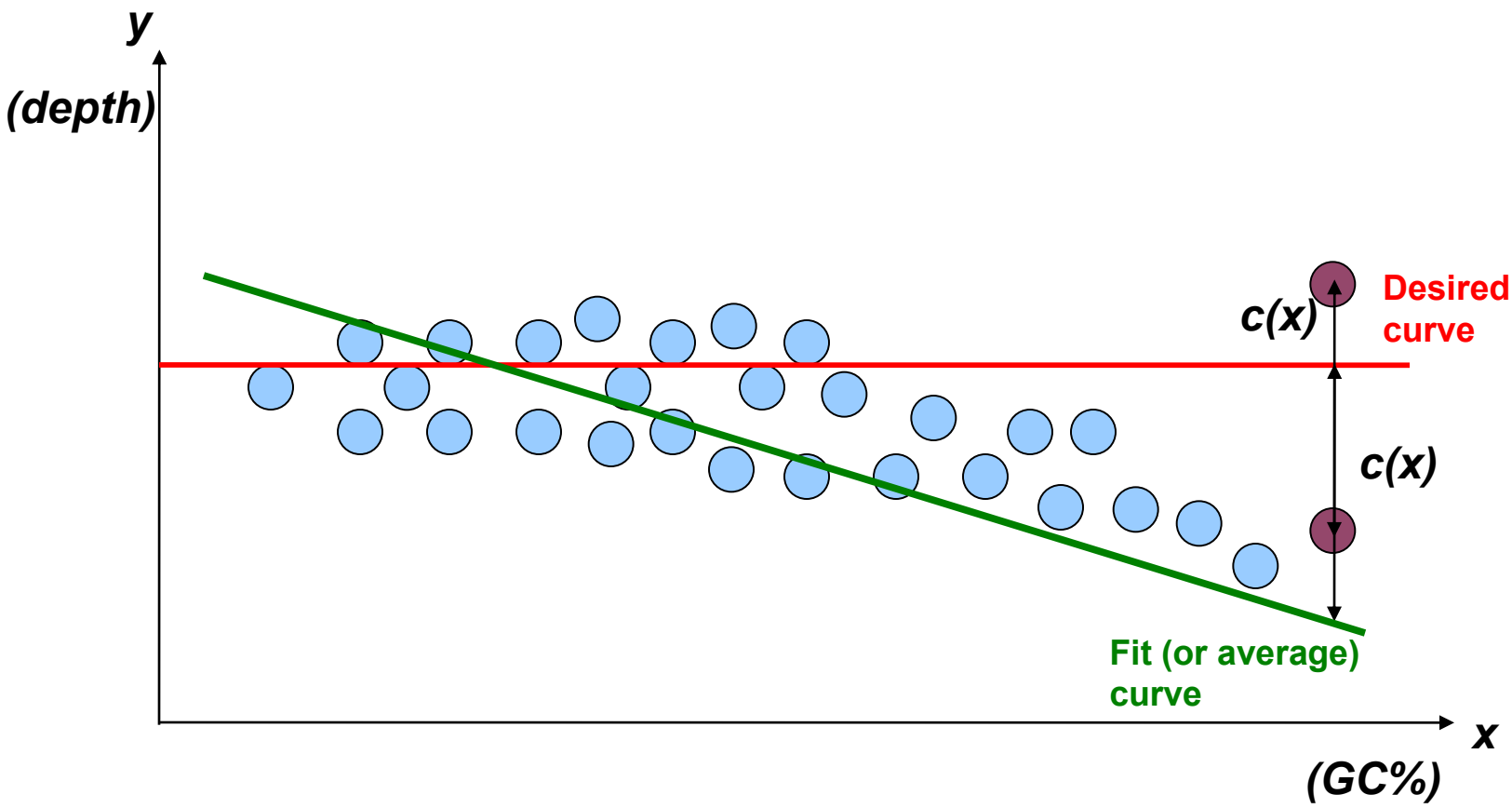
- Read depth doesn't really follow Poisson distribution
  - Biases against high and low GC %







GC% correction: LOESS



$y' = y - c(x)$

$c(x) = f(x) - e(x)$

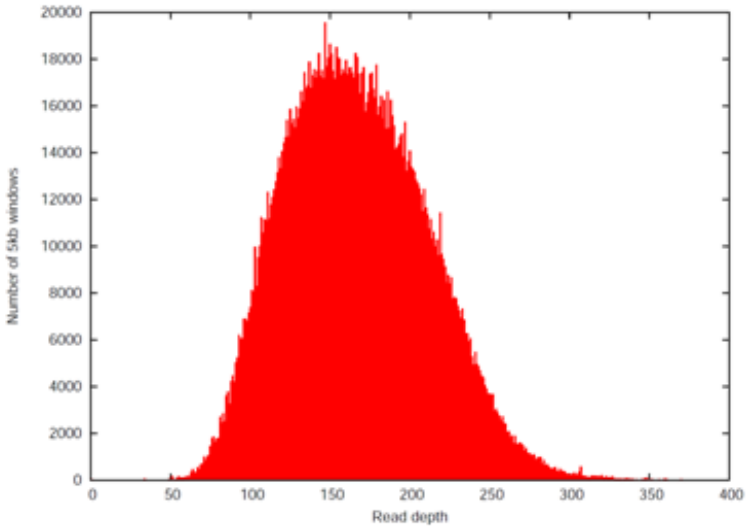
GC% correction (modified LOESS)

$$k_{gc} = \mu_{total}/\mu_{gc}$$

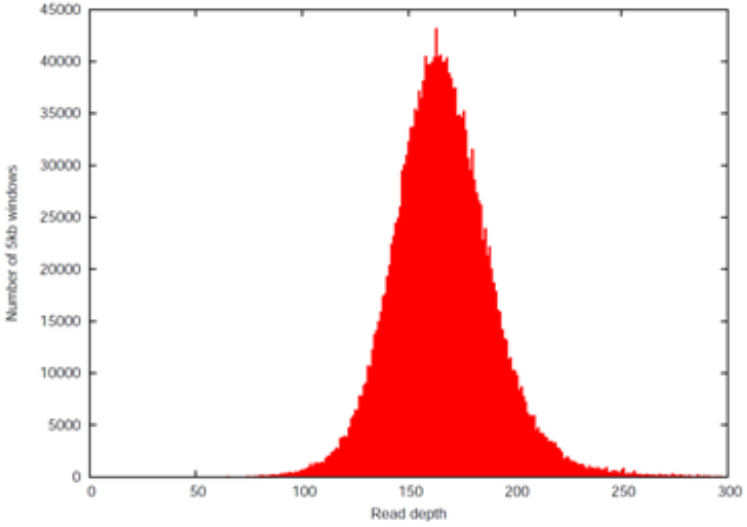
$$d'_{gc} = d_{gc}k_{gc}$$

The version in SegSeq and CNVnator

GC% correction

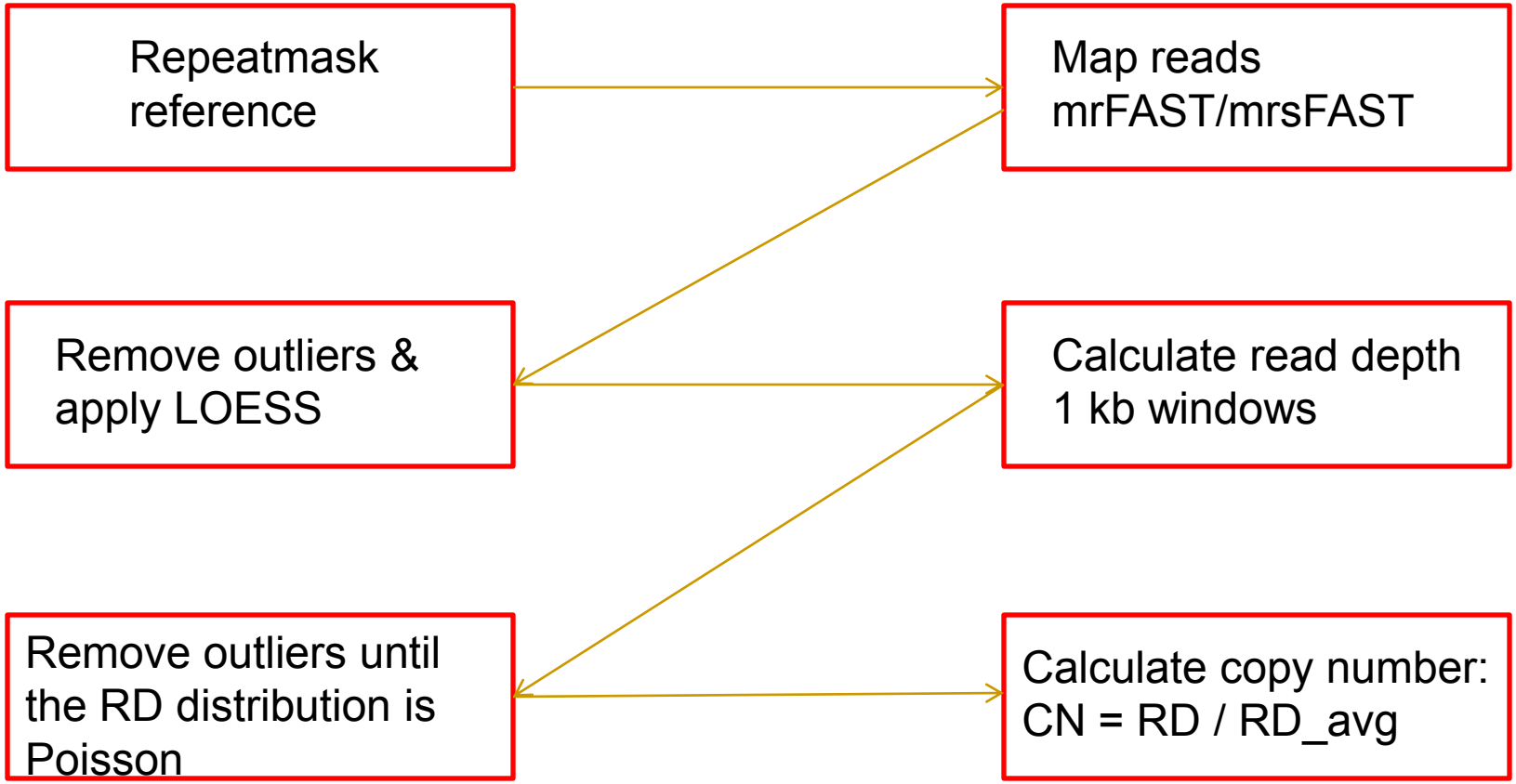


Before GC correction



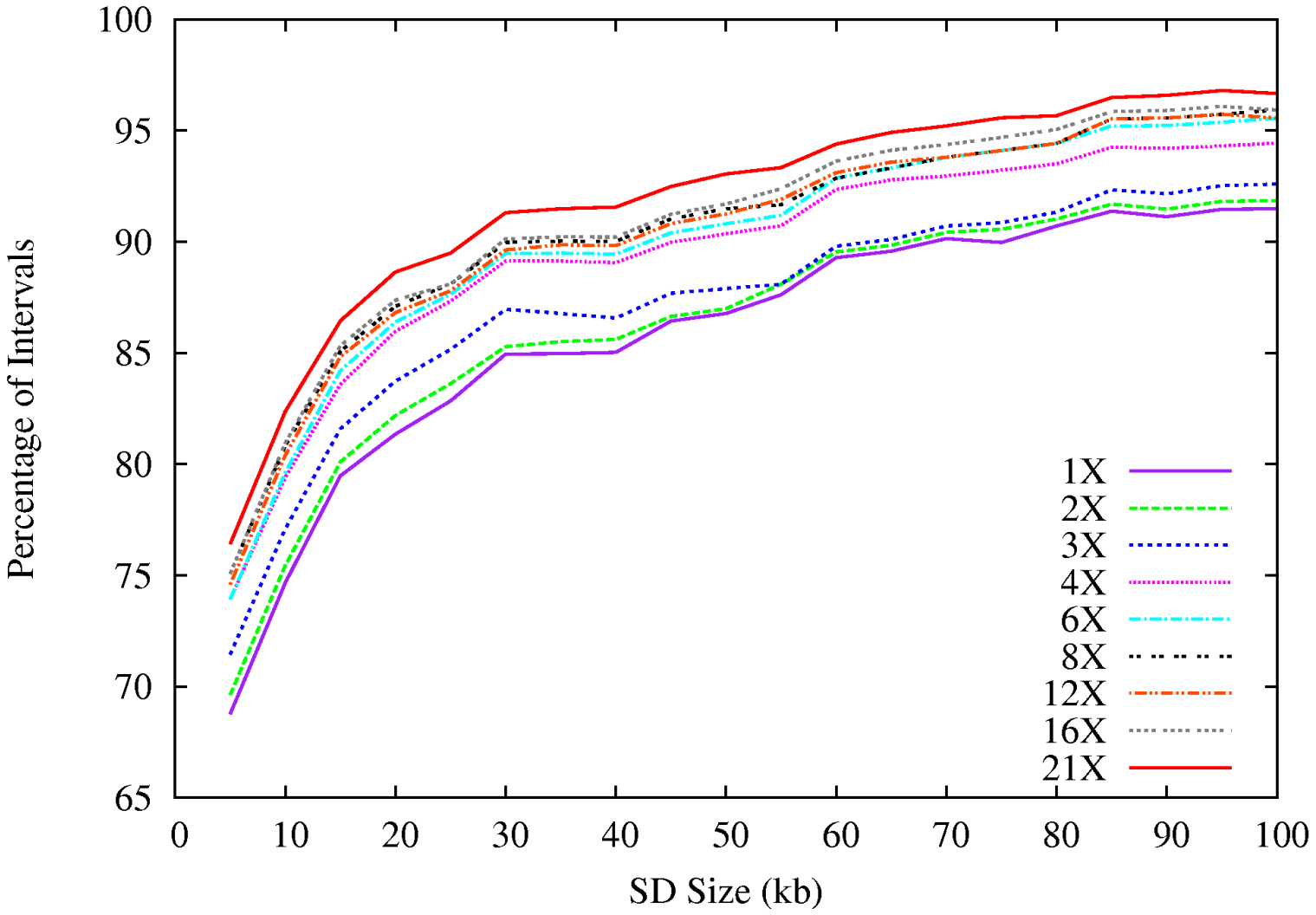
After GC correction

WSSD workflow

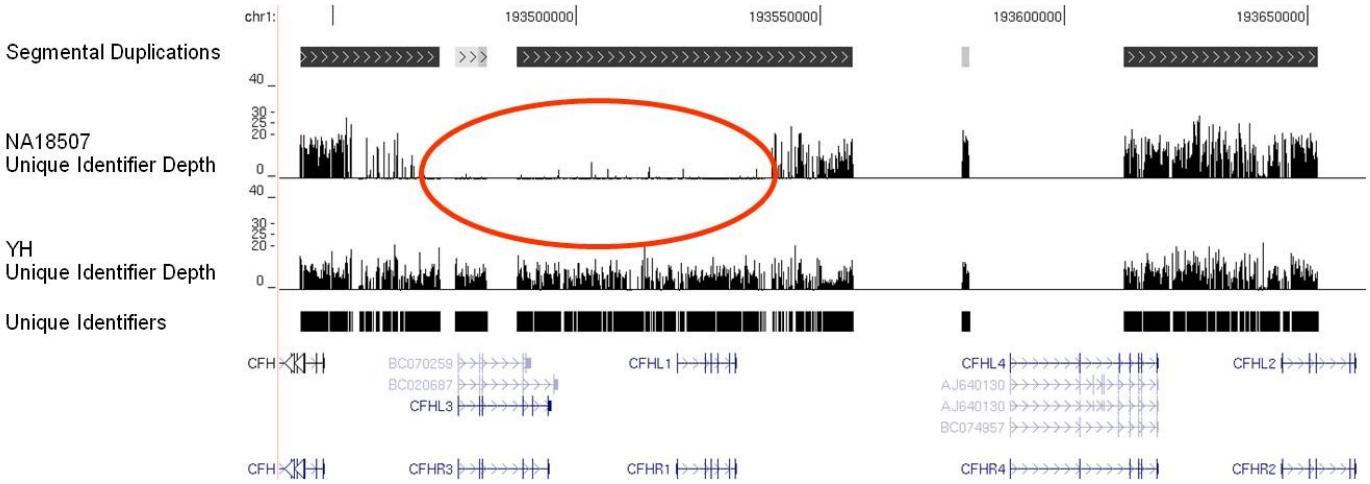


Alkan et al., Nat Genet, 2009

Sequence coverage and detection power

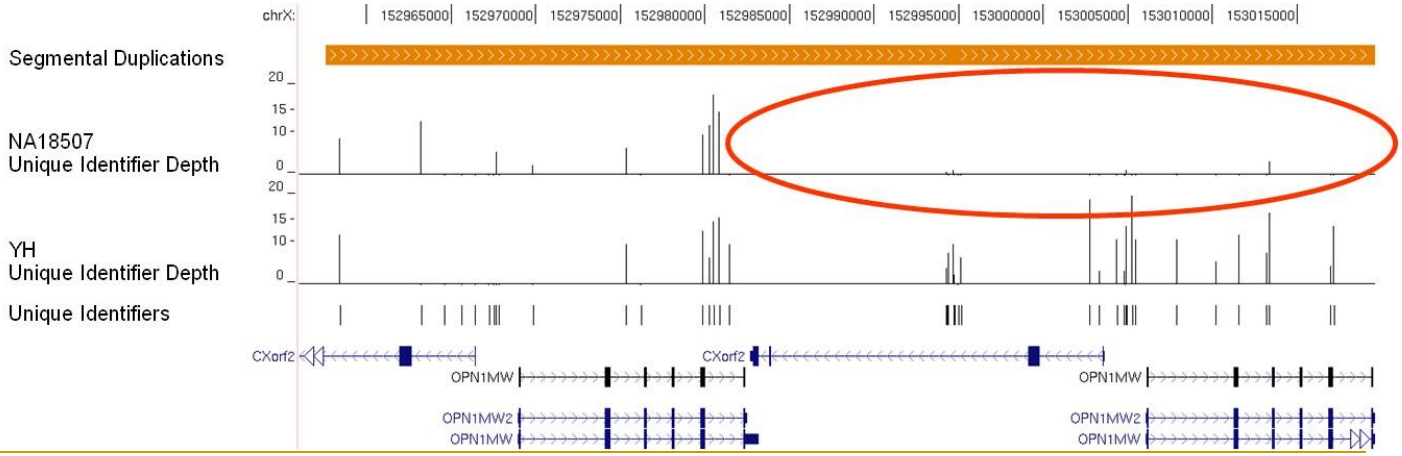


Differentiating Paralogous Genes



Associated with psoriasis an Crohn's disease

CFHR



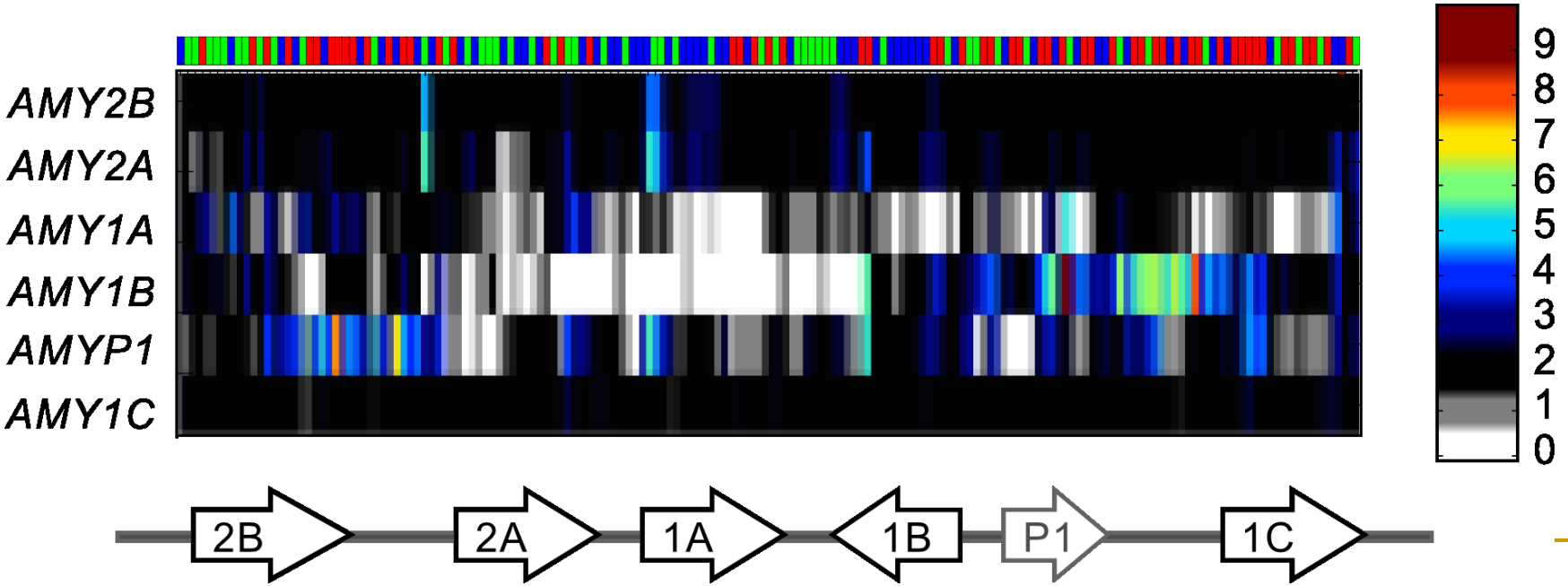
Associated with color blindness

opsin

Alkan et al., Nature Genetics, 2009

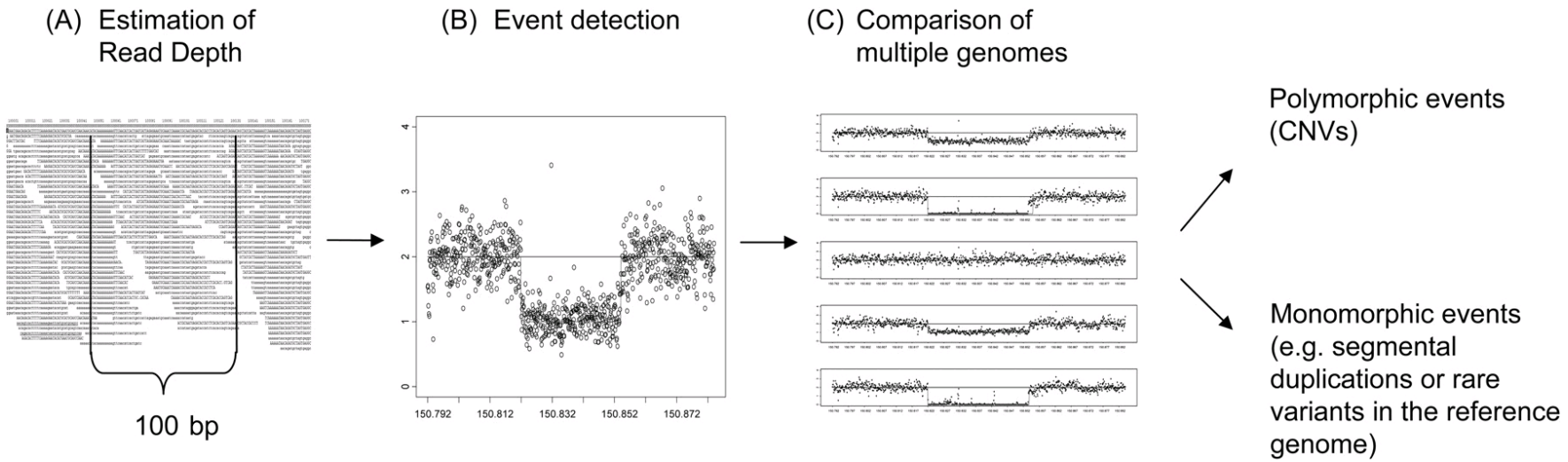
Singly Unique Identifiers (SUNs)

```
Copy 1  ATACTAGGCATATAATATCCGACGATATACATATAGATGTTAG
Copy 2  ATGCTAGGCATGTAATATCCGACGACATACATATACATGTTAG
Copy 3  ATACTAGGCATATAACATCCGACGATATACATATACATGTTAG
Copy 4  ATGCTACGCATATAATATCCCACGATATACATATACATGTTAG
Copy 5  ATGCTACGCATATAATATCCGACGATATACATATACATGATAG
Copy 6  ATACTAGGCATGTAATATCCGACGATATAC- -ATACATGTTAG
```



Sudmant et al., Science, 2010

Event-Wise Testing (EWT)



- Unique mappings are used
- No masking
- Window size 100 bp
- Probabilistic analysis

Yoon et al. Genome Research, 2009

Event-Wise Testing (EWT)

- Read counts are converted to Z score:
  - $z_i = (RC_i - \mu_i) / \sigma_i$
- Upper and lower tail probabilities
  - $p_i^U = P(Z > z_i)$
  - $p_i^L = P(Z < z_i)$
- Unusual events for interval *A*, *I* = |*A*|; *L* number of windows in chromosome; FPR: false positive rate

$$\max\{ p_i^U \mid i \in \quad \} < \left(\frac{FPR}{L/l} \right)^1$$

Duplication

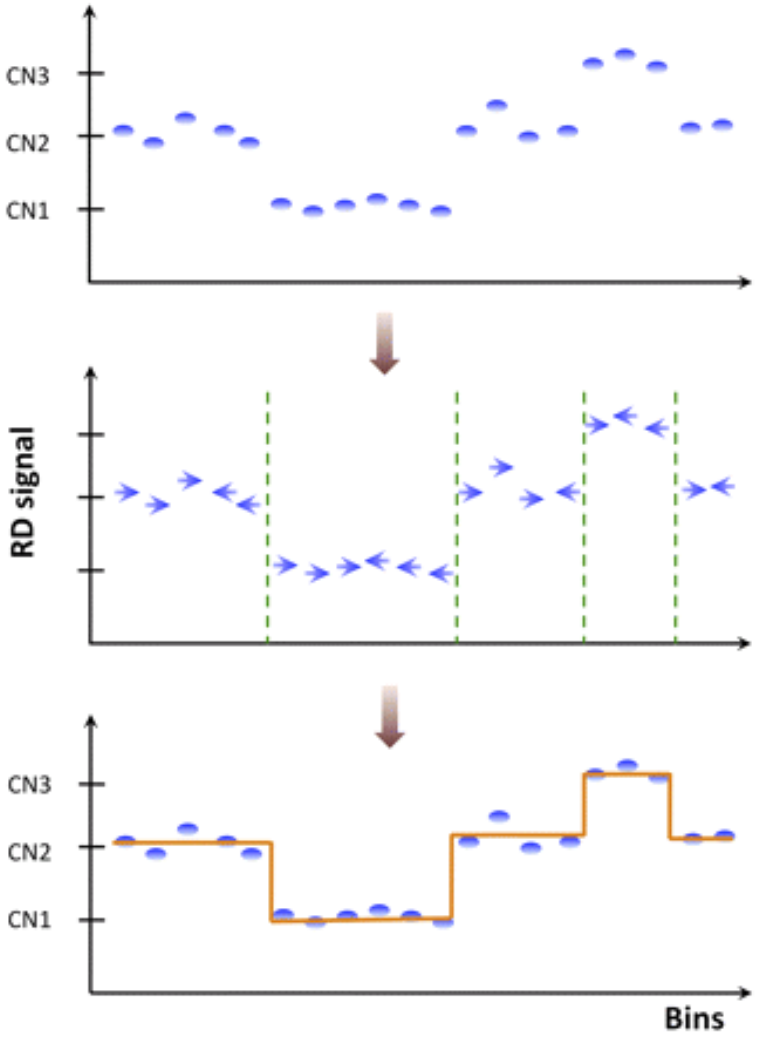
$$\max\{ p_i^L \mid i \in \quad \} < \left(\frac{FPR}{L/l} \right)^1$$

Deletion

Yoon et al. Genome Research, 2009

CNVnator

- Unique mappings
- Mappings with low MAPQ are discarded
- Partitioning is based on mean-shift technique developed for image processing



Abyzov et al. Genome Research, 2011

CNVs with exome sequencing

- Exome sequencing: capture only coding exons from DNA and sequence
  - 1% of total genome
  - Good for protein coding variants but misses regulatory sequence, introns, etc.
- Whole genome sequencing generates random data, but exome does not
- Capture efficiency changes for *every* exon (n~200,000)
- CNVs from exons: ExomeCNV

Open problems (read depth)

- Deletions are the most studied, but still not perfect:
  - Many FPs and FNs
  - Breakpoint resolution is often poor
  - Different algorithms capture different CNVs
  - Overlap with other experimental methods is poor
- Duplications are studied in lesser detail
- Exome read depth analysis
  - Very poor results due to differences in capture efficiency

NEXT: READ PAIRS + SPLIT READS