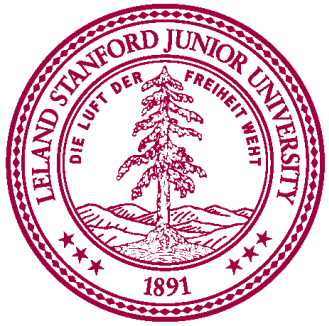


Algorithmic developments in genome structural variation detection

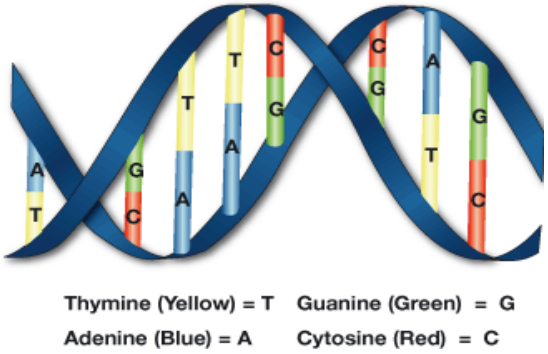
Iman Hajirasouliha

Postdoctoral Scholar, Batzoglou Lab

www.imanh.org @hajirsaouliha imanh@stanford.edu

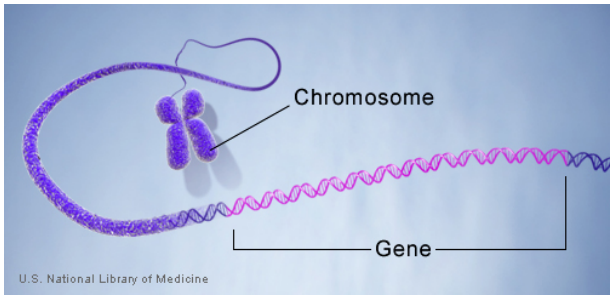
The human genome

The human genome contains 3 billion letters (bases) from {A, C, G, T}.



Divided into 23 Chromosome pairs.

Chromosomes contain genes.



The Human Genome Project made a reference human genome available (2003).

Human genome variation

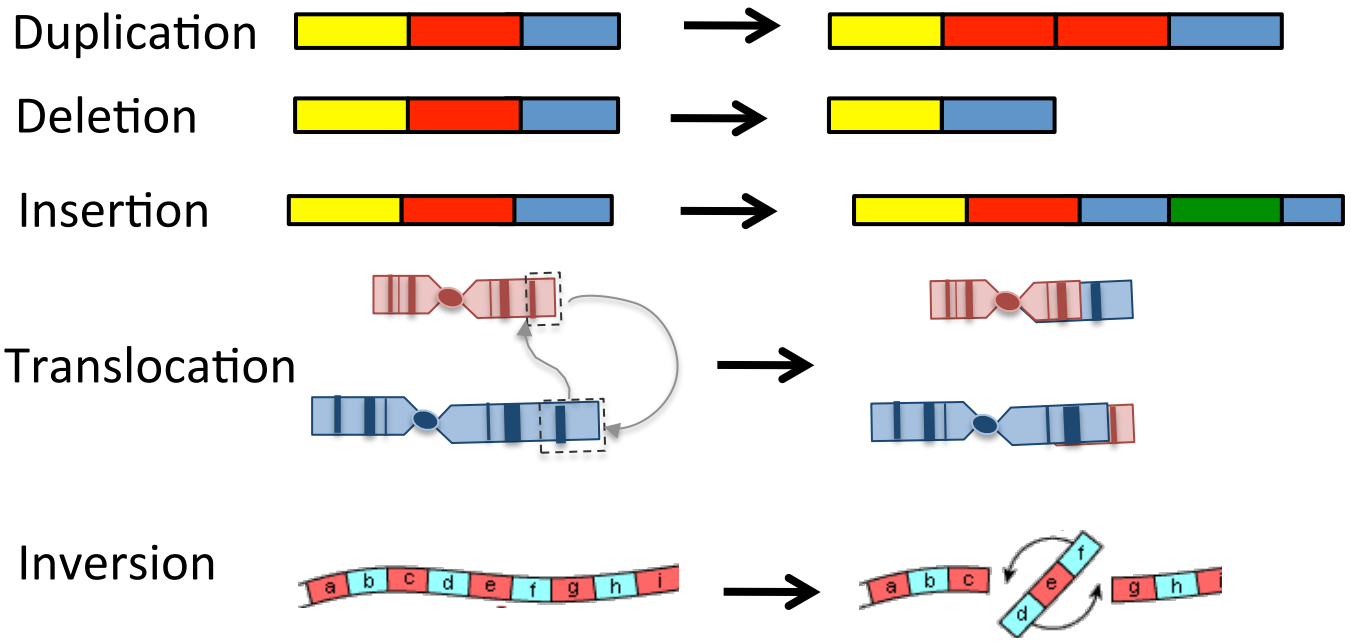
Single Nucleotide Variant (SNV)

…ACCT**G**GTA… → …ACCT**C**GTA…

Small insertions or deletions (Indels)

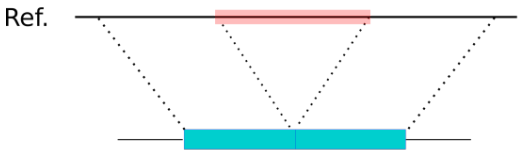
…ACCTGTA… → …ACCT**CT**GTA…

Structural Variations



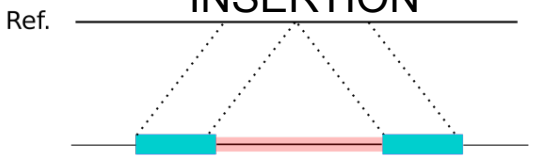
Structural Variation Classes

DELETION

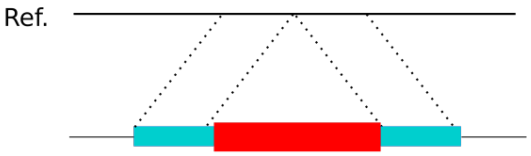


Autism, mental retardation, Crohn's

NOVEL SEQUENCE INSERTION



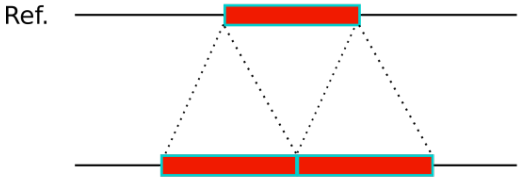
MOBILE ELEMENT INSERTION



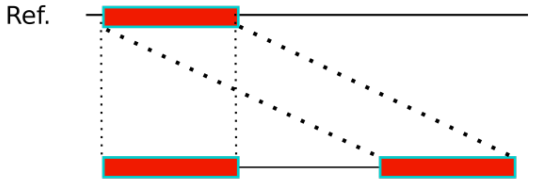
Alu/L1/SVA

Haemophilia

TANDEM DUPLICATION



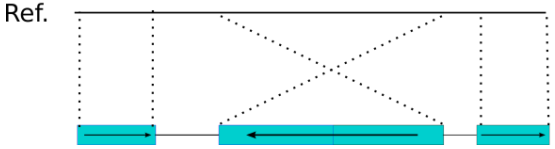
INTERSPERSED DUPLICATION



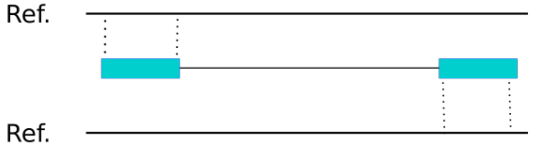
Schizophrenia, psoriasis

CNV: Copy number variants

INVERSION



TRANSLOCATION

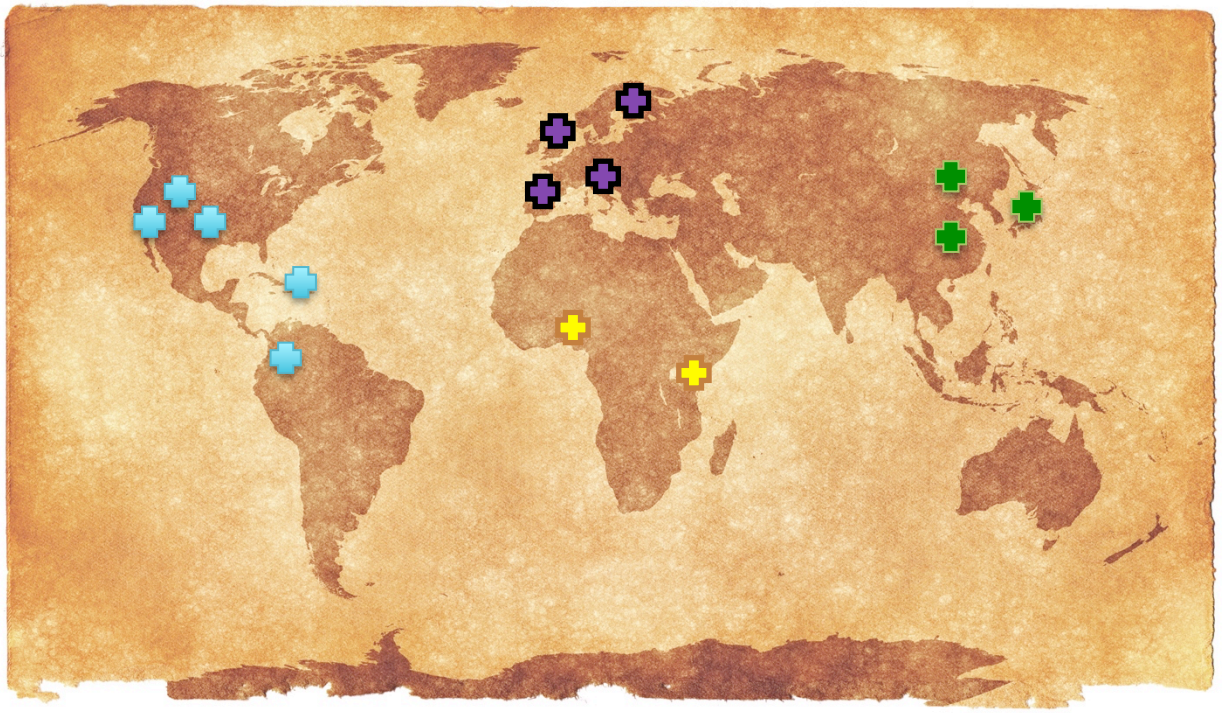


Chronic myelogenous leukemia

Balanced rearrangements

Analyzing 1092 genomes and counting

[1000 Genomes Project (GP), *Nature* 2010, *Nature* 2011, *Nature* 2012]



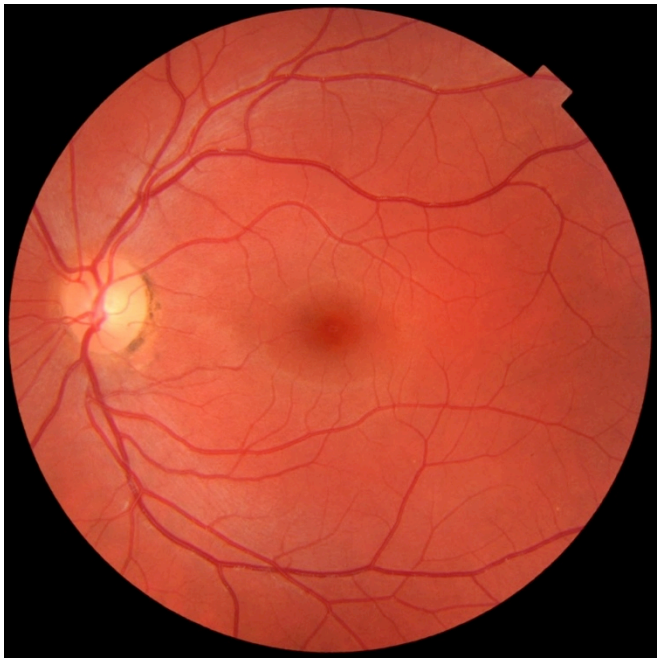
38 million SNVs

1.4 million small insertions and deletions

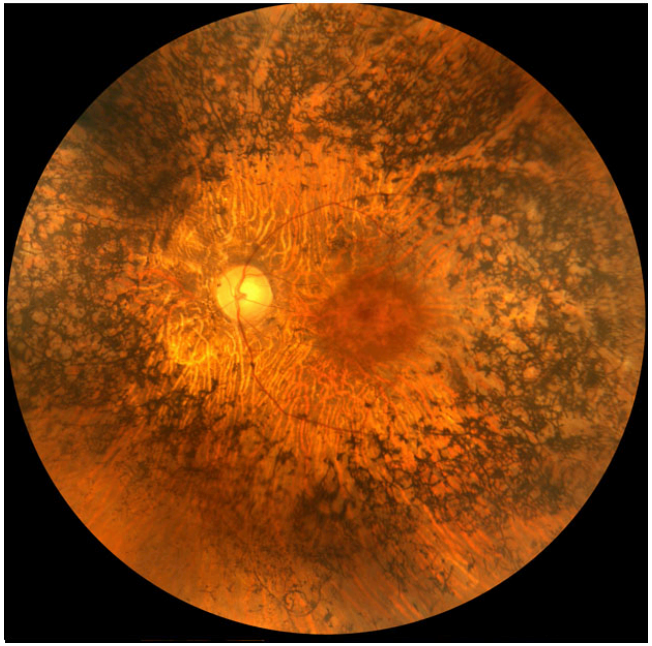
20,000 larger structural variations

Including 14,000 deletions

Structural Variations and Diseases

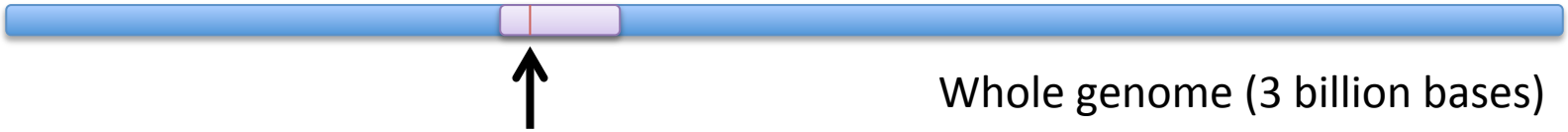


Fundus photo of a normal left eye



Patient with RP (retinitis pigmentosa)

Disease caused by an insertion of a 353 bases sequence in MAK gene.
(Tucker et al. 2011)



Next Generation Sequencing (NGS)

BIG amount of sequencing DATA

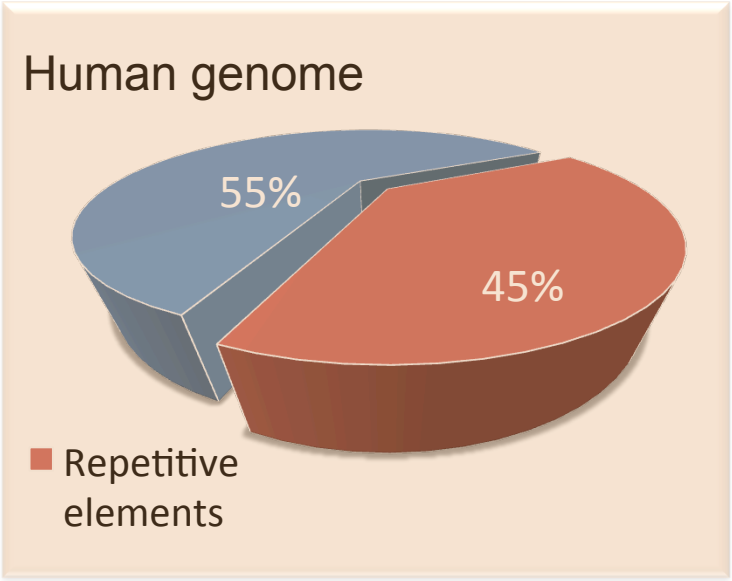
Terabyte per day for Illumina/HiSeq 2500

Fast and cheap!

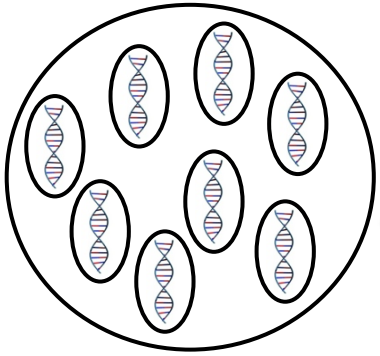
Limitations of NGS technologies

Short reads (e.g. 35bp to 150bp)

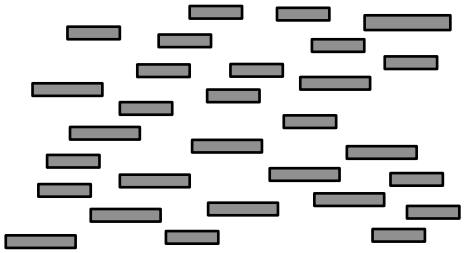
The human genome is repetitive!



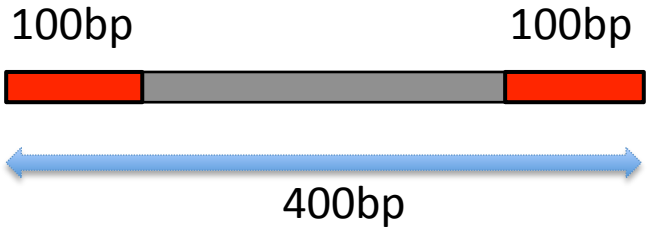
Paired-end Sequencing



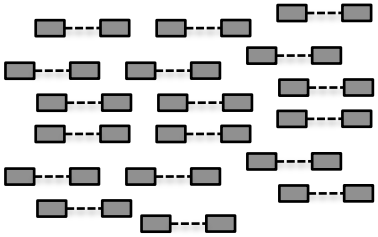
Cells



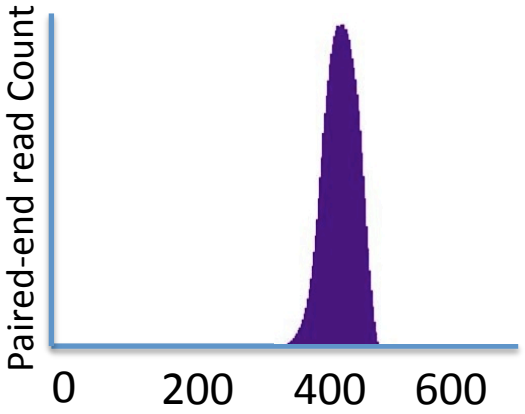
DNA is fragmented



Paired-end sequences of a fragment

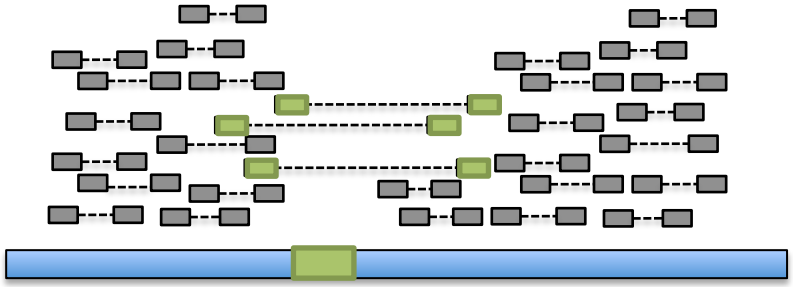


Paired-end reads



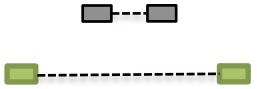
Insert size distribution

Paired-end Sequencing and SV

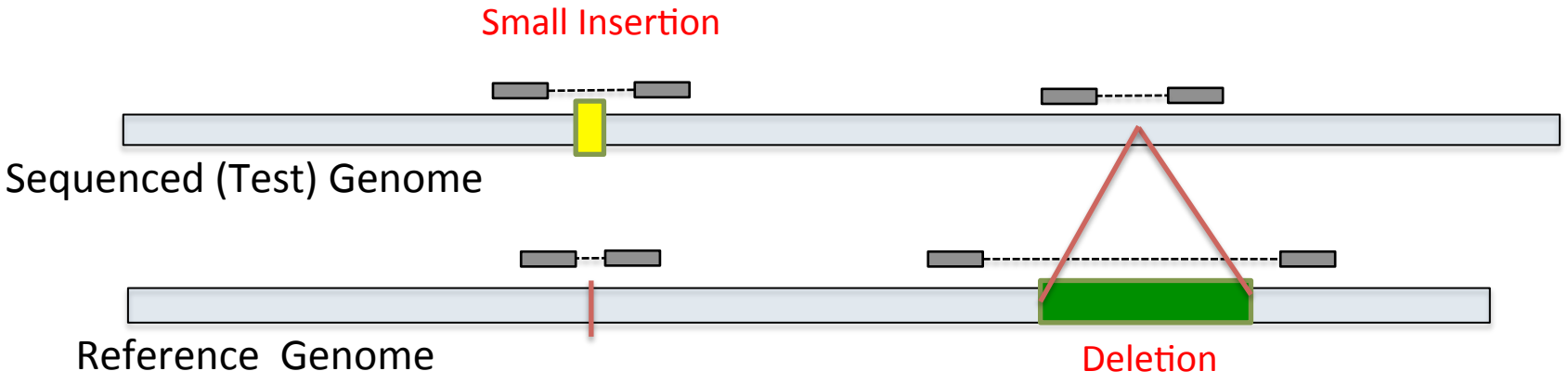


Concordant alignment

Discordant alignment



Paired-end reads are aligned to the reference

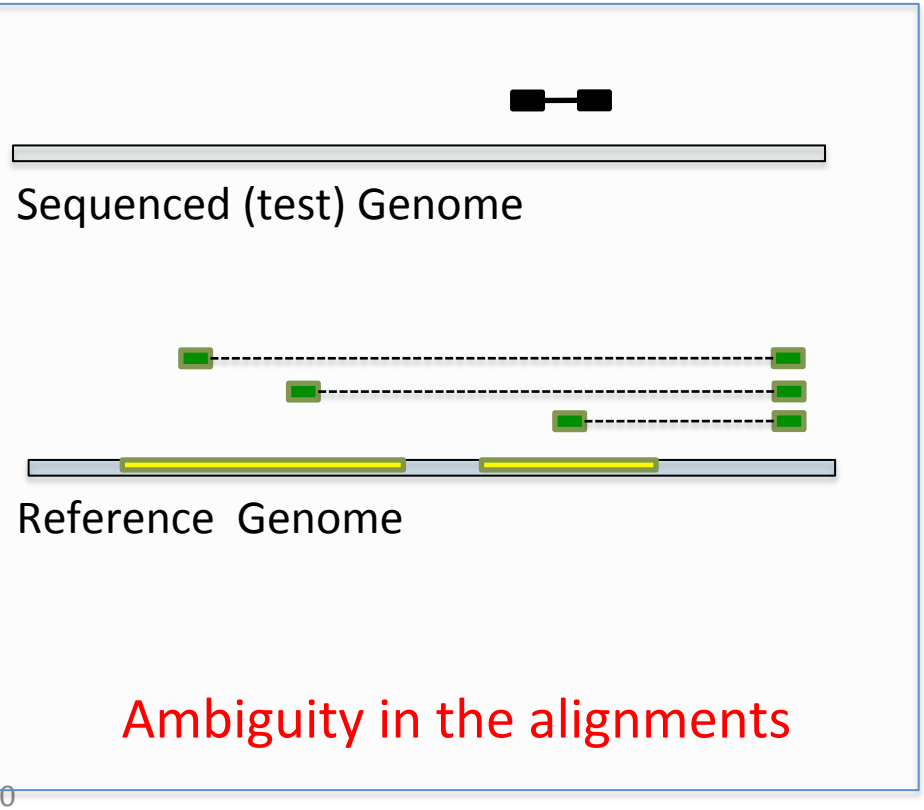


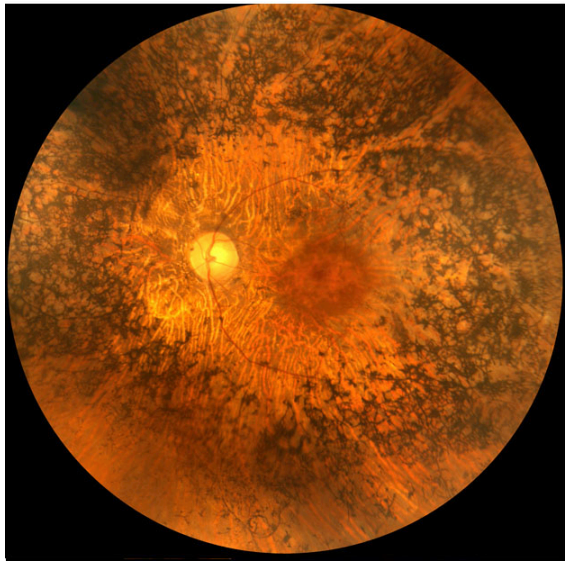
Korbel et al. 2007, Kidd et al. 2008, Hormozdiari et al. 2009

Why is it challenging to discover SVs?

Repeats have always caused problems!

Especially in next-generation sequencing projects.





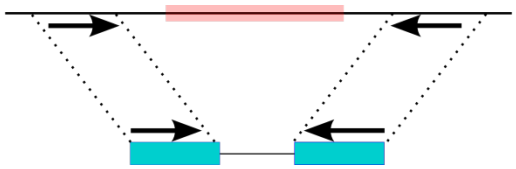
The software used for aligning short reads to the genome trimmed off repeat sequences and MAK initially appeared normal. [Tucker et al. 2011]

Why is it challenging to discover SVs?

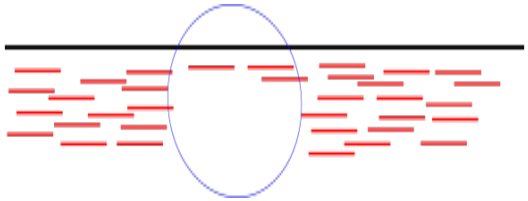
- Most SVs are embedded within or around segmental duplications or long repeats
  - If you use unique mapping, you will lose sensitivity
  - Ambiguous mapping of reads will increase false positives
  - Reference genome is incomplete; missing portions are duplications which cause more problems in accurate 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

Sequence signatures of structural variation

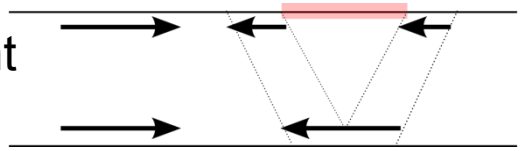
- Read pair analysis
  - Deletions, small novel insertions, inversions, transposons
  - Size and breakpoint resolution dependent to insert size



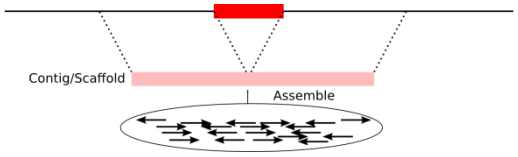
- Read depth analysis
  - Deletions and duplications only
  - Relatively poor breakpoint resolution



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



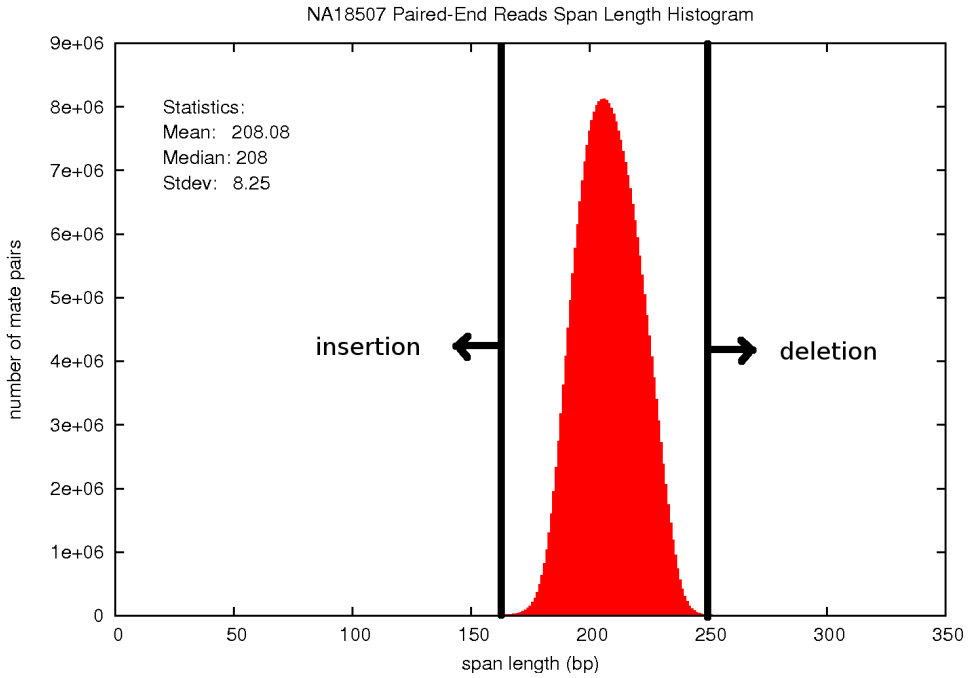
- Local and *de novo* assembly
  - SV in unique segments
  - 1bp breakpoint resolution



READ PAIR ANALYSIS

Span size distribution

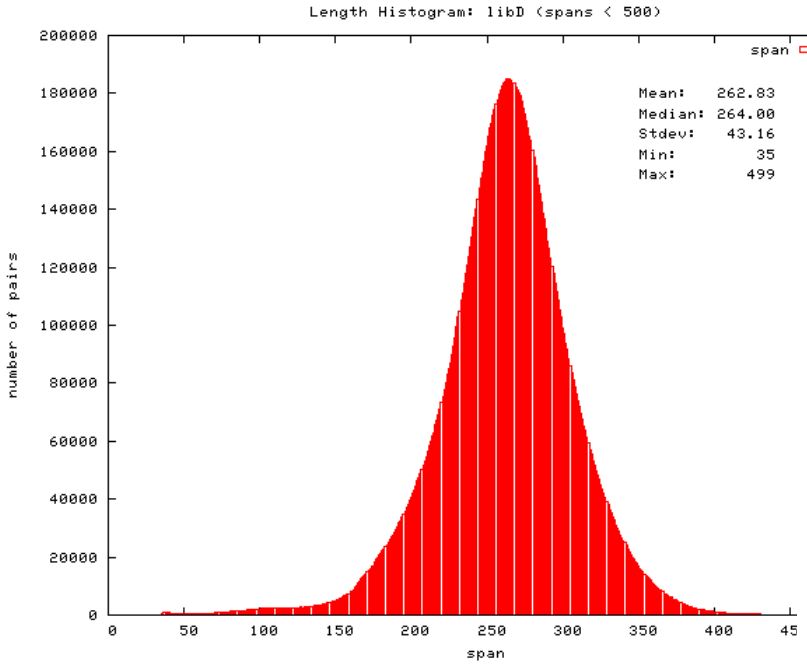
Key Idea: Observing discordances in paired-end alignments.

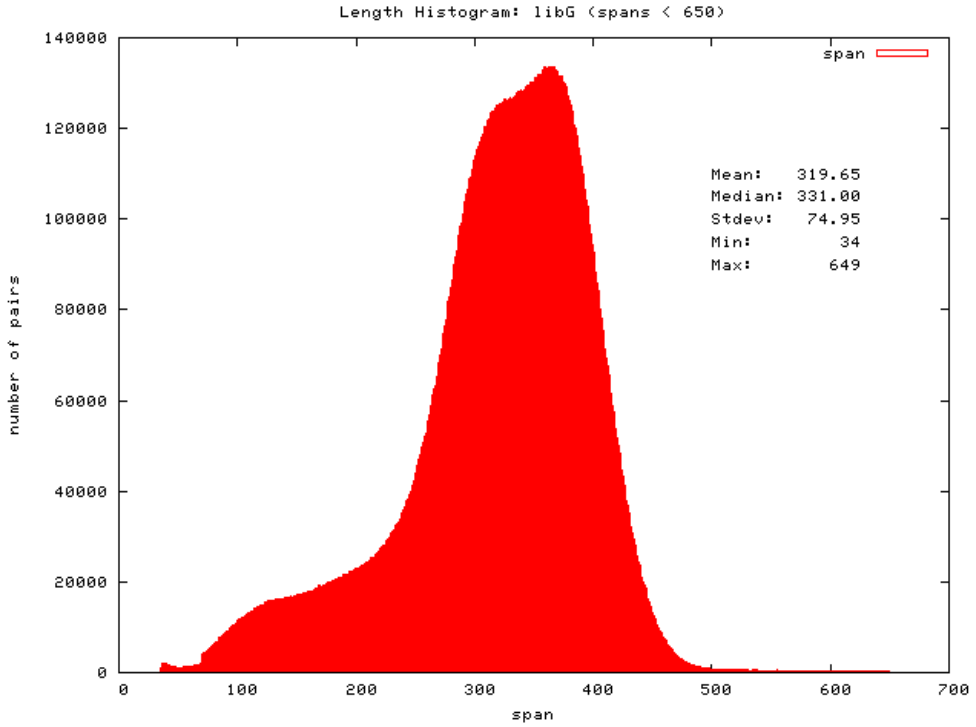


Span size = fragment length = insert size

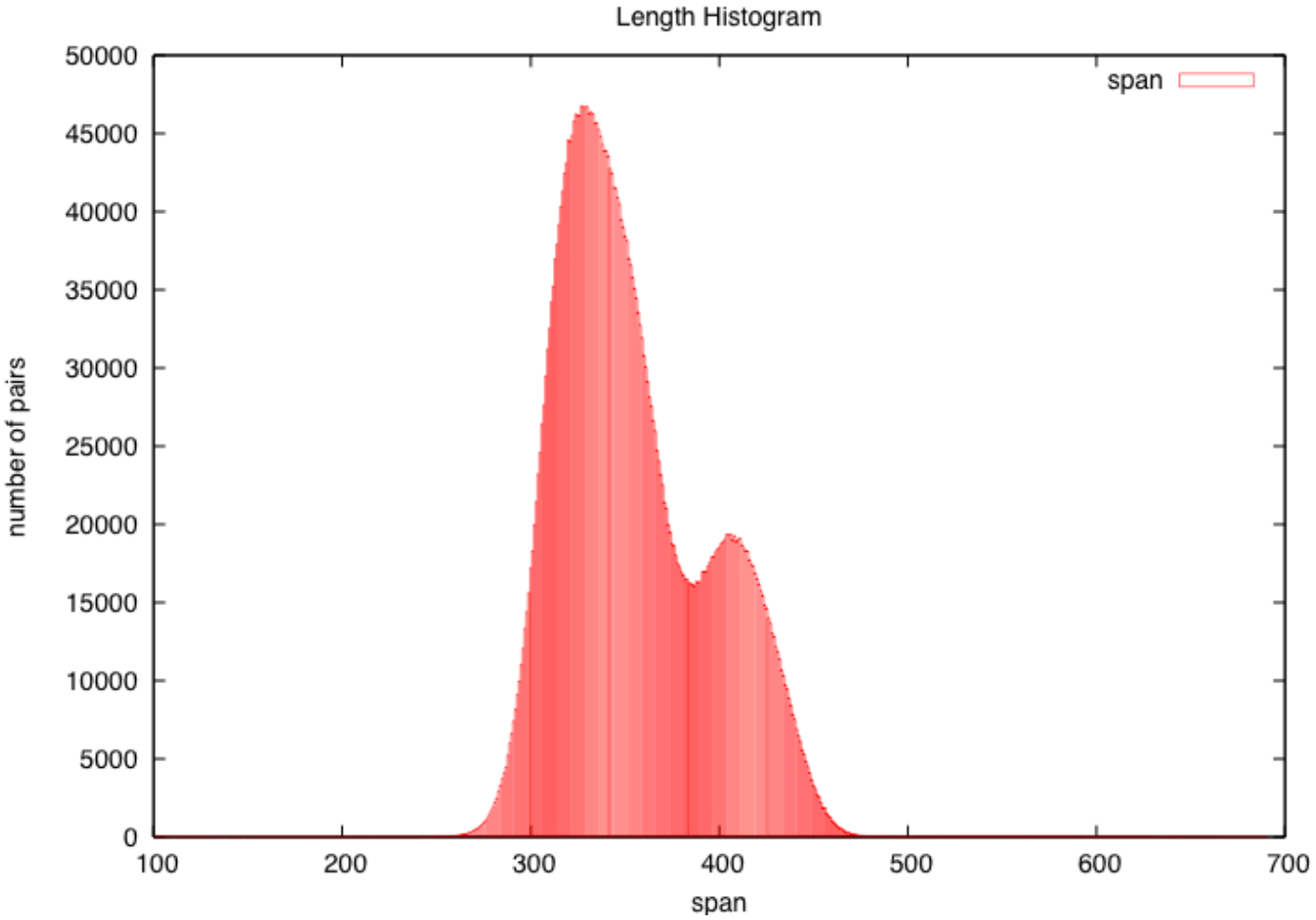
Concordant = read pairs that map in expected orientation & size
Discordant = read pairs that map different than what is expected

Span size distribution: not-so-good

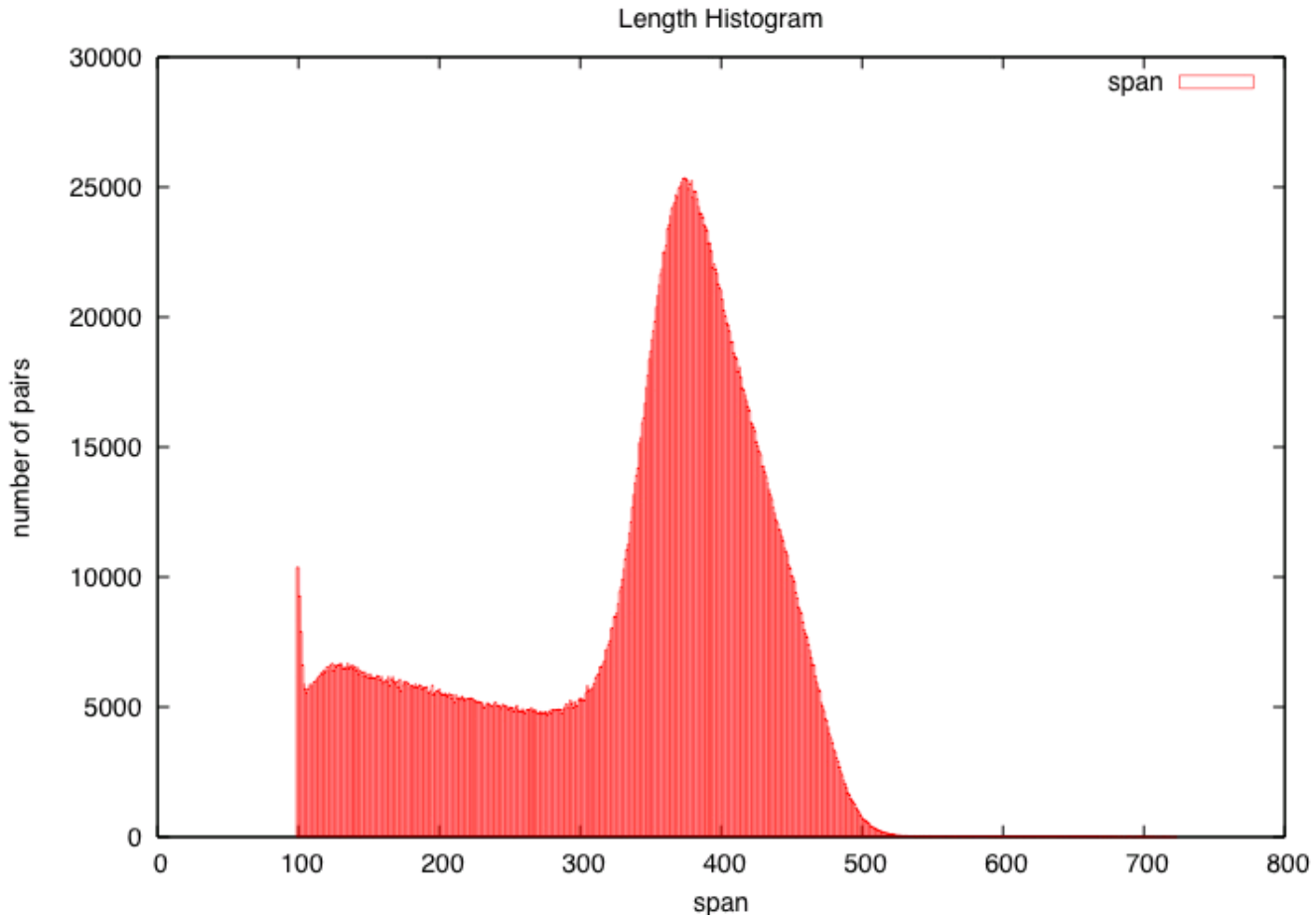




Span size distribution: bad

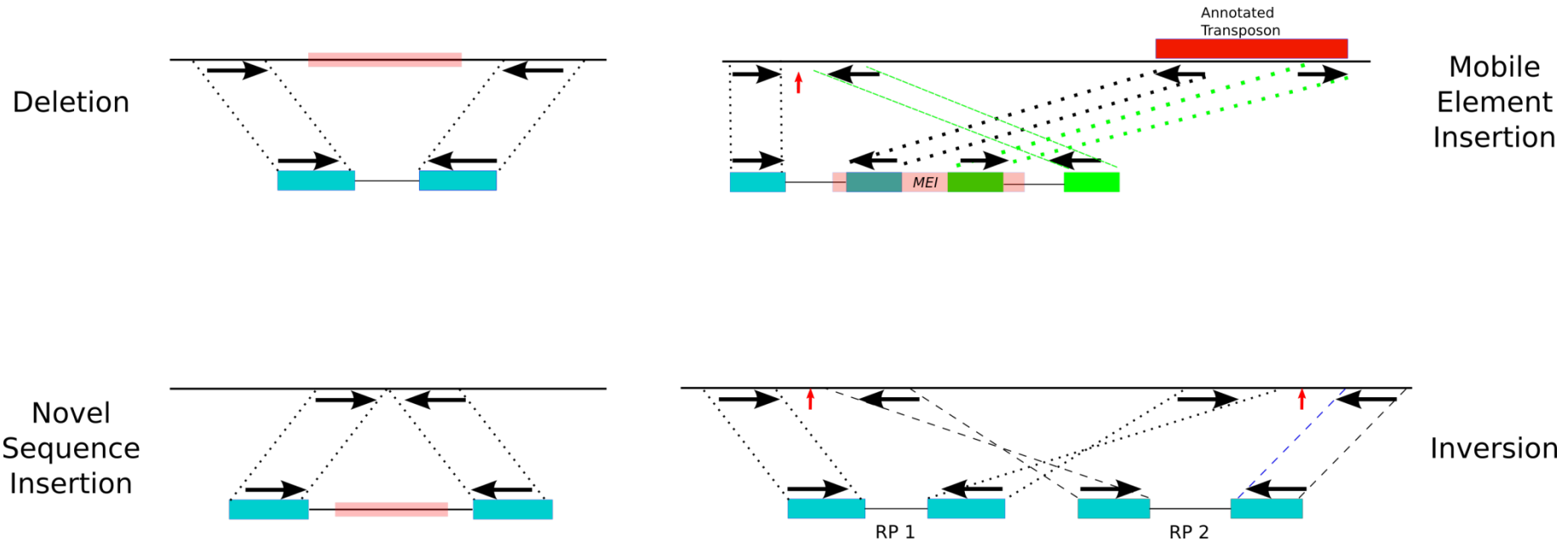


Span size distribution: bad



VariationHunter

- **VariationHunter: Maximum parsimony approach; using all discordant map locations; finds an optimal set of SVs through a combinatorial algorithm based on *set-cover***

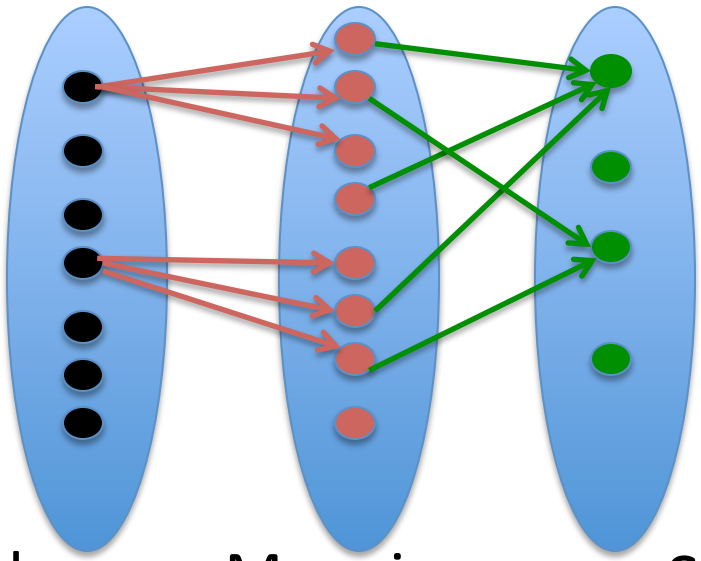


Hormozdiari, Alkan, et al, Genome Res. 2009

Maximum Parsimony SV Detection

(Hormozdiari et al 2009, Ritz et al 2010)

Objective: To pick minimum number of SV, to cover all the paired-end reads

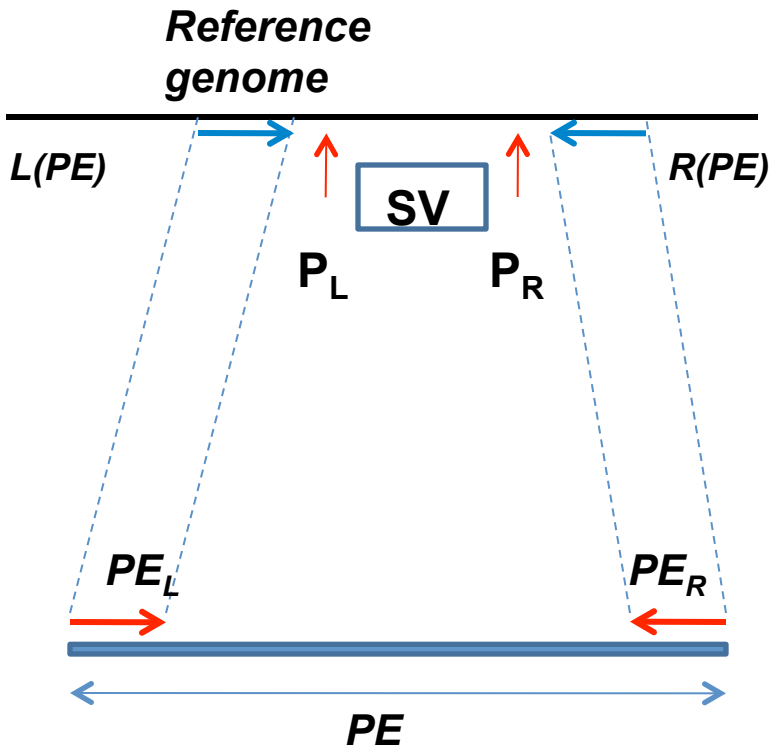


Valid Cluster: Every two or more discordant paired-end read alignment that can support the same potential SV.

Maximal Valid Clusters: number of valid clusters is exponential, while maximal valid cluster is polynomial (Hormozdiari et al 2009, Sindi et al 2009)

Set-Cover Approach: Having the maximal valid clusters, we can find maximum parsimony solution with approximation of *O(log n)*

Definitions



Paired-end read

PE:= (PE_L, PE_R)

PE-Alignment

(PE, L(PE), R(PE), O(PE))

O(PE): mapping orientation:

- "+/-": normal
- "+/+" or "-/-": inversion
- "-/+": tandem duplication

SV = (P_L, P_R, L_{min}, L_{max})

Hormozdiari, Alkan, et al, Genome Res. 2009

Mathematical model

Let L_{min}, L_{max} be *minimum* and *maximum* size of the predicted variant

A Structural Variation is defined by event:

$$SV = (P_L, P_R, L_{min}, L_{max})$$

A PE-Alignment APE=(PE, L(PE), R(PE), O(PE)) supports an insertion $SV = (P_L, P_R, L_{min}, L_{max})$ if:

$$L(PE) \leq P_L$$

$$R(PE) \geq P_R$$

$$L_{min} \geq \Delta_{min} - (R(PE) - L(PE))$$

$$L_{max} \leq \Delta_{max} - (R(PE) - L(PE))$$

Hormozdiari, Alkan, et al, Genome Res. 2009

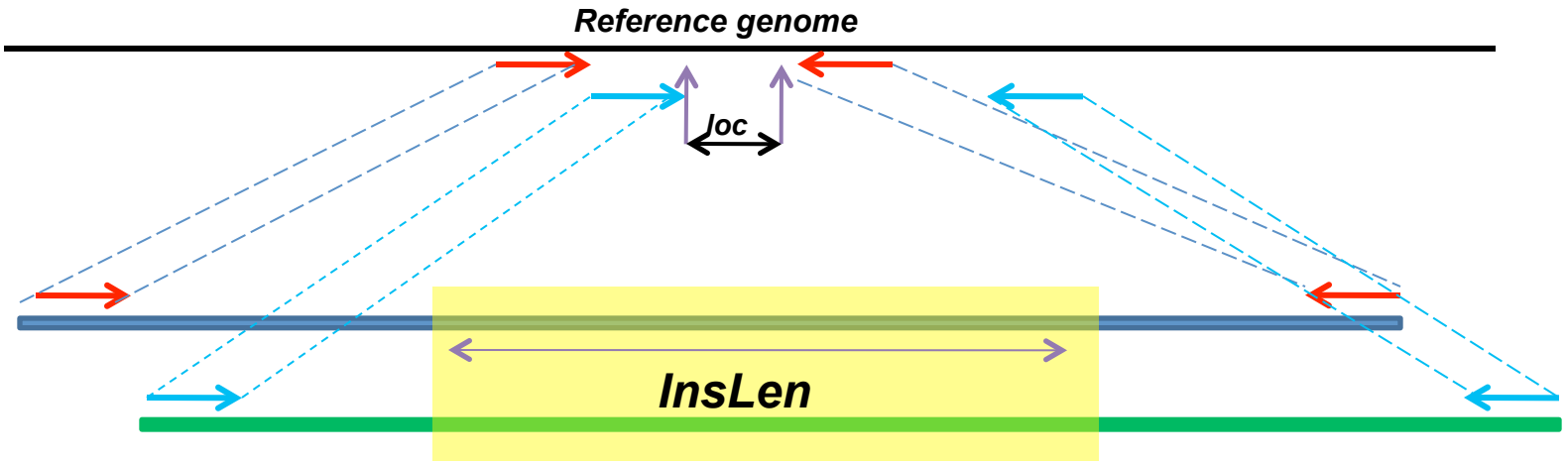
Valid clusters

A set of PE-Alignments that support the same structural variation event SV

A cluster C is a *valid cluster* supporting insertions if:

$$\exists loc, \forall APE \in C : L(APE) < loc < R(APE)$$

$$\exists InsLen, \forall APE \in C : \Delta_{\min} - (R(APE) - L(APE)) < InsLen < \Delta_{\max} - (R(APE) - L(APE))$$



Hormozdiari, Alkan, et al, Genome Res. 2009

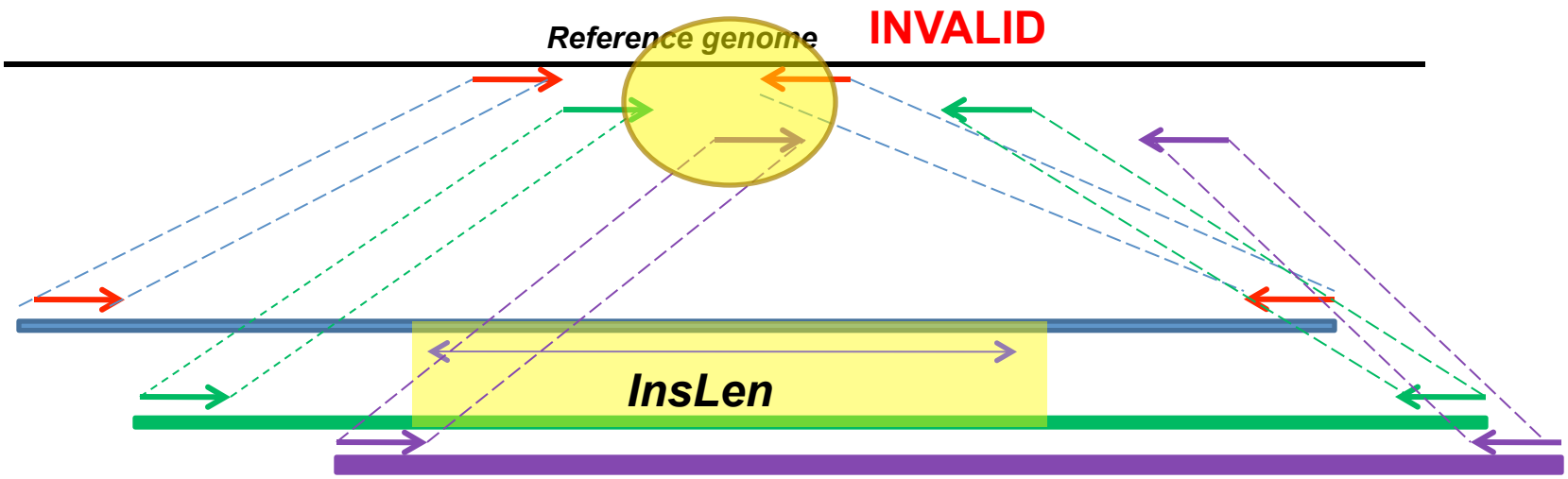
Valid clusters

A set of PE-Alignments that support the same structural variation event SV

A cluster C is a *valid cluster* supporting insertions if:

$$\exists loc, \forall APE \in C : L(APE) < loc < R(APE)$$

$$\exists InsLen, \forall APE \in C : \Delta_{\min} - R(APE) + L(APE) < InsLen < \Delta_{\max} - R(APE) + L(APE)$$



Hormozdiari, Alkan, et al, Genome Res. 2009

Maximal Valid Clusters for Insertions

A *Maximal Valid Cluster* is a valid cluster that no additional APE can be added without violating the validity of the cluster

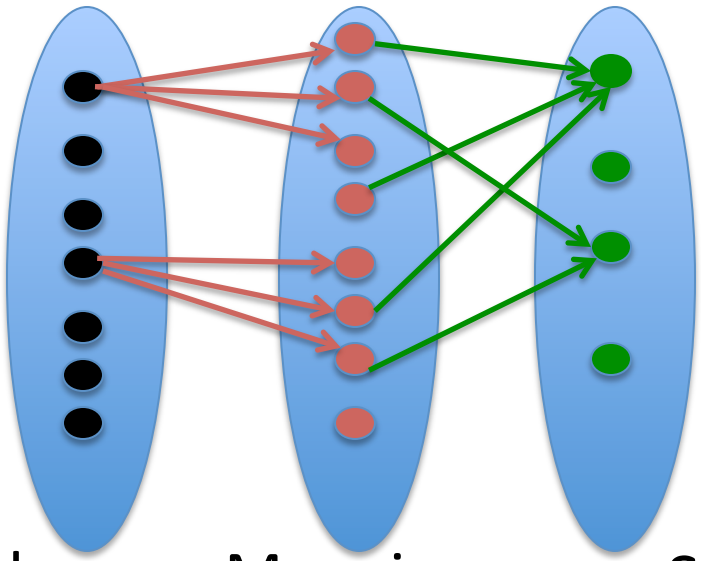
1. Find all the Maximal sets of overlapping paired-end alignments
2. For each maximal set S_k found in Step 1, find all the maximal subsets s_i in S_k that the insertion size (*InsLen*) they suggest is overlapping
3. Among all the sets s_i found in Step 2, remove any set which is a proper subset of another chosen set

Hormozdiari, Alkan, et al, Genome Res. 2009

Maximum Parsimony SV Detection

(Hormozdiari et al 2009, Ritz et al 2010)

Objective: To pick minimum number of SV, to cover all the paired-end reads

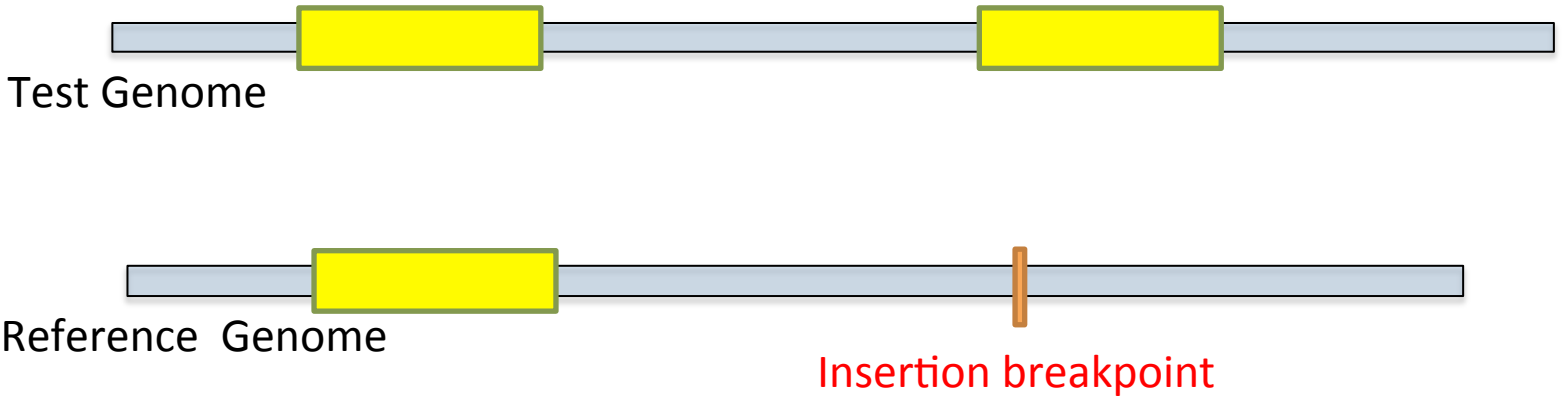


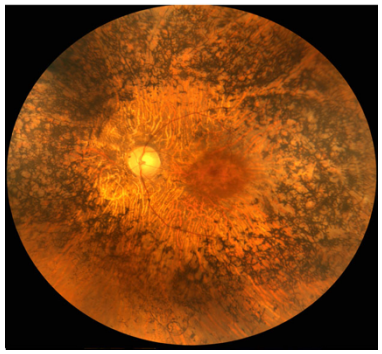
Valid Cluster: Every two or more discordant paired-end read alignment that can support the same potential SV.

Maximal Valid Clusters: number of valid clusters is exponential, while maximal valid cluster is polynomial (Hormozdiari et al 2009, Sindi et al 2009)

Set-Cover Approach: Having the maximal valid clusters, we can find maximum parsimony solution with approximation of *O(log n)*

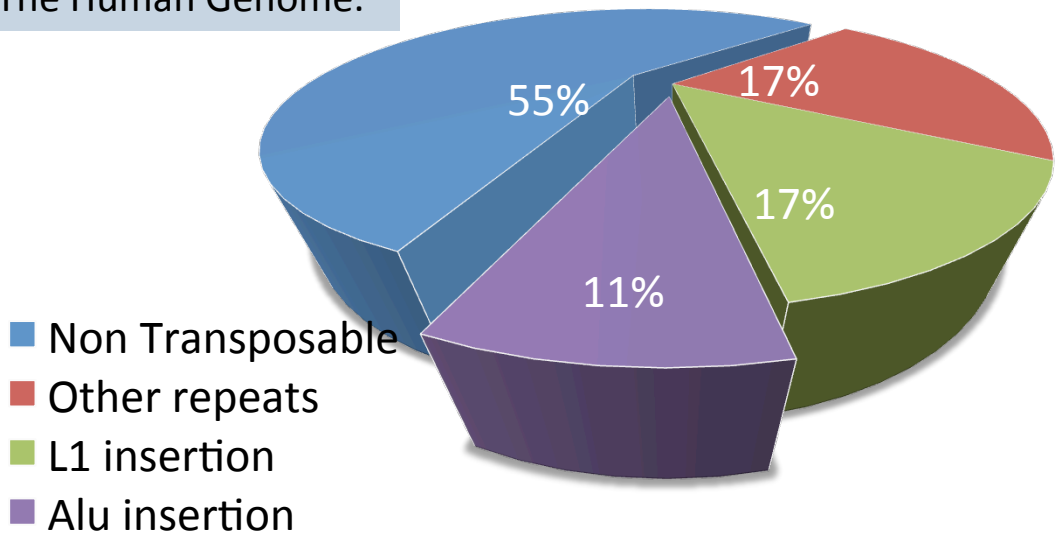
Transposon insertions (transposable elements)



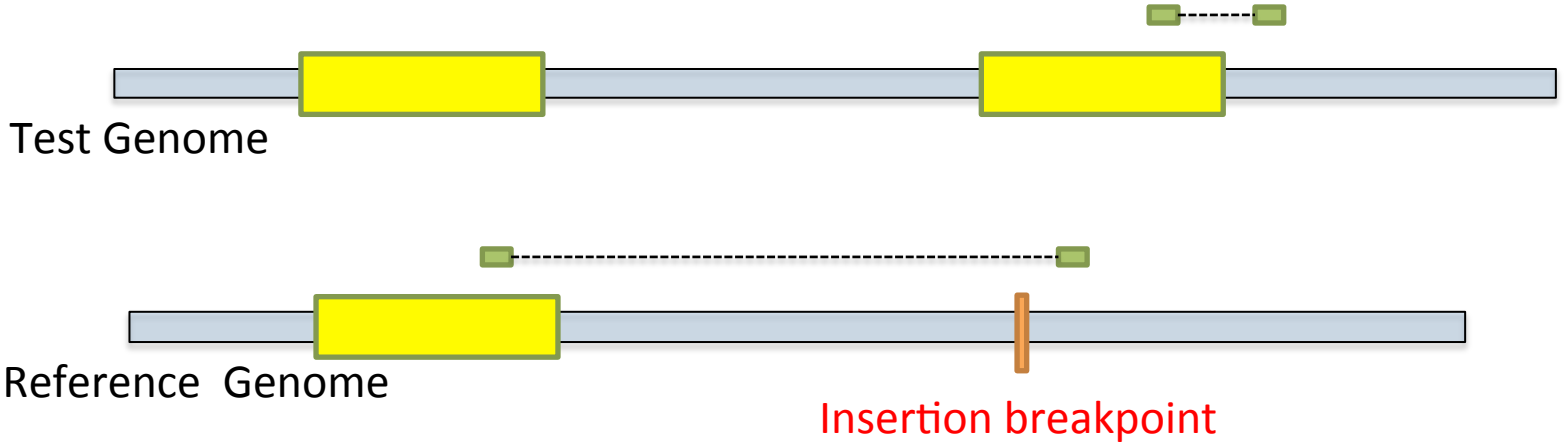


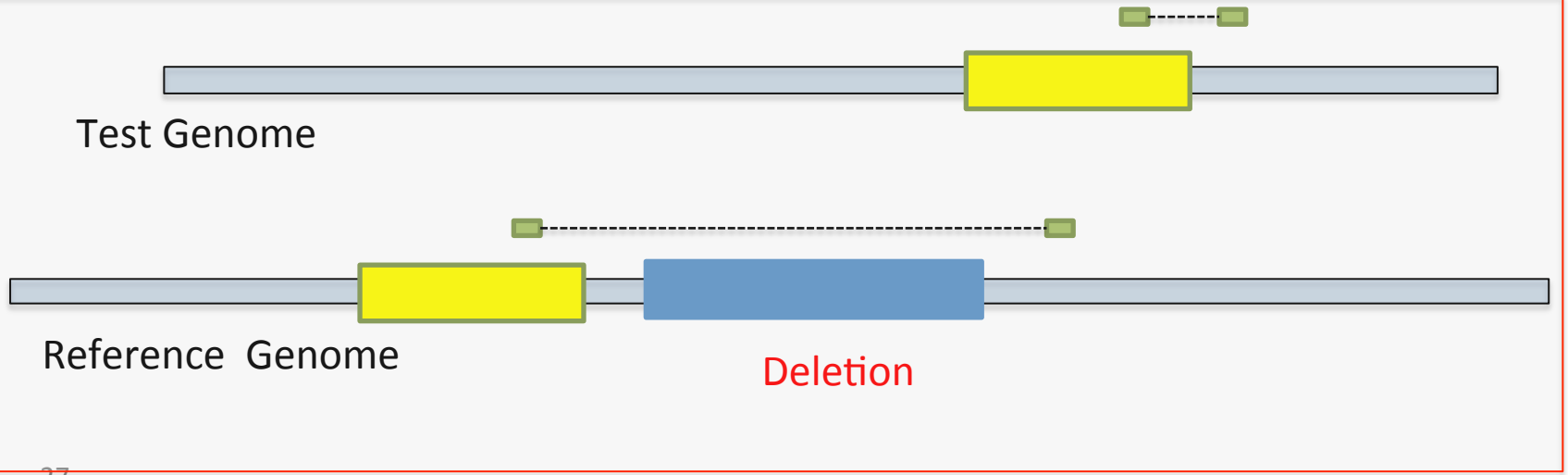
Note: Insertion of an Alu in MAK caused RP.

The Human Genome.

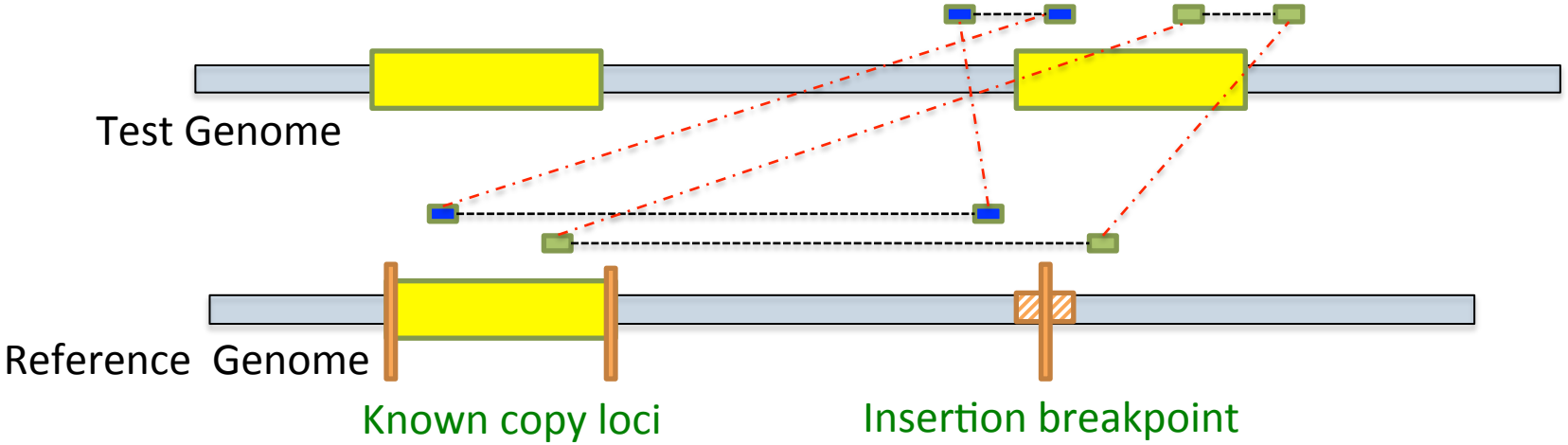


Transposon insertions



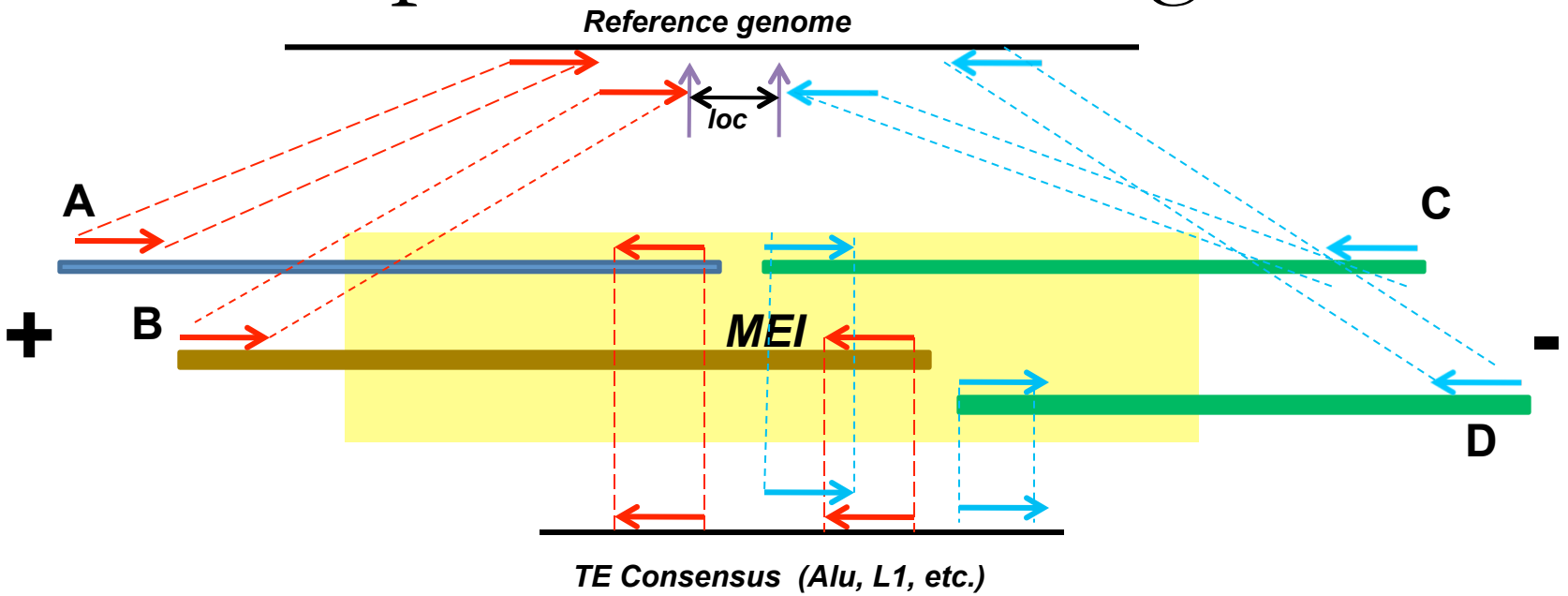


How to cluster transposon insertions



Based on discordant alignments, the insert size, and the loci of copied element, we guess a breakpoint interval.

Transposon insertion signature

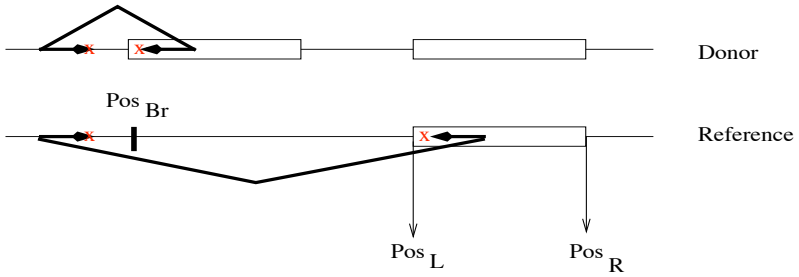


- Strand rules: MEI-mapping "+" reads and MEI mapping "-" reads should be in different orientations:
  - +/- and -/+ clusters; or +/+ and -/- clusters (inverted MEI)
- Span rules: A=(A1, A2); B=(B1, B2); C=(C1, C2); D=(D1, D2)
  - |A1-B1| ~ |A2-B2| and |C1-D1| ~ |C2-D2| (simplified; we have 8 rules)
- Location and 2-breakpoint rule:

$$\exists loc, \forall PE : RightMost(+) < loc < LeftMost(-)$$

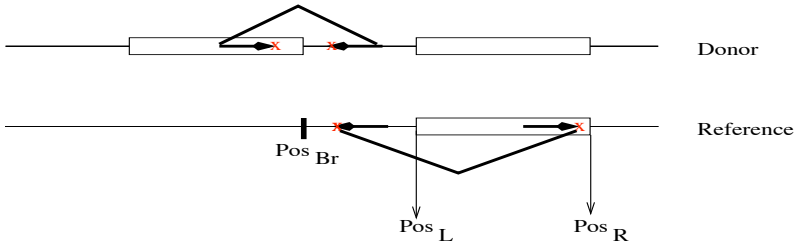
Hormozdiari et al., Bioinformatics 2010

Paired-end Reads and Transposition



$$(pos_{Br} < Pos_L \; and \quad or(ape) = +-)$$

$$\Delta_{\min} < Pos_{Br} - L_r(ape) + R_l(ape) - Pos_L < \Delta_{\max}$$



$$(pos_{Br} < Pos_L \; and \quad or(ape) = -+)$$

$$\Delta_{\min} < L_l(ape) - Pos_{Br} - R_r(ape) + Pos_r < \Delta_{\max}$$

Maximal Valid Cluster For Transposons

- Having set of all subsequences from genome which can be copied/transposed (e.g. Alu, L1, or SVA), we find all the maximal valid cluster in *O(n log n).*
- At the core of finding maximal valid clusters is finding the maximal intersection of breakpoint intervals of paired-end reads.

Maximal Intersecting Intervals

- Given a set of *m* intervals, we can find all the maximal intersecting intervals in *O(m log m).*

- Maximal intersecting intervals has a lower bound of *O(m lg m)*:
  - *Maximum clique in interval graphs has a lower bound of O(m log m)*

Problem and Solutions

Problem: Among all the maximal valid clusters, which ones are correct?
Aim: Assign a single PE-Alignment to all paired-end reads

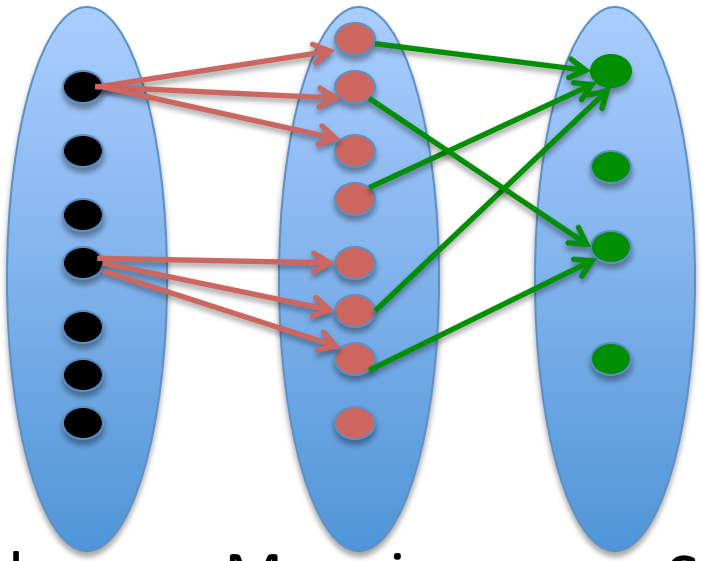
- Maximum Parsimony Structural Variation
  - Find a *minimum* number of *SVs* such that all the paired-end reads are covered
    - Similar to SET-COVER problem
    - Greedy algorithm. Approximation factor ***O(log(n))***

Hormozdiari, Alkan, et al, Genome Res. 2009

Maximum Parsimony SV Detection

(Hormozdiari et al 2009, Ritz et al 2010)

Objective: To pick minimum number of SV, to cover all the paired-end reads



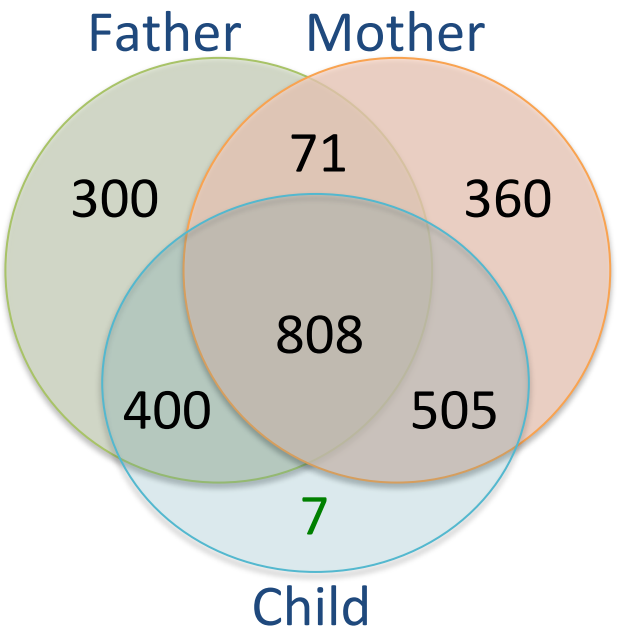
Valid Cluster: Every two or more discordant paired-end read alignment that can support the same potential SV.

Maximal Valid Clusters: number of valid clusters is exponential, while maximal valid cluster is polynomial (Hormozdiari et al 2009, Sindi et al 2009)

Set-Cover Approach: Having the maximal valid clusters, we can find maximum parsimony solution with approximation of *O(log n)*

Alu discovery within human genomes

Individual	Population	Sequence Coverage	# *Alu*
NA18506	YRI (child)	40.1x	1720
NA18507	YRI (father)	27.1x	1579
NA18508	YRI (mother)	37x	1744
NA10851	CEU	22x	1282
AK1	Korean	22.5x	909
YH	Chinese	11.4x	1160
KB1	Khoisan	21x	457
HGDP01029	Khoisan	4x	307
YRI trio			2451
Total			4342



Wet lab validation confirmed the potential de novo insertions were transmitted from one of the parents.

Handling multiple genomes

Classic way:

1. Detect SVs in individuals independently.
2. Check whether the genomes agree or disagree on SV.

New way:

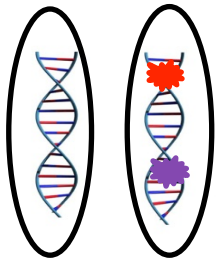
All genomes are compared with the reference simultaneously.

Simultaneous structural variation discovery in multiple paired-end sequenced genomes

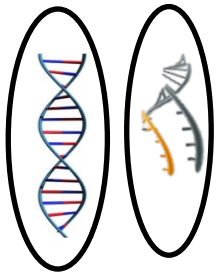
RECOMB/Genome Research 2011 *Hormozdiari, *Hajirasouliha et al.

*Joint First Authors

Genomes of closely related individuals



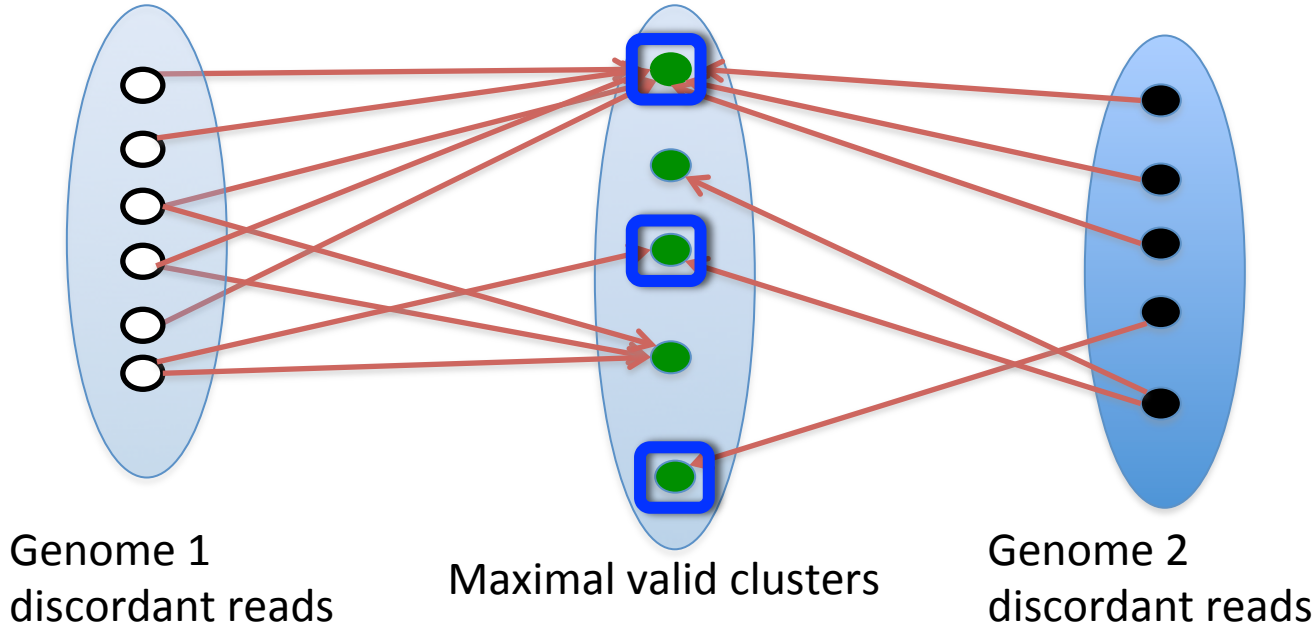
Normal/tumor pair



Genome/transcripts

An assignment problem

Generalize maximum parsimony clustering approach to multiple genomes, via a combinatorial assignment problem.



ω_{black} : Weight of clusters with only black reads

ω_{white} : Weight of clusters with only white reads

$\omega_{black\&white}$: Weight of clusters with both black and white reads

Simultaneous Structural Variation (SSV)

SSV Problem: Given a collection of discordant reads from a set of genomes, find a unique assignment of each discordant read to a maximal SV cluster such the following cost is minimized:

$$COST = \sum_{s \in S} I_s . \omega_{C(s)} \Delta_s$$

S : The set of all maximal valid clusters

s : A maximal valid cluster

I_s : An indicator variable

$\omega_{c(S)}$: Weight associated with the subset of genomes in S

Δ_s : Weight associated with cluster itself

Complexity of the SSV Problem

We proved the SSV problem is NP-hard.

There exists a constant c such that SSV cannot be approximated within

$$c\frac{\omega_{\max}}{\omega_{\min}}\log n$$

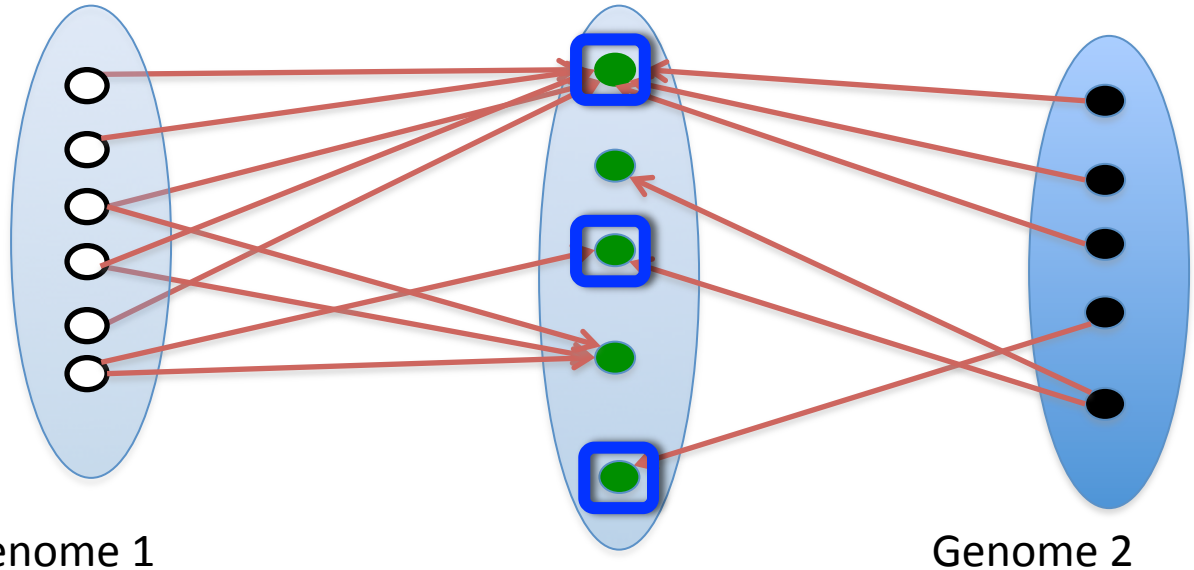
unless P = NP.

n is the total number of discordant reads. $\omega_{\max}$ and $\omega_{\min}$ are the maximum and minimum possible weights for of the SV events.

Greedy algorithm utilizing weights

Selects the SV clusters based on their "cost-effectiveness" value in each iteration.

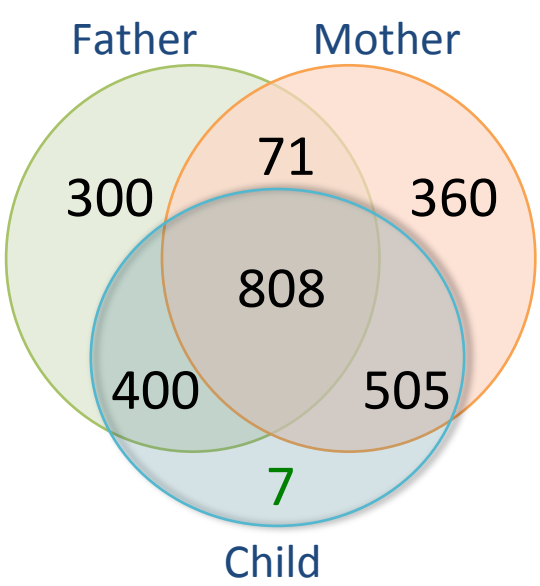
Cost-effectiveness = Weight / Number of newly covered read

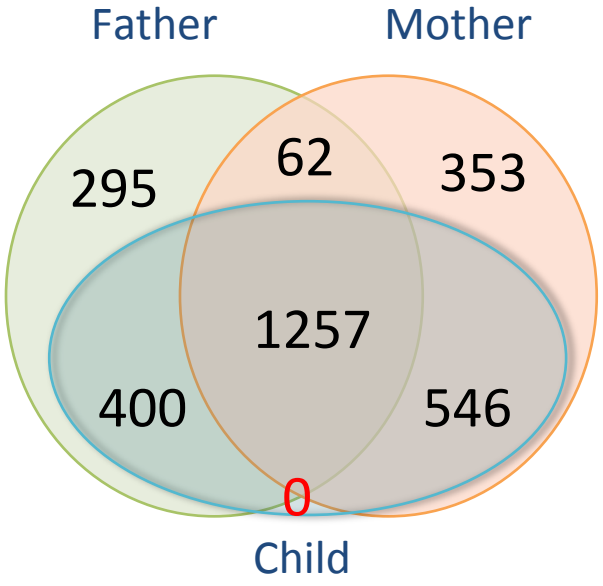


$$COST = \frac{\omega_{black\&white}}{8} + \frac{\omega_{black\&white}}{2} + \omega_{black}$$

ω: Weights associated with subsets of genomes

De novo and common *Alu* insertion analysis of the YRI trio



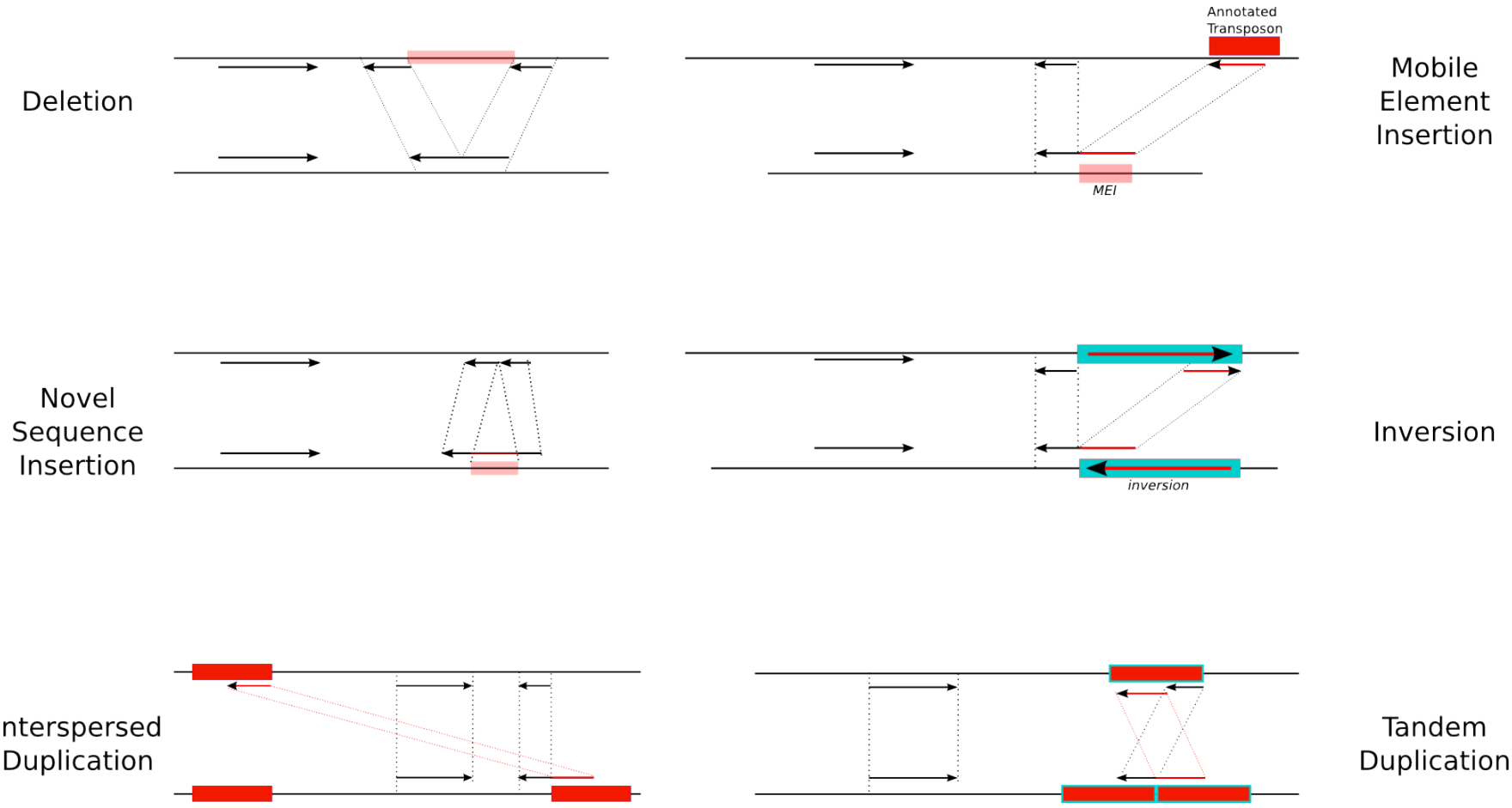


Earlier analysis of the YRI trio.

Hormozdiari*, Hajirasouliha* et al. ***RECOMB/Genome Research*** 2011
*Joint First Authors

SPLIT READ

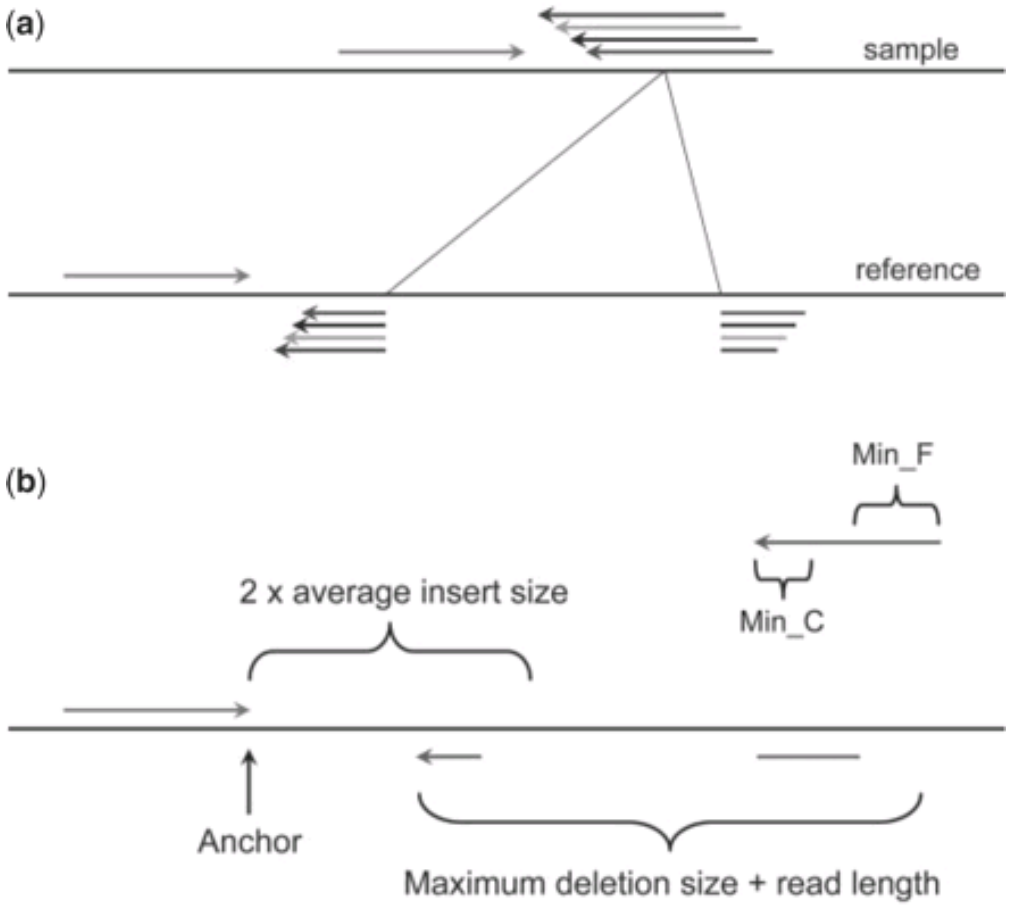
Split Read analysis



Split Read based algorithms

- Unique mapping:
  - Pindel (Ye et al. Bioinformatics, 2009)
  - SRiC (for the 454 platform; Zhang et al., BMC Bioinformatics, 2011)
- Multiple mapping:
  - SPLITREAD (Karakoc et al., Nature Methods, 2012)
- Specialized for RNA alternative splicing:
  - TopHat (Trapnell et al., Bioinformatics, 2009)

Pindel: pattern growth approach



Ye et al. Bioinformatics, 2009

Pattern growth

S = ATCAAGTATGCTTAGC
P = ATGCA

Search **A:**
ATCAAGTATGCTTAGC

Projected database of **A:**
1,4,5,8,14

Search **T** in Projected Database of **A**:
ATCAAGTATGCTTAGC

Projected database of **AT**:
1,8

Search **G** in Projected Database of **AT**:
ATCAAGTATGCTTAGC

Projected database of **ATG**:
8

ATG appears only once: minimum unique substring of pattern P

Search **C** in Projected Database of **ATG**:
ATCAAGTATGCTTAGC

Projected database of **ATGC**:
8

No **ATGCA**. Therefore, ATGC is the maximum unique substring of pattern P

Ye et al. Bioinformatics, 2009

Pindel

1. Read in the location and the direction of the mapped read from the mapping result obtained in the preprocessing step;
2. Define the 3′ end of the mapped read as anchor point;
3. Use pattern growth algorithm to search for minimum and maximum unique substrings from the 3′ end of the unmapped read within the range of two times of the insert size from the anchor point;
4. Use pattern growth to search for minimum and maximum unique substrings from the 5′ end of the unmapped read within the range of read length +*Max_D_Size* starting from the already mapped 3′ end of the unmapped read obtained in step 3;
5. Check whether a complete unmapped read can be reconstructed combining the unique substrings from 5′ and 3′ ends found in steps 3 and 4. If yes, store it in the database *U*. Note that exact matches and complete reconstruction of the unmapped read are required so that neither gap nor substitution is allowed.

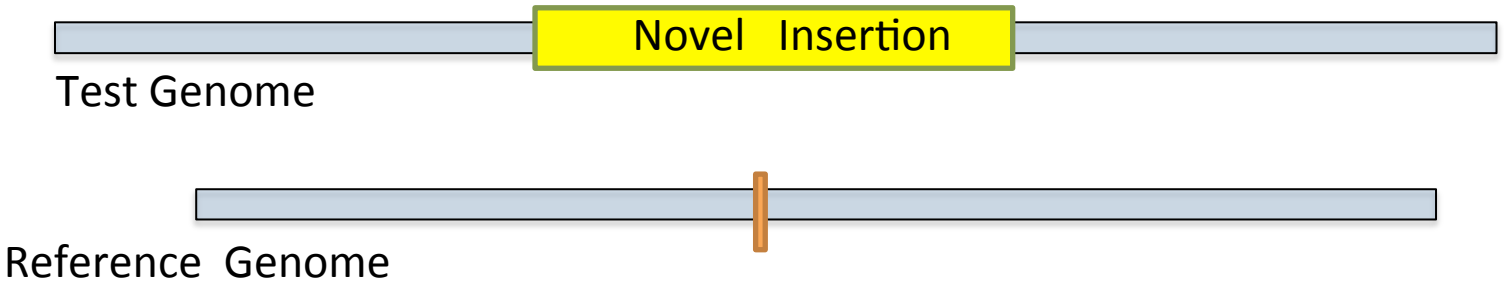
- Large Max_D_Size -> slow execution

Ye et al. Bioinformatics, 2009

DENO ASSEMBLY BASED ALGORITHMS

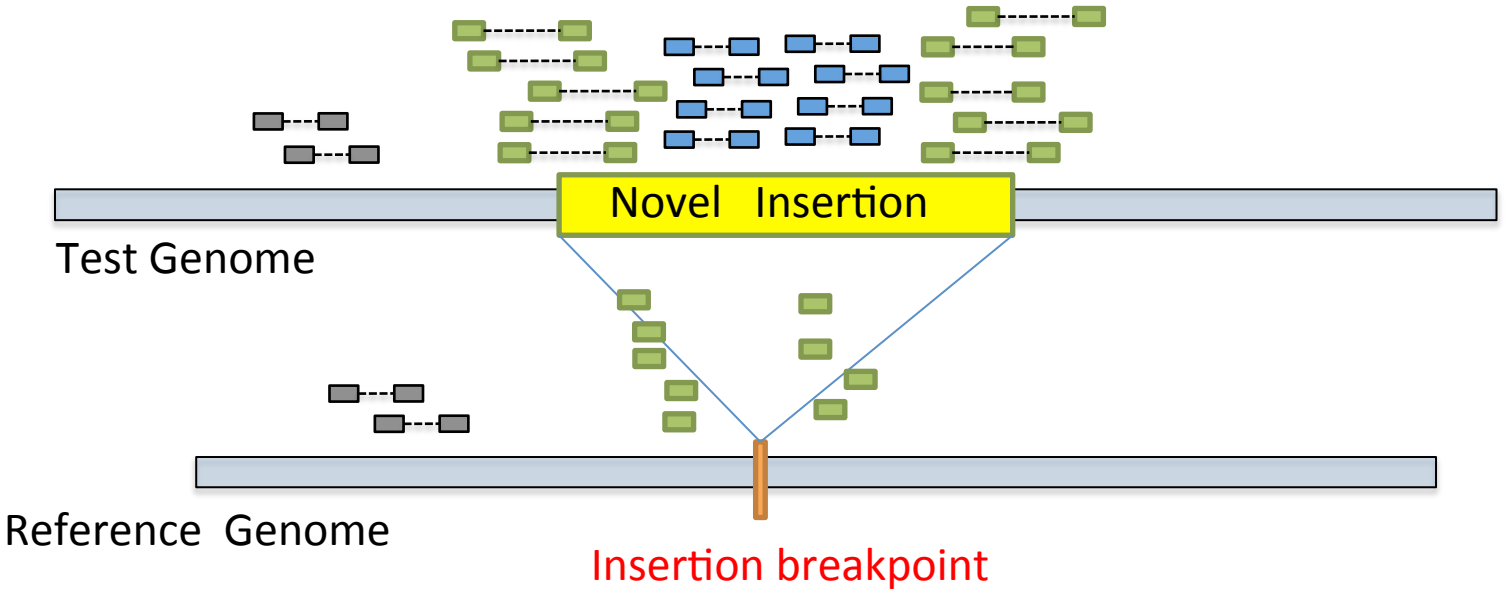
Novel sequence insertions

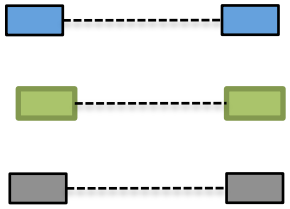
An insertion of a sequence into a genome where no similar sequence exists in the reference genome.



- The missing sequences can harbor undiscovered sequences of functional importance.
- The discovery helps us build comprehensive reference genomes.

Novel sequence insertions



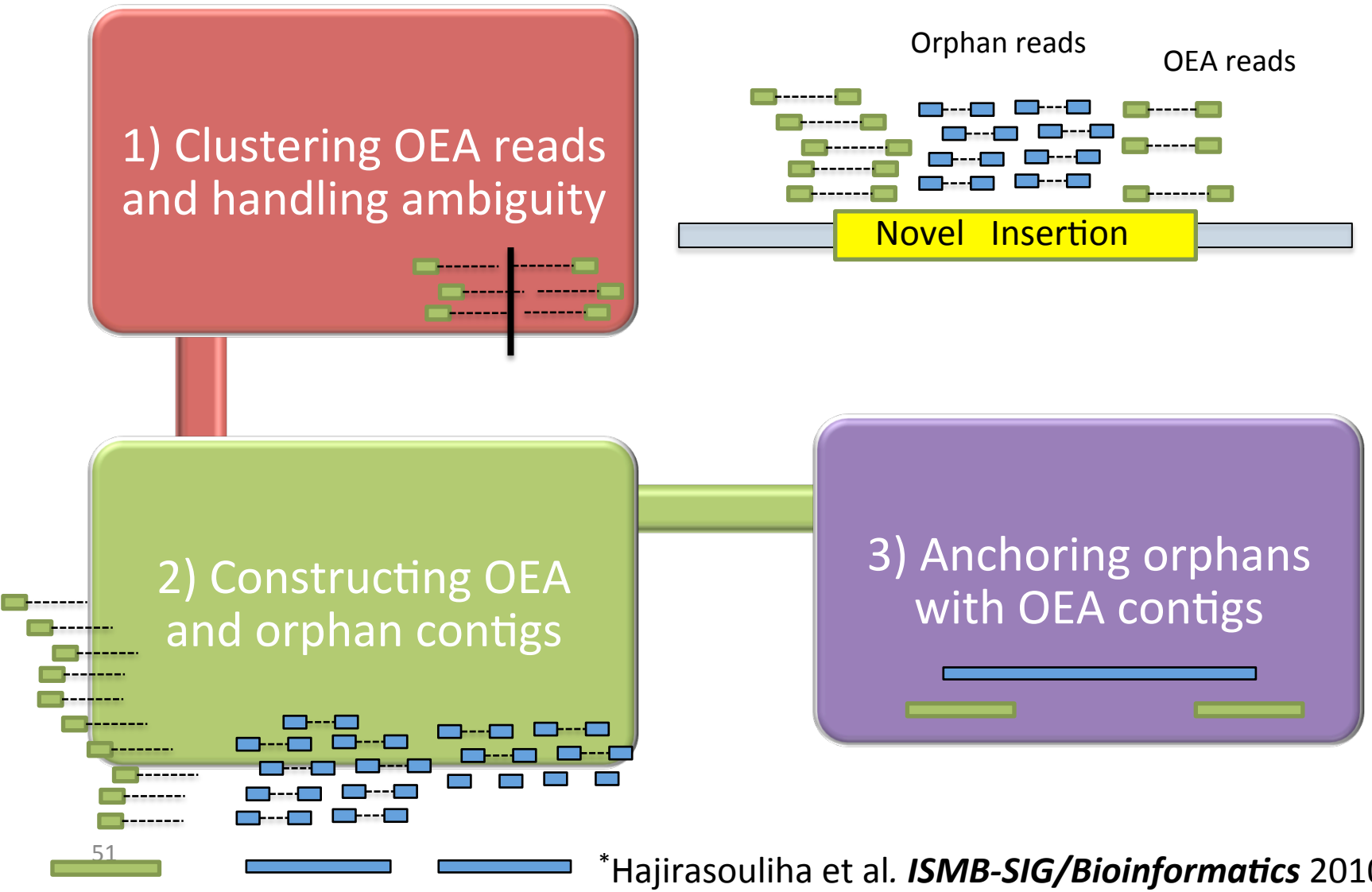


Orphan reads

One-end anchor (OEA) reads

Concordant reads

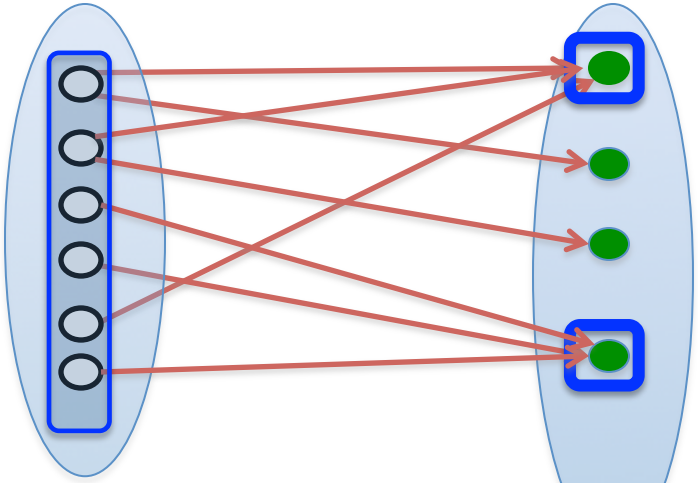
Novelseq algorithms*



*Hajirasouliha et al. ***ISMB-SIG/Bioinformatics*** 2010

Clustering OEA reads and handling ambiguity

Objective: find the minimum number of novel insertion breakpoints that explains all OEA reads.

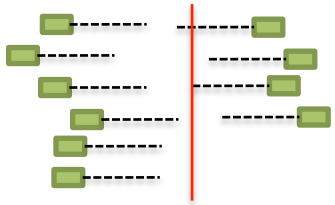


OEA reads

Maximal valid clusters

Maximal Valid Cluster:

All OEA read alignments that support the same breakpoint.

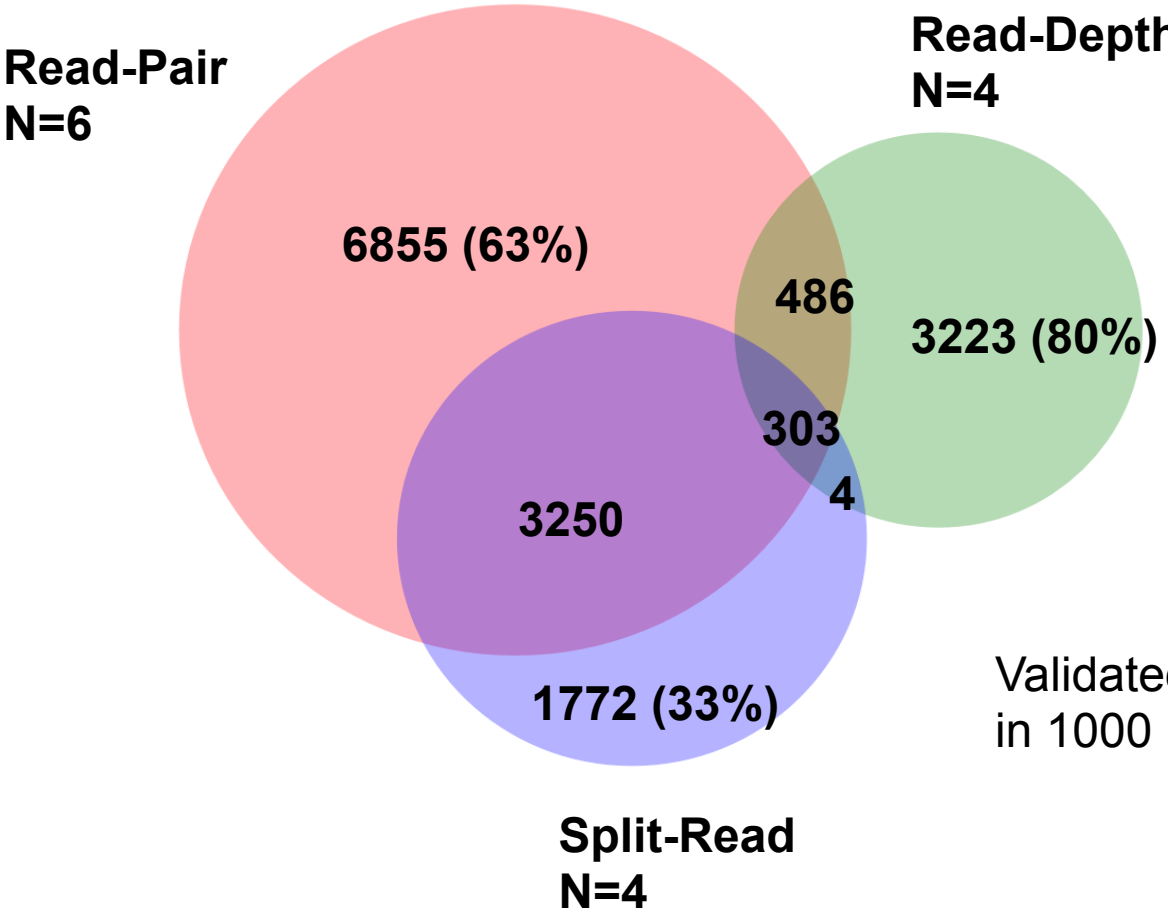


Minimum Set Cover:

Finds an approximated solution.

OPEN PROBLEMS?

No method is comprehensive



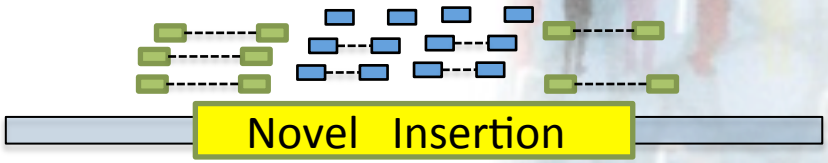
Validated deletions
in 1000 Genomes data

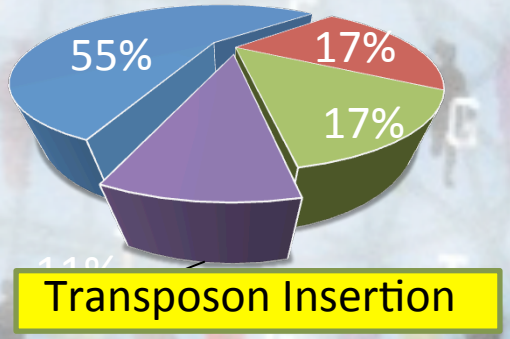
Kidd et al., Cell, 2010

Open problems

- Identify ***inversions*** and ***translocations***
- Discover SVs in repeat- and duplication-rich regions
- Accurate & comprehensive detection of CNVs with a *single* algorithm
  - High sensitivity
  - High specificity

Summery

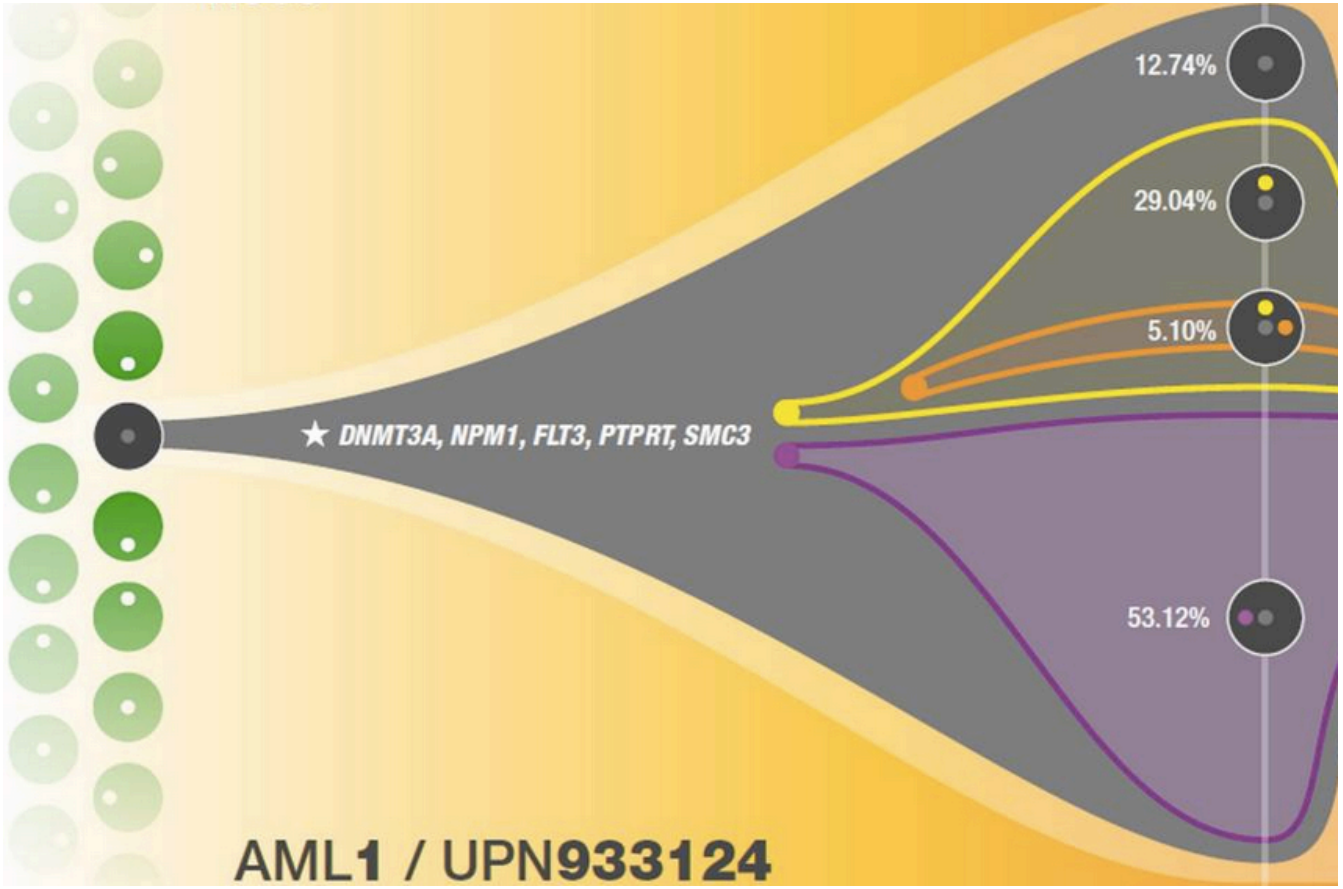




- Variationhunter
- Pindel
- CommonLAW
- NovelSeq

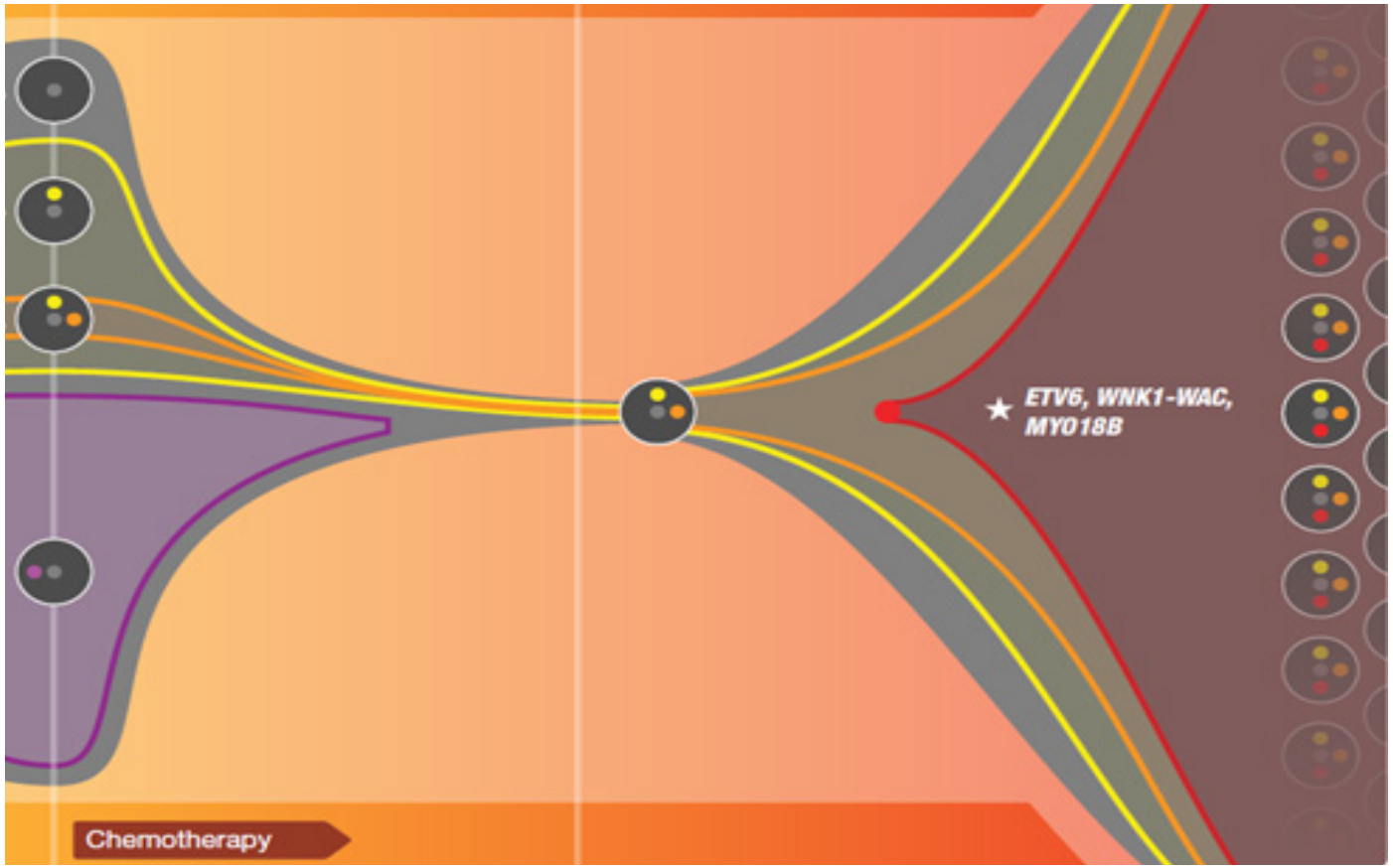
SV IN CANCER GENOMES

Cancer Evolution*



*Ding et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012

Evolution in relapsed AML*

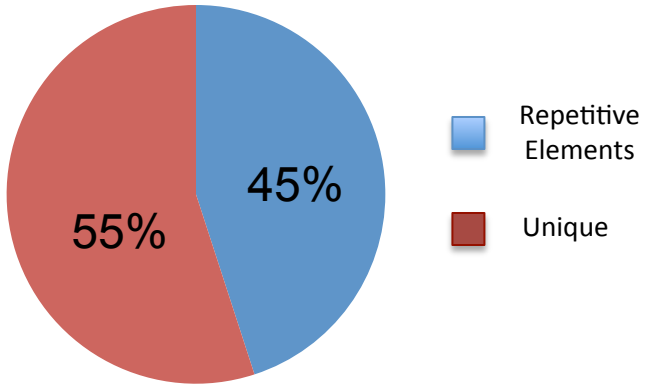


*Ding et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012

Structural Variations in Cancer

- Not very well studied but there are a lot of evidences in recently sequenced cancer genomes.

- SVs are in general harder to detect, using short reads due to repetitive DNA.



SV in Cancer Genomes

- Often lower depth of coverage in WGS
- Mixture of tumor and normal tissue
- Cancer Heterogeneity:
  - Different cells accumulate different sets of somatic mutations
  - We normally sequence mixture of cells.

THANK YOU!!