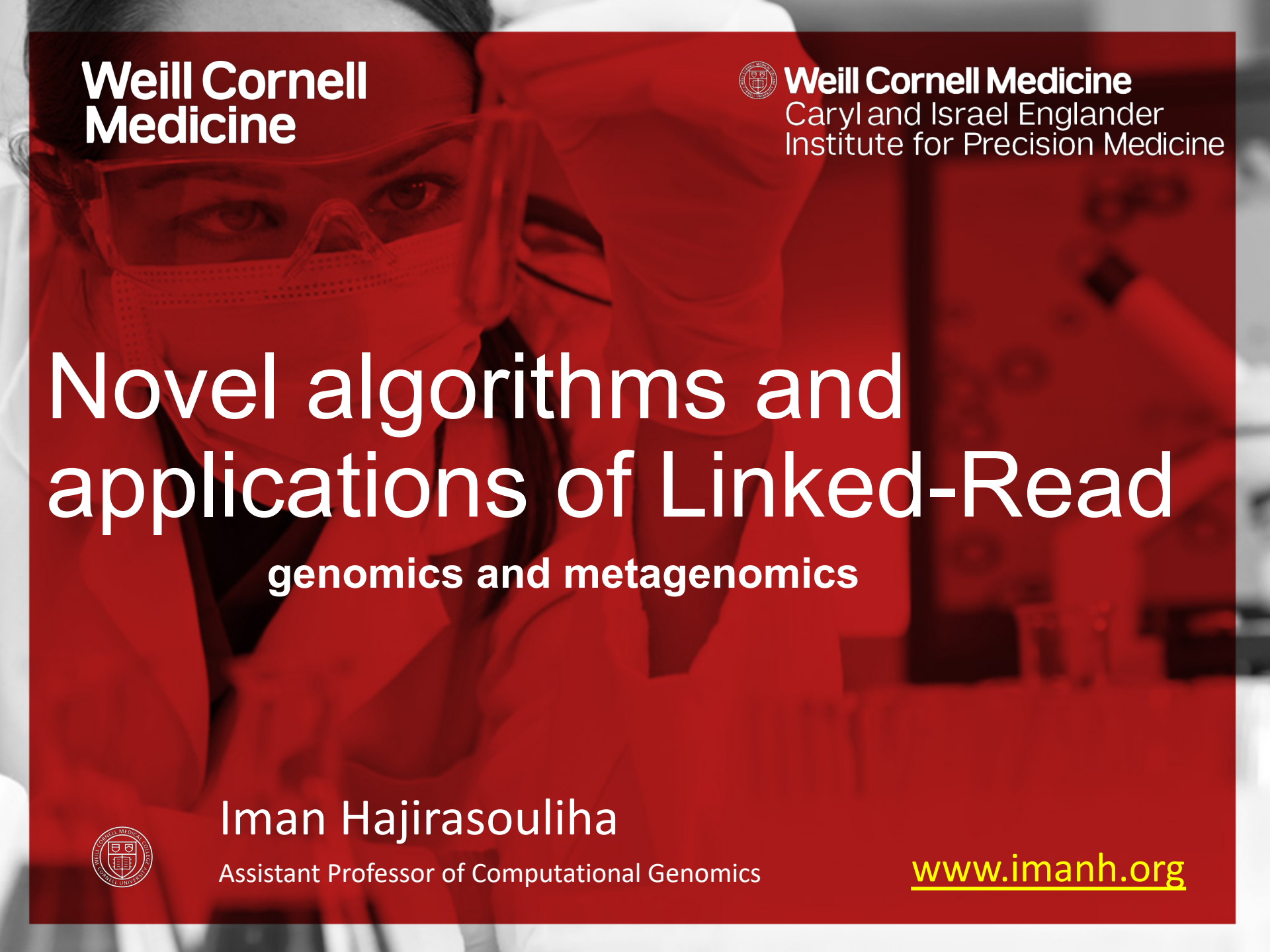


Weill Cornell Medicine

Weill Cornell Medicine
Caryl and Israel Englander
Institute for Precision Medicine

Novel algorithms and applications of Linked-Read

genomics and metagenomics

Iman Hajirasouliha

Assistant Professor of Computational Genomics

www.imanh.org

Human genome variation

Single Nucleotide Variant (SNV)

…ACCT**G**GTA… → …ACCT**C**GTA…

Small insertions or deletions (Indels) [1-50bp]

…ACCTGTA… → …ACCT**CT**GTA…

Structural Variations [>50bp]

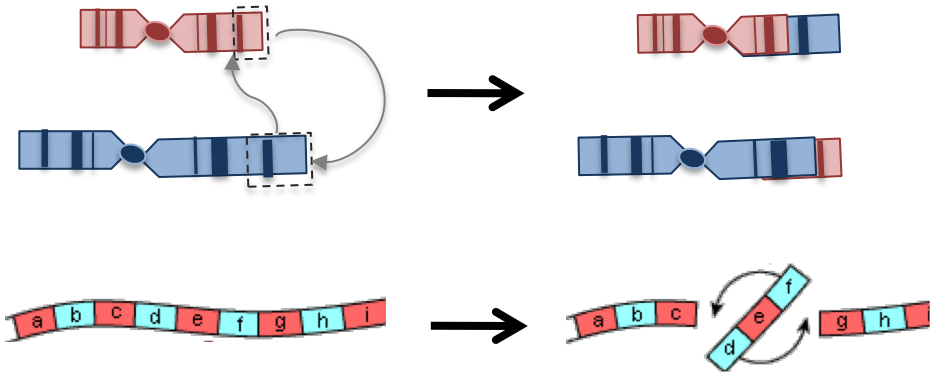
Duplication

Deletion

Insertion

Translocation

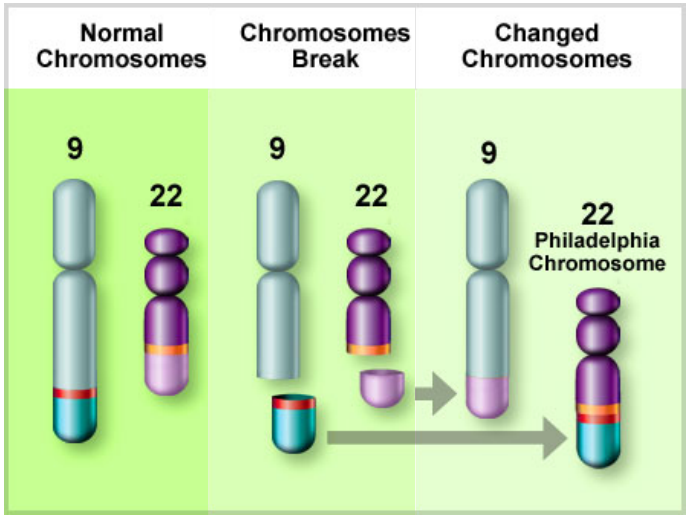
Inversion



Structural Variations and Diseases

Majority of Chronic Myelogenous Leukemia (CML) caused by a single SV.

A piece of chromosome 9 and a piece of chromosome 22 break off.



The pieces on different chromosomes trade places.

Structural Variations and Diseases

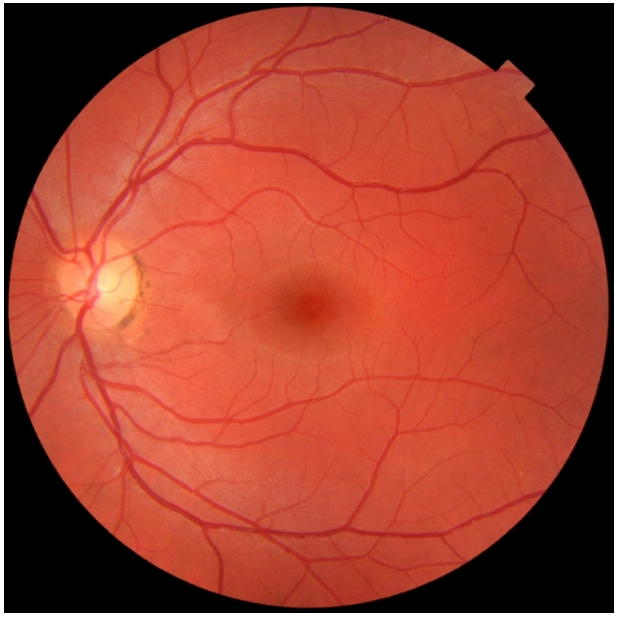
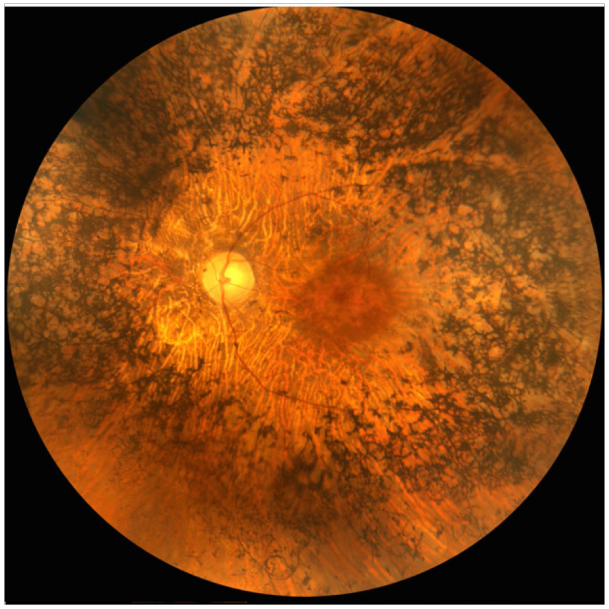
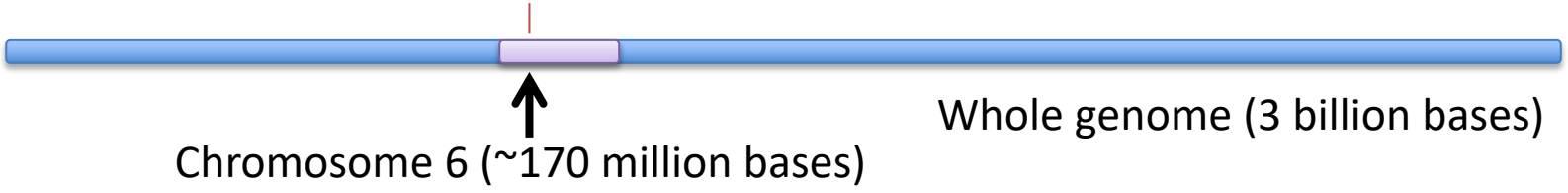


Photo of a normal left eye



Patient with retinitis pigmentosa (RP)

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!

International Cancer Genome Consortium

AGRE Autism Genetic Resource Exchange — A SCIENCE PROGRAM OF AUTISM SPEAKS

Autism Genome 10K

华大基因 BGI

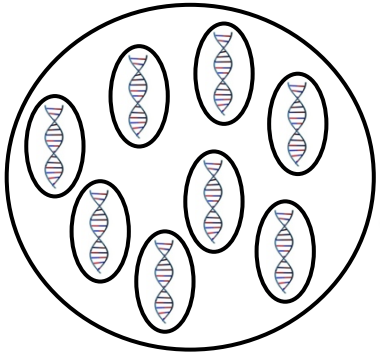
1000 Genomes
A Deep Catalog of Human Genetic Variation

Go NL
GENOMEoftheNETHERLANDS

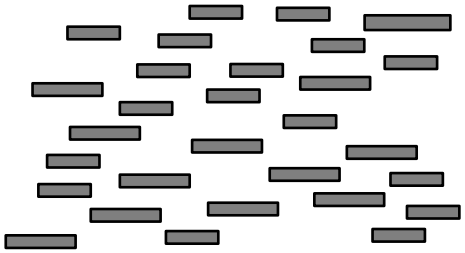
UK 10K
RARE GENETIC VARIANTS IN HEALTH AND DISEASE

Genomics england

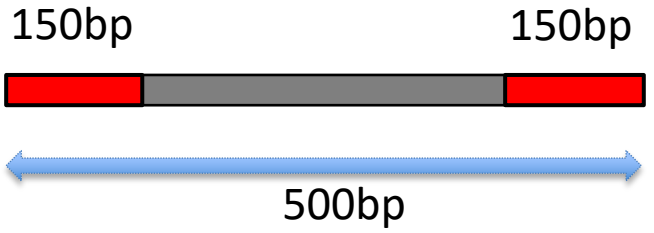
Standard short-read sequencing



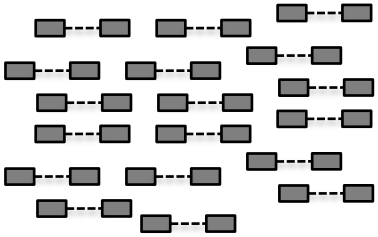
Cells



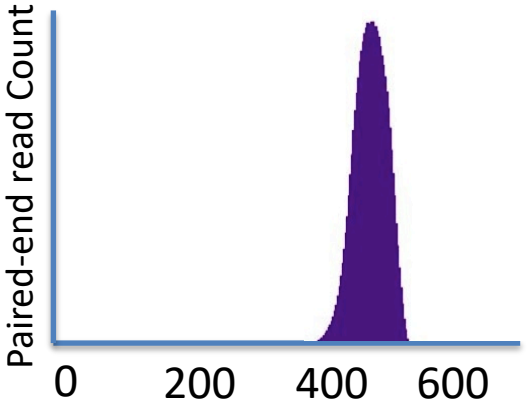
DNA is fragmented



Paired-end sequences of a fragment



Paired-end reads

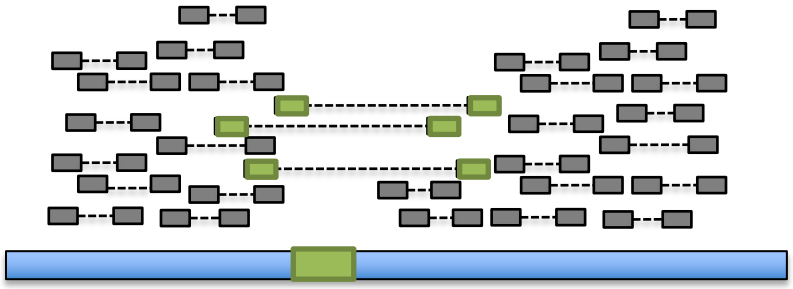


Insert size distribution

Determining Sequenced Genomes

1) Reference based methods

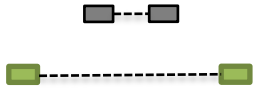
2) *De novo* assembly of short-reads



Paired-end reads are mapped to the reference

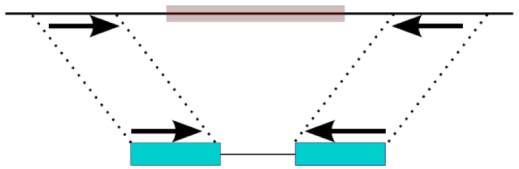
Concordant mapping

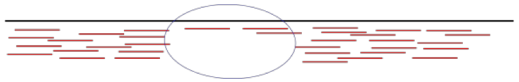
Discordant mapping

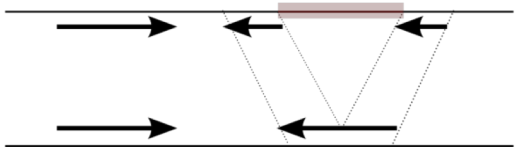


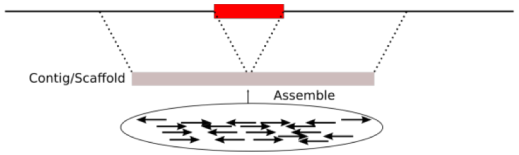
Determining Sequenced Genomes

- **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





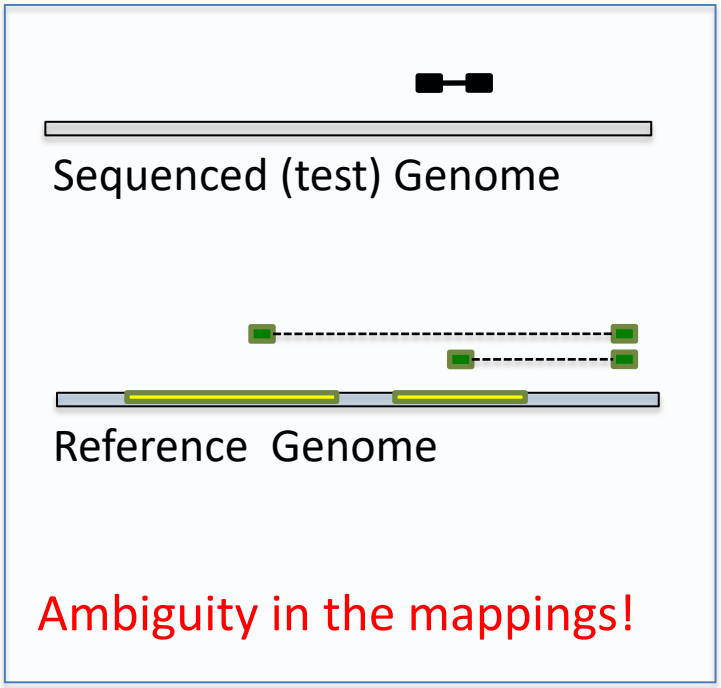


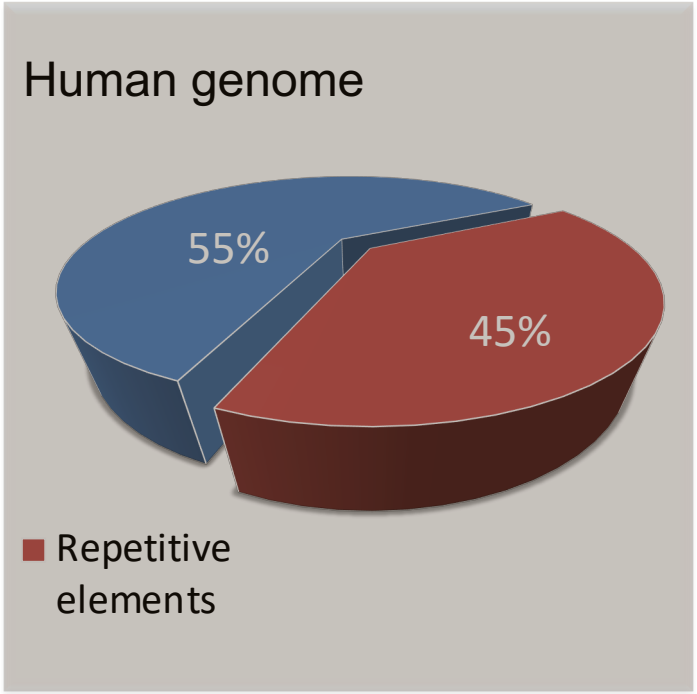


Limitations of short-read sequencing

NGS produce “short reads” (e.g. 50bp to 150bp)

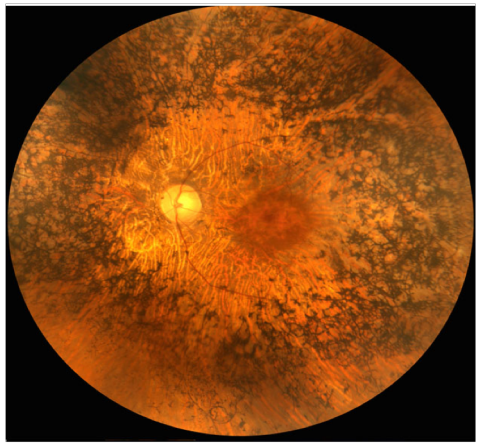
The human genome is repetitive!





Ignoring repeats is not an option!

Cause of the disease was an insertion of a repetitive element.



Patient with retinitis pigmentosa

*The software used for mapping short reads to the genome trimmed off repeat sequences and MAK initially appeared normal.**

* Todd J. Treangen & Steven L. Salzberg, ***Nature Review Genetics*** 2011

Challenges to determine sequenced genomes and metagenomes

Structural Variations, including those within repetitive regions or complex events.

The reference genome is incomplete or often nonexistent for metagenomes.

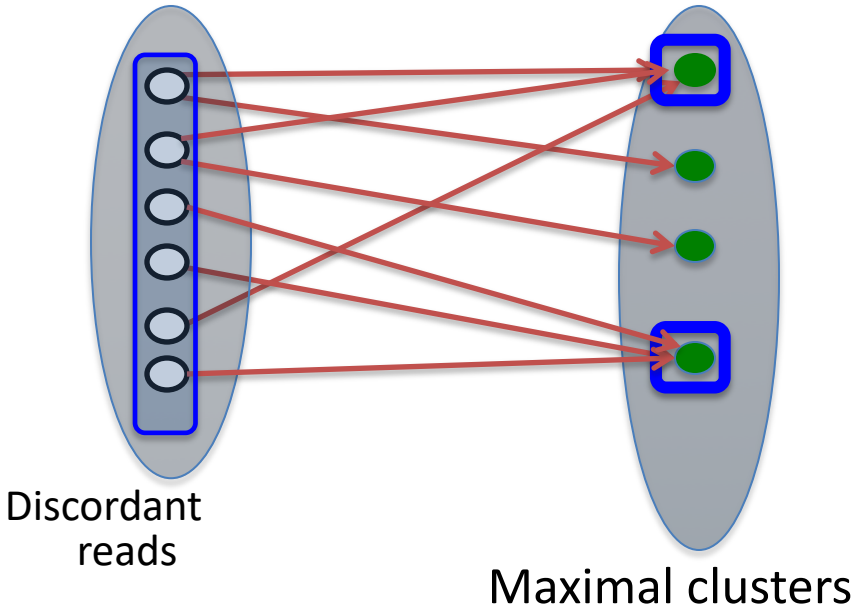
Early contributions in repetitive SV discovery using multi-mapped short-reads

Based on a combinatorial approach (i.e. maximum parsimony*) for handling repeats and ambiguity in mappings

Objective: find the minimum number of breakpoints that explains all discordant reads.

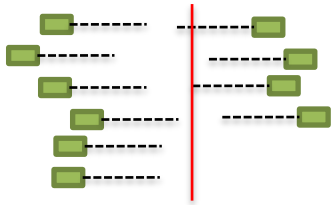
Clustering discordant reads while allowing multi-mapped reads

Objective: find the minimum number of breakpoints that explains all discordant reads.



Maximal Cluster:

All discordant read mapping that support the same breakpoint.

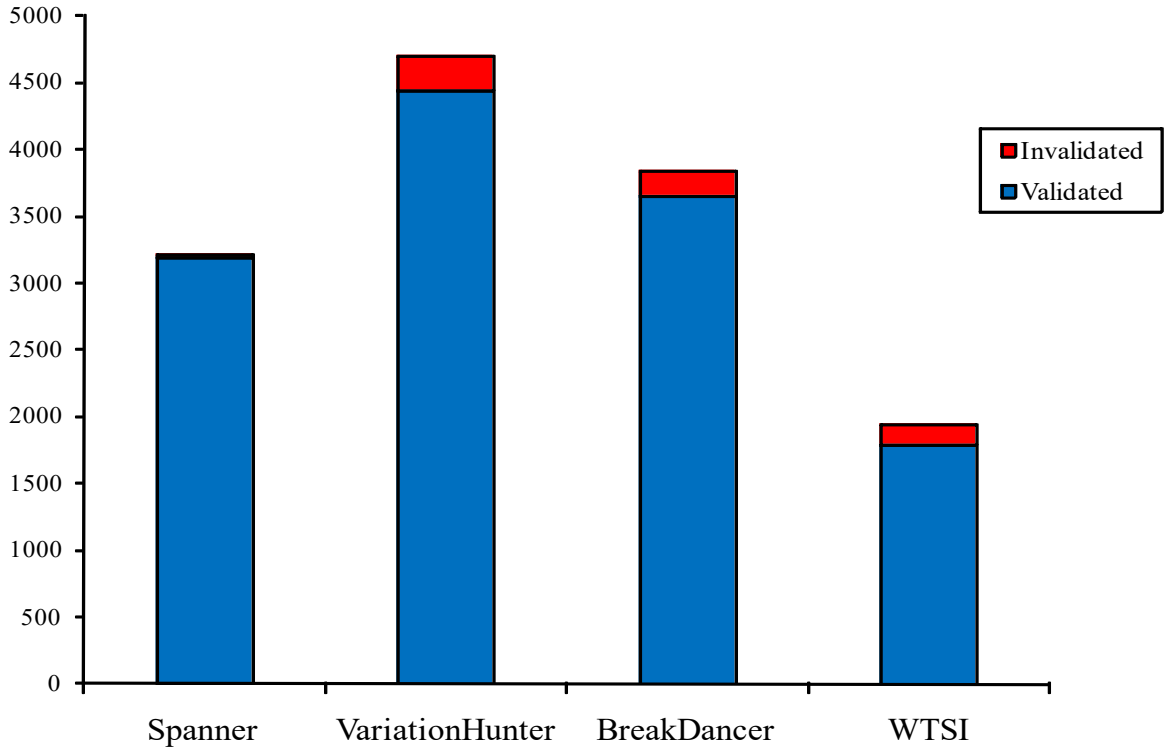


Minimum Set Cover:

Finds an approximated solution.

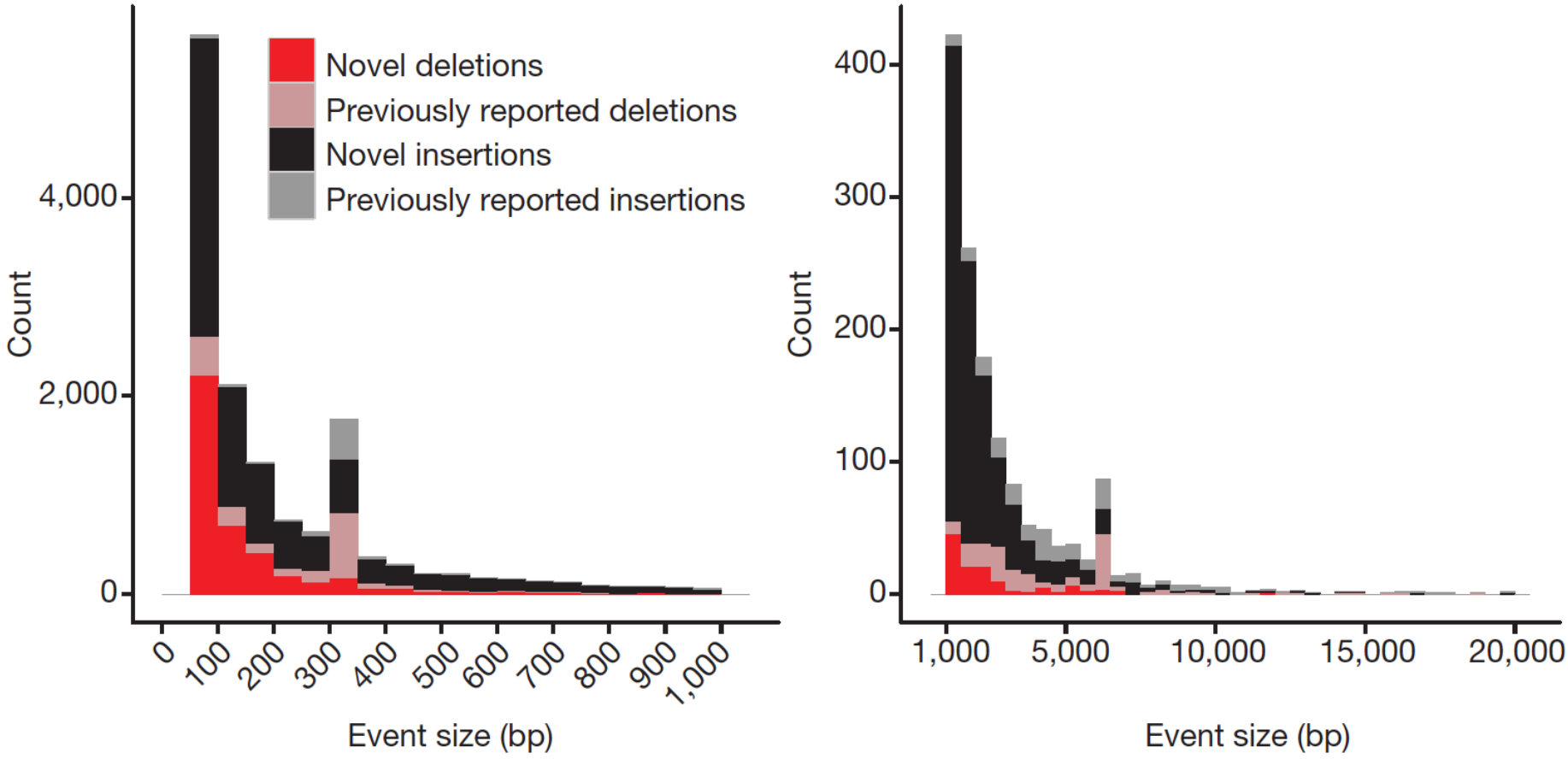
Hormozdiari et al. **Genome Research** 2009, *Hajirasouliha* et al. **Bioinformatics** 2010, *Hormozdiari, Hajirasouliha et al*. **Genome Research** 2011

The 1000 Genomes Yoruban Trio (validated/unvalidated deletions)



Higher sensitivity than unique-location based methods
Higher false discovery rate

Majority of SVs detected by PacBio long-reads are **novel**



Chaisson et al. **Nature** 2015

Beyond short-read sequencing

Long Read:

- PacBio
- ONT

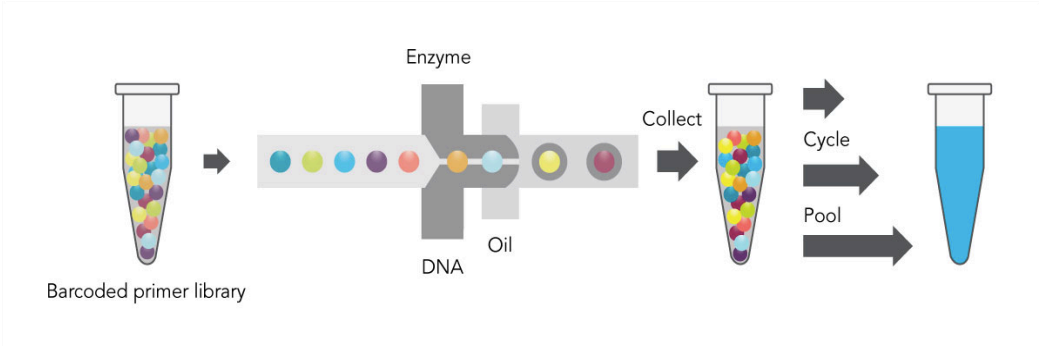
Expensive, low throughput, high DNA input
But they are real long-reads!

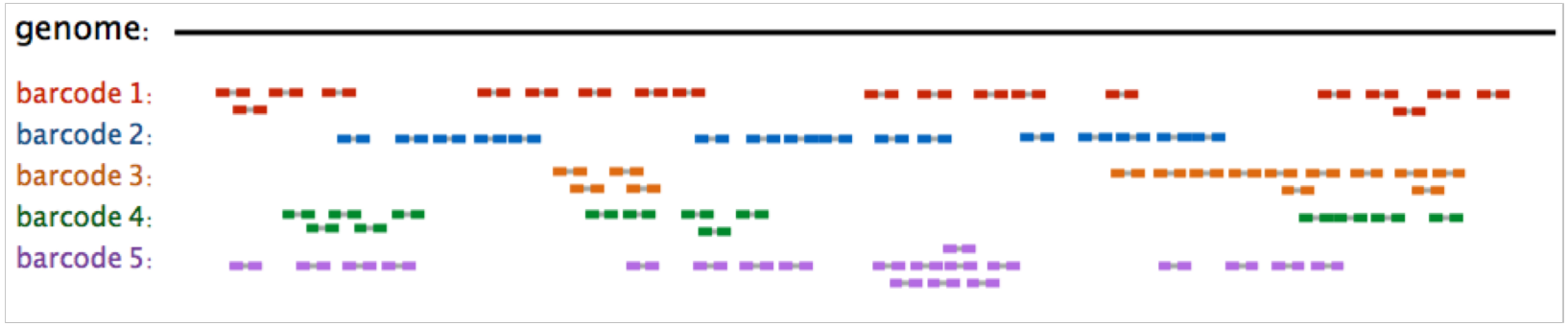
Linked-Read (or read cloud technologies):

- Moleculo (Illumina)
- 10x Genomics

Cheaper, high throughput, low DNA input
But they are synthetic long-reads!

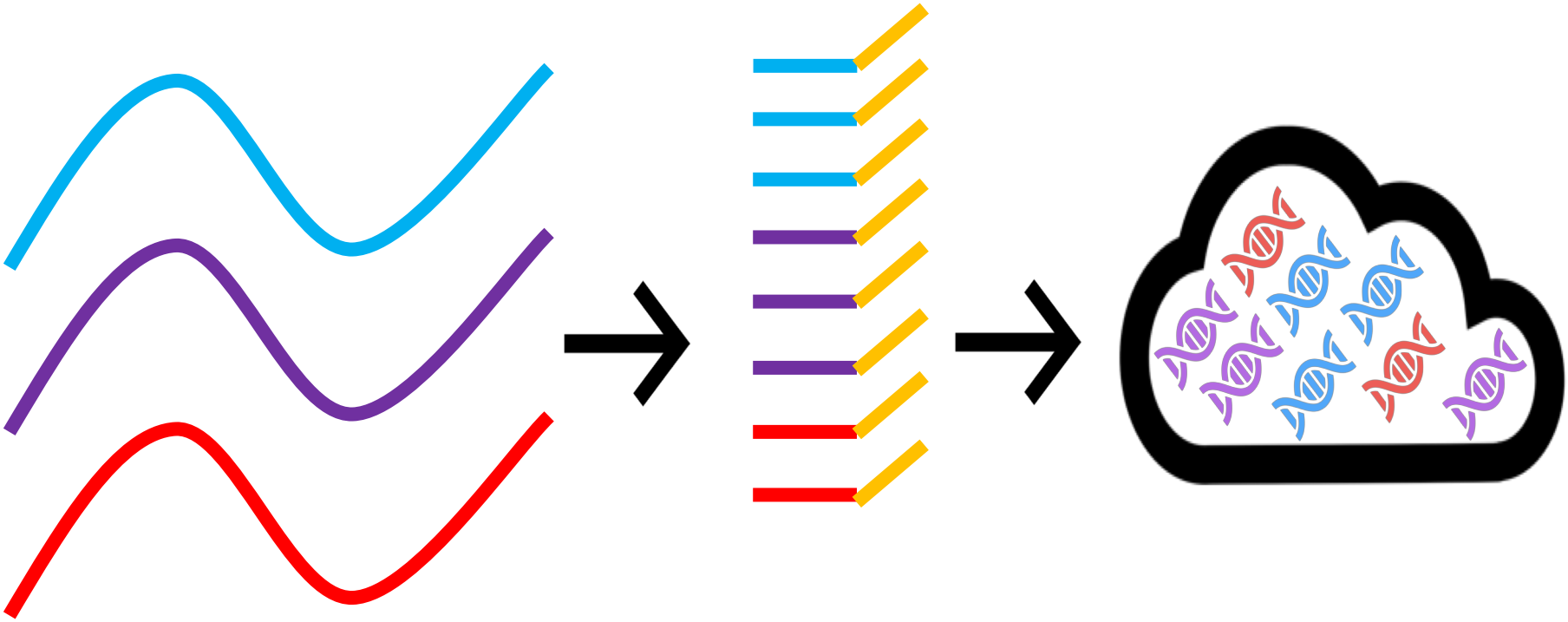
Linked-Read Technologies (e.g. 10x Genomics)





Knowing that the reads "should" form clusters, can we handle ambiguity in read mappings and SV detection better?

Linked Read Sequencing



10-100 kbp
Molecules/Fragments

Reads +
BCs

Read
Cloud

10x Genomics model

Long molecules / fragments:

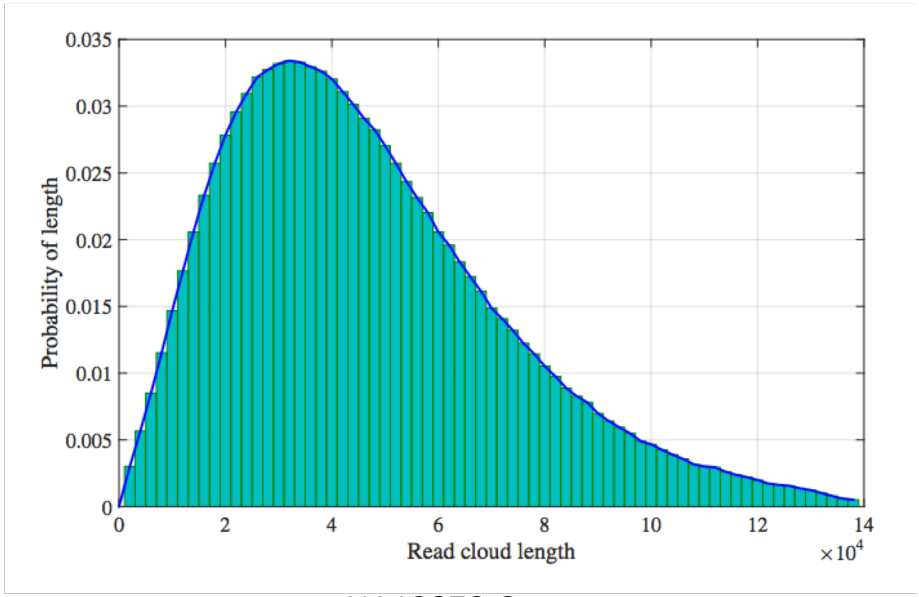
1. coverage C_F
2. mean length: ~10-100Kb

Short reads:

1. coverage C_R
2. length:150 bps

Barcodes:

1. # useful barcode ~ 1M
2. Distribution of barcode: Poisson



NA12878 Genome

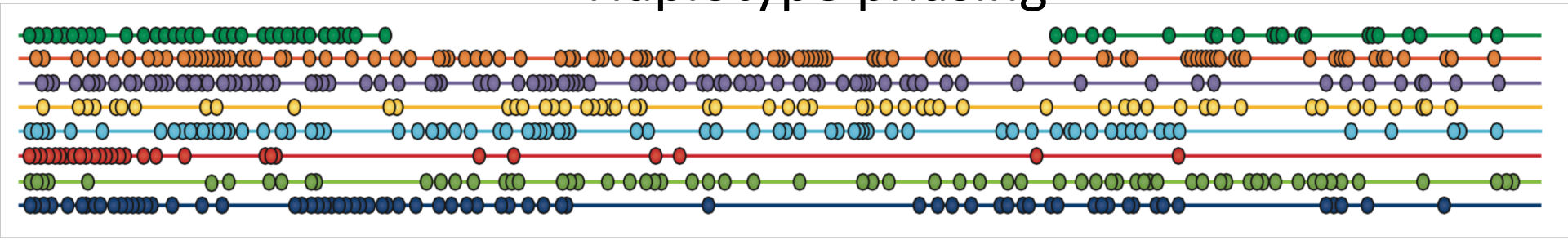
A new set of algorithmic challenges

1. Typically each barcode matches reads from 2-20 long fragments of DNA.

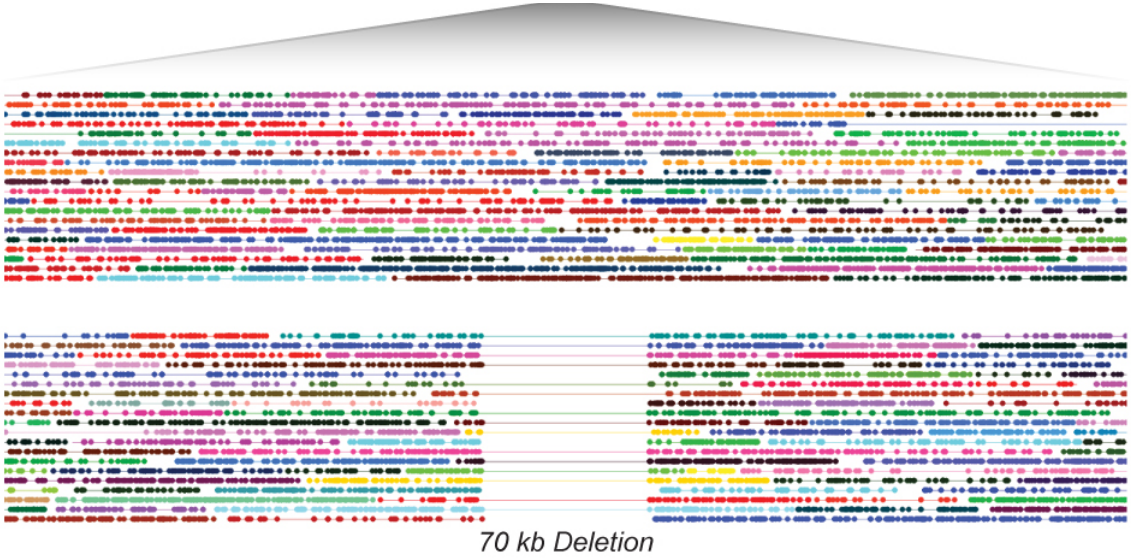
2. Each long fragment of DNA is covered only sparsely by short reads.

10x Genomics application

Haplotype phasing



Large structural variation calling



Part 1: SV detection in whole genome Linked-Read data with $VALOR_2$

Inversions and Duplications

New Results

Characterization of segmental duplications and large inversions using Linked-Reads

Fatih Karaoglanoglu, Camir Ricketts, Marzieh Eslami Rasekh, Ezgi Ebren, Iman Hajirasouliha, Can Alkan

doi: https://doi.org/10.1101/394528

Linked-Read SV Detection - $VALOR_2$

- Detection of balanced SVs with no gain or loss of genomic segment (e.g. inversions) is particularly a challenging task.

- Novel algorithm to characterize large (>40Kbp) interspersed segmental duplications and (>80Kbp) inversions

Definitions

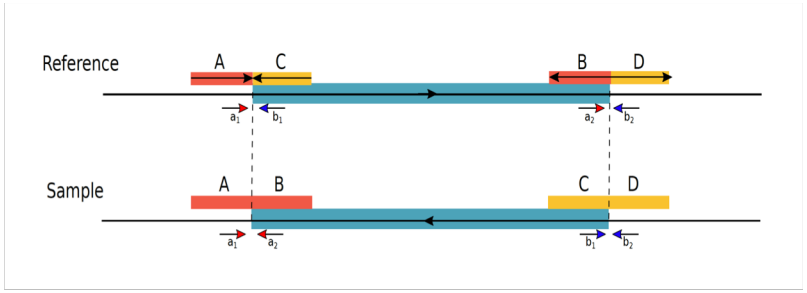
- Molecule/long fragment: a large molecule (10-100 Kbp) that was barcoded and pooled using the 10x Genomics platform. Here we refer to the physical entity.

- Submolecule: a molecule identified in silico by the VALOR$_2$ algorithm by analyzing read map locations.

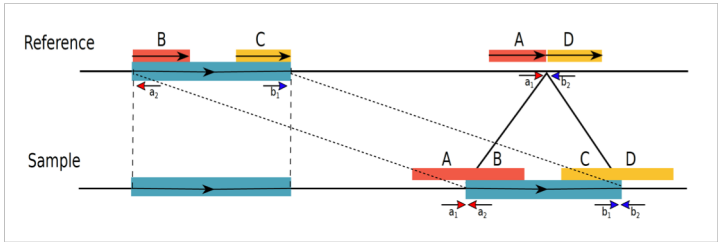
- Candidate split: a pair of submolecules with the same barcode that potentially signal a SV event.

- Split molecule pair: a pair of candidate splits with different barcodes that potentially signal the same SV event.

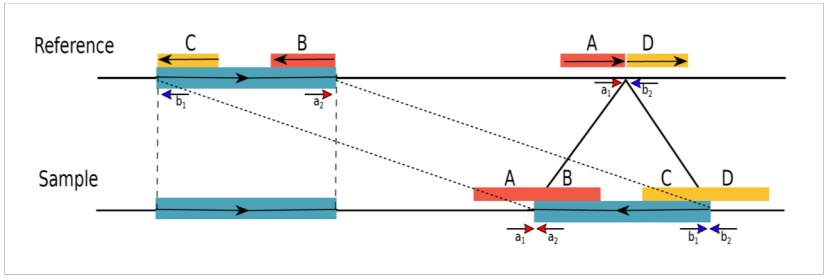
Split molecule signatures



INVERSION

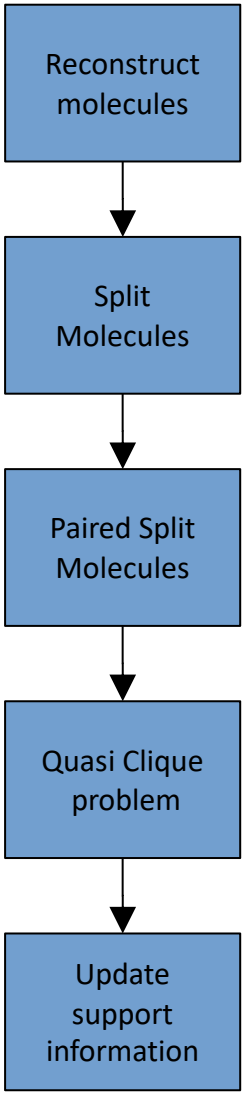


INTERSPERSED DUPLICATION



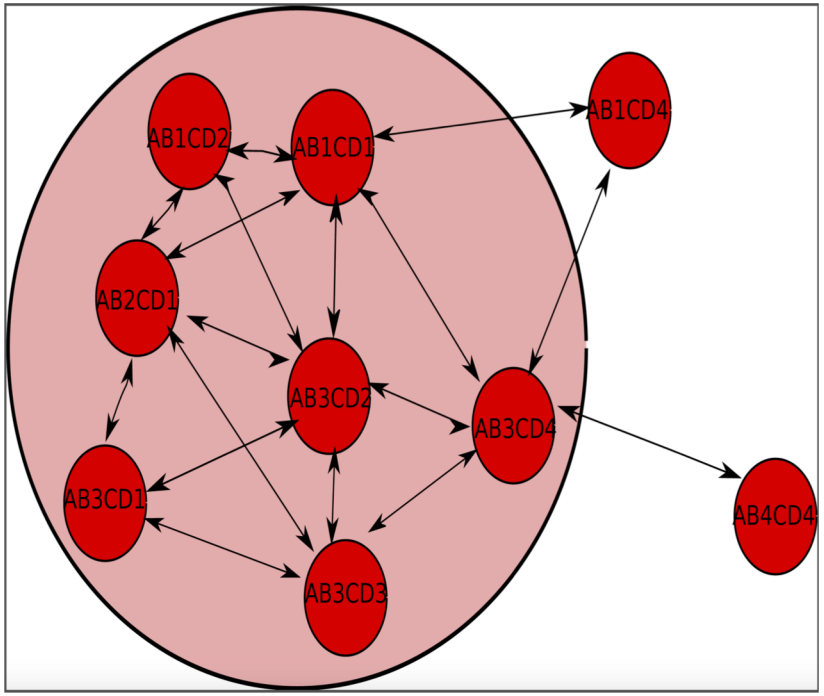
INVERTED DUPLICATION

VALOR$_2$:



- Reconstruct molecule locations
- Find pairs of recovered molecules with same barcode which are shorter than average size, but at expected length when combined.
- Find pairs of compatible split molecules with different barcodes.
- Find maximal quasi cliques on the graph with nodes of split molecule pairs and edges between compatible molecules.
- For each found clique update support information using read pairs and split molecule count.

Valor$_2$: Maximal quasi clique problem



- Here a quasi clique is defined as an approximate clique with V vertices and $\gamma \cdot \binom{|V|}{2}$ edges

Simulation experiments

Table 1. Prediction performance evaluation using simulated structural variants.

Variant	Tool	# Simulated	# Predicted	True	False	Precision	Recall
Duplication (direct)	$VALOR_2$	78	66	61	5	0.92	0.78
Duplication (inverted)	$VALOR_2$	56	51	49	2	0.96	0.88
Inversion	$VALOR_2$	94	65	64	1	0.98	0.76
	LUMPY	94	42	44	4	0.90	0.47
	DELLY	94	896	79	761	0.15	0.84
	Long Ranger	94	92	68	27	0.71	0.72

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Large Inversions - NA12878

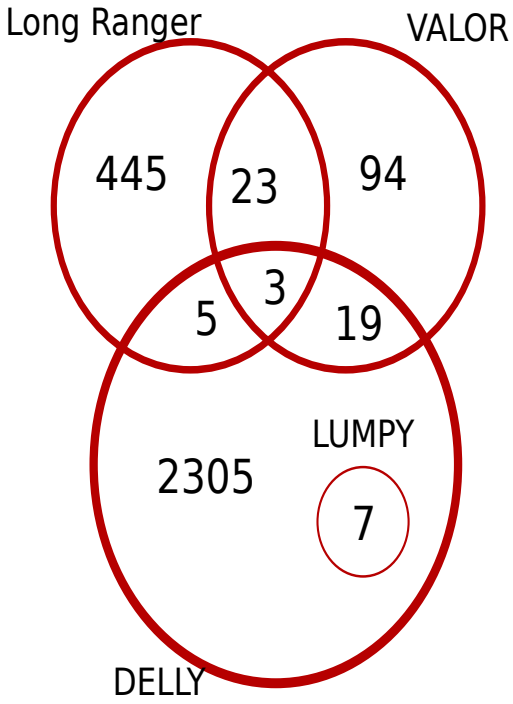


Table 3. Inversion prediction performance evaluation in the NA12878 genome using InvFEST database.

	Called	InvFEST-Valid.	InvFEST-Pred.	InvFEST-All
$VALOR_2$	135	6	5	17
Long Ranger	476	1	10	14
LUMPY	7	0	0	0
DELLY	2,340	1	6	24

Segmental Duplications – NA12878

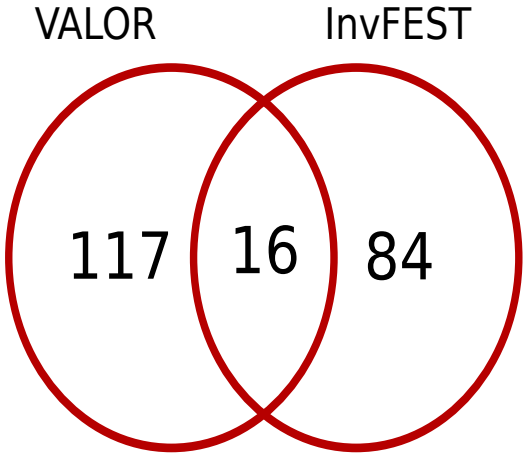
Chr	Start	End	Type	Target	No. of genes
1	120,600,786	120,692,870	Direct	1q21.1	1
1	144,832,884	145,751,706	Direct	1p22.3	25
1	145,062,336	145,116,024	Direct	1p11.2	
16	86,451,165	86,498,200	Direct	16q11.2	
17	21,522,544	21,551,840	Direct	17p11.2	
1	17,019,657	17,111,181	Inverted	1q42.3	4
1	145,983,326	146,027,347	Inverted	1p22.3	3
4	15,160	67,199	Inverted	4q35.2	2
8	2,189,297	2,290,508	Inverted	8p23.2	
10	46,965,140	47,022,150	Inverted	10q11.22	2
11	4,250,956	4,331,367	Inverted	11p15.4	
16	21,542,145	21,593,639	Inverted	16p12.2	
16	22,543,245	22,709,969	Inverted	16p12.2	2
X	153,423,995	153,485,001	Inverted	Xq28	3

The CHM1 genome

- Using a haploid human genome cell line (CHM1)

Overall, VALOR2 characterized

- 133 inversions (>80 Kbp)
- 14 inverted segmental duplications
- 22 direct segmental duplications (>40 Kb).



Part 2: Metagenomics using Linked-Read data

A new set of algorithmic challenges

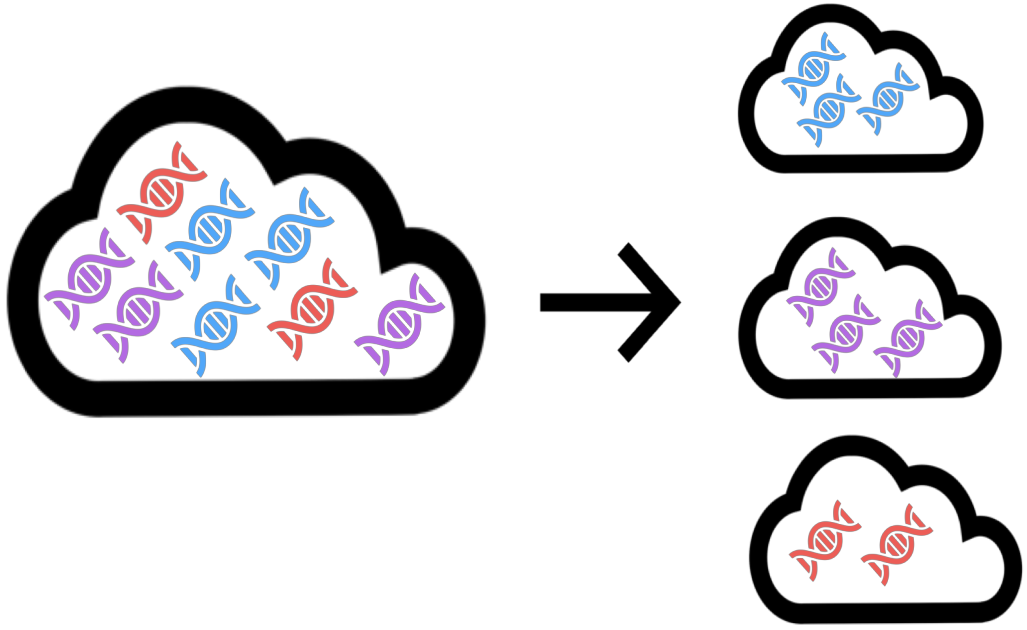
1. Typically each barcode matches reads from 2-20 long fragments of DNA.

2. Each long fragment of DNA is covered only sparsely by short reads.

Problem: Linked-Read Deconvolution

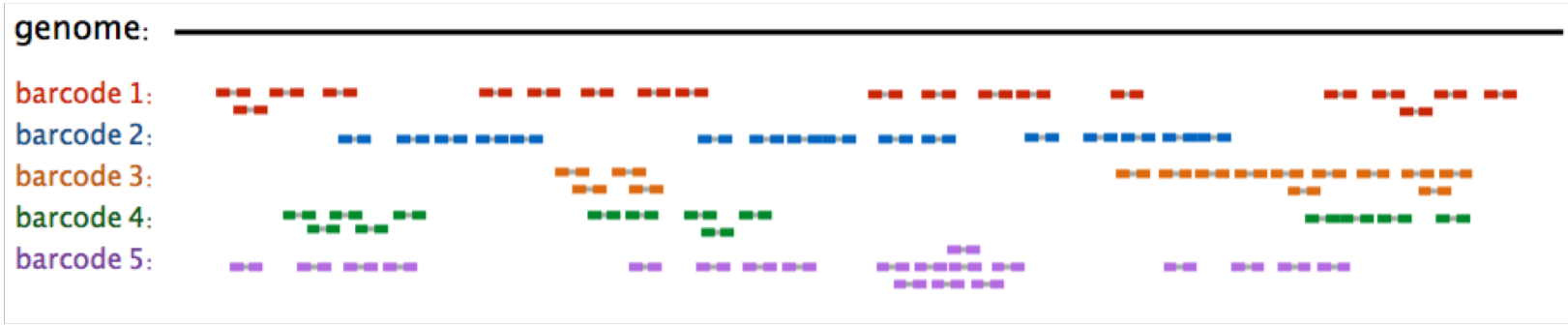
The deconvolution of reads with a single barcode into clusters that correspond to a single long fragment of DNA.

Any idea?



Problem: Linked-Read Deconvolution

Linked-Read Deconvolution when a reference is available



Linked-read Deconvolution when a reference is not available (metagenomics application?)

CSH PRESS GENOME RESEARCH

Minerva: an alignment- and reference-free approach to deconvolve Linked-Reads for metagenomics

David C. Danko[1,2], Dmitry Meleshko[1,2], Daniela Bezdan[2], Christopher Mason[2,3] and Iman Hajirasouliha[2,4]

OPEN ACCESS ARTICLE

This Article

Published in Advance December 6, 2018, doi: 10.1101/gr.235499.118

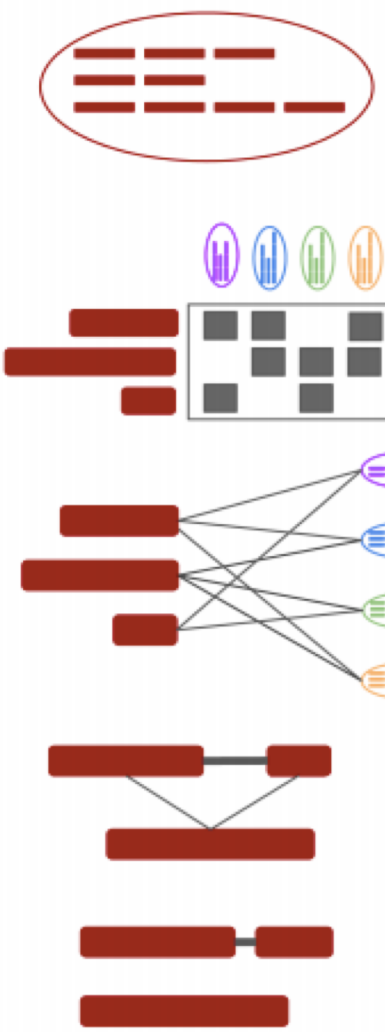
Genome Res. 2019. 29: 116-124

- A new graph-based algorithm for an approximate solution.
- Our approach also further uses some techniques from the field of topic modeling in Natural Language Processing (NLP).

Our graph based method

Key Observation: reads from the same fragment would tend to overlap with similar sets of reads that had different barcodes.

Our graph based method



We obtain reads with the same barcode grouped into a read-cloud

For each read cloud reads are mapped to other read-clouds

A bipartite graph is constructed between reads and read-clouds

A graph between reads is constructed

Reads are clustered into groups

primary real data sets from two microbial mock communities

- **Dataset 1:** 5 bacterial species: *E. coli, Enterobacter cloacae, Micrococcus luteus, Pseudomonas antarctica,* and *Staph. epidermis*.

- **Dataset 2:** 8 bacterial species and 2 fungi: *Bacillus subtilis, Cryptococcus neoformans, Enterococcus faecalis, E. coli, Lactobacillus fermentum, Listeria monocytogenes, Psuedomonas aeruginosa, Sachharomyces cerevisiae, Salmonella enterica,* and *Staphylococcus aureus*.

- Roughly 1ng of high molecular weight, processed using a 10x Chromium instrument, sequenced on an Illumina Hiseq with 2x150 paired-end reads.

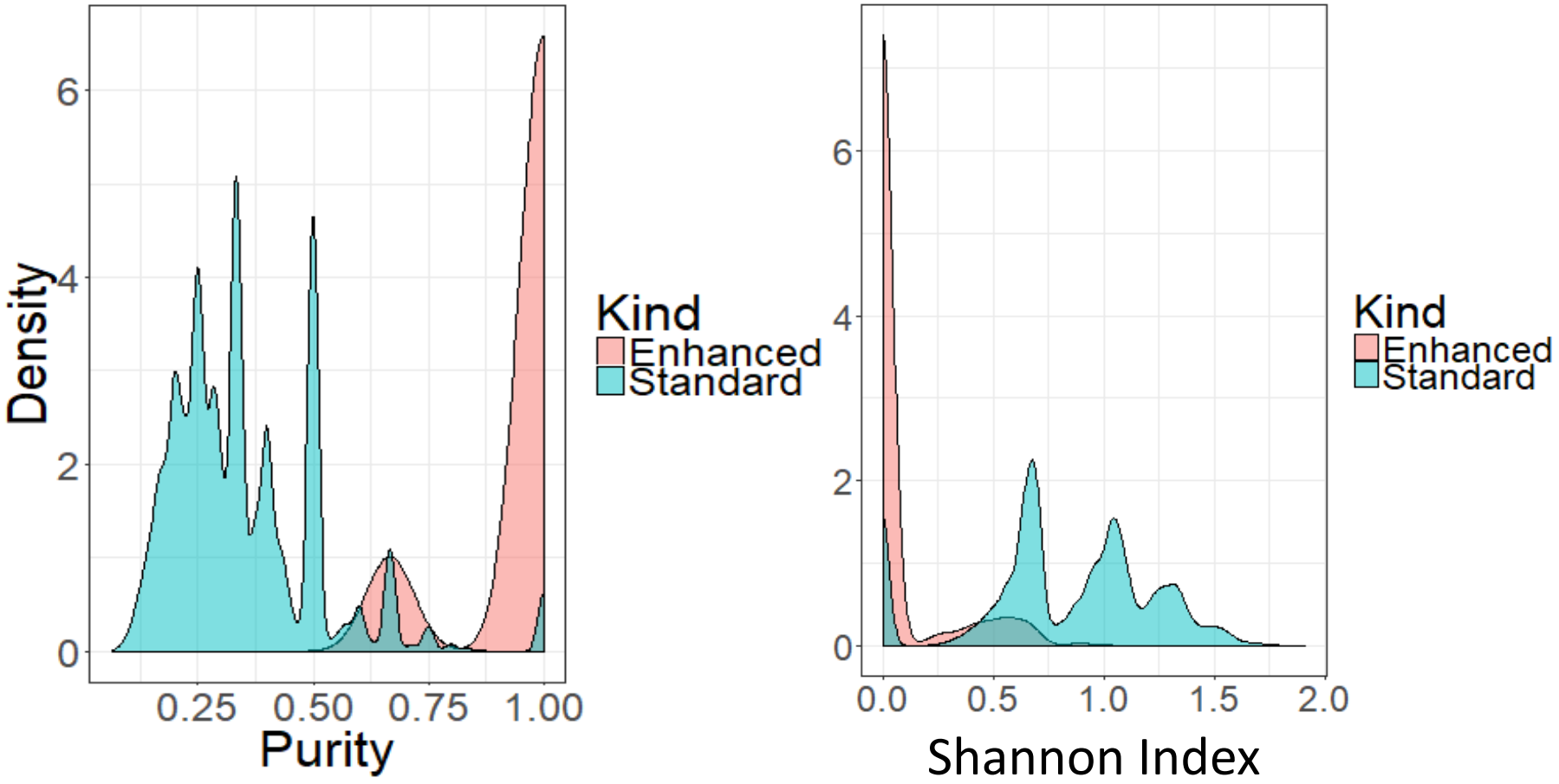
Experimental Results

- Minerva was able to identify subgroups in barcodes that largely corresponded to individual fragments of DNA. i.e. Enhanced Barcodes.

- We quantified this using two measures:
  - Shannon diversity index $H = \sum p_i \log p_i$
  - Purity $P = \max(\vec{p})$

where p_i indicates the proportion of an enhanced barcode that belongs to each fragment.

Minerva deconvolves barcodes



(Left) Purity for enhanced and standard barcodes
(Right) Shannon index in dataset one for enhanced and standard barcodes

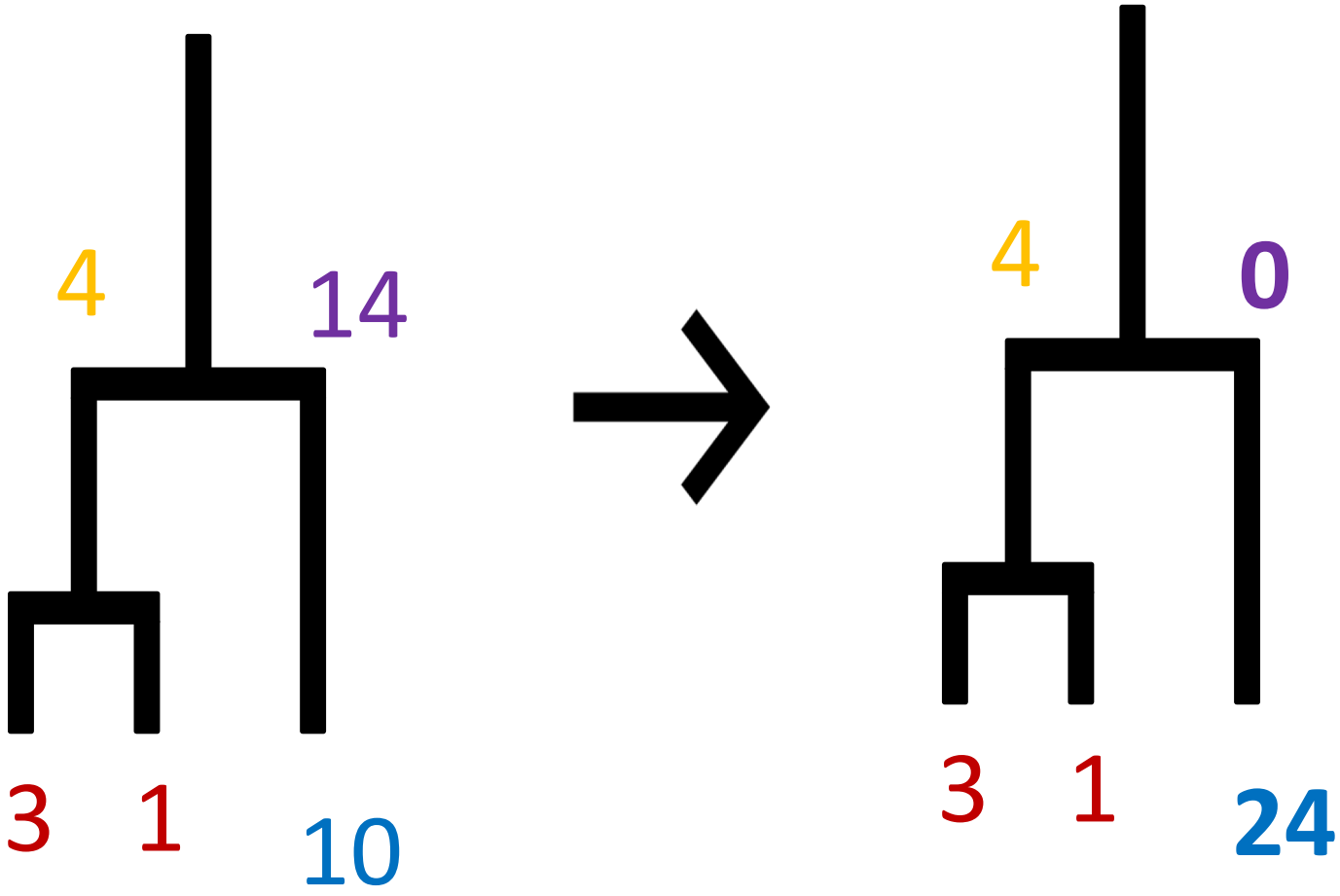
Minerva improves taxonomic assignments

- Minerva can improve the specificity of short read taxonomic assignments obtained from Kraken, a popular tool.

- All reads from the same long-fragment must have the same taxonomic rank!

- We were able to rescue a large number of reads from unspecific taxonomic assignments.

Minerva improves taxonomic assignments



Read Clouds Enable Improved Taxonomy

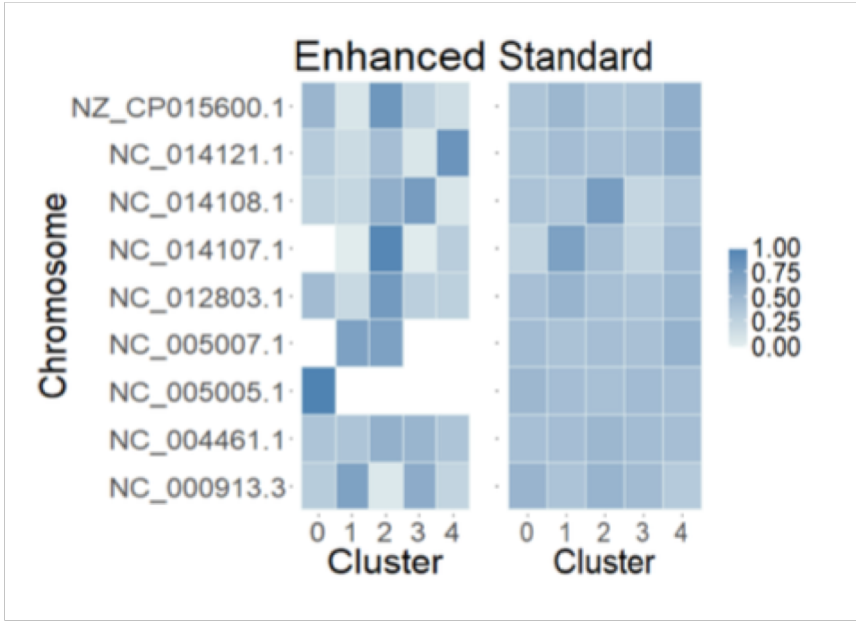
Original Rank	Promoted Rank	Enhanced	Standard	Difference	Ratio
Bacteria	*Enterobacter cloacae*	3	2	1	1.5
Proteobacteria	*Enterobacter cloacae*	24	17	7	1.41
Gammaproteobacteria	*Enterobacter cloacae*	21	13	8	1.62
Enterobacterales	*Enterobacter cloacae*	87	72	15	1.21
Enterobacteriaceae	*Enterobacter cloacae*	765	642	123	1.19
Bacteria	*Escherichia coli*	9	6	3	1.5
Proteobacteria	*Escherichia coli*	8	7	1	1.14
Enterobacterales	*Escherichia coli*	17	13	4	1.31
Enterobacteriaceae	*Escherichia coli*	9221	7846	1375	1.18
Escherichia	*Escherichia coli*	201	198	3	1.02
Gammaproteobacteria	*Pseudomonas antarctica*	3	2	1	1.5
Pseudomonas	*Pseudomonas antarctica*	256	200	56	1.28

Applications of Enhanced Barcodes (future work)

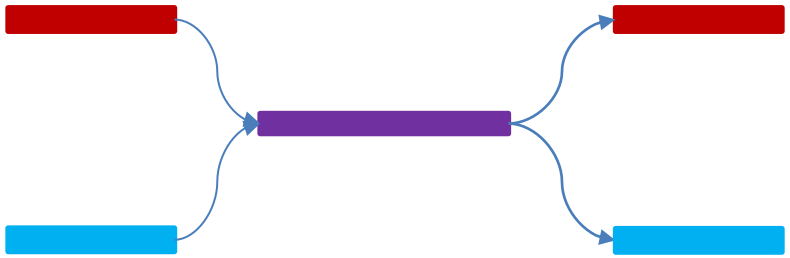
1. It is useful to group enhanced barcodes that likely came from the same genome.

   We used a clustering algorithm based on Latent Dirichlet Allocation (LDA), a classic model in NLP.

2. This technique can be also used to improve de novo assembly algorithms.

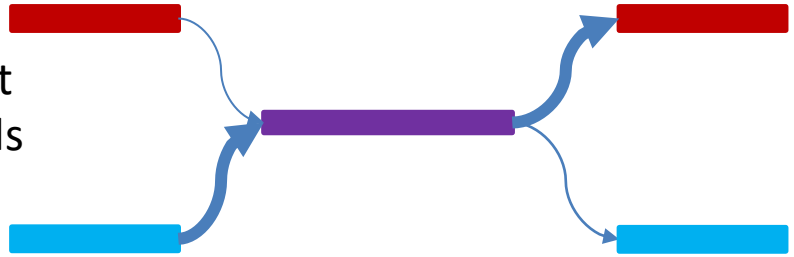


Applications of Enhanced Barcodes (future work)



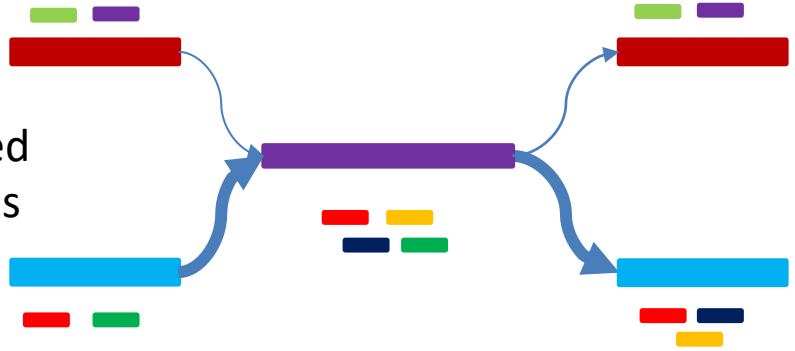
Compacted de Bruijn Graph

Short Reads



Heuristic for path extension, often read depth

Linked Reads



Extend paths based on barcode overlap

Thank you!

Postdoc, graduate student, internship positions available!

Weill Cornell Medicine
Camir Ricketts (Tri-CBM)
David Danko (Tri-CBM)
Dmitrii Meleshko (Tri-CBM)
Chris Mason
Daniela Bezdan

10x Genomics
Stephen Williams
Patrick Marks

Bilkent
Can Alkan
Fatih Karaognaloglu

The EIPM Team
Alicia Alonso
Olivier Elemento
Rob Kim
Andrea Sboner
David Wilkes