Accurate detection of complex structural variations using PacBio

Fritz Sedlazeck

June, 27, 2017

Baylor College of Medicine

HGSC
HUMAN GENOME SEQUENCING CENTER

Scientific interests

Mapping/ Assembly reads

NextGenMap-LR
(in preparation)

Falcon Unzip
Chin et.al. (2016)

NextGenMap
Sedlazeck et.al. (2013)

Detection of Variants

Sniffles
(in preparation)

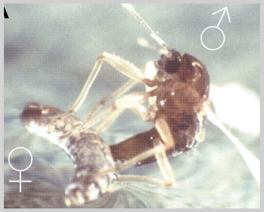
SURVIVOR
Jeffares et. al. (2017)

BOD-Score
Sedlazeck et.al.(2013)

Benchmarking

Teaser
Smolka et.al. (2015)

Sequencing
Jünemann et.al. (2013)



Applications

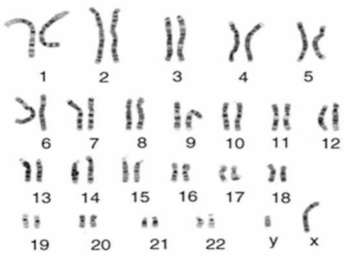
Model organisms:

-Cancer (SKBR3) (in preparation)
-miRNA editing (Vesely et.al. 2012)

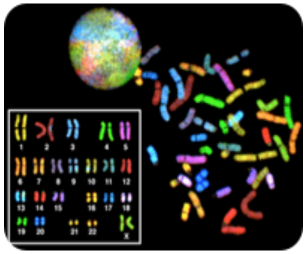
Non Model organisms:

-Cottus transposons (Dennenmoser et. al. 2017)
-Clunio (Kaiser et. al. 2016)
-Seabass (Vij et.al. 2016)
-Pineapple (Ming et.al. 2015)

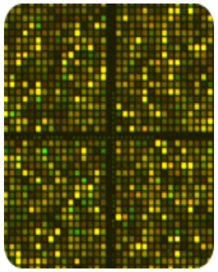
Our understanding of structural variation is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA



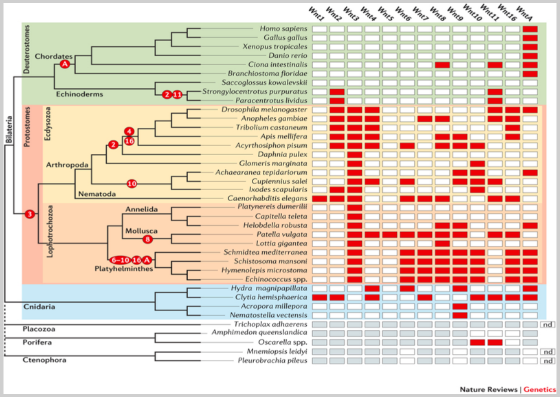
2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

High throughput
DNA sequencing

Single molecule
sequencing

Structural Variations

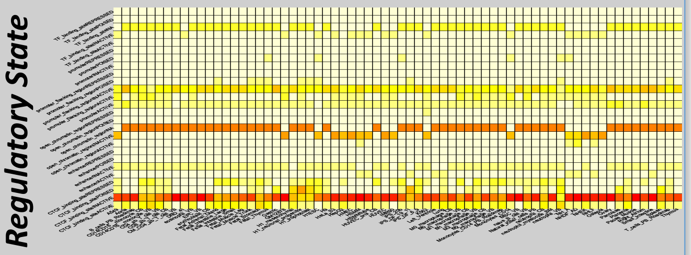
Evolution



Genomic Disorders

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X

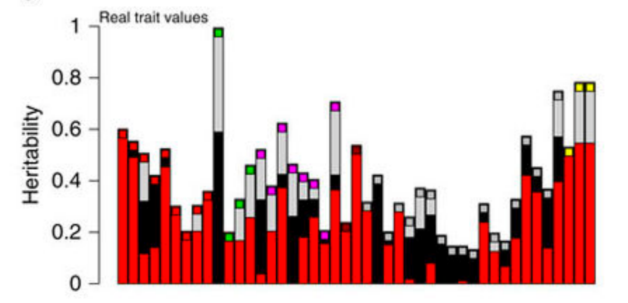
Impact on regulation



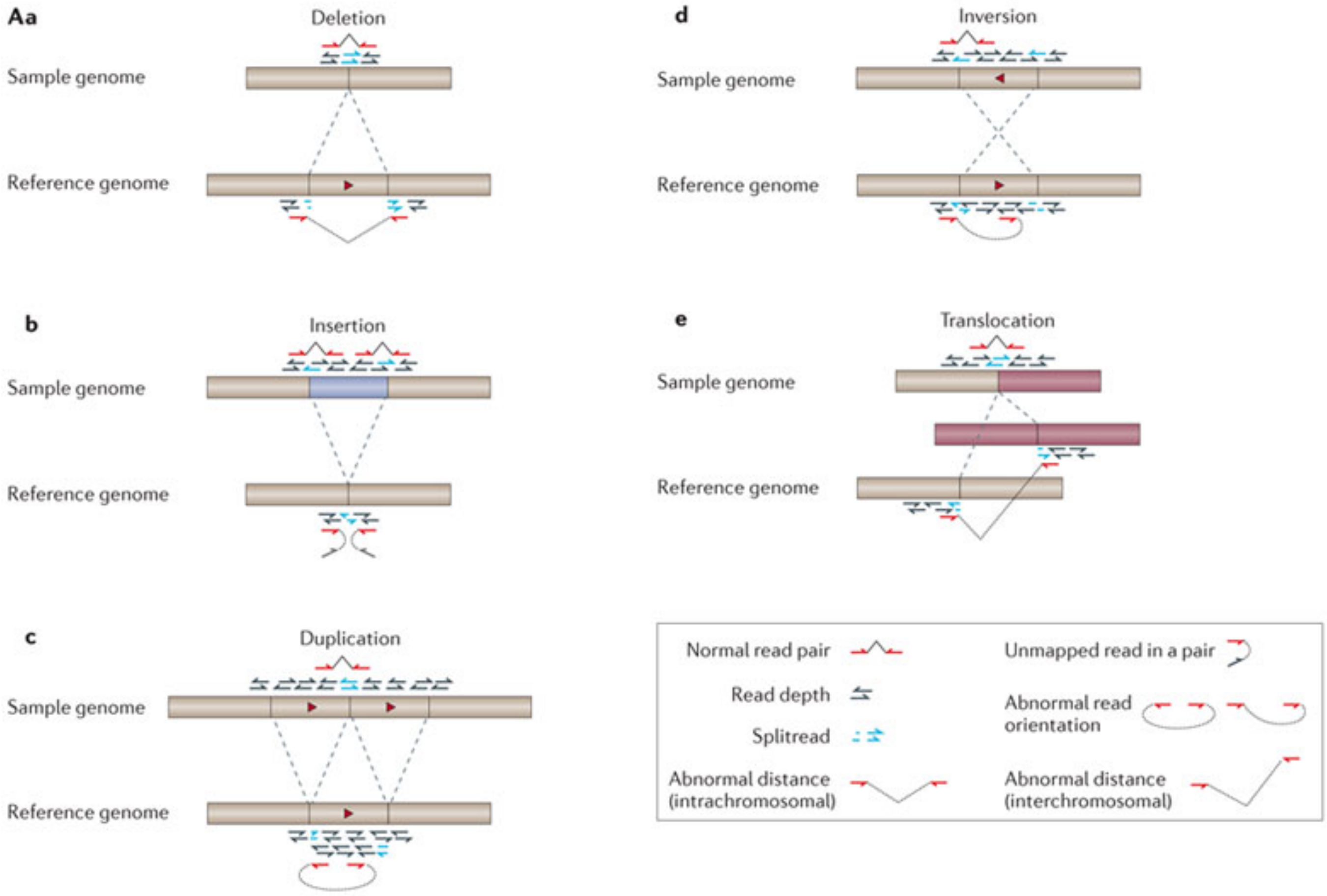
ENCODE

GTEx

Impact on phenotypes



How to detect Structural Variations



Long Read Technologies

- (+) SVs in repetitive regions
- (+) Span SVs
- (+) Uniform coverage
- (+) Can identify more complex SVs

- (-) Higher seq. error rate
- (-) Hard to align

How can we fully leverage this technology?

1. Improvement in mapping (NGMLR)

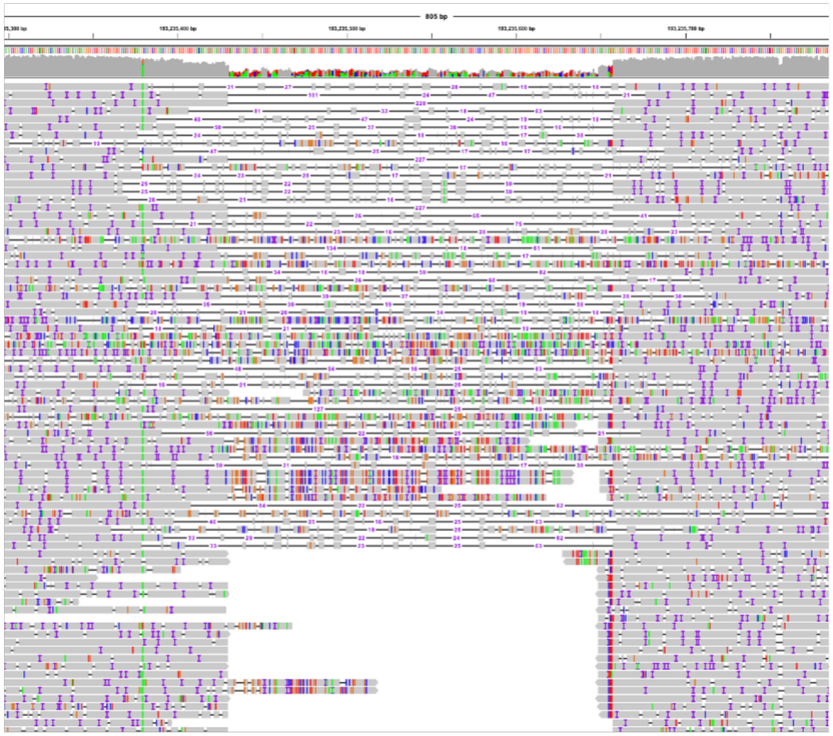
2. Improvement in SV calling (Sniffles)

3. Evaluation + results

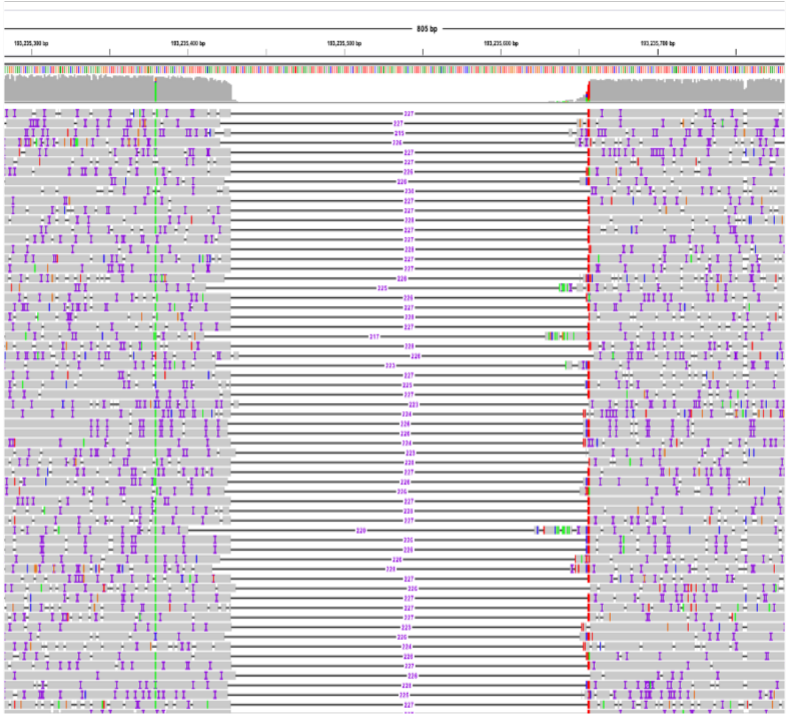
14
15
16
17
18
19
20
21
22
X
1
2
3
4
5
6
7
8
9
10
11
12
13

Why another mapper?

BWA-MEM:

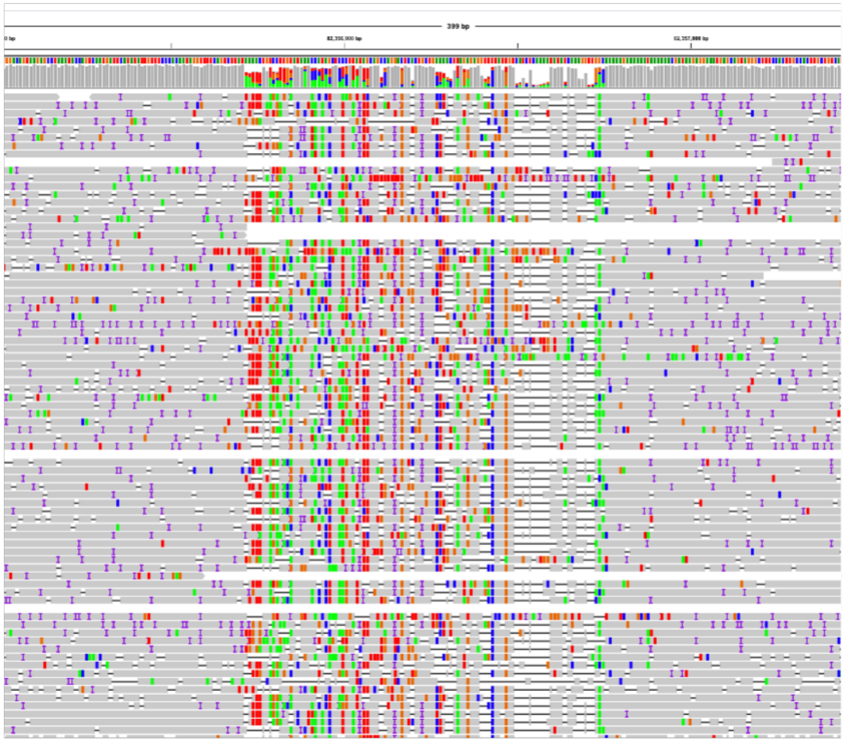


NGMLR:

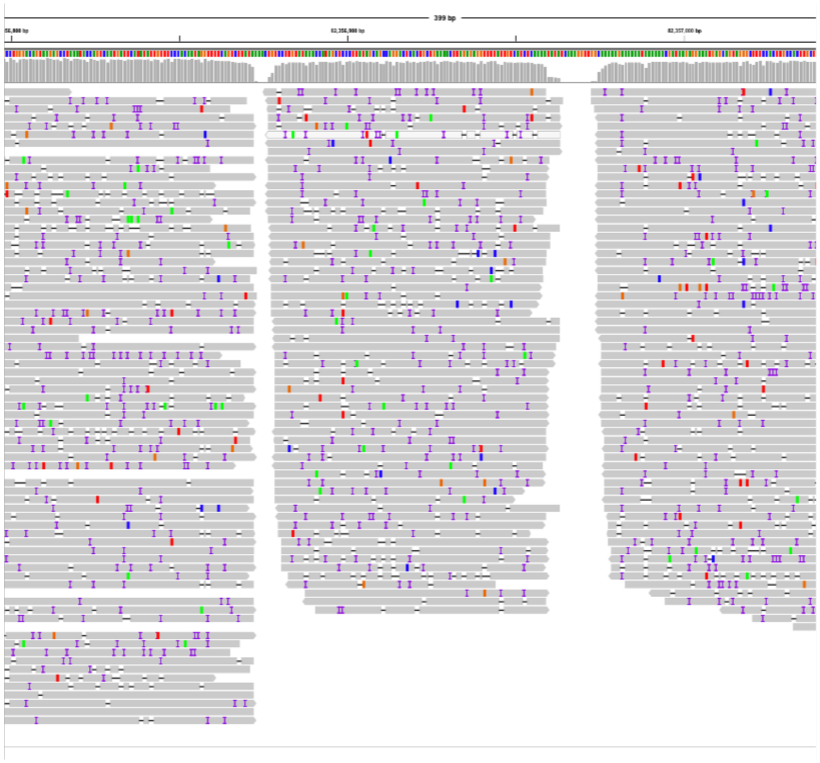


Why another mapper?

BWA-MEM:



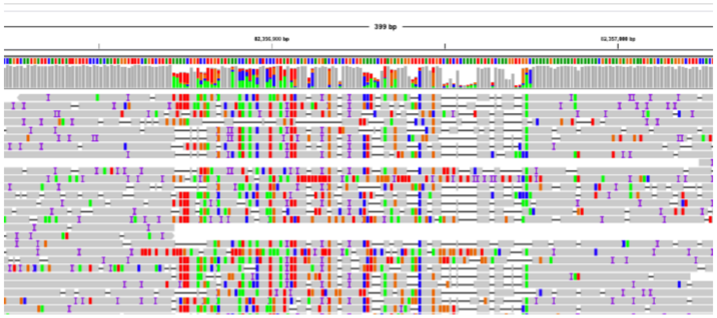
NGMLR:



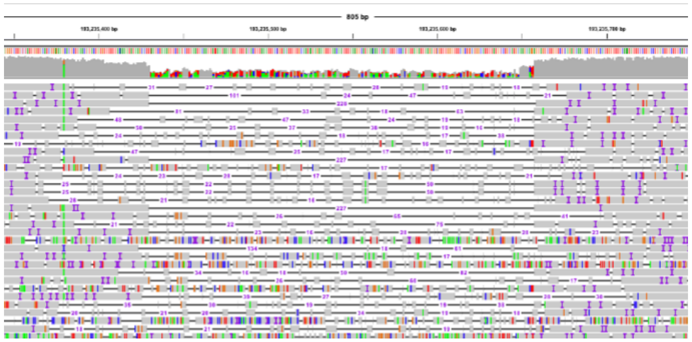
1. Improving long read alignment

Philipp Rescheneder

1. Split the reads:
    - Translocations
    - Inversions
    - Duplications

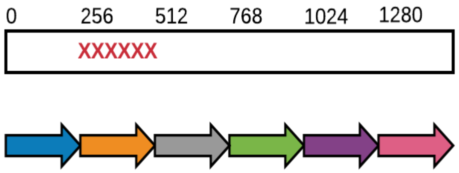


2. Improve alignment:
    - Insertions
    - Deletions

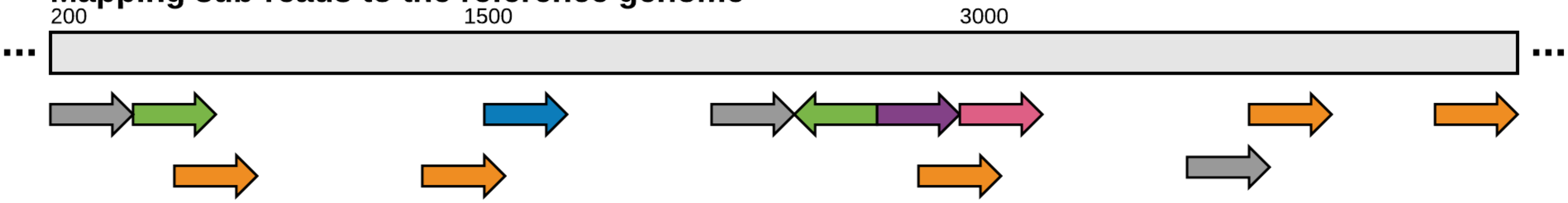


1.1 NGMLR: Split reads

Splitting read into sub-reads

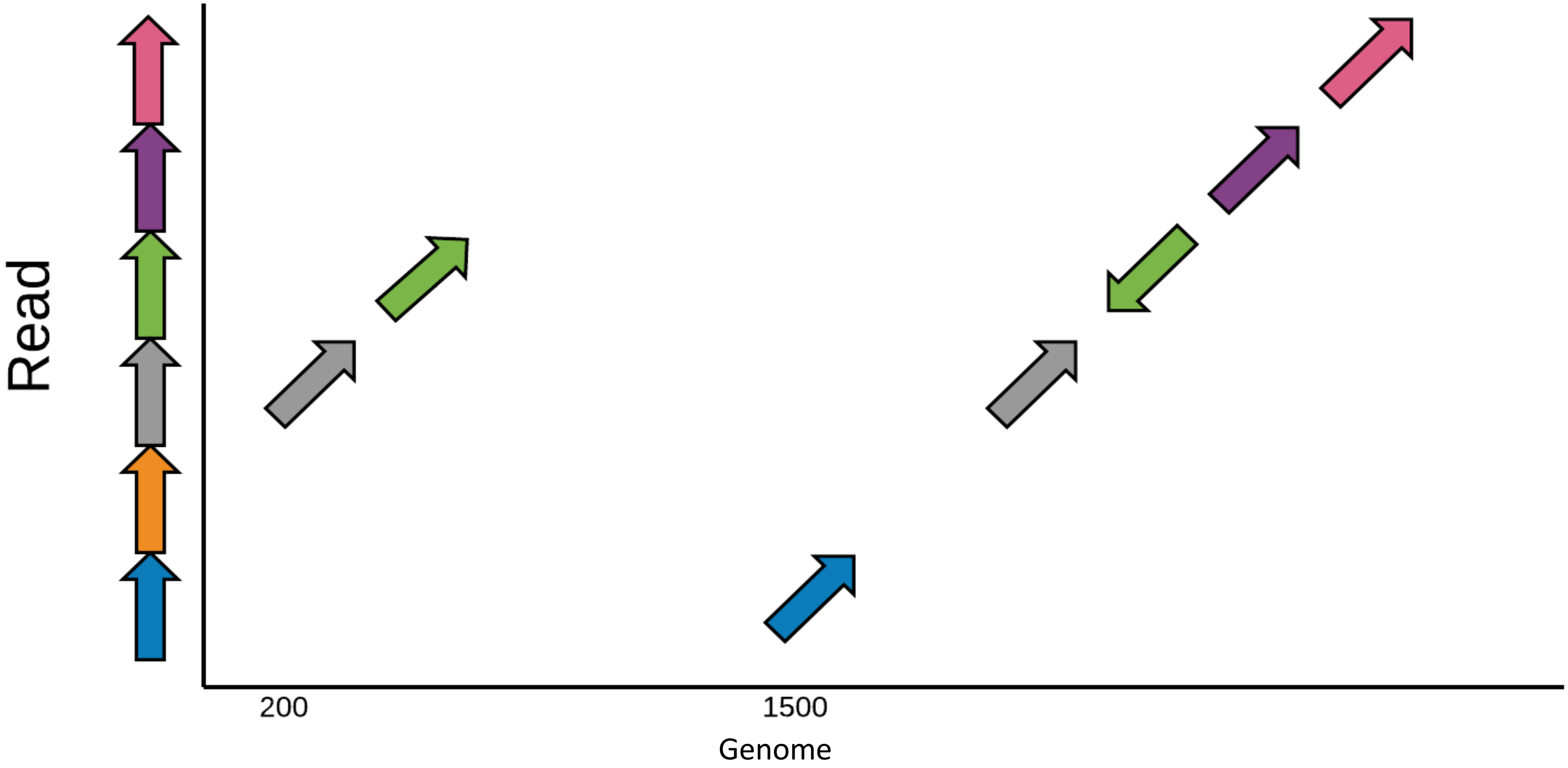


Mapping sub-reads to the reference genome

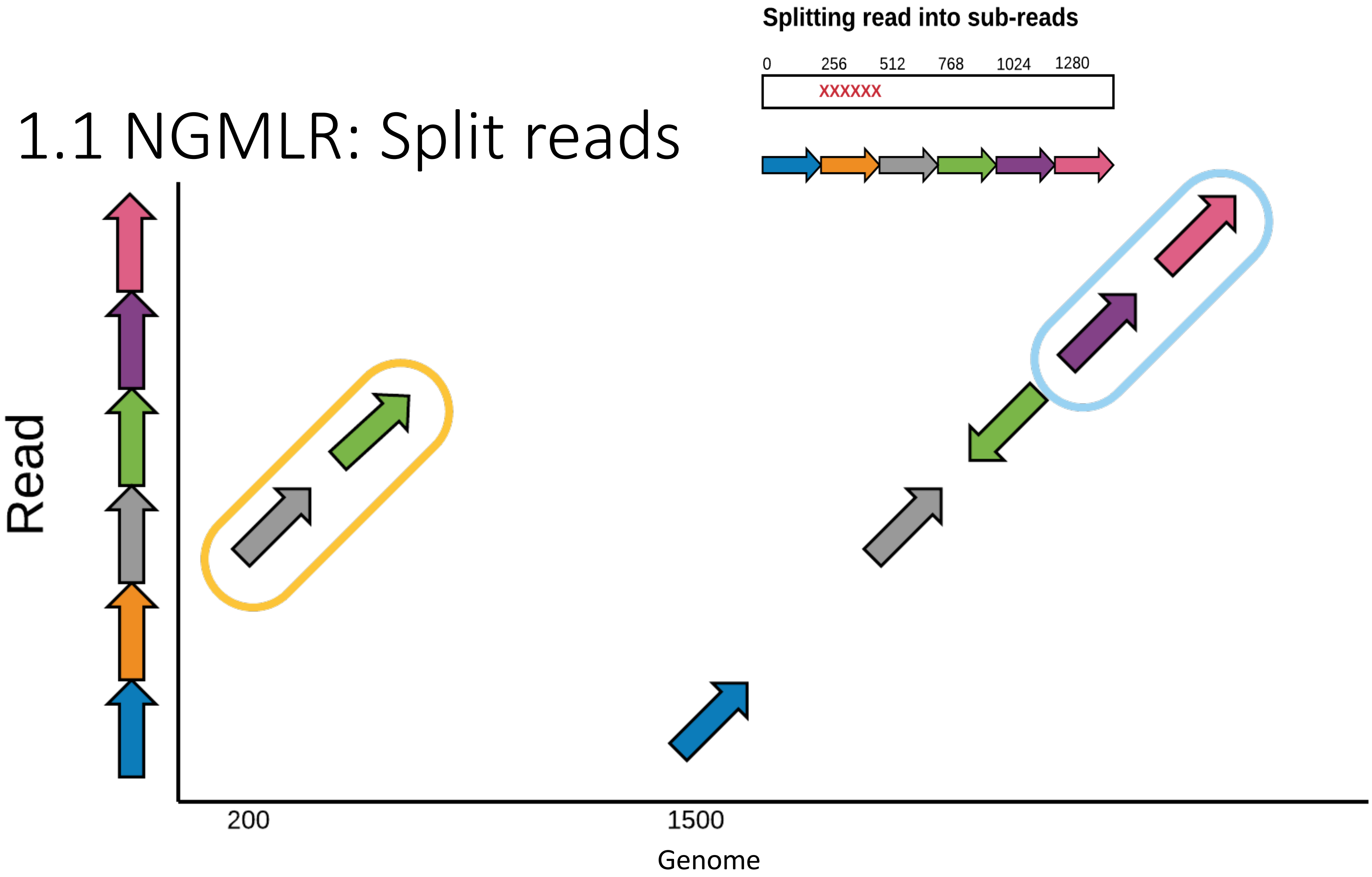


Score / Location:	1	2	3	4	5
🟦	99 / 1500				
🟧	~~56 / 521~~	~~56 / 1350~~	~~54 / 2930~~	~~52 / 4200~~	~~52 / 4697~~
⬜	110 / 2008	105 / 200	~~99 / 3989~~		
🟩	103 / 2260	101 / 450			
🟪	100 / 2740				
🟥	105 / 2998				

1.1 NGMLR: Split reads

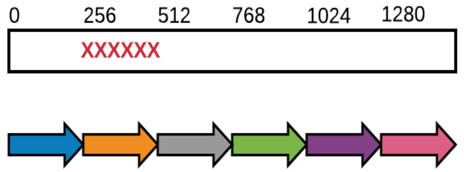


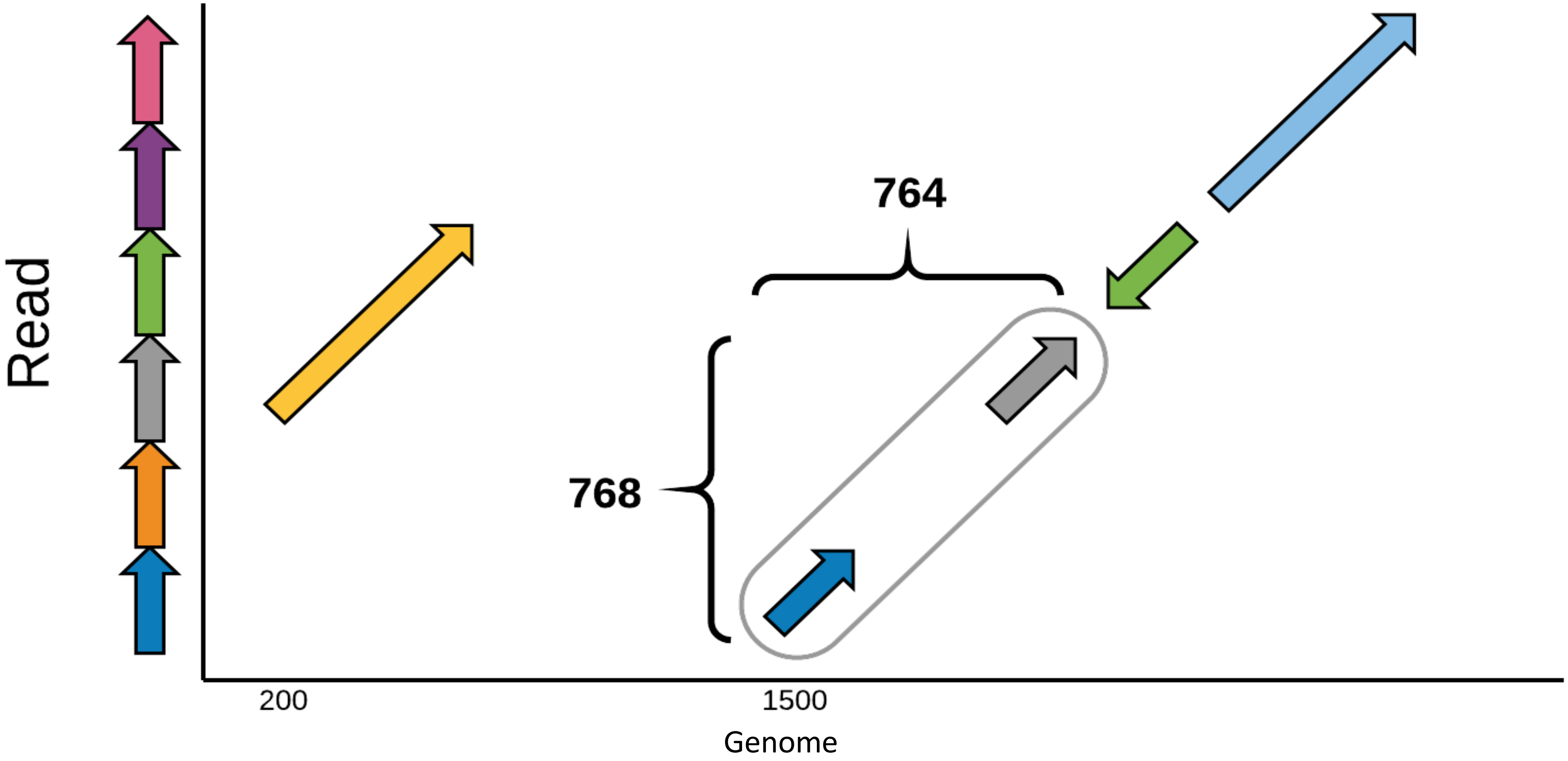
1.1 NGMLR: Split reads



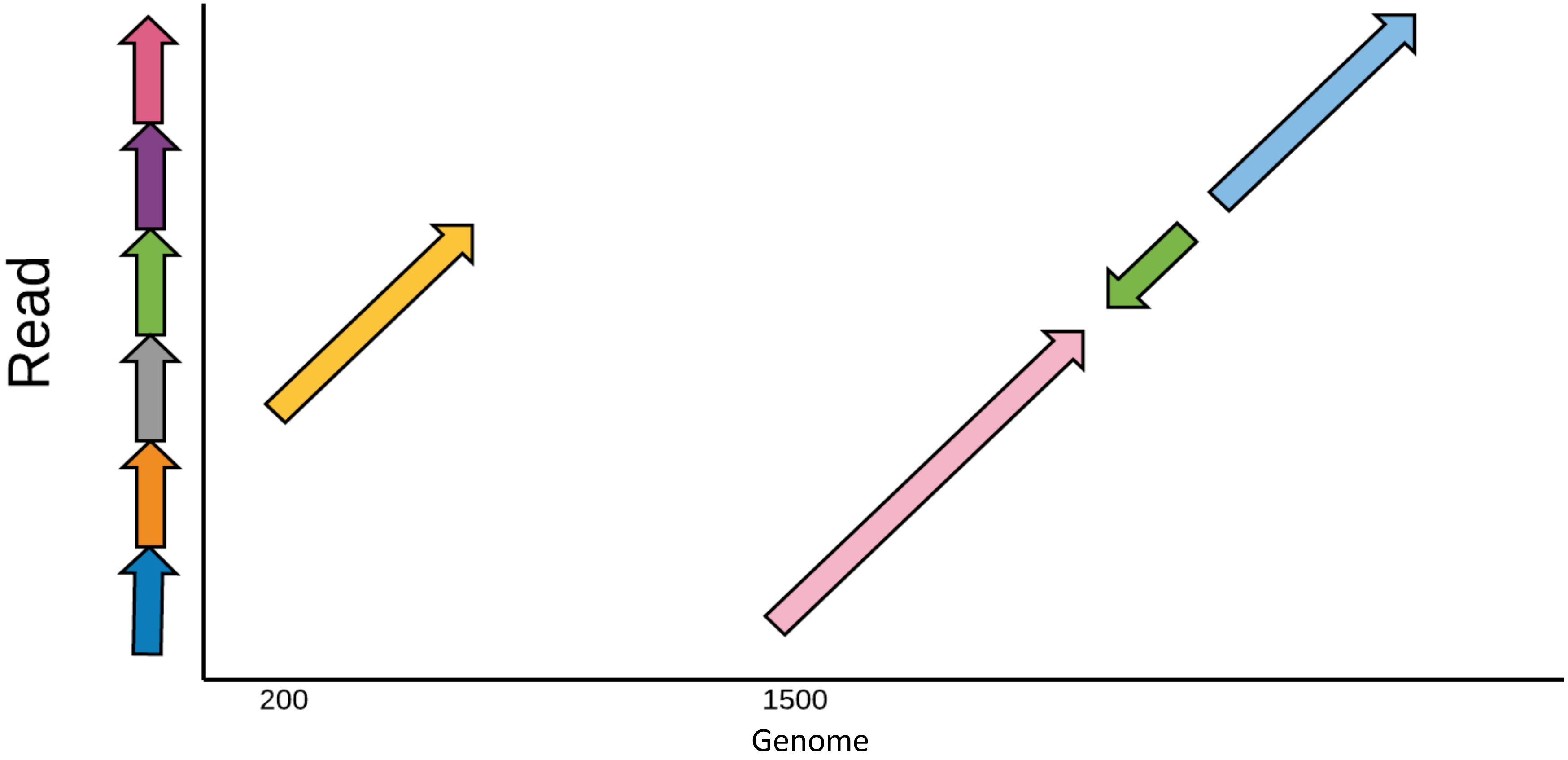
1.1 NGMLR: Split reads

Splitting read into sub-reads

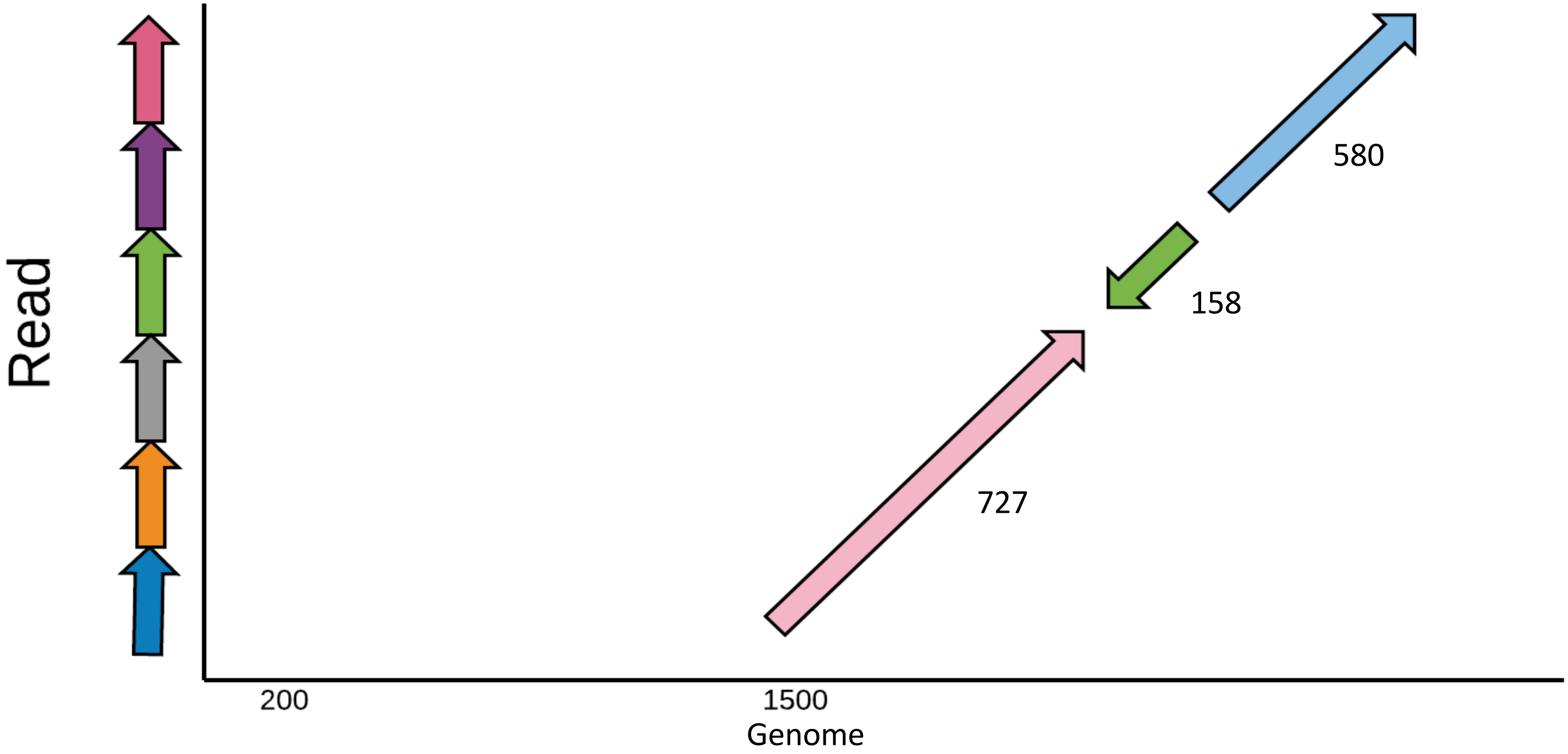




1.1 NGMLR: Split reads



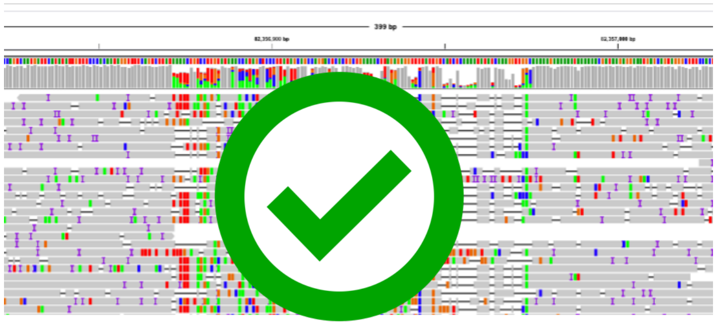
1.1 NGMLR: Split reads



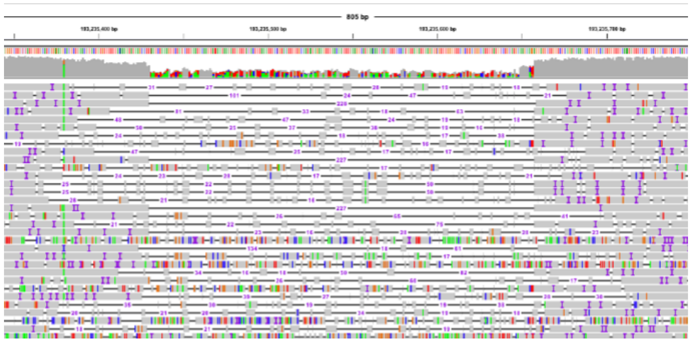
1. Improving long read alignment

Philipp Rescheneder

1. Split the reads:
    - Translocations
    - Inversions
    - Duplications

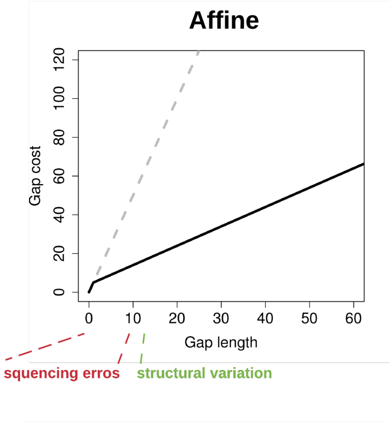


2. Improve alignment:
    - Insertions
    - Deletions



1.2 NGMLR: Alignments

- **Linear**: gap cost always the same
- **Affine**: separate penalties for opening and extending a gap
- **Convex**: initially similar to affine, but becomes proportionally less costly for larger gaps

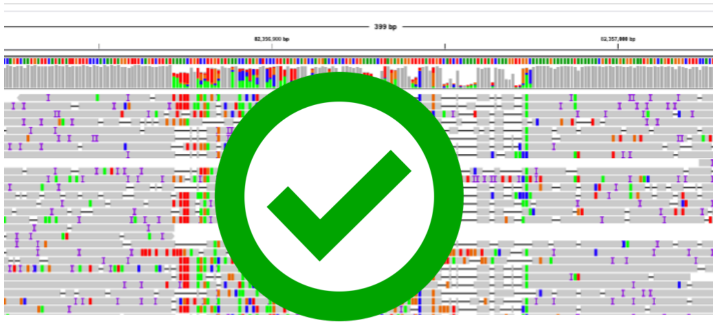


```
AA-GAATTCATAAGCAAACACTGG-TAAACTACT-C
AAAGA-T-CA----------CTGGGTA-ACTACTAC
                  =
AA-GAATTCATAAGCAAACACTGG-TAAACTACT-C
AAAGA-----T---CA----CTGGGTA-ACTACTAC
```

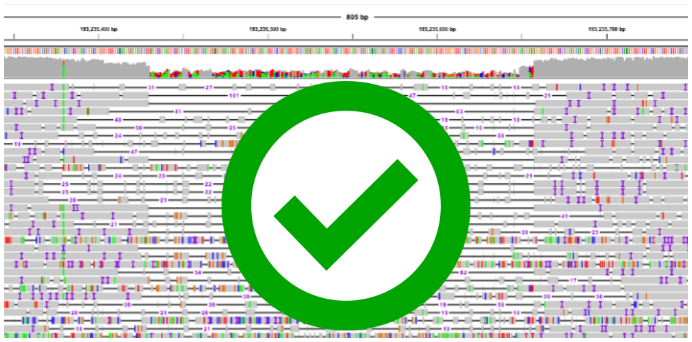
1.Improving long read alignment

Philipp Rescheneder

1. Split the reads:
   - Translocations
   - Inversions
   - Duplications

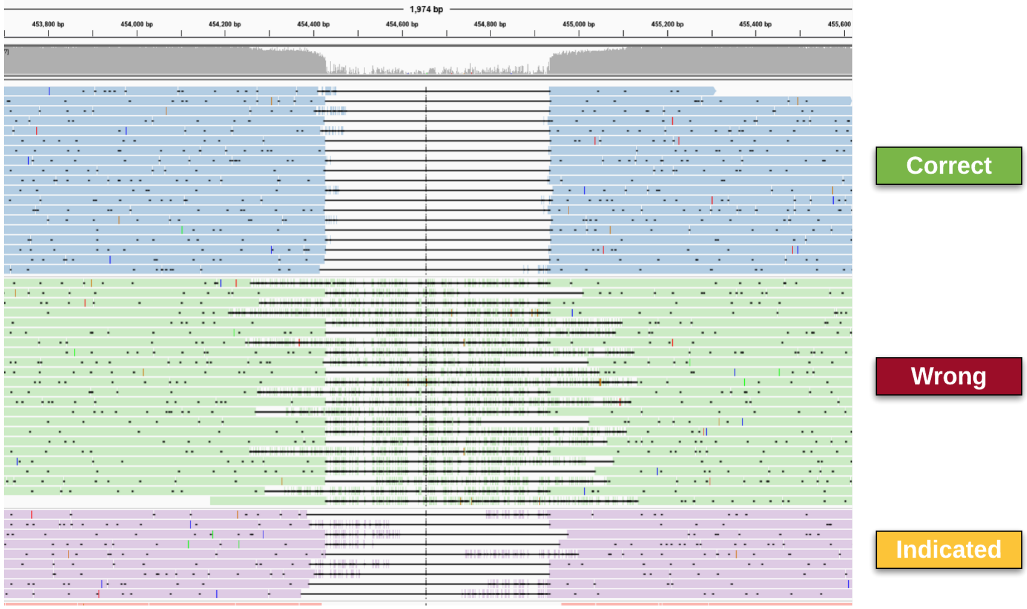


2. Improve alignment:
   - Insertions
   - Deletions



1.3 Simulations/ Evaluation

- Simulate 20 SVs of each type using SURVIVOR
- Simulate Pacbio like reads
- Evalutated:
  - BlasR
  - BWA-MEM
  - Graphmap
  - NGMLR



Correct

Wrong

Indicated

1.3 Results

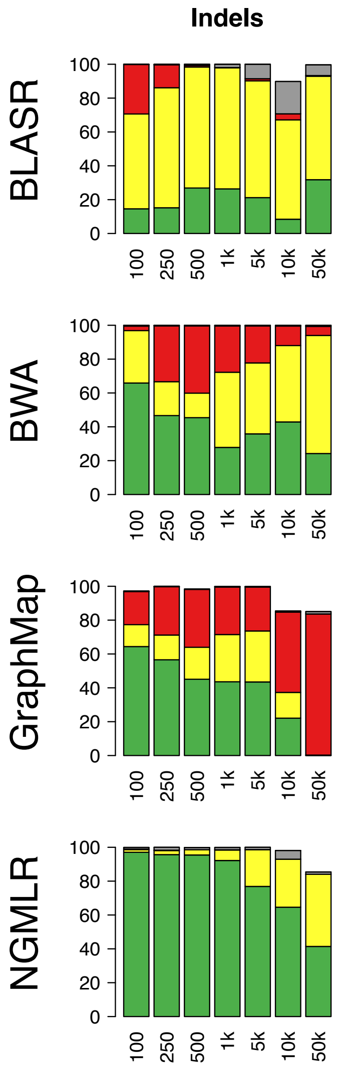
🟩 Precise

🟨 Indicated

🟥 Wrong

▩ Alignment stopped prior

□ Not aligned



How can we fully leverage this technology?

1. Improvement in mapping (NGMLR)
2. Improvement in SV calling (Sniffles)
3. Evaluation + results

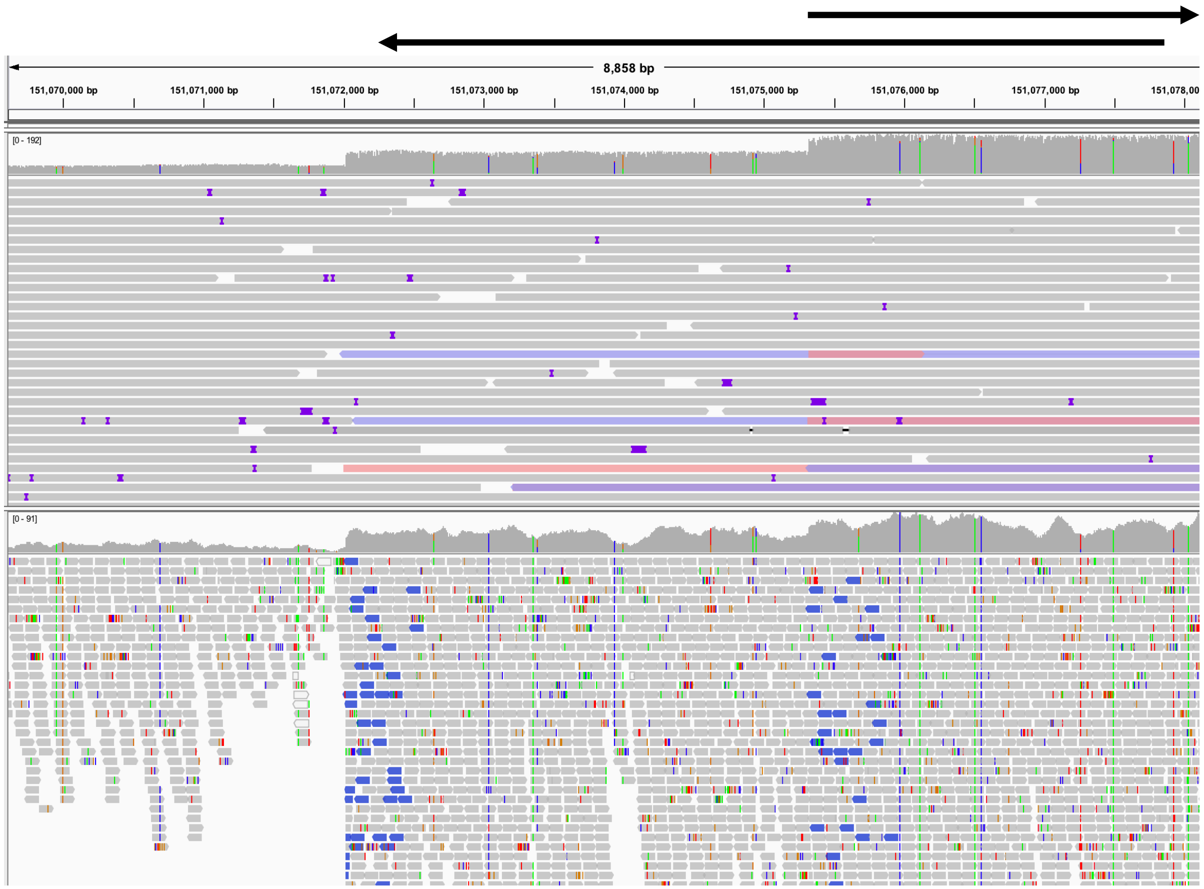
Why another SV caller?

Leverage technology:

- All types of SV:
  - DEL, DUP, INS, INV, TRA
- Cope with artifacts

Other types of variations:

- Inverted tandem duplication:
  - Pelizaeus-Merzbacher disease
  - MECP2
  - VIPR2
- Inversion flanked by deletions:
  - Haemophilia A



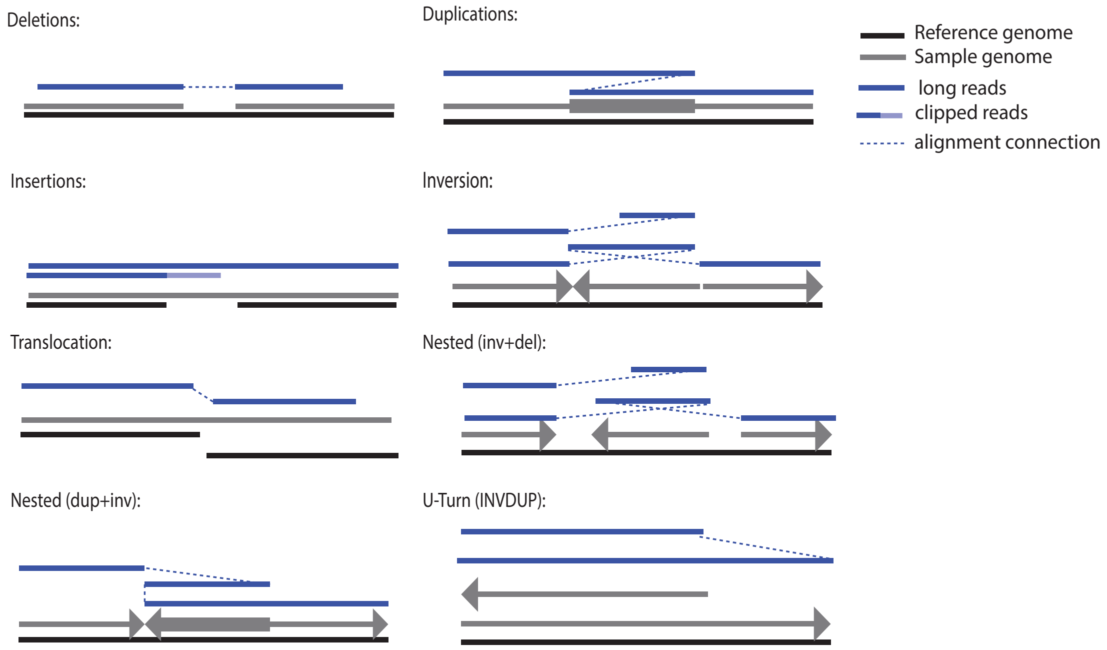
2. Sniffles

- Analyzing:
  - split reads
  - alignment events
  - noisy regions
- Parameter estimation
- Detect sequencing artifacts

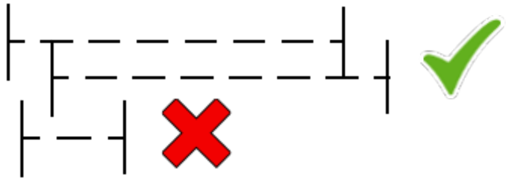
- Optional:
  - Genotype estimation
  - Clustering/phasing of SVs

2.1 Sniffles: Detection of SVs

Split the reads:



2.2 Sniffles: Clustering of SVs



When are two events the same?

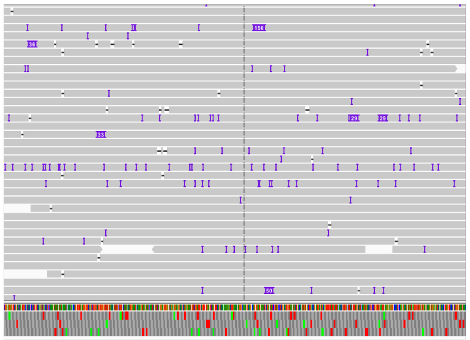
- Allowed distance depending on the size of the event

Detecting clustering of noise?

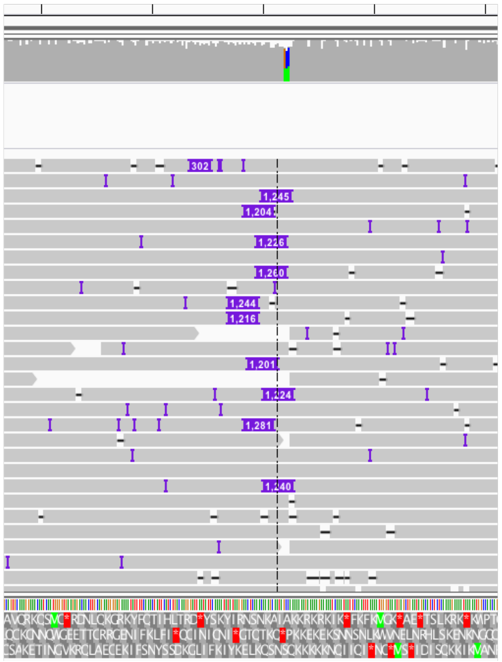
- Random appearance of Insertions (5-100bp)

- Standard deviation
    - Higher noise -> more likely artifact

Phantom insertion events:



Scattered events:



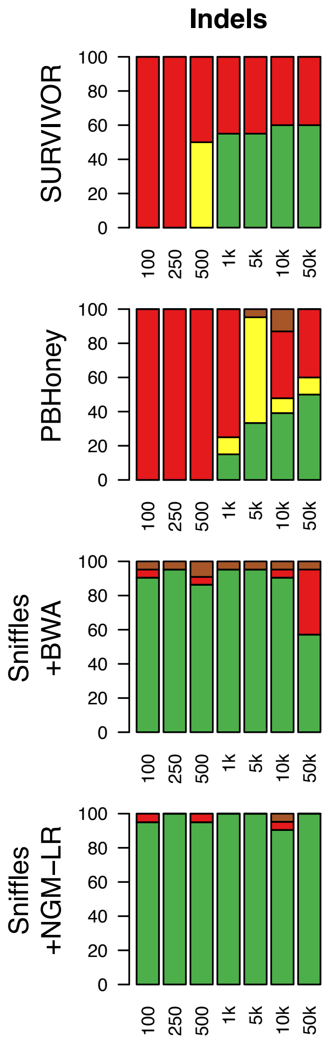
2.3 Results

🟩 Precise

🟨 Indicated

🟥 Not found

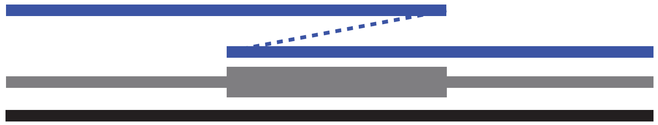
🟫 Additional events



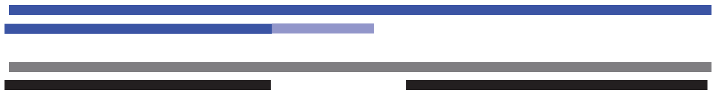
2.3 Results: Insertion vs. Duplication

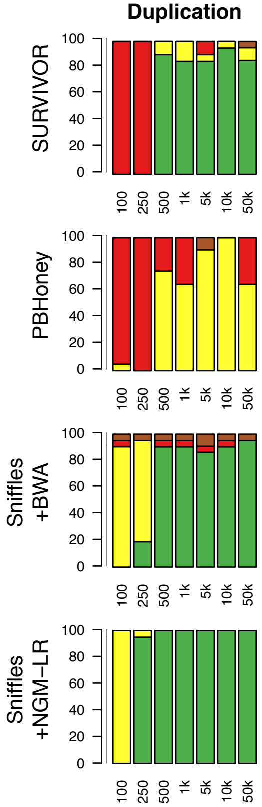
Tandem duplications are a insertion of the same sequence next to its original location

Duplications:



Insertions:





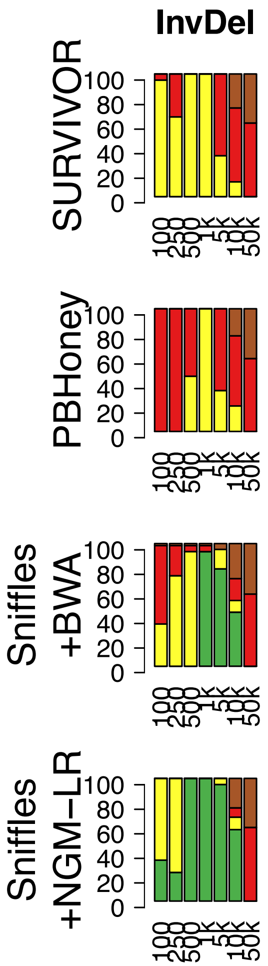
2.3 Results

🟩 Precise

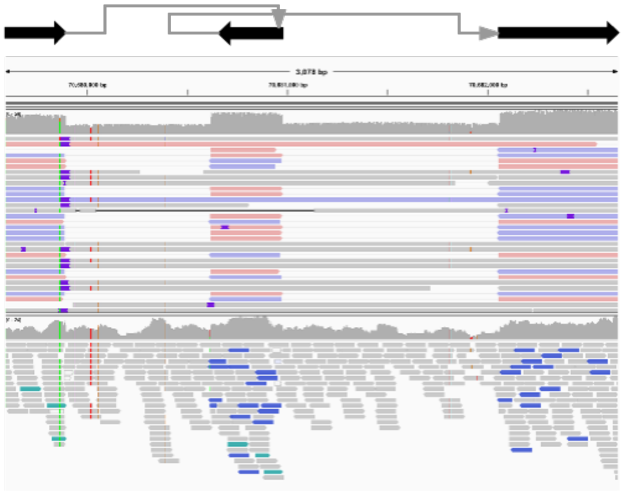
🟨 Indicated

🟥 Not found

🟫 Additional events



INVDEL



How can we fully leverage this technology?

1. Improvement in mapping (NGMLR)
2. Improvement in SV calling (Sniffles)
3. Evaluation + results

3.1 SURVIVOR

- Toolkit for:
    - Simulation + Evaluation SVs
    - Comparison + merging of multiple SVs call sets (vcf)
        - Consensus calling for short read data
    - Summarization of SVs calling + comparisons

Jeffares et. al. 2017

3.2 Arabidopsis trio

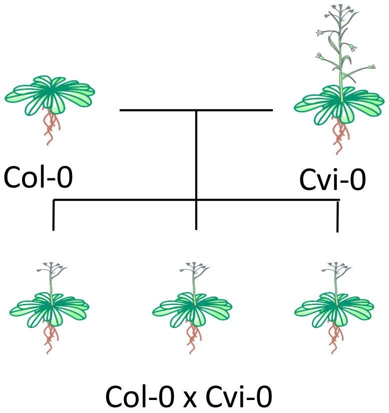


Image credits:
Pajoro, et al, Trends in plant science 21.1 (2016): 6-8.

Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
Col-0	127x	6,482	456	83	68	63	191	51
CVI	123x	6,073	15,966	6,922	421	416	6,496	1,711
COL-0 x CVI F1	155x	11,206	16,145	6,889	571	582	6340	1,763

3.2 Arabidopsis trio: Col-0 vs. F1

- 57 (Col-O) SVs homozygous

- 4 SVs initially missing:
  - 1 INS (47bp vs. 53bp) + 1 DEL (48bp vs. 53bp)
  - 1 Del + 1 DUP supported by only 4 reads

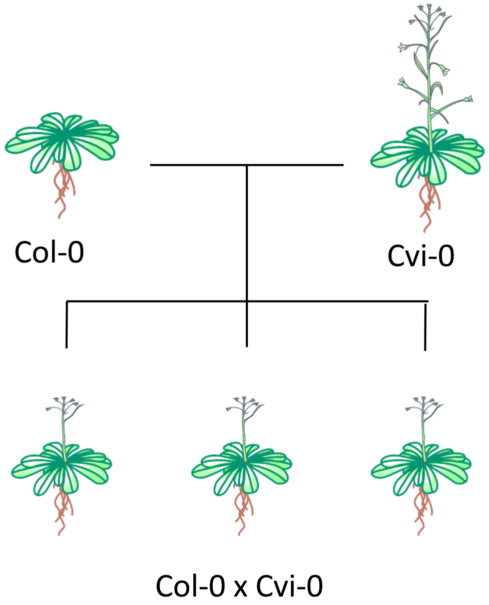


Image credits:
Pajoro, et al, Trends in plant science 21.1 (2016): 6-8.

3.2 Arabidopsis trio: CVI vs. F1

- 10,288 (CVI) SVs homozygous

- 370 (3.62%) initially missing:
  - 159 supporting read
  - 101 size threshold
  - 43 different types (e.g. transposons)
  - 50 COL unique region

- only 17 (0.17%) SVs could not be found!

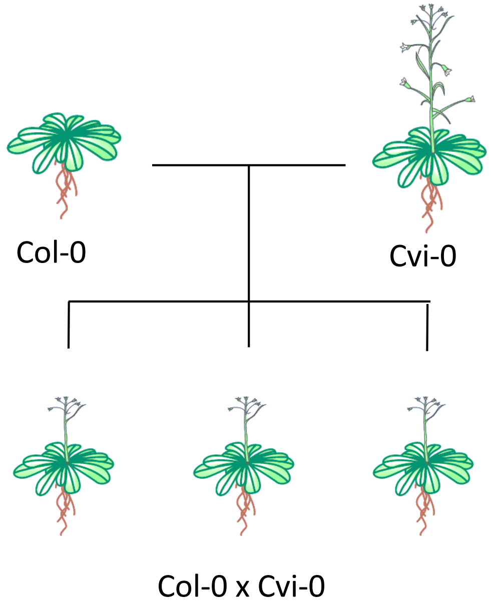


Image credits:
Pajoro, et al, Trends in plant science 21.1 (2016): 6-8.

3.3 NA12878

- Healthy female

- Gold standard in genomics

- Sequenced with many technologies independently:
  - Illumina, PacBio, Oxford Nanopore

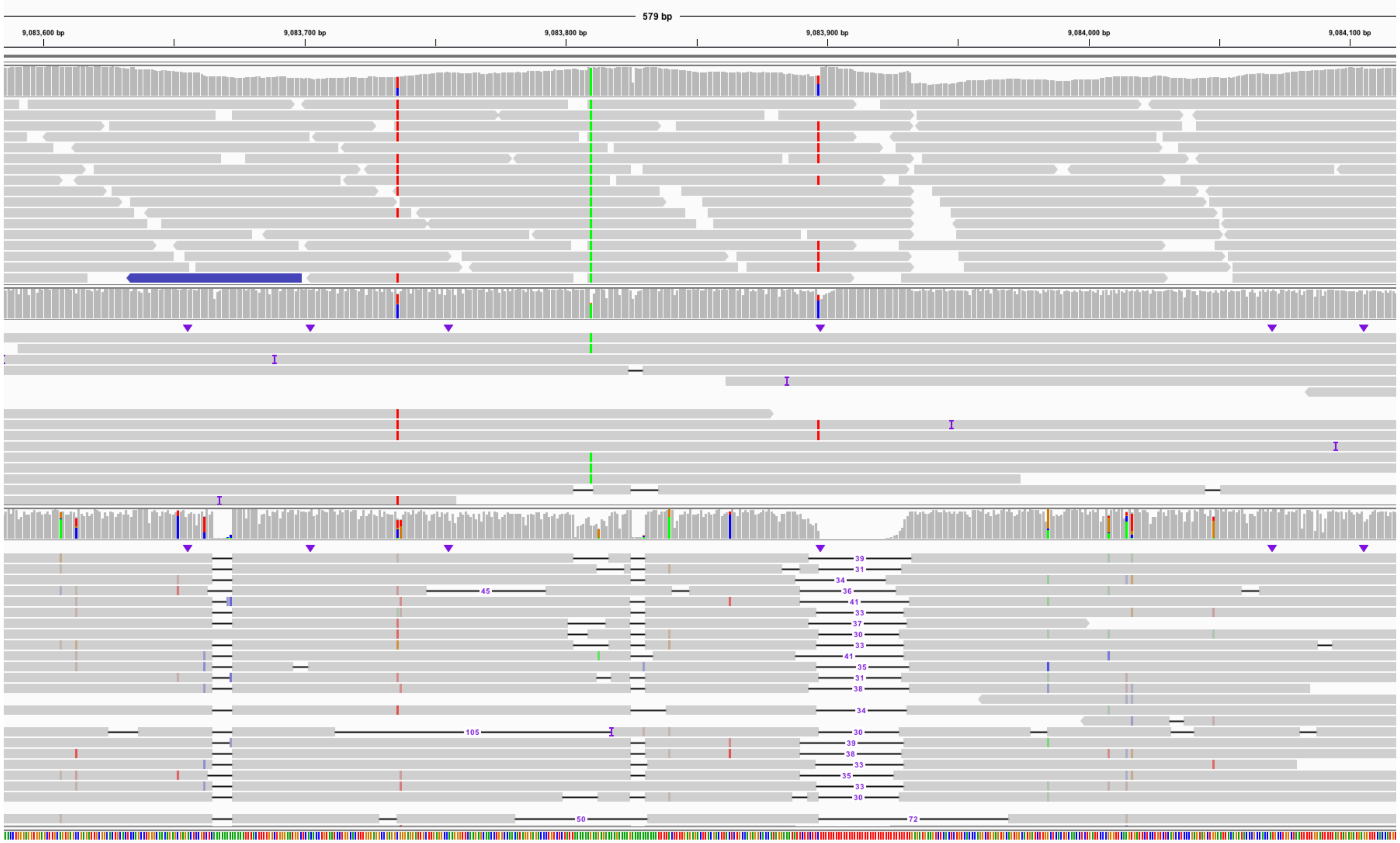
3.3 NA12878: SVs calling

Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	43
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

3.3 NA12878: SVs calling

Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	119
Oxford Nanopore	28x	6,432	32,409	**27,147**	87	323	4,809	43
Illumina	50x	2 x 101	7,275	3,744	731	553	0	2,247

3.3 Oxford Nanopore deletions!



illumina

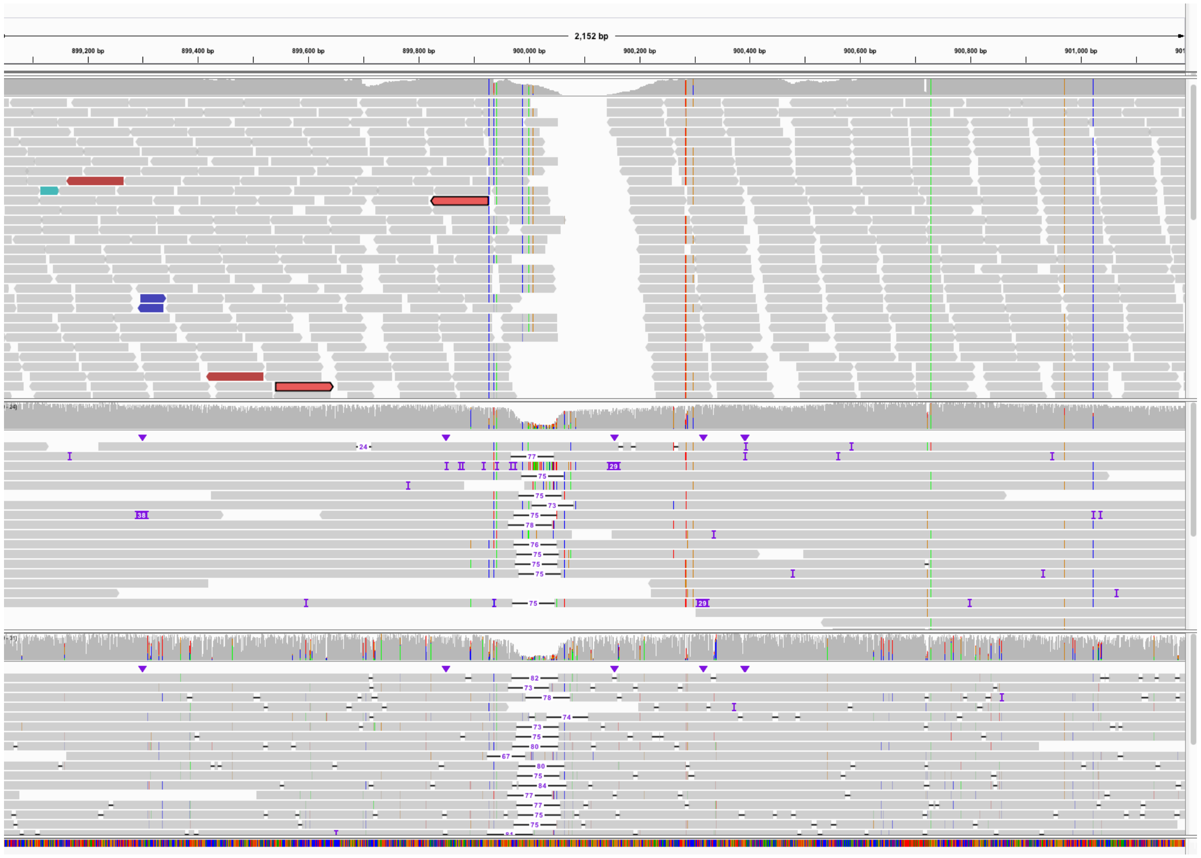
PacBio

Oxford Nanopore

3.3 NA12878: SVs calling

Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	**9,933**	162	611	**12,052**	119
Oxford Nanopore	28x	6,432	32,409	**27,147**	87	323	**4,809**	43
Illumina	50x	2 x 101	7,275	**3,744**	731	553	**0**	2,247

3.3 NA12878: Test for alterations in illumina



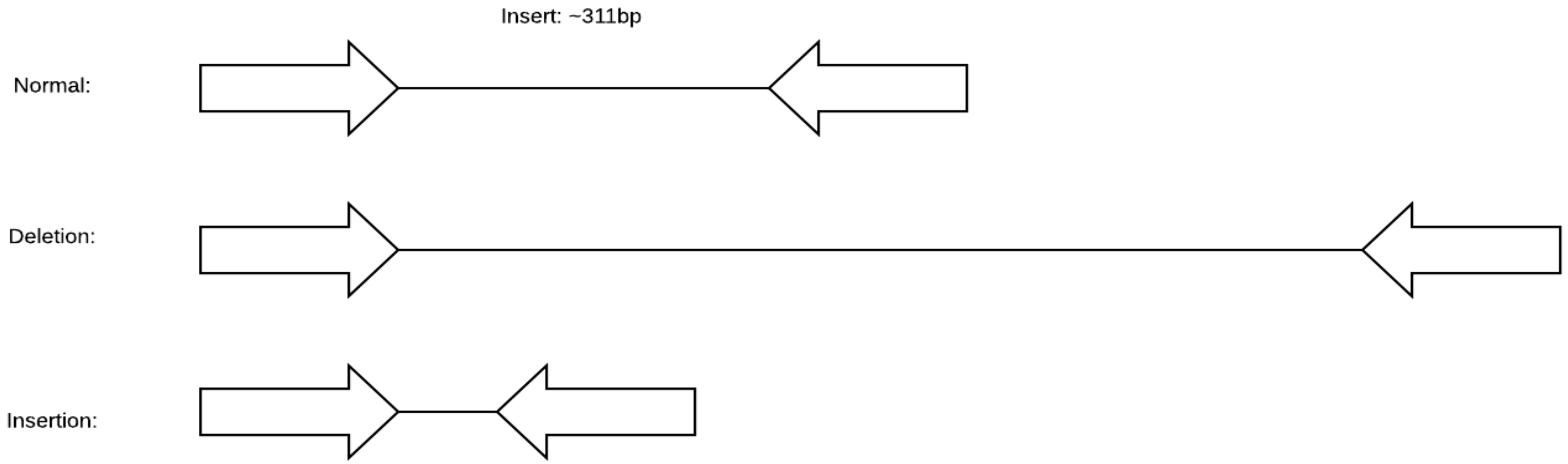
Illumina

PacBio

Nanopore

3.3 NA12878: Test for alterations in illumina

1. Measure insert sizes



2. Test for significant alterations (two sided T-test)
   - Deletions: 50bp -3kb
   - Insertions: 50bp-300bp

3.3 NA12878: Test for alterations in illumina

Tech.	Cov	DEL	INS	DEL (50bp-3kb)	INS (50bp-300bp)	Significant DEL	Significant INS
PacBio	55x	9,933	12,052	6,399	5,786	3,415	2,685
Oxford Nanopore	28x	27,147	4,809	12,045	3,488	3,879	1,703
Illumina	50x	3,744	0	3,102		1,873	

Significant: p<0.01

3.3 NA12878: Test for alterations in illumina

Tech.	Cov	DEL	INS	DEL (50bp-3kb)	INS (50bp-300bp)	Significant DEL	Significant INS
PacBio	55x	9,933	12,052	6,399	5,786	3,415	2,685
Oxford Nanopore	28x	27,147	4,809	12,045	3,488	3,879	1,703
Illumina	50x	3,744	0	3,102		1,873	

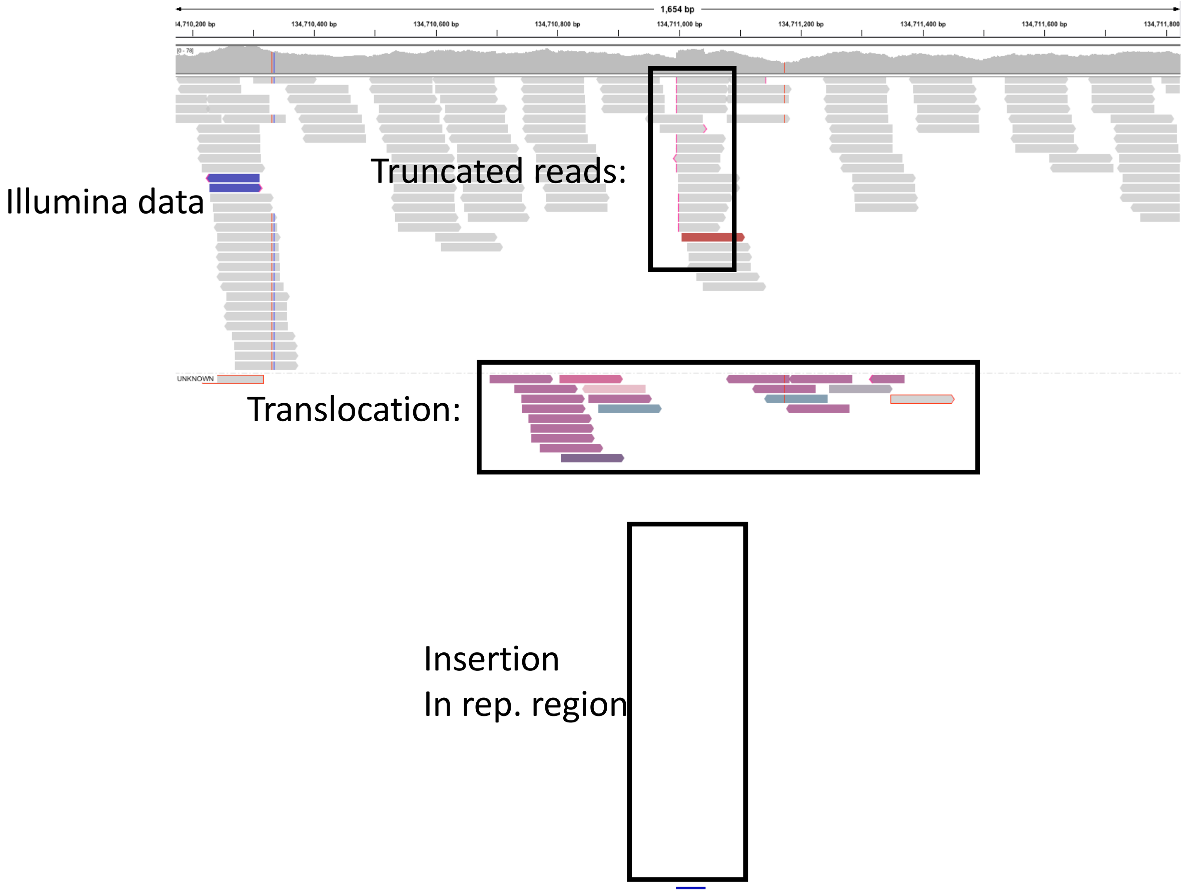
5,383 (84.12%) deletions and 2,719 (46.99%) insertions are supported by PacBio + Nanopore.

3.3 NA12878: SVs calling

Tech.	Cov.	Avg len	SVs	DEL	DUP	INV	INS	TRA
PacBio	55x	4,334	22,877	9,933	162	611	12,052	**119**
Oxford Nanopore	28x	6,432	32,409	27,147	87	323	4,809	**43**
Illumina	50x	2 x 101	7,275	3,744	731	553	0	**2,247**

3.3 NA12878: check 2,247 vs 119 TRA

Overlap	**Illumina TRA(%)**
Translocations	7.74
Insertions	53.05
Deletions	12.06
Duplications	0.57
Nested	0.31
High coverage	1.87
Low complexity	9.79
Explained	**85.40**



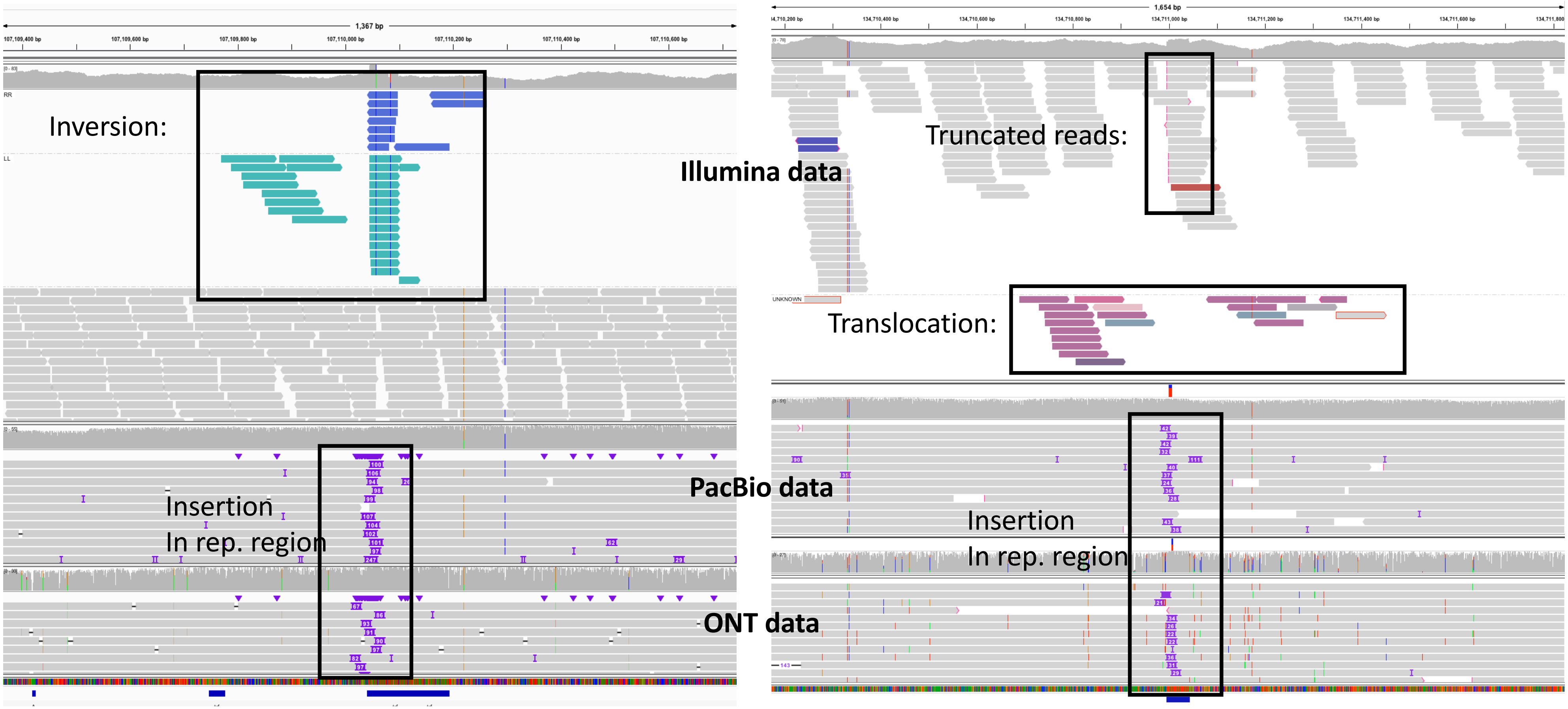
Illumina data

Truncated reads:

Translocation:

Insertion
In rep. region

3.3 NA12878: check 2,247 vs 119 TRA



Inversion:

Illumina data

Truncated reads:

Translocation:

PacBio data

Insertion

In rep. region

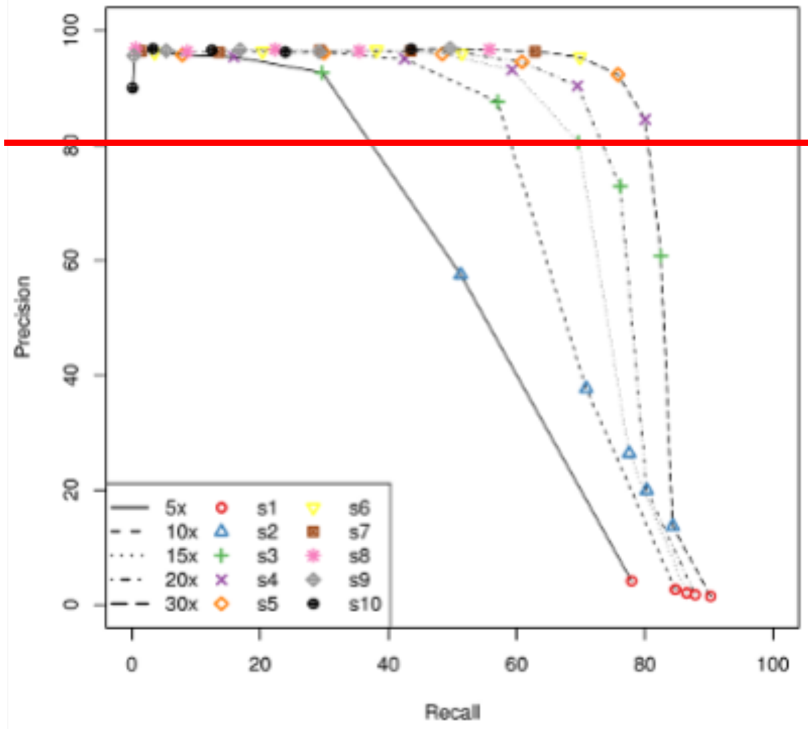
ONT data

Insertion

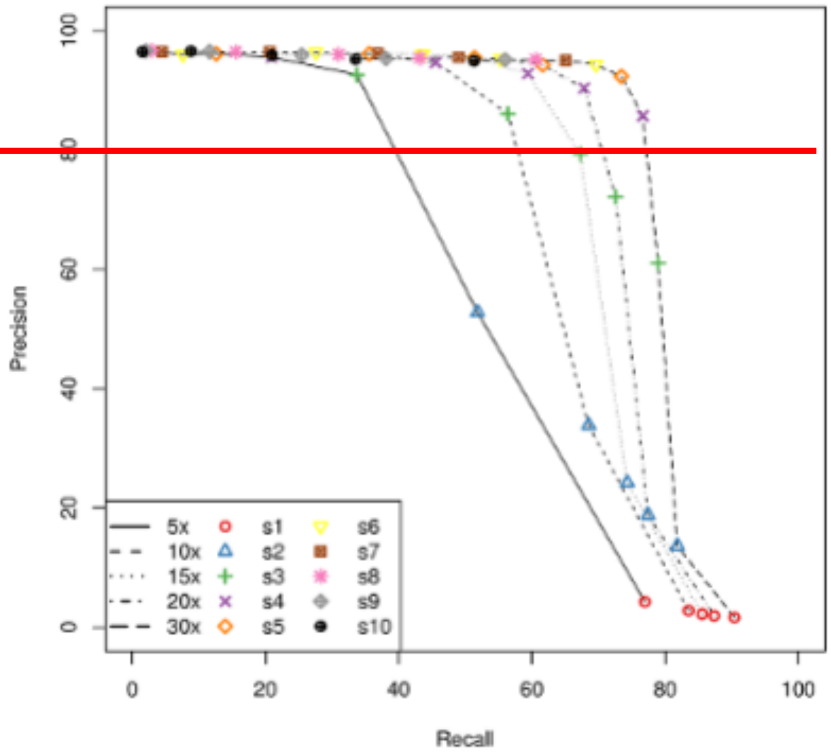
In rep. region

3.4 How much coverage do we need?

NA12878 (55x original)



SKBR3 (69x original)



Summary

- **My 3 wishes:**
  - **Don’t just pick subset of SV types**
  - **PacBio more + longer reads for less money**
  - **PacBio base calling**

- **Take home massage**
  - We can detect more small SVs and complex types
  - Biases in short read data + ONT
  - NGMLR + Sniffles: increase sensitivity, reduce FDR and required coverage

Methods

NextGenMap-LR:

- Long read mapper
- Manuscript in preparation
- Available:
  github.com/philres/nextgenmap-lr

NextGenMap

- Short read alignment
- Published: Bioinformatics (2013)
- Available:
  github.com/cibiv/NextGenMap

Sniffles:

- SVs detection for long reads
- Also nested SV
- Manuscript in preparation
- Available:
  github.com/fritzsedlazeck/Sniffles

SURVIVOR:

- Tool kit for SVs
- Published: Nature Communications (2017)
- Available:
  github.com/fritzsedlazeck/SURVIVOR

Acknowledgments

CSH Cold Spring Harbor Laboratory

Maria Nattestad

Han Fang

UCL

Daniel Jeffares
Jürg Bähler
Christophe Dessimoz

universität wien

Philipp Rescheneder
Moritz Smolka
Arndt von Haeseler

JOHNS HOPKINS UNIVERSITY

Michael Schatz
Schatz lab