How to sequence and analyze 100 tomato genomes in 100 days?

Fritz Sedlazeck

Jan, 15, 2019

Emerging technologies improve genomes

REVIEWS

COMPUTATIONAL TOOLS

Piercing the dark matter: bioinformatics of long-range sequencing and mapping

Fritz J. Sedlazeck[1], Hayan Lee[2], Charlotte A. Darby[3] and Michael C. Schatz[3,4]*

BRIEF REPORT | Genetics inMedicine

Long-read genome sequencing identifies causal structural variation in a Mendelian disease

Jason D. Merker, MD, PhD[1,2], Aaron M. Wenger, PhD[3], Tam Sneddon, DPhil[2],

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Detection of GBA missense mutations and other variants using the Oxford Nanopore MinION

Melissa Leija-Salazar, Fritz J Sedlazeck, Katya Mokretar, Stephen Mullin, Marco Toffoli, Maria Athanasopoulou,

New Results

Multi-platform discovery of haplotype-resolved structural variation in human genomes

Mark J.P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J.

New Results

Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line

Maria Nattestad, Sara Goodwin, Karen Ng, Timour Baslan, Fritz J. Sedlazeck, Philipp Rescheneder, Tyler Garvin,

BRIEF COMMUNICATIONS

Detecting DNA cytosine methylation using nanopore sequencing

Jared T Simpson[1,2], Rachael E Workman[3], P C Zuzarte[1], Matei David[1], L J Dursi[1] & Winston Timp[3]

single-molecule detection[5]. Enzymatic treatment to convert 5-mC to 5-carboxylcytosine via tet methylcytosine dioxygenase 1 (Tet1) can enhance this signal[6]; however, others have found this conversion to be incomplete and context dependent[7].

Previous work has shown that 5-mC can be distinguished from cytosine by careful analysis of the electrical current signals measured by nanopore-based sequencing devices[8,9]. Here we describe the development of a method to directly detect DNA modifications, in this case 5-mC, based on only the electrical readout of

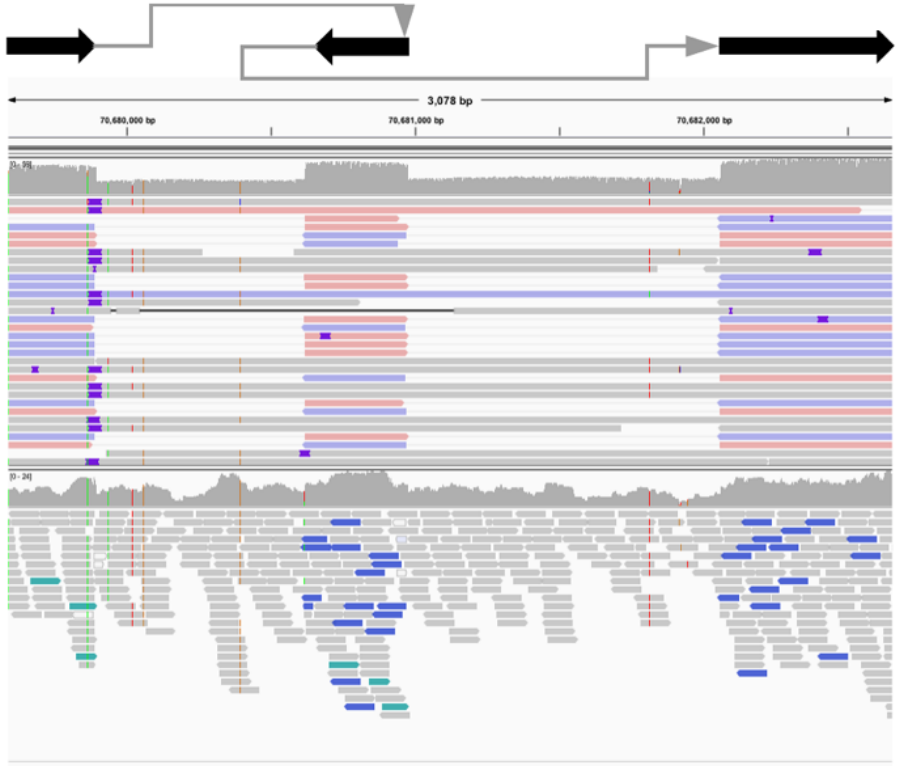
Highly parallel direct RNA sequencing on an array of nanopores

Daniel R Garalde[1], Elizabeth A Snell[1], Daniel Jachimowicz[1], Botond Sipos[1], Joseph H Lloyd[1], Mark Bruce[1], Nadia Pantic[1], Tigist Admassu[1], Phillip James[1], Anthony Warland[1], Michael Jordan[1], Jonah Ciccone[1], Sabrina Serra[1], Jemma Keenan[1], Samuel Martin[1], Luke McNeill[1], E Jayne Wallace[1], Lakmal Jayasinghe[1], Chris Wright[1], Javier Blasco[1], Stephen Young[1], Denise Brocklebank[1], Sissel Juul[2], James Clarke[1], Andrew J Heron[1] & Daniel J Turner[1]

Identification of complex SVs with long reads

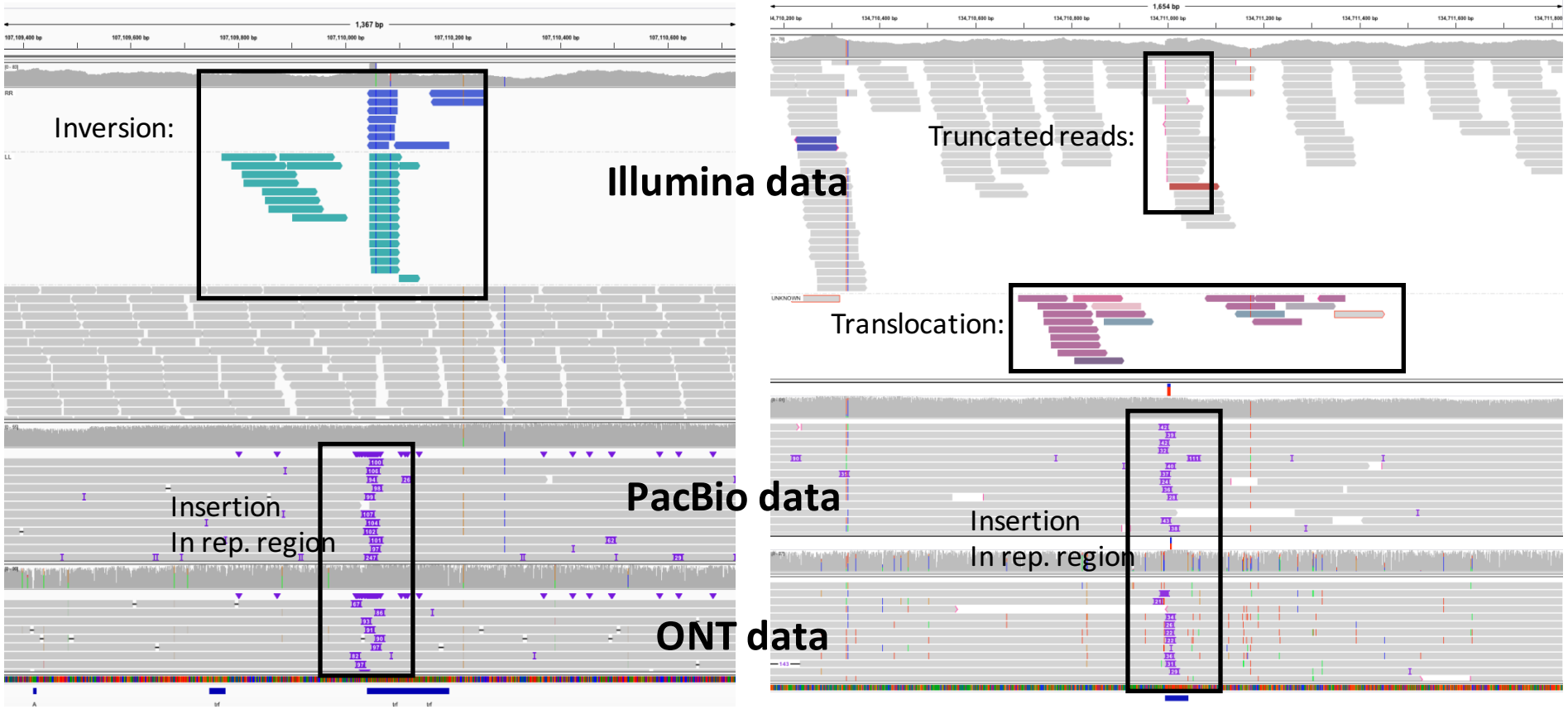
Inversion flanked by deletions:

- Haemophilia A
- Only found over long range PCR! (2007)



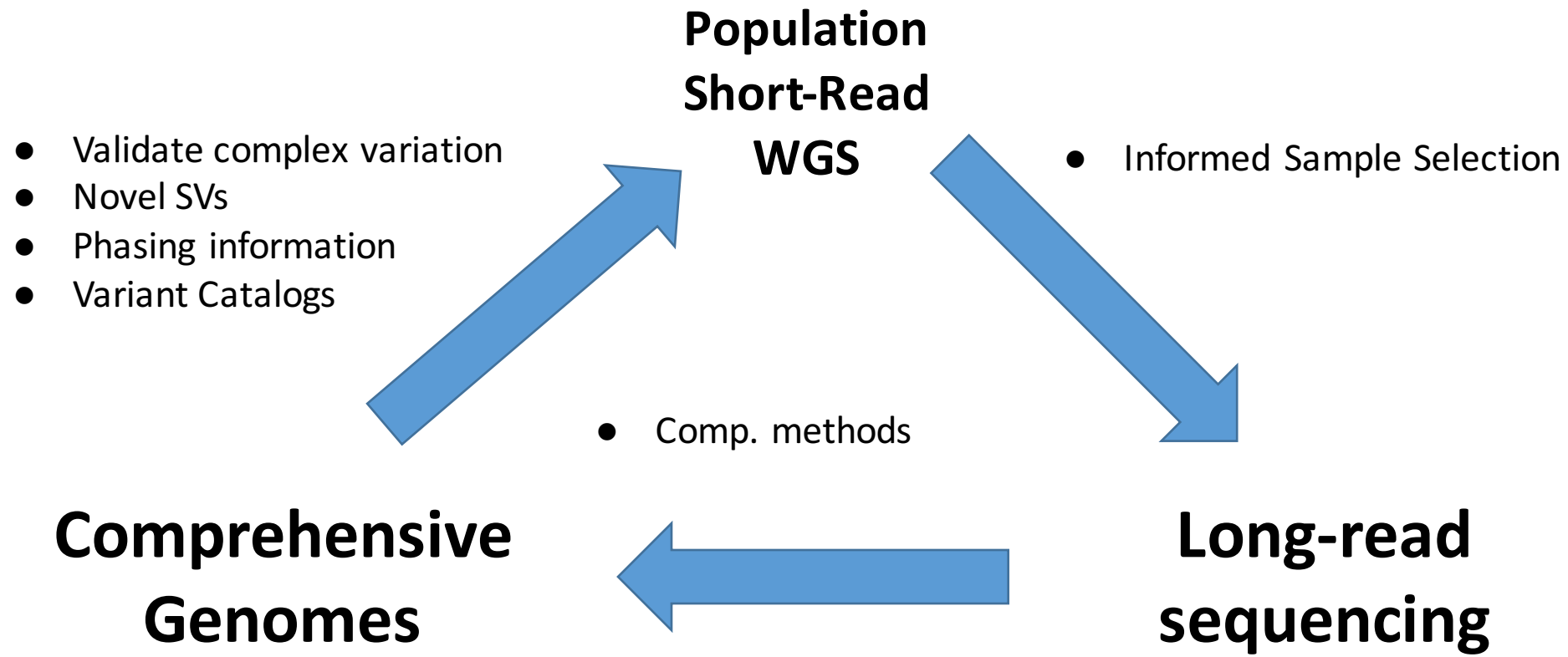
Sedlazeck et al. Nature Methods (2018)

Short-read validation / False Positives



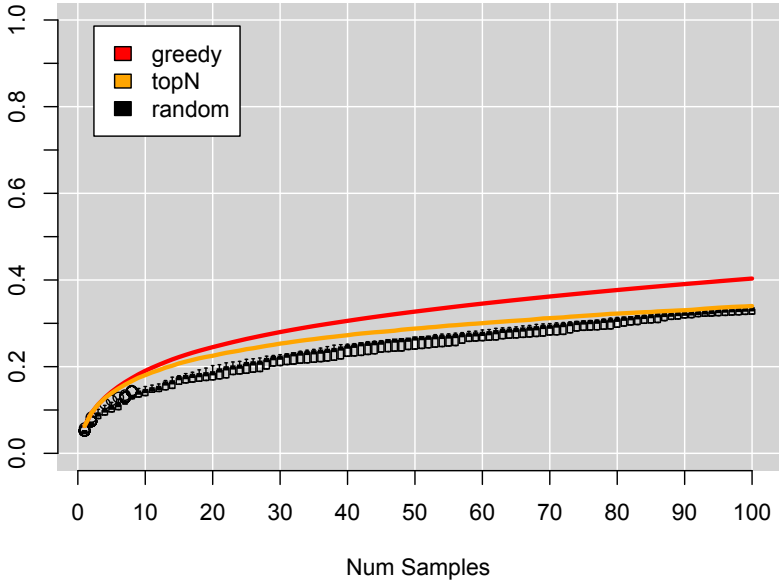
Sedlazeck et al. Nature Methods (2018)

How can we leverage these technologies in larger cohorts?



4. How to select samples: SVCollector

1000 Genomes



Population	TopN	SVCollector
AFR	99	60
SAS	0	16
EAS	0	14
EUR	0	6
AMR	1	4
Subpopulation	30.77%	96.15%

Sedlazeck et al. (bioarchive)

SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	1	0	0	1
2	0	1	1	1
3	1	1	0	1
4	0	0	1	0
5	1	1	0	1
6	0	0	1	0
7	0	1	1	0
8	0	0	1	1
9	1	1	0	1
10	1	1	0	0
11	1	0	0	1

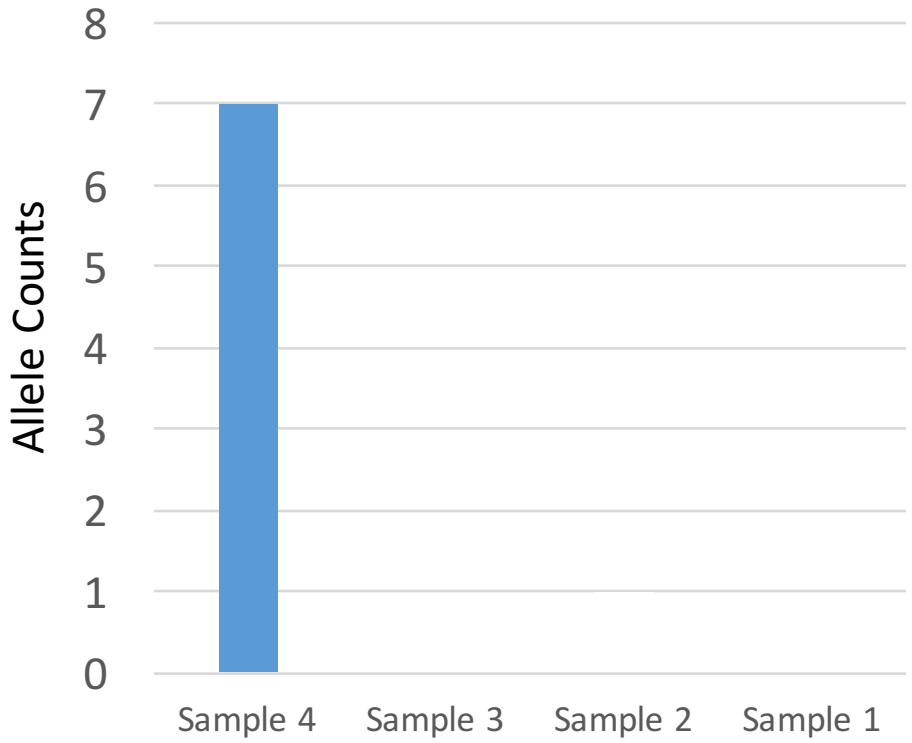
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	1	0	0	1
2	0	1	1	1
3	1	1	0	1
4	0	0	1	0
5	1	1	0	1
6	0	0	1	0
7	0	1	1	0
8	0	0	1	1
9	1	1	0	1
10	1	1	0	0
11	1	0	0	1

SUM:	6	6	5	7

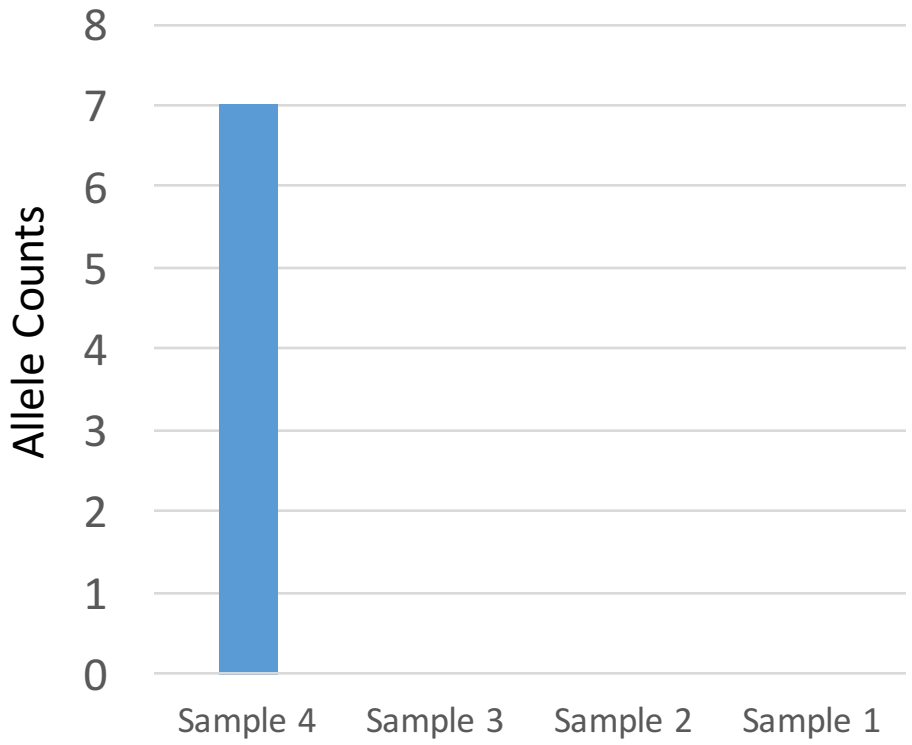
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	1	0	0	1
2	0	1	1	1
3	1	1	0	1
4	0	0	1	0
5	1	1	0	1
6	0	0	1	0
7	0	1	1	0
8	0	0	1	1
9	1	1	0	1
10	1	1	0	0
11	1	0	0	1
SUM:	**6**	**6**	**5**	**7**



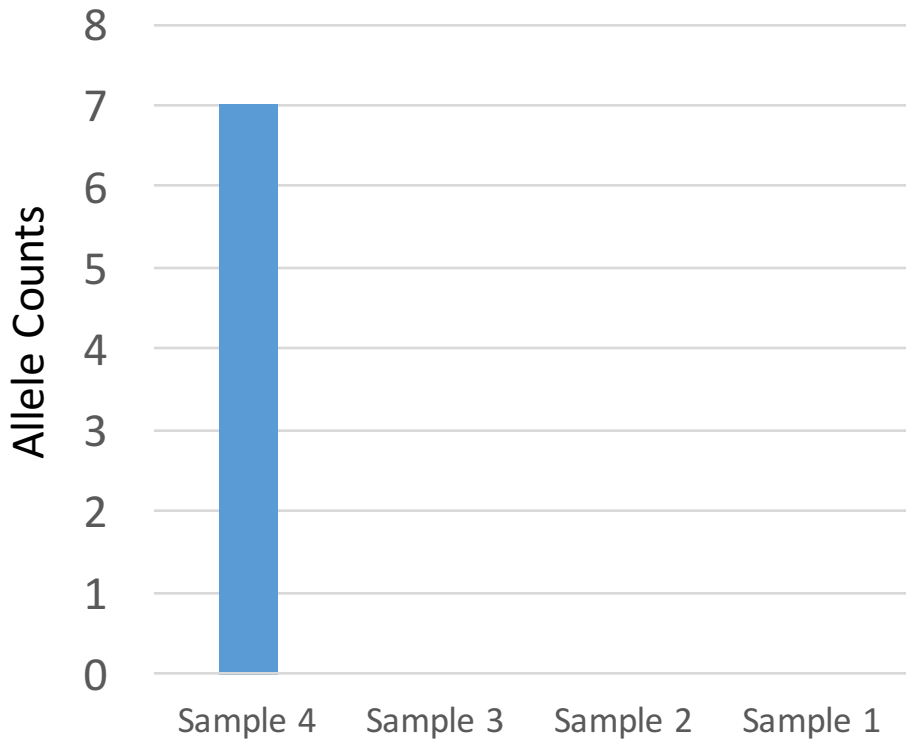
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	1	0	0	1
2	0	1	1	1
3	1	1	0	1
4	0	0	1	0
5	1	1	0	1
6	0	0	1	0
7	0	1	1	0
8	0	0	1	1
9	1	1	0	1
10	1	1	0	0
11	1	0	0	1
SUM:	**6**	**6**	**5**	**7**



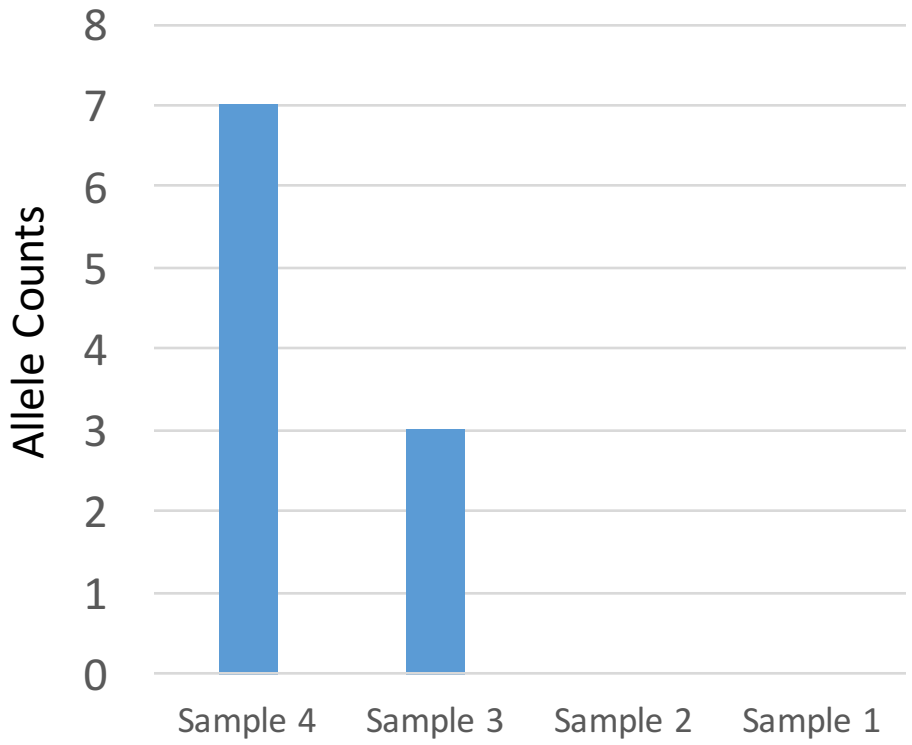
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	-	0	0	1
2	0	-	-	1
3	-	-	0	1
4	0	0	1	0
5	-	-	0	1
6	0	0	1	0
7	0	1	1	0
8	0	0	-	1
9	-	-	0	1
10	1	1	0	0
11	-	0	0	1
SUM:	**6**	**6**	**5**	**7**



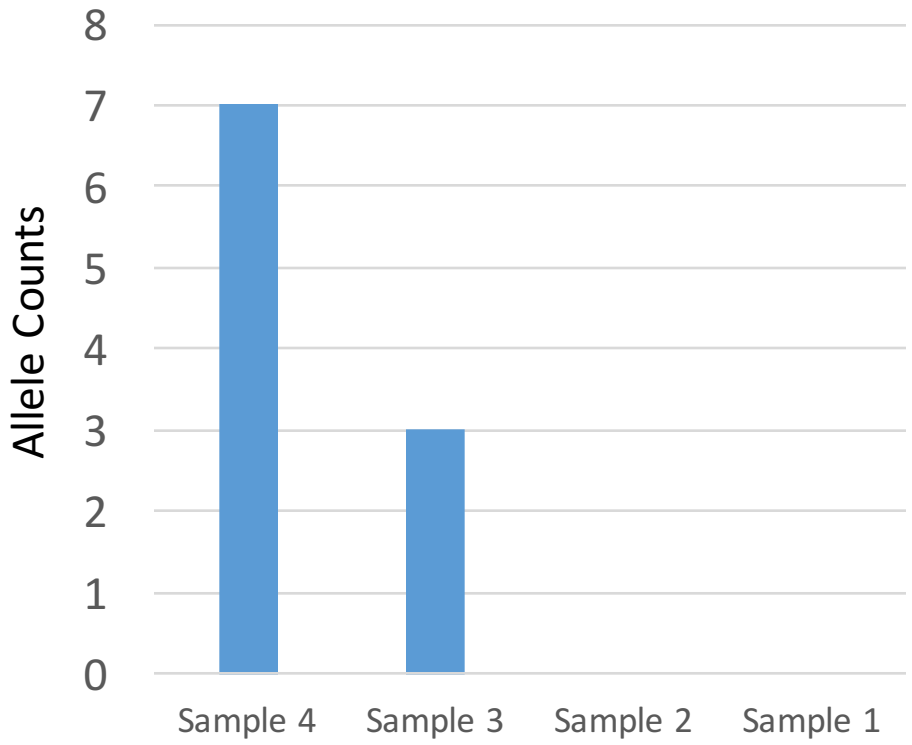
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	-	0	0	1
2	0	-	-	1
3	-	-	0	1
4	0	0	1	0
5	-	-	0	1
6	0	0	1	0
7	0	1	1	0
8	0	0	-	1
9	-	-	0	1
10	1	1	0	0
11	-	0	0	1
SUM:	**1**	**2**	**3**	**0**



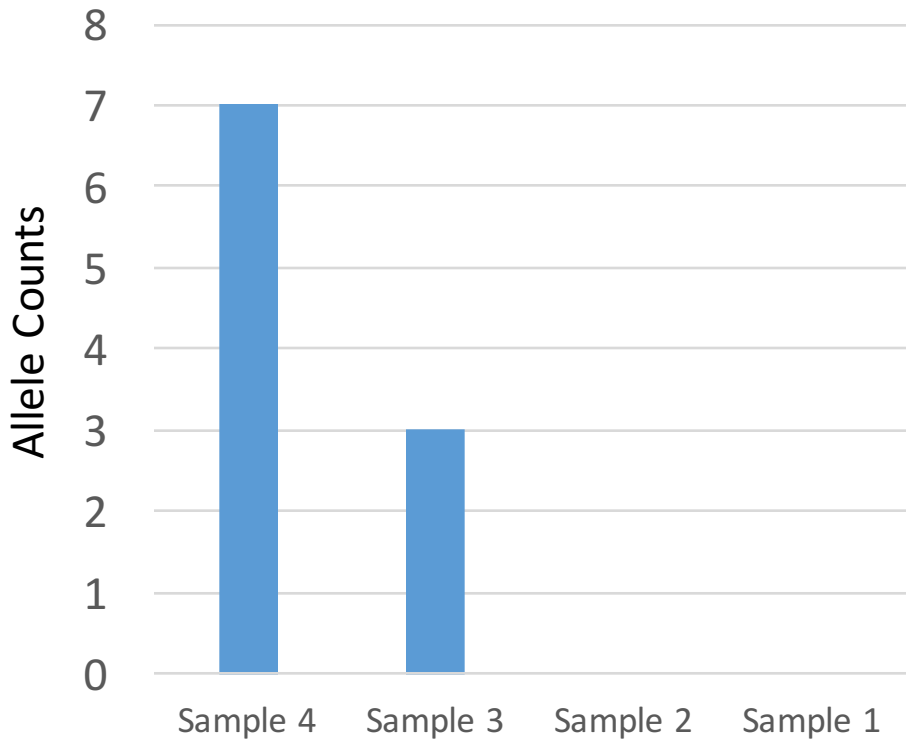
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	-	0	0	1
2	0	-	-	1
3	-	-	0	1
4	0	0	1	0
5	-	-	0	1
6	0	0	1	0
7	0	1	1	0
8	0	0	-	1
9	-	-	0	1
10	1	1	0	0
11	-	0	0	1
SUM:	**1**	**2**	**3**	**0**



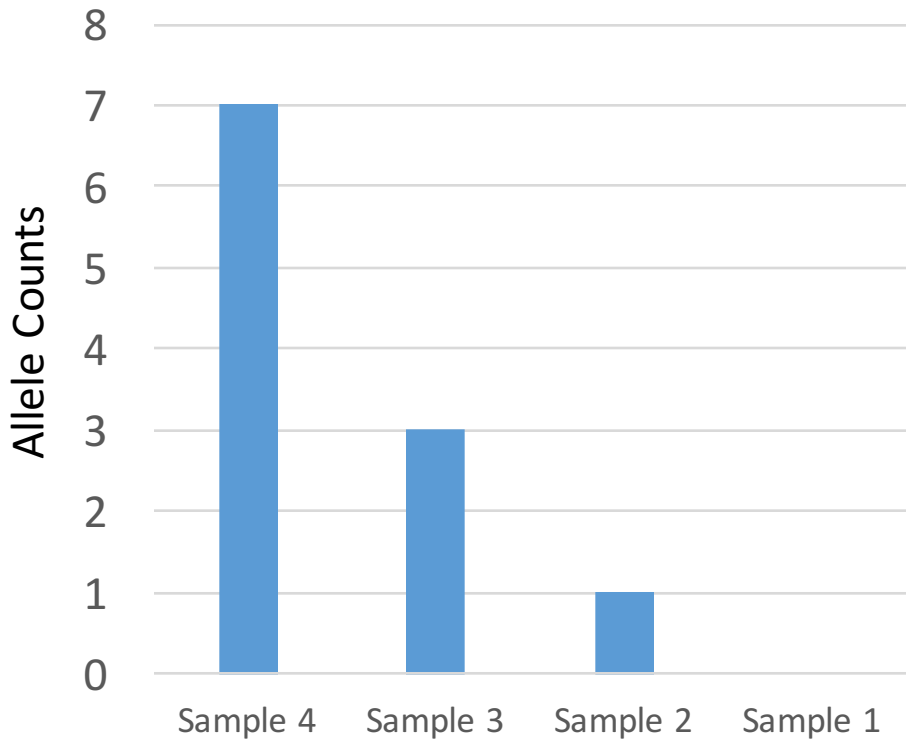
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	-	0	0	1
2	0	-	-	1
3	-	-	0	1
4	0	0	1	0
5	-	-	0	1
6	0	0	1	0
7	0	-	1	0
8	0	0	-	1
9	-	-	0	1
10	1	1	0	0
11	-	0	0	1
SUM:	**1**	**2**	**3**	**0**



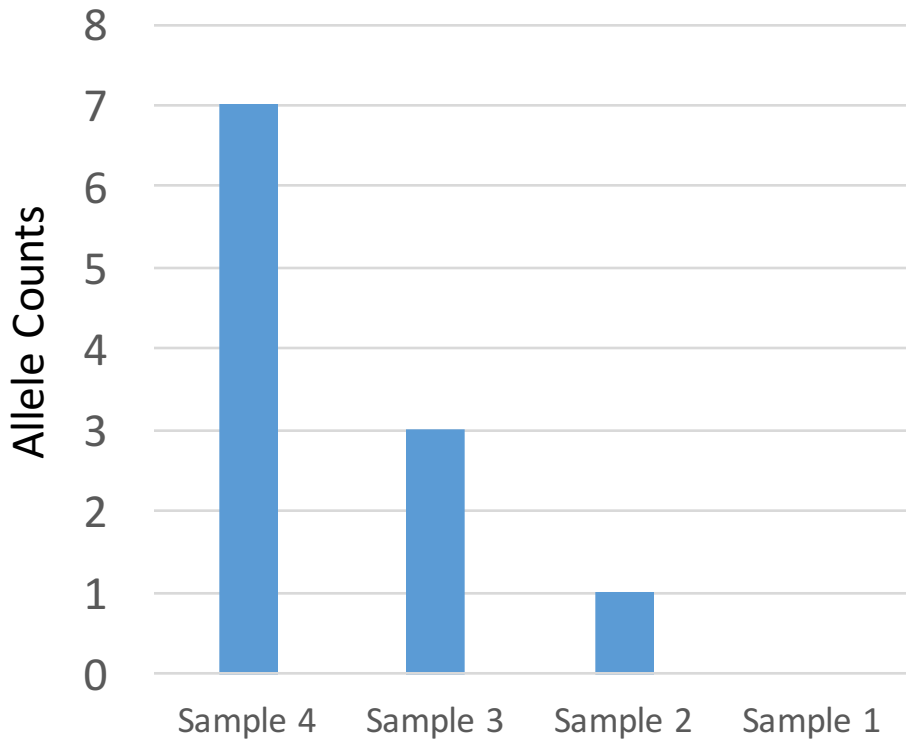
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	-	0	0	1
2	0	-	-	1
3	-	-	0	1
4	0	0	1	0
5	-	-	0	1
6	0	0	1	0
7	0	-	1	0
8	0	0	-	1
9	-	-	0	1
10	1	1	0	0
11	-	0	0	1
SUM:	**1**	**1**	**0**	**0**



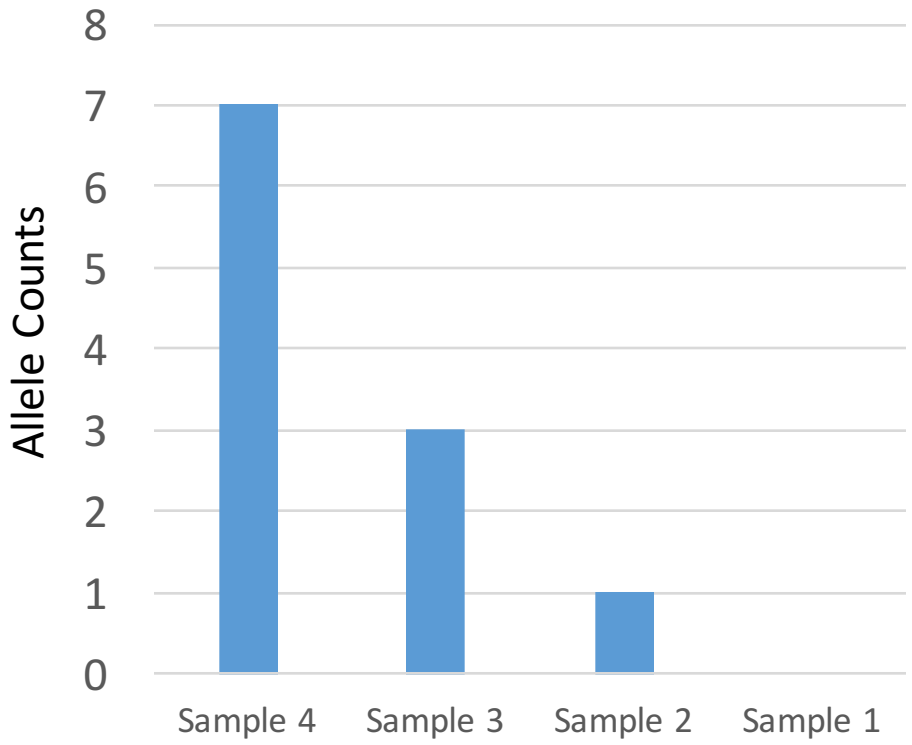
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	-	0	0	1
2	0	-	-	1
3	-	-	0	1
4	0	0	1	0
5	-	-	0	1
6	0	0	1	0
7	0	-	1	0
8	0	0	-	1
9	-	-	0	1
10	1	1	0	0
11	-	0	0	1
SUM:	**1**	**1**	**0**	**0**



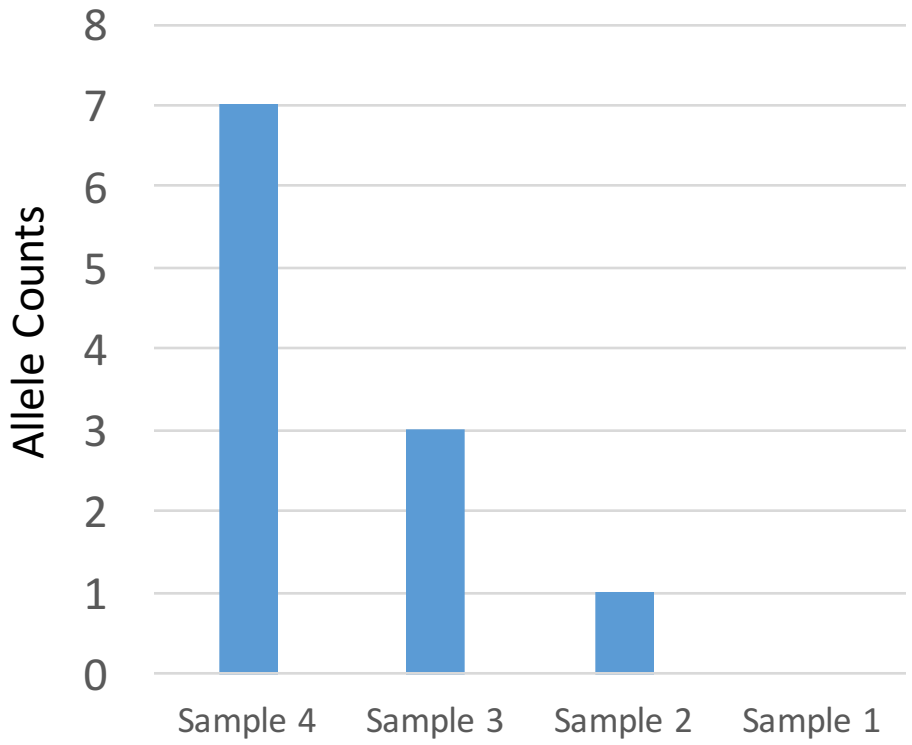
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	-	0	0	1
2	0	-	-	1
3	-	-	0	1
4	0	0	1	0
5	-	-	0	1
6	0	0	1	0
7	0	-	1	0
8	0	0	-	1
9	-	-	0	1
10	-	1	0	0
11	-	0	0	1
SUM:	**1**	**1**	**0**	**0**



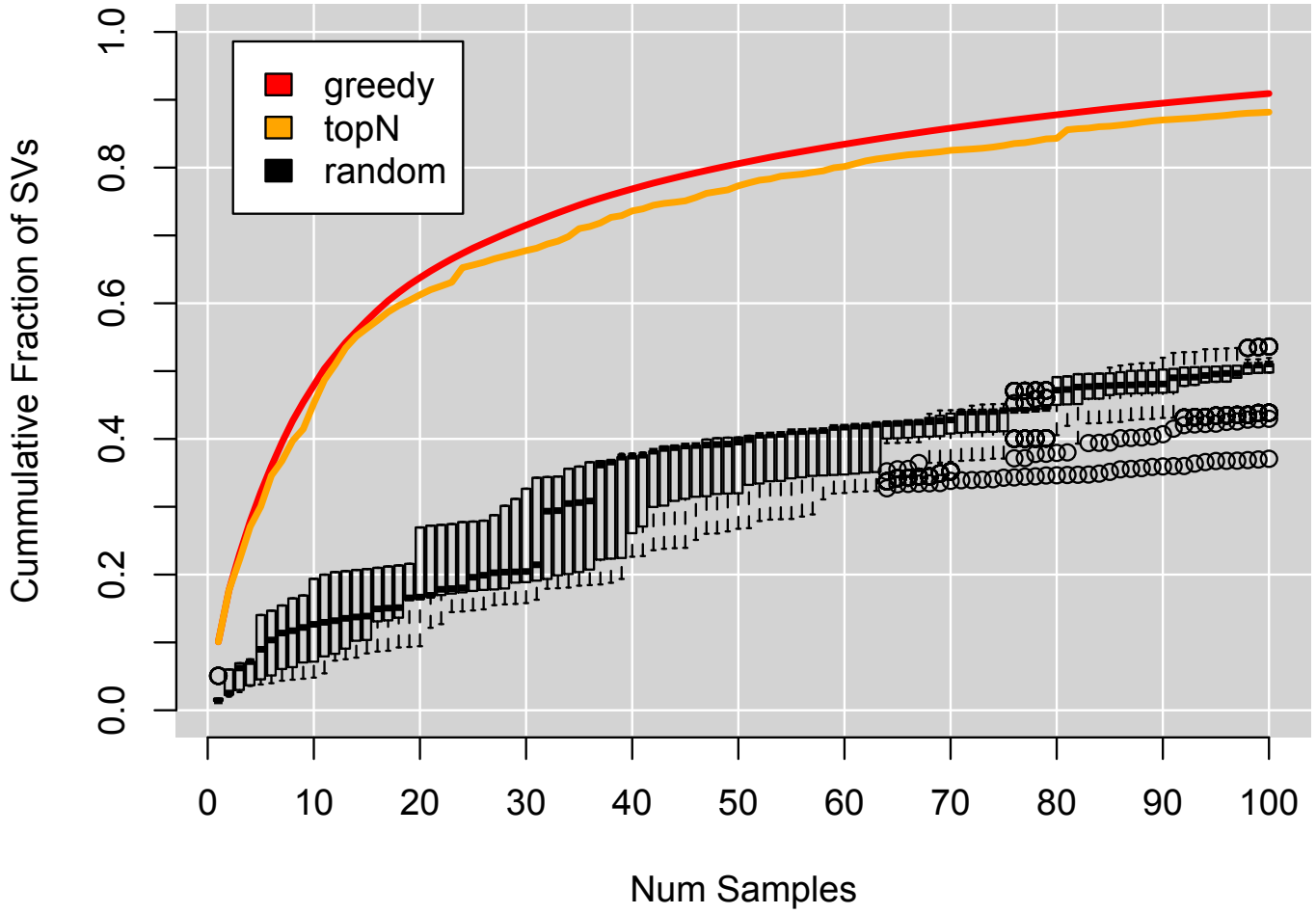
SVCollector: How it works?

Variant	Sample 1	Sample 2	Sample 3	Sample 4
1	-	0	0	1
2	0	-	-	1
3	-	-	0	1
4	0	0	1	0
5	-	-	0	1
6	0	0	1	0
7	0	-	1	0
8	0	0	-	1
9	-	-	0	1
10	-	1	0	0
11	-	0	0	1
SUM:	**0**	**0**	**0**	**0**



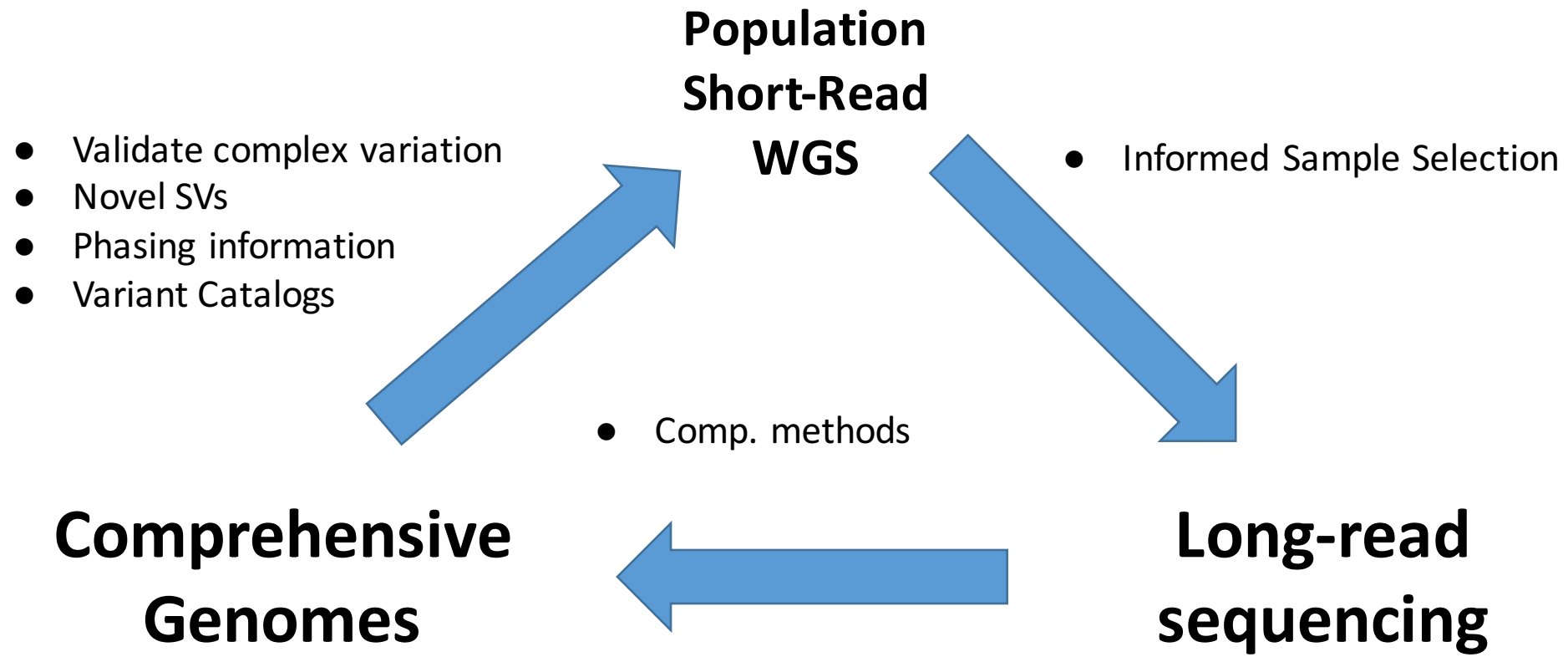
SVCollector: Tomato

- Based on 354 Illumina WGS
- 201,201 SV called using SURVIVOR based on:
  - DELLY
  - Lumpy
  - Manta

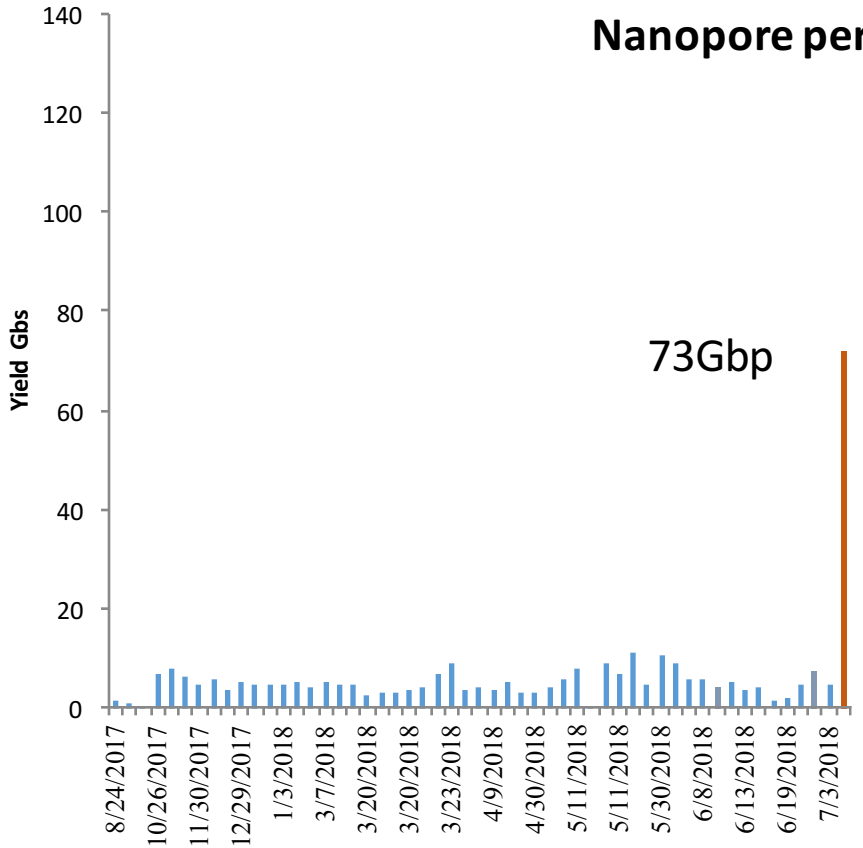


Sedlazeck et al. (bioRxiv)

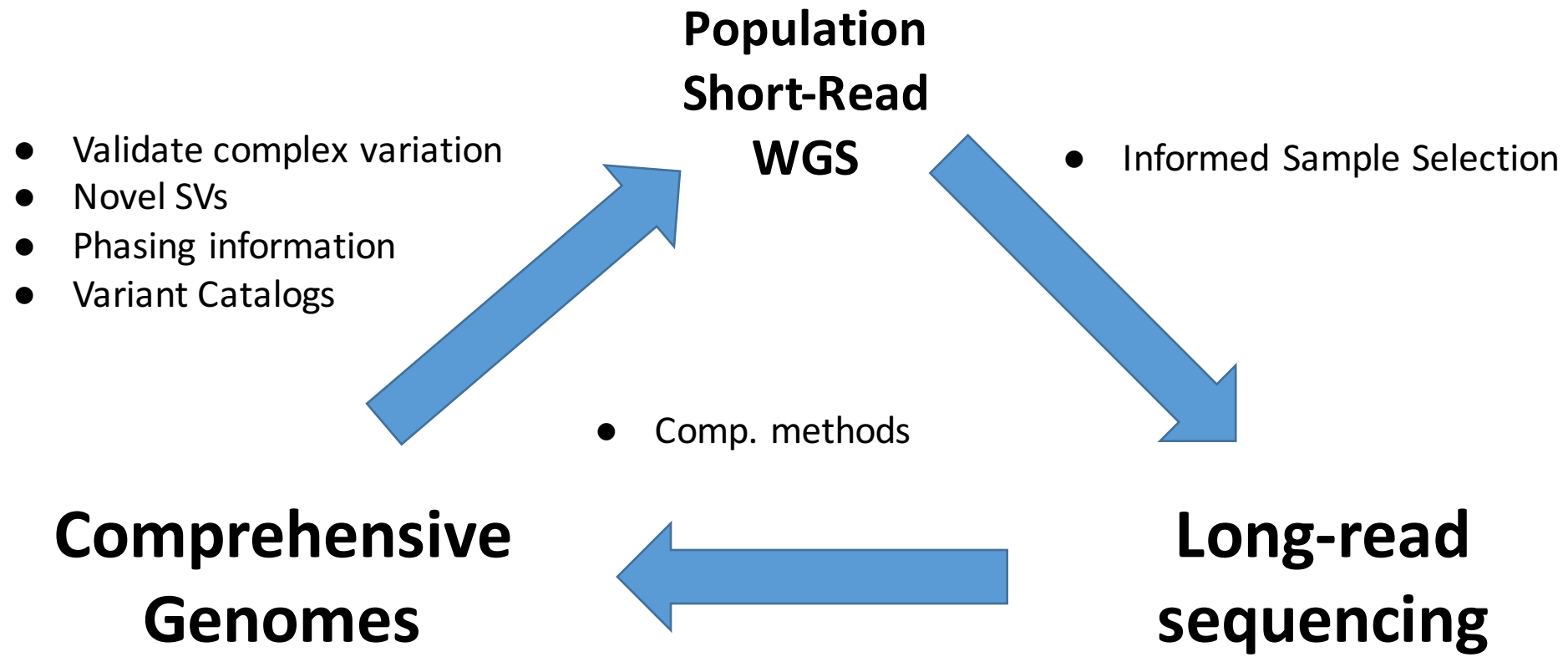
How can we leverage these technologies in larger cohorts?



ONT process on Tomato population



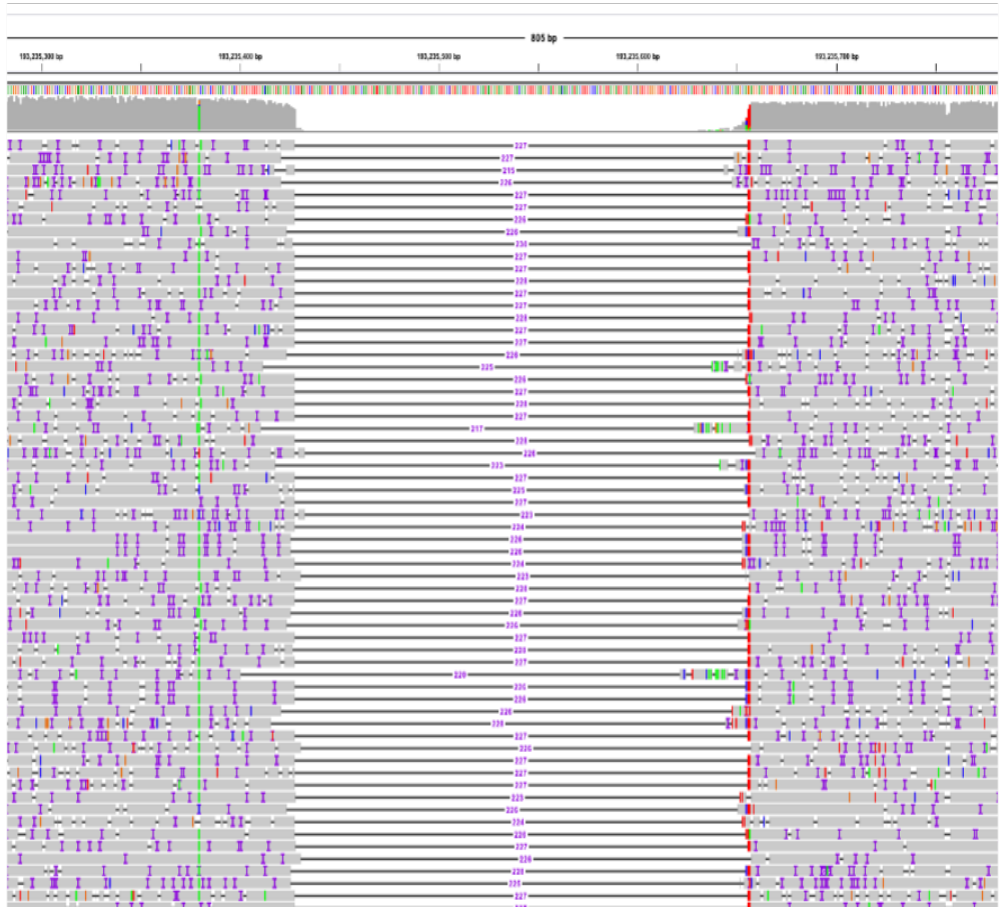
How can we leverage these technologies in larger cohorts?



Why another mapper?

BWA-MEM:

NGMLR:



Why another mapper?

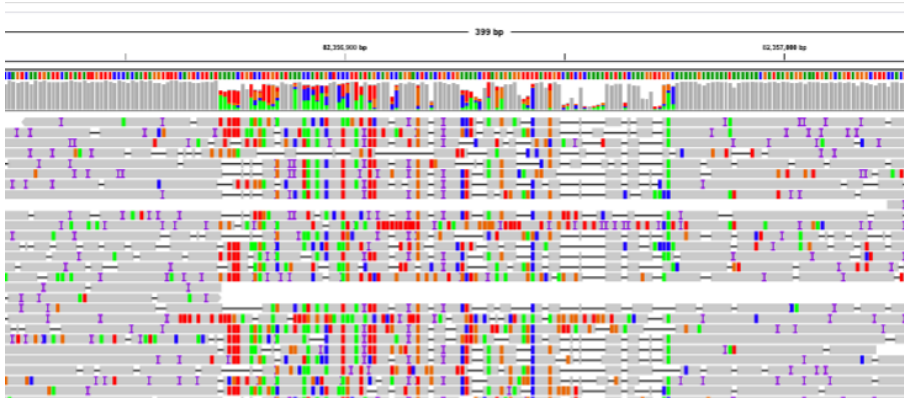
BWA-MEM:

NGMLR:

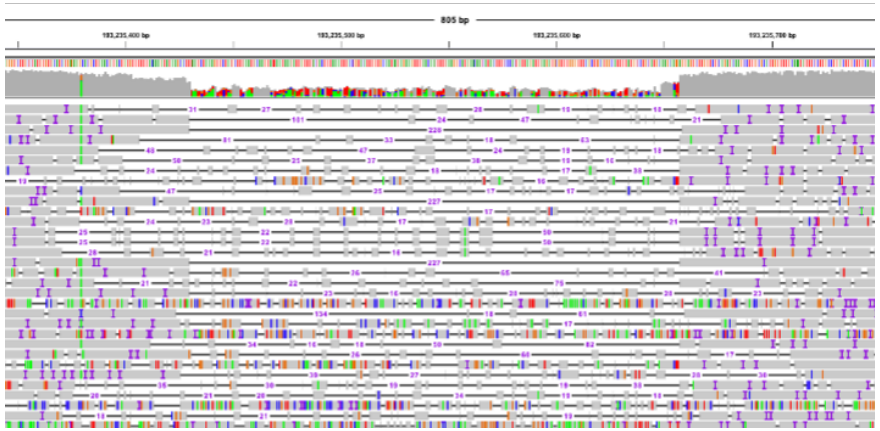
Improving long read alignment: NGMLR

Philipp Rescheneder

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

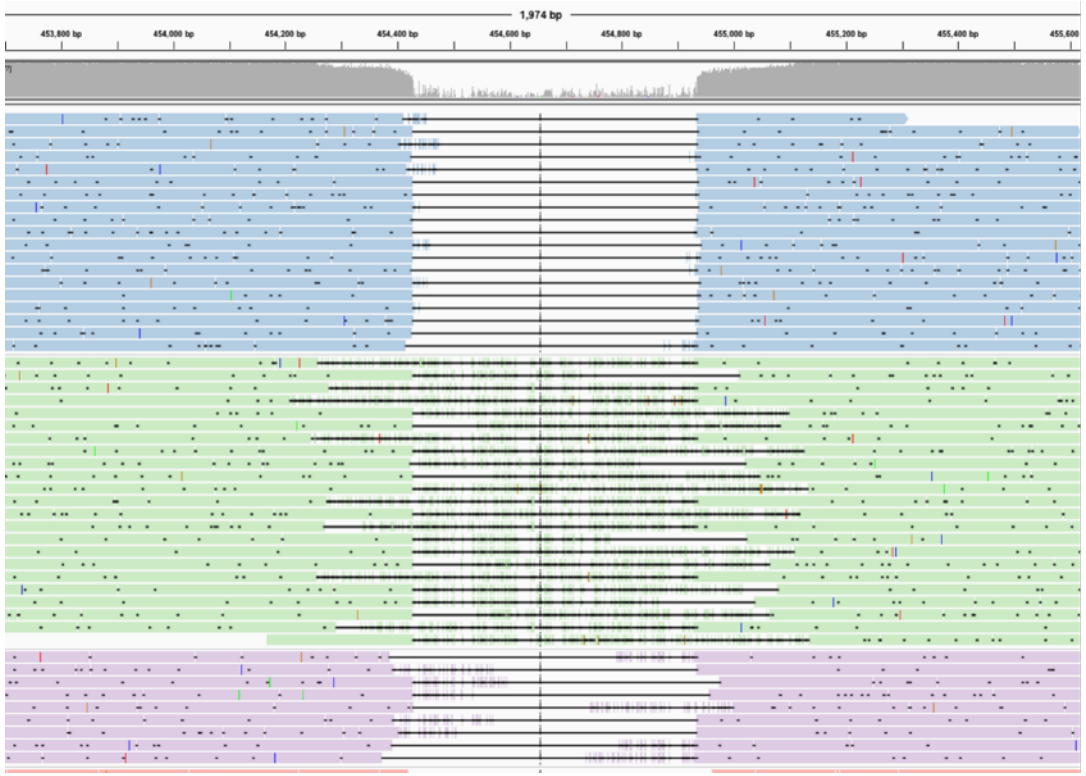


2. Improve alignment:
   - Insertions
   - Deletions



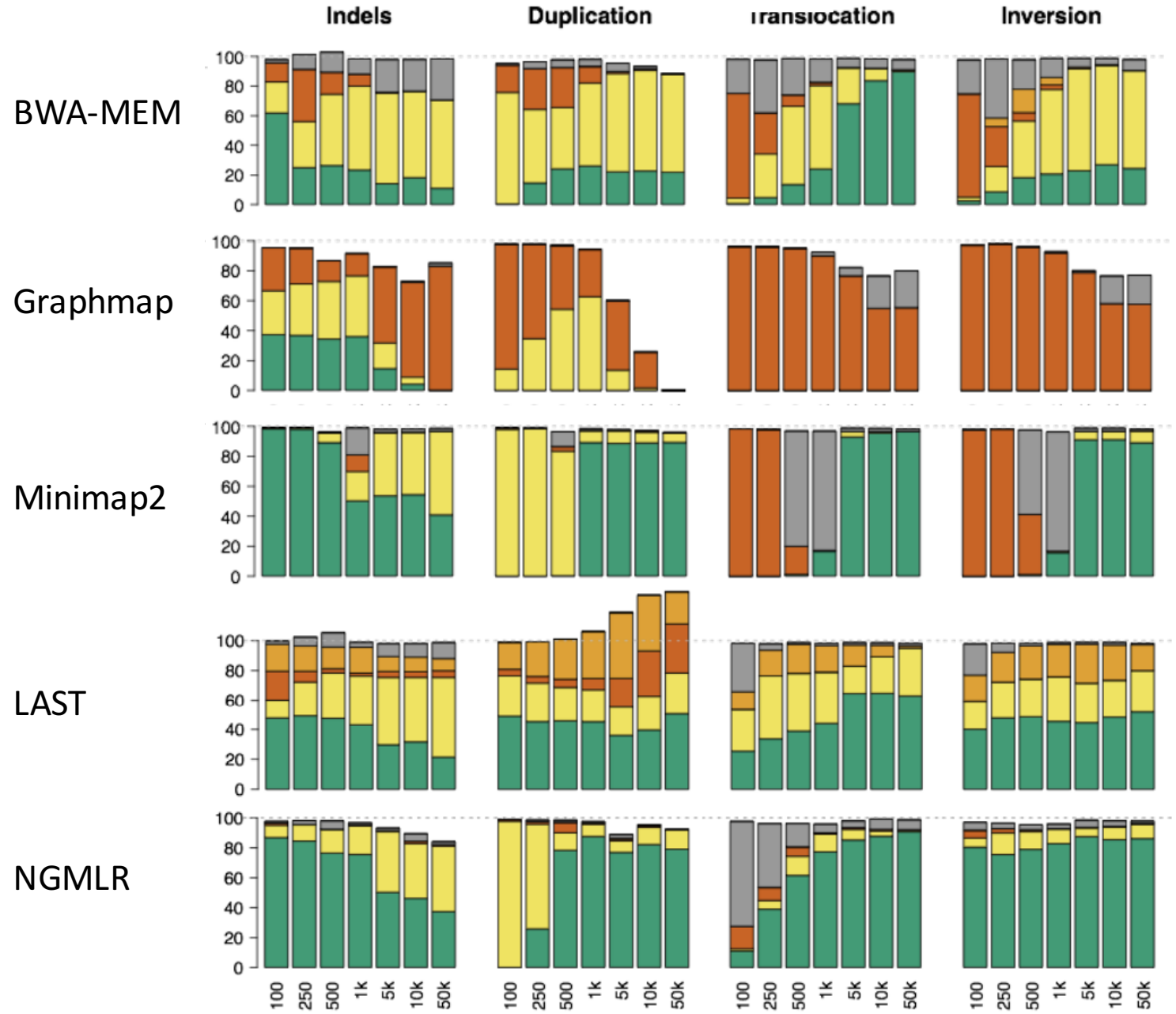
Simulations/ Evaluation

- Simulate 20 SVs of each type using SURVIVOR
- Simulate ONT like reads
- Evalutated:
  - BWA-MEM
  - Graphmap
  - Minimap2
  - LAST
  - NGMLR



Results

- Precise
- Indicated
- Wrong
- Alignment stopped prior
- Fragmented



Why Sniffles?

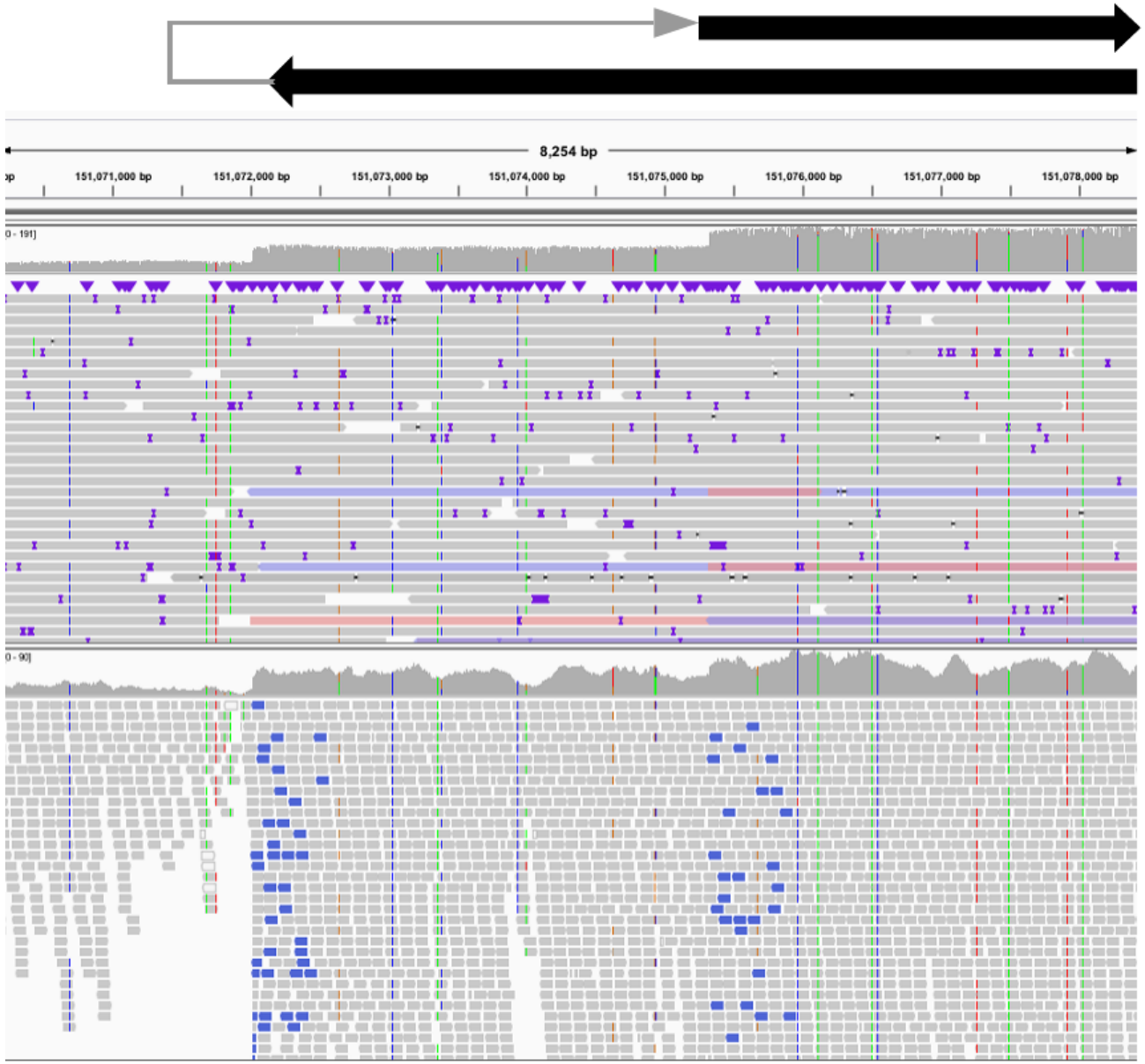
Leverage technology:

- 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



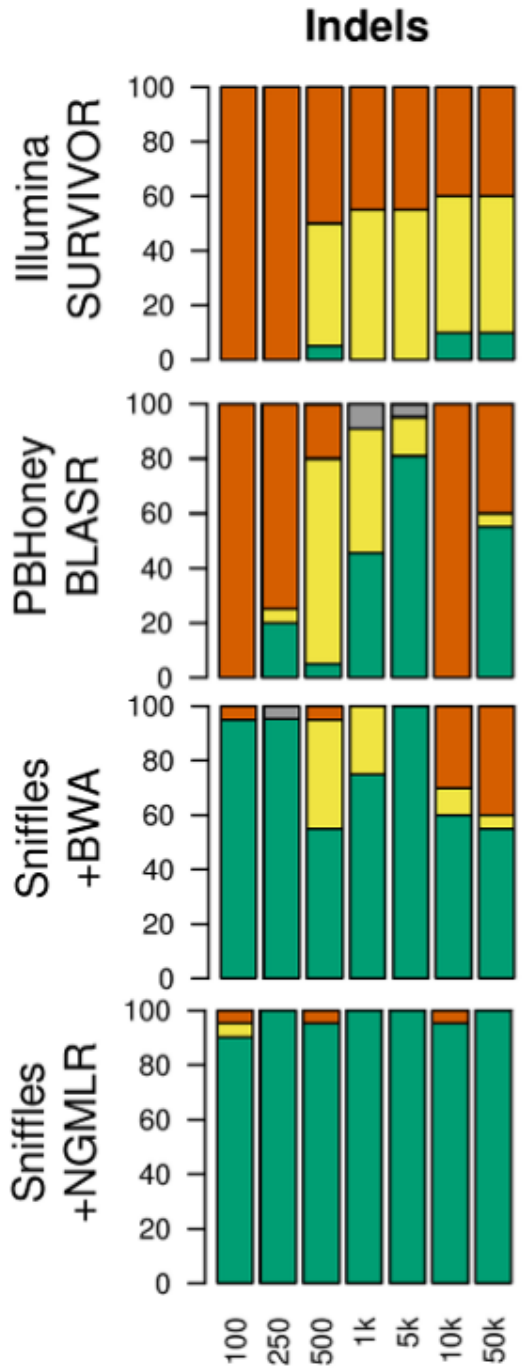
Sniffles

- Parameter estimation
- Detecting all SV types
- Detect sequencing artifacts

- Optional:
  - Genotype estimation
  - Clustering/phasing of SVs
  - Reports sequence resolved indels

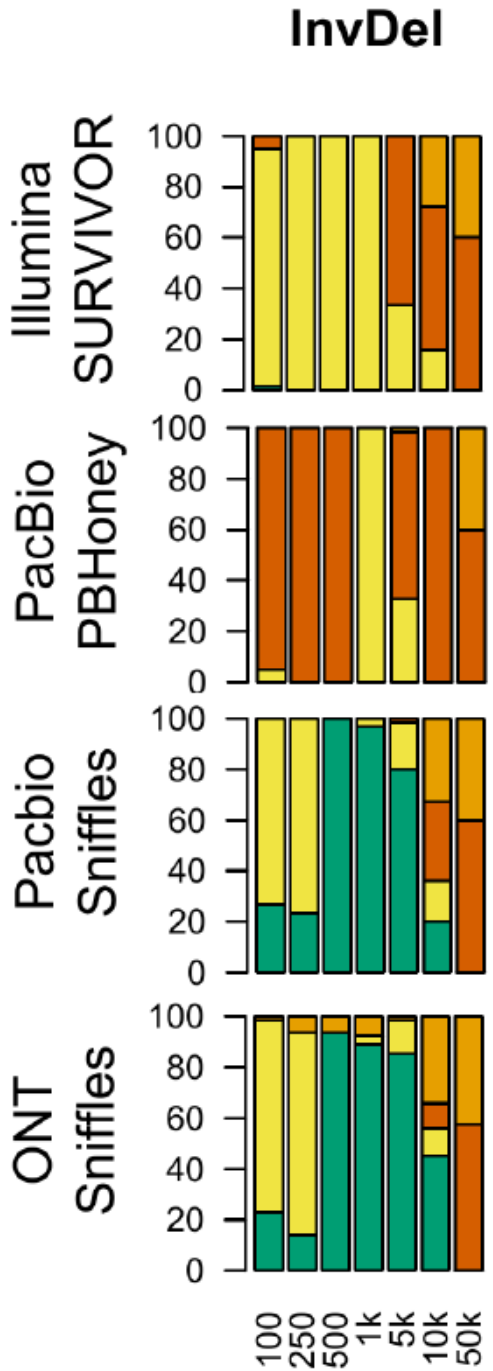
Results

- Precise
- Indicated
- Not found
- Additional events

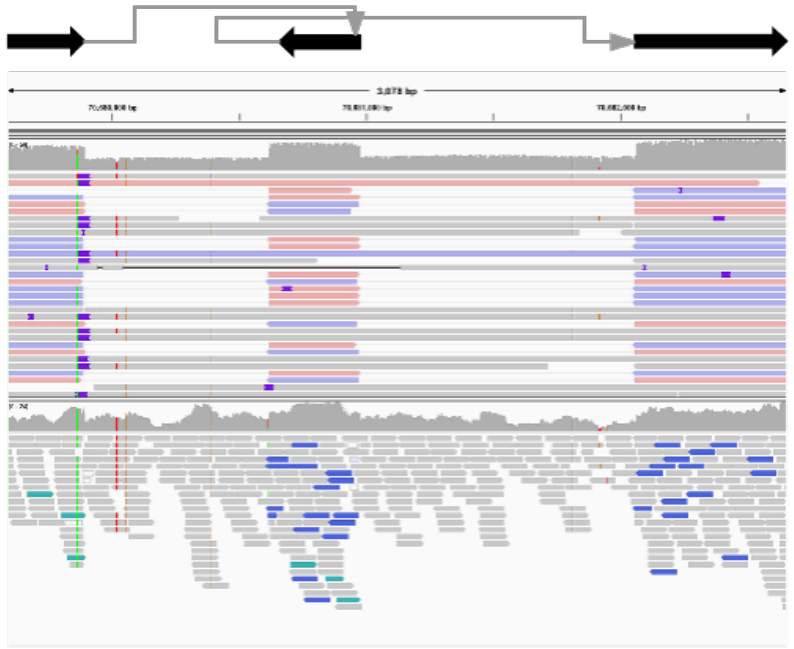


Results

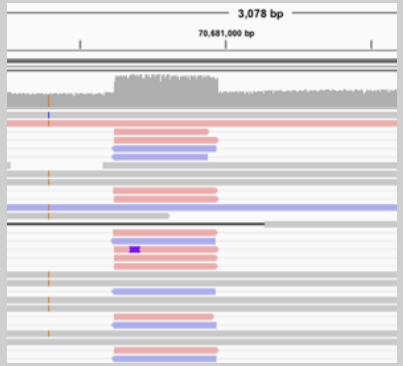
- Precise
- Indicated
- Not found
- Additional events



InvDel



Software



NGMLR
Sedlazeck et. al. (2018)

- Convex gap model
- Improved split reads
- Oxford Nanopore + PacBio
- github.com/philres/ngmlr

Sniffles
Sedlazeck et. al. (2018)

- Detection of all SV types
- Also nested/adjacent events
- Oxford Nanopore+ PacBio
- github.com/fritzsedlazeck/Sniffles

Clairvoyante
Lou et. al. (in review)

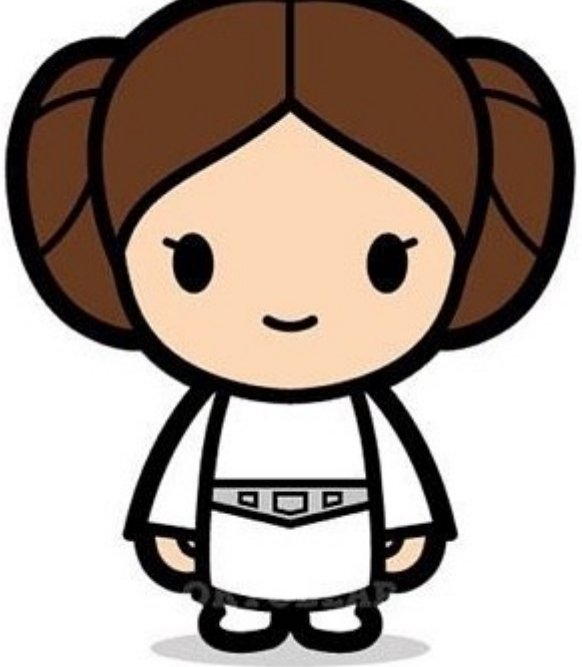
- Neural Network based SNP caller
- Illumina+ Oxford Nanopore + PacBio
- github.com/aquaskyline/Clairvoyante

Crosstich
Kirsche et.al.(in prep)

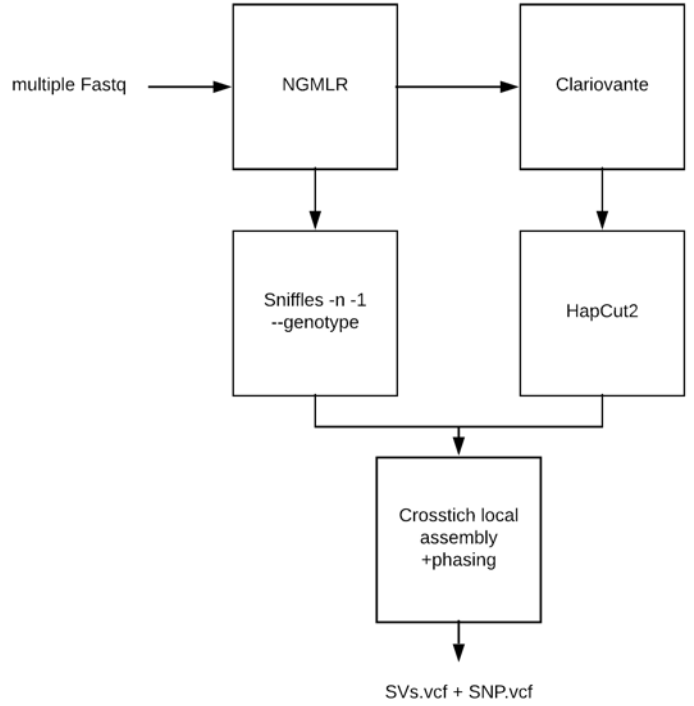
- Phases SV + SNPs
- Local assembly
- Uses HapCut2 (SNPs) + Sniffles(SVs)
- github.com/schatzlab/crossstitch

Upcoming: Princess

- Snakemake pipeline to:
  - Call SNPs
  - Call SVs
  - Phase SNPs + SVs
- Includes helpful options
  - Parameter estimation
  - Parental phasing
  - Summary statistics and plots
    - Read length
    - Mapping statistics
    - Variation statistics
    - Phasing length statistics



Medhat Mahmoud



Acknowledgments

Medhat Mahmoud
Richard Gibbs

Sara Goodwin
Sebastian Soyk
Xingang Wang
Zachary Lemmon
Zachary Lippman

Michael Schatz
Michael Alonge
Srividya Ramakrishnan
Michael Kirsche
Schatz lab

Philipp Rescheneder