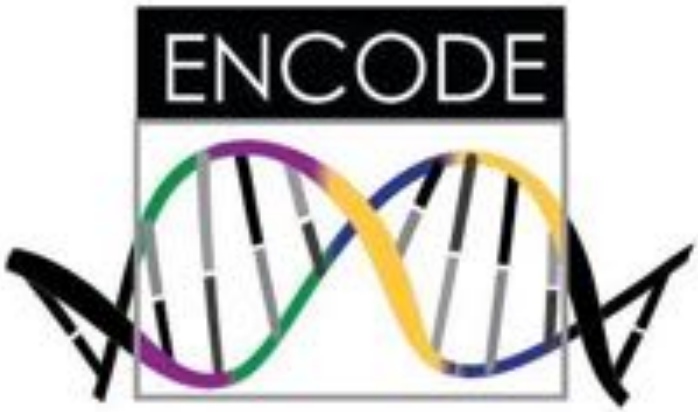
Personalized Phased Diploid Genomes of the EN-TEx Samples

Michael Schatz, Fritz Sedlazeck, Han Fang, Maria Nattestad, Ruibang Luo, Srividya Ramakrishnan, Charlotte Darby, Philipp Rescheneder, Alex Dobin, Carrie Davis, Ashwin Prakash, Anna Vlasova, Alessandra Breschi, Roderic Guigo, Tom Gingeras

Feb 15, 2017

AGBT Informatics



Catalog “functional elements” through large-scale RNA-seq, ChIP-seq and other assays in many tissue types

No DNA sequencing,
Mostly cell lines

Analyze how genomic variations impacts expression in many tissue types in many people

No regulatory assays,
Mostly unphased SNP analysis

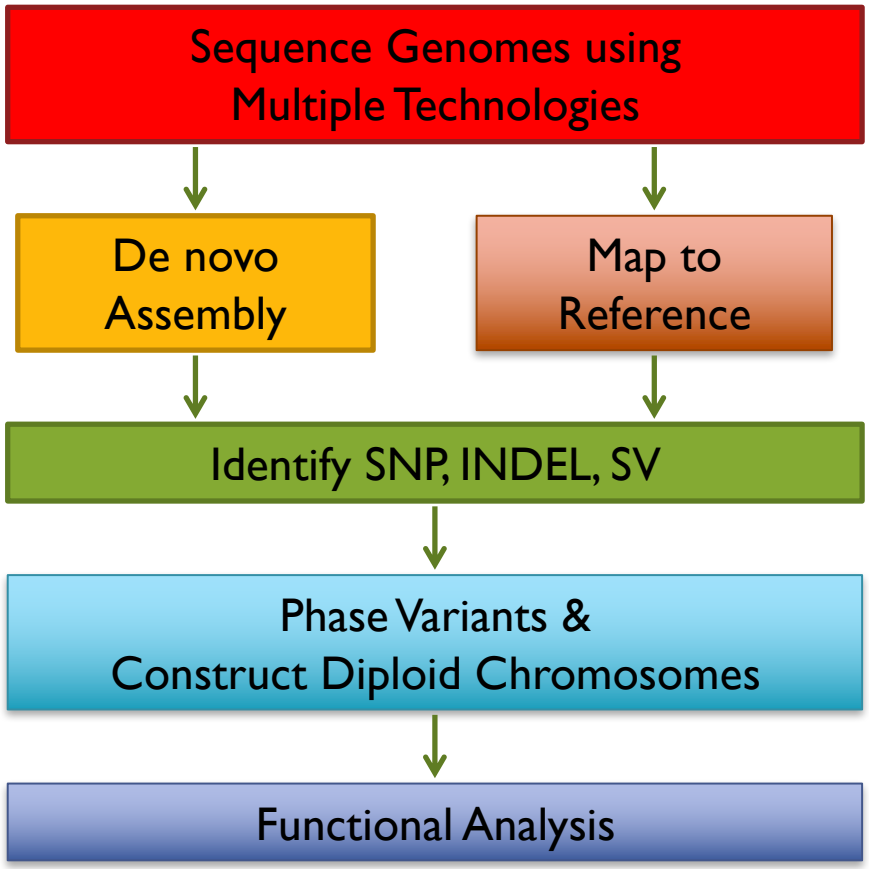
http://encodeproject.org

EN-TEx: Expression & Regulation Analysis of Personalized Genomes

	ENC-001	ENC-002	ENC-003	ENC-004
Age	37	54	53	51
Sex	Male	Male	Female	Female
Cause of Death	Anoxia	Anoxia	Cerebral Vascular Accident	Cerebral Vascular Accident
Total Libraries	319	299	488	299

- Sequenced the genomes for 2 male and 2 female samples using transverse colon tissue
- Large number of ChIP-seq, RNA-seq, ATAC-seq, DNase-seq, and other functional datasets available in dozens of tissues

Assembling and Analyzing Personal Genomes

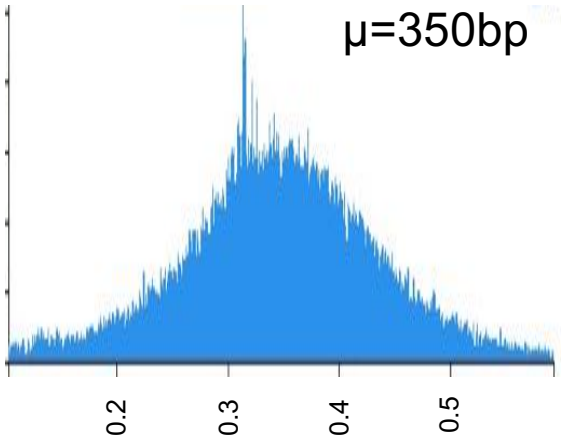


Goals

1. **What are the most effective biotechnologies for sequencing?**
2. **What do we learn from a personalized genome instead of the reference?**
3. **Can we use the genomic variants as natural perturbations of the encyclopedia elements?**

Genomic Sequencing Data

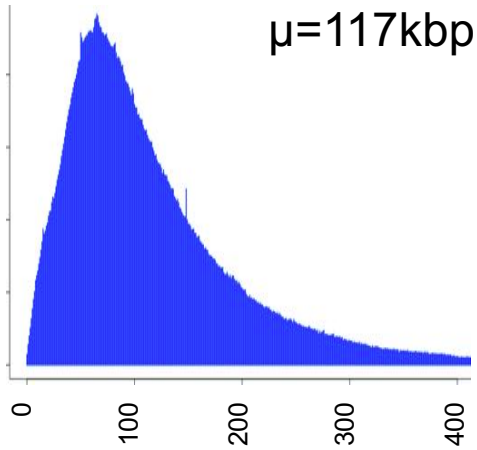
Illumina



Fragment Length (kbp)

60x Paired End
All 4 samples

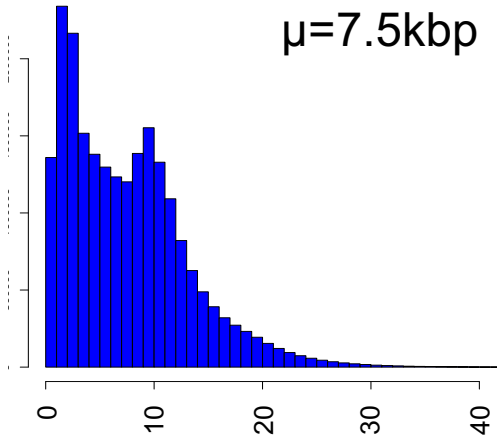
10X Genomics



Molecule Length (kbp)

35x Linked Reads
All 4 samples

PacBio

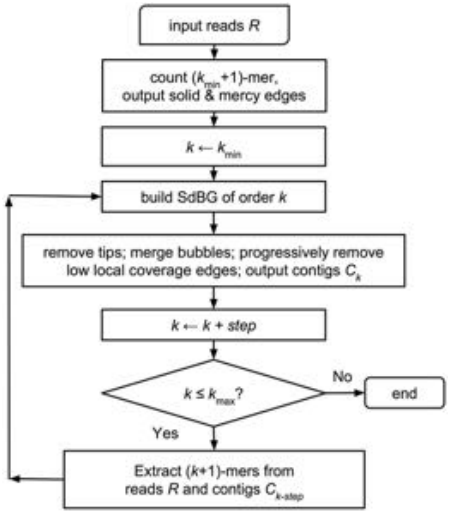


Read Length (kbp)

55x Long Reads
****Only ENC-002***

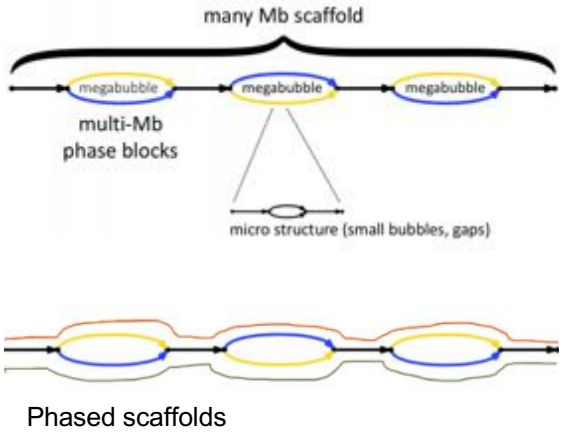
De Novo Assembly

Illumina



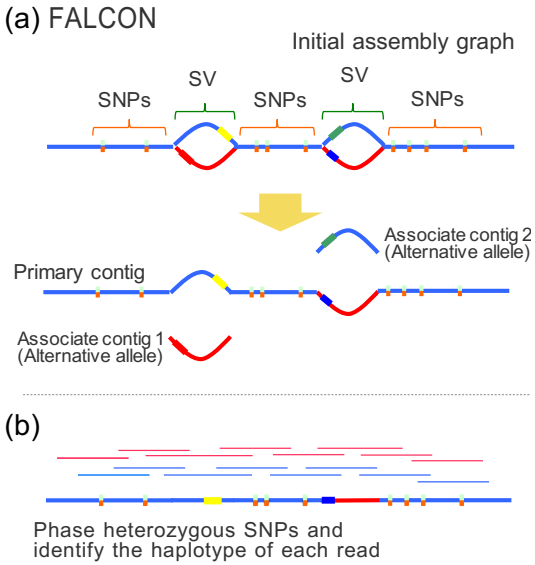
MegaHit
(Li et al, 2015)

10X Genomics



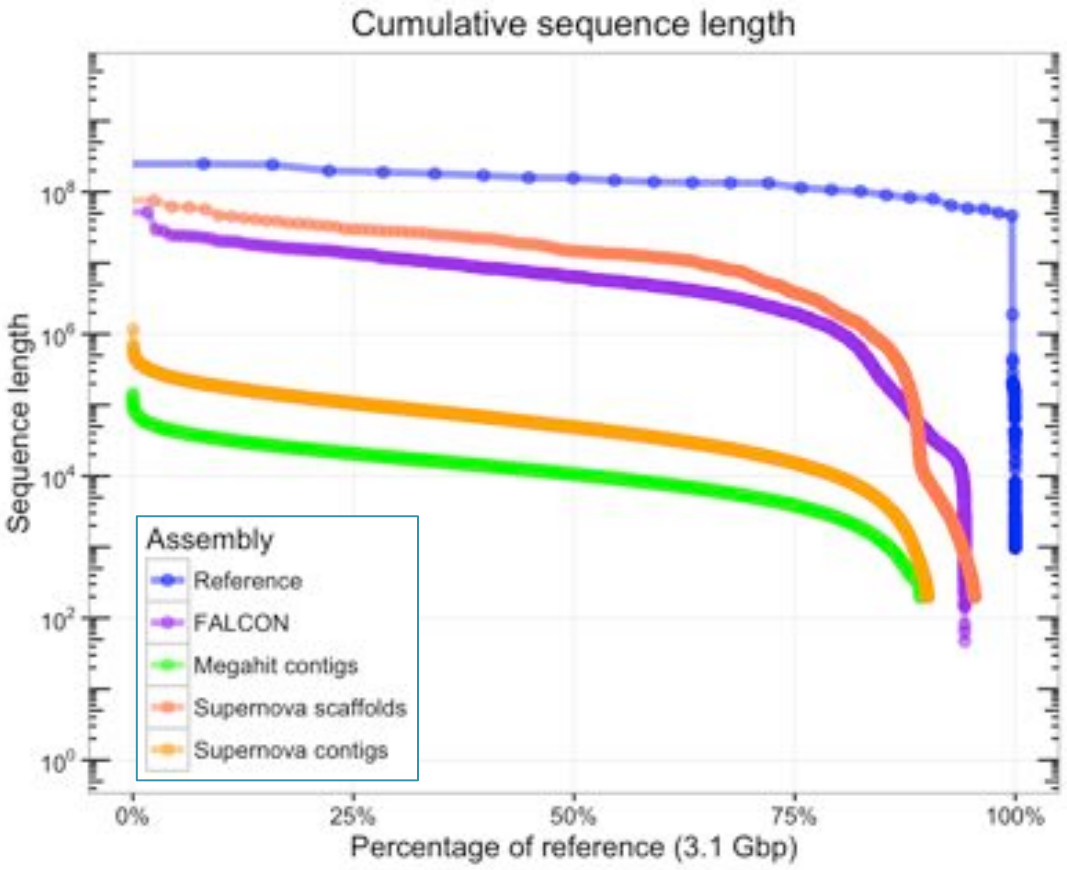
SuperNova
(Weisenfeld et al, 2016)

PacBio



FALCON-unzip
(Chin et al, 2016)

Assembly Contiguity



GRC38 Reference
- **Includes alt sequences**

10X Genomics/SuperNova
- 21 Mbp scaffold N50
- 162 Mbp in scaffold gaps

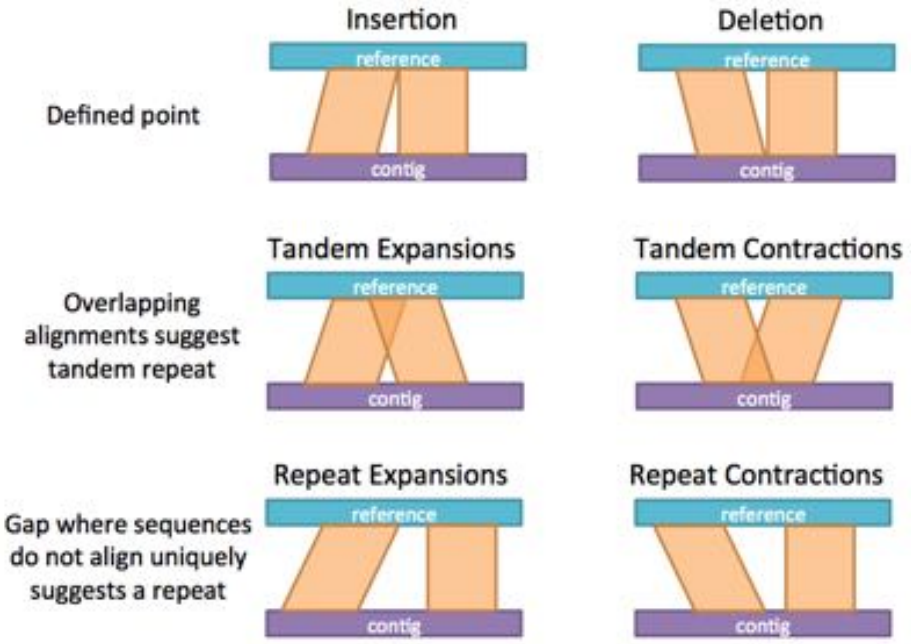
PacBio/Falcon-unzip
- 7.0 Mbp contig N50

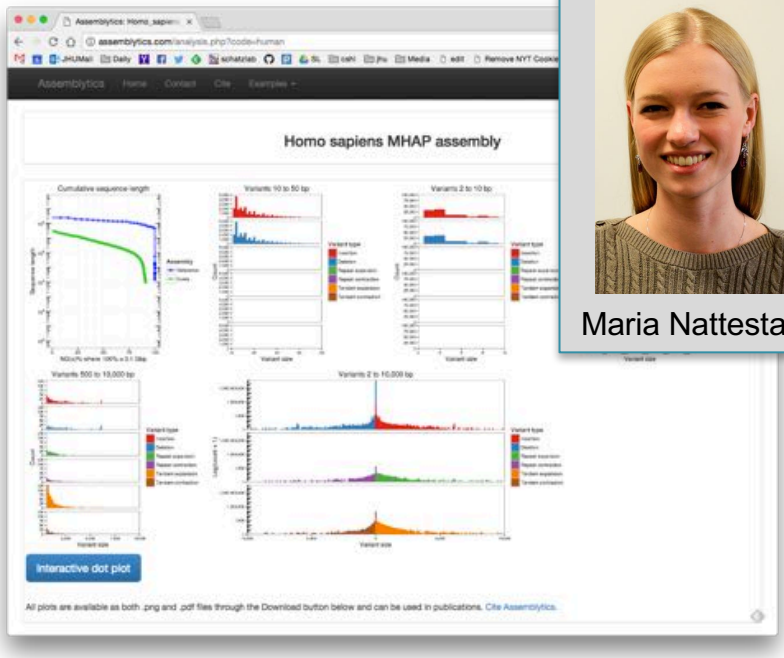
10X Genomics/Supernova
- 50 kbp contig N50

Illumina/MegaHit
- 13 kbp contig N50

Assemblytics: Assembly-Based Variant-Caller

http://assemblytics.com



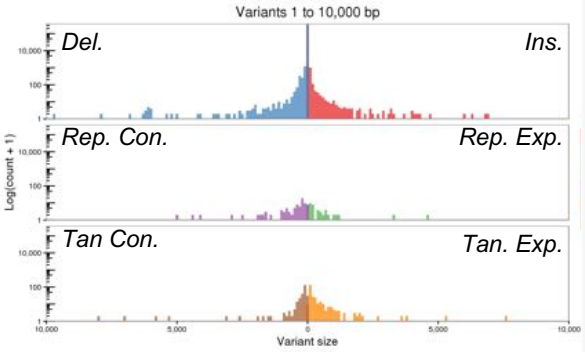


Maria Nattestad

Assemblytics: a web analytics tool for the detection of variants from an assembly
Nattestad, M, Schatz, MC (2016) Bioinformatics doi: 10.1093/bioinformatics/btw369

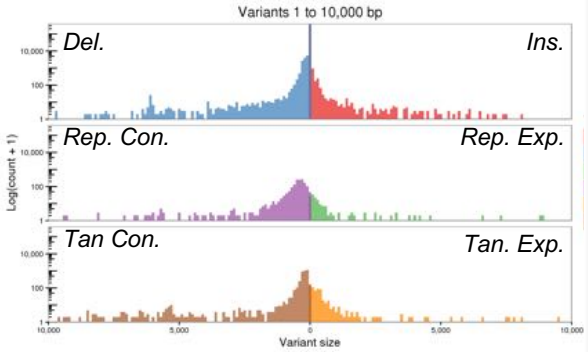
Structural Variations vs GRCh38

Illumina



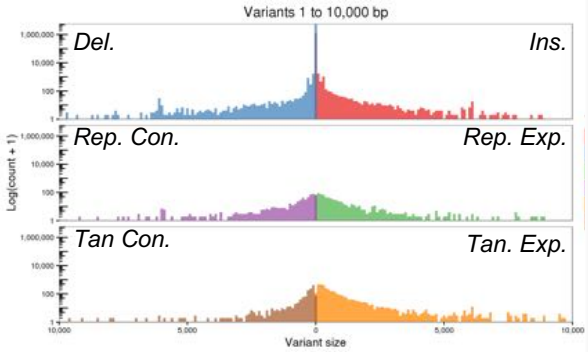
SVs (50bp – 10kb)
Count: 3,997
Bases: 1.11 Mbp

10X Genomics



SVs (50bp – 10kbp)
Count: 18,025
Bases: 6.13 Mbp

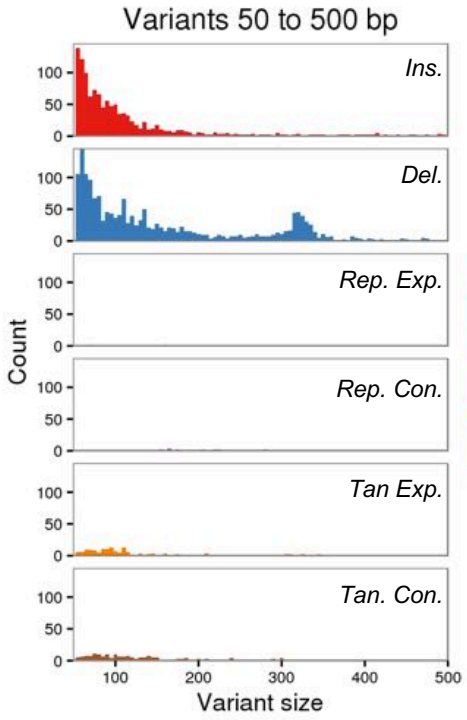
PacBio



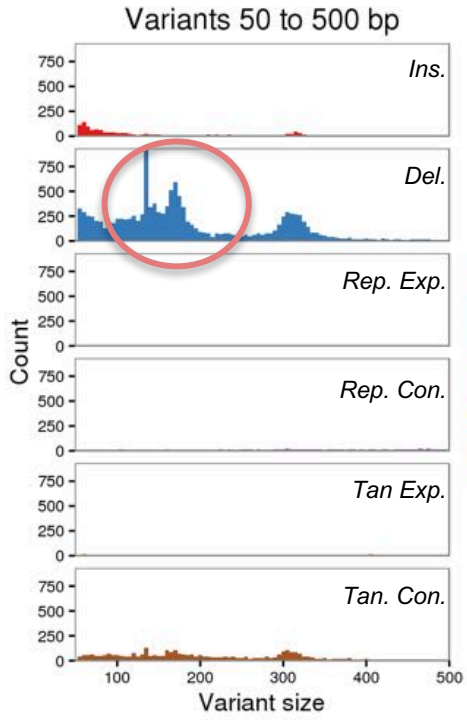
SVs (50bp-10kbp)
Count: 12,965
Bases: 8.13 Mbp

Missing Insertions from Short and Linked Read?

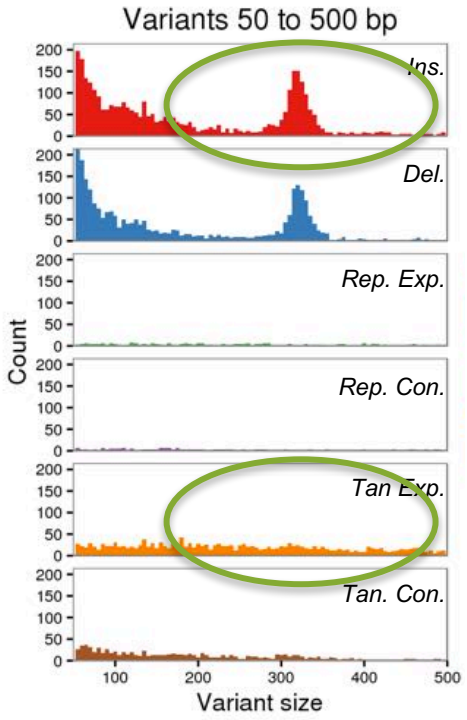
Illumina



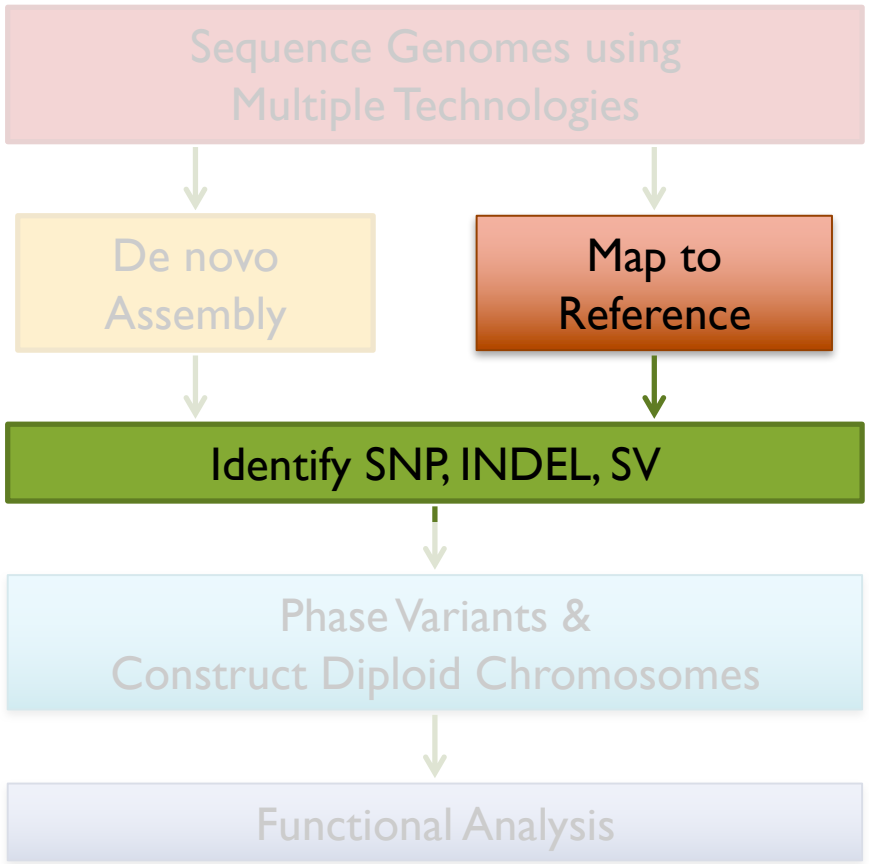
10X Genomics



PacBio



Assembling and Analyzing Personal Genomes



Goals

1. What are the most effective biotechnologies for a sequencing?
2. What do we learn from an personalized genome instead of the reference?
3. Can we use the genomic variants as naturals perturbations of the encyclopedia elements?

Small Variant Analysis

- Mapped Illumina PE reads using BWA-MEM (Li, 2013)
- Identified 3.7M SNPs using GATK (Van der Auwera *et al.* 2013)
- Identified 700k indels using Scalpel (Fang *et al.*, 2016)
- Annovar (Wang *et al.*, 2010) characterization of variants

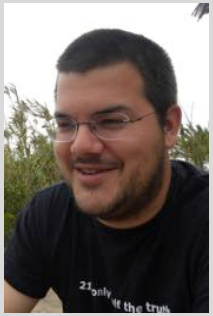

Han Fang

	ENC-001	ENC-002	ENC-003	ENC-004
Synonymous SNP	12,007	12,249	12,524	12,172
Non Syn. SNP	11,507	11,816	12,078	12,009
Frameshift Indel	304	344	344	322
Stop Gain + Loss	113 / 27	120 / 33	136 / 25	135 / 28
Splicing SNP + Indel	109 / 41	102 / 43	117 / 55	111 / 50

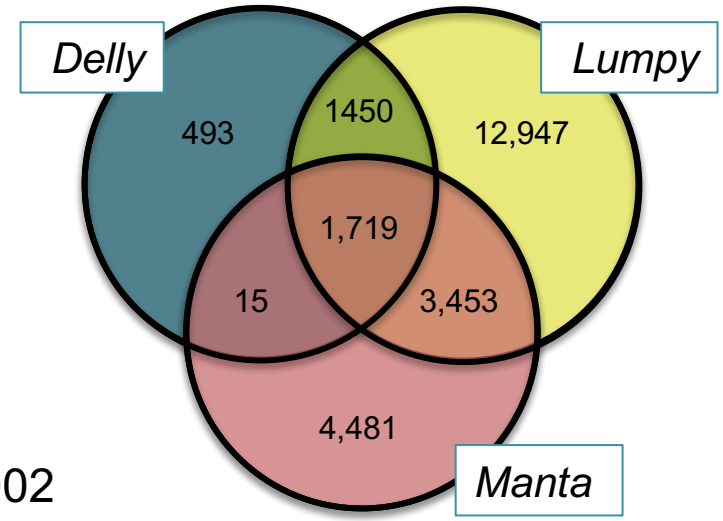
Establishes a catalog of variation, heterozygous positions informative for phasing

Consensus SV Analysis with SURVIVOR

https://github.com/fritzsedlazeck/SURVIVOR


Fritz Sedazeck

- Analyzed the Illumina PE sequence data using 3 different algorithms that use split-reads and discordant pairs to identify SVs: Manta (Chen *et al.*, 2015), Delly (Rausch *et al.*, 2012), and Lumpy (Layer *et al.*, 2014)
- Use SURVIVOR (Jeffares *et al.*, 2016) to improve accuracy by excluding variants identified by only 1 method



ENC-002

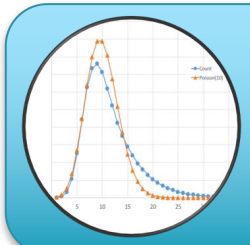
Type	SURVIVOR2
Deletions	3,692
Duplications	1,144
Insertions	253
Inversions	602
Translocations	676
All	***6,367***

10X LongRanger Variant Calls


Han Fang

- Recent versions of LongRanger report SVs in addition to phasing
- Two classes: large_svs.vcf (>30kbp) and dels.vcf (40bp to 30kbp)

	Large SV	DEL	DUP	INV	Small Del.	Del. Span	Del. Mean
ENC-001	96	21	7	7	4,022	3.77 Mbp	937 bp
ENC-002	194	36	6	6	3,796	3.21 Mbp	852 bp
ENC-003	96	32	2	8	4,055	3.74 Mbp	927 bp
ENC-004	103	33	1	4	4,294	3.43 Mbp	805 bp



LRSim: Linked Read Simulator

Lead Author: Ruibang Luo
https://github.com/aquaskyline/LRSIM
bioRxiv: https://doi.org/10.1101/103549

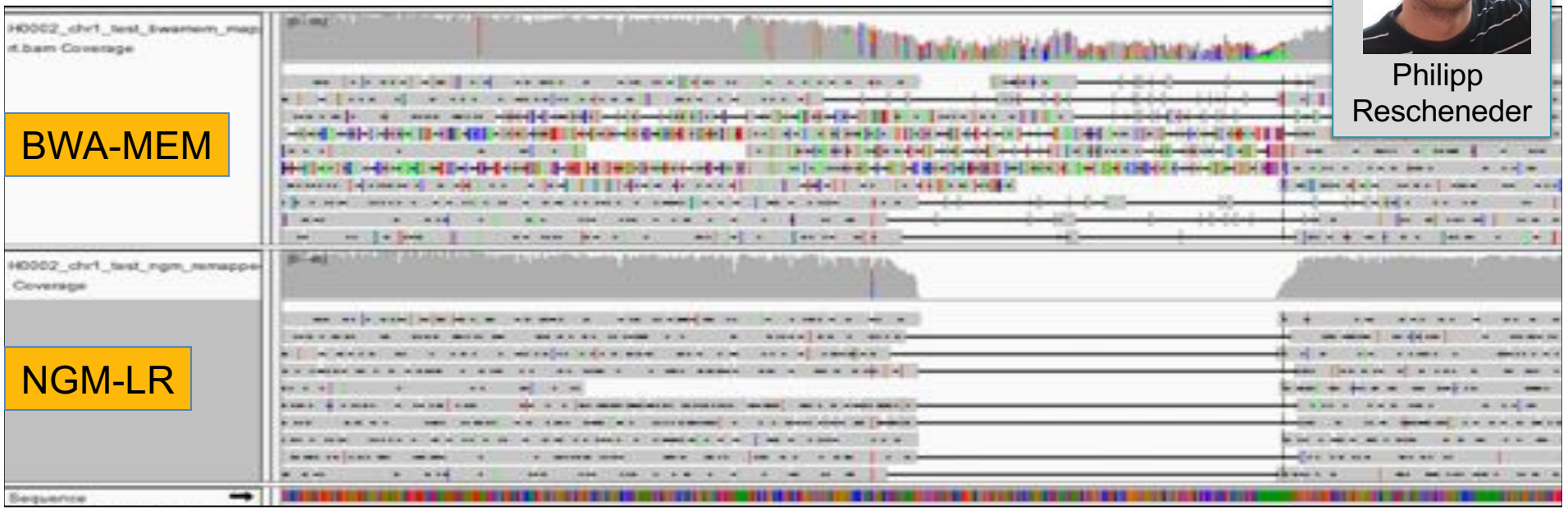
TopSorter: 10X SV Analysis

Lead Author: Han Fang
https://github.com/hanfang/Topsorter
Preprint: Coming soon

NGM-LR + Sniffles: PacBio SV Analysis Tools

https://github.com/philres/ngmlr & https://github.com/fritzsedlazeck/Sniffles

Philipp Rescheneder



Improved SV Variant Detection with long reads

1. NGM-LR: Improve mapping of noisy long reads: improved seeding, convex gap scoring

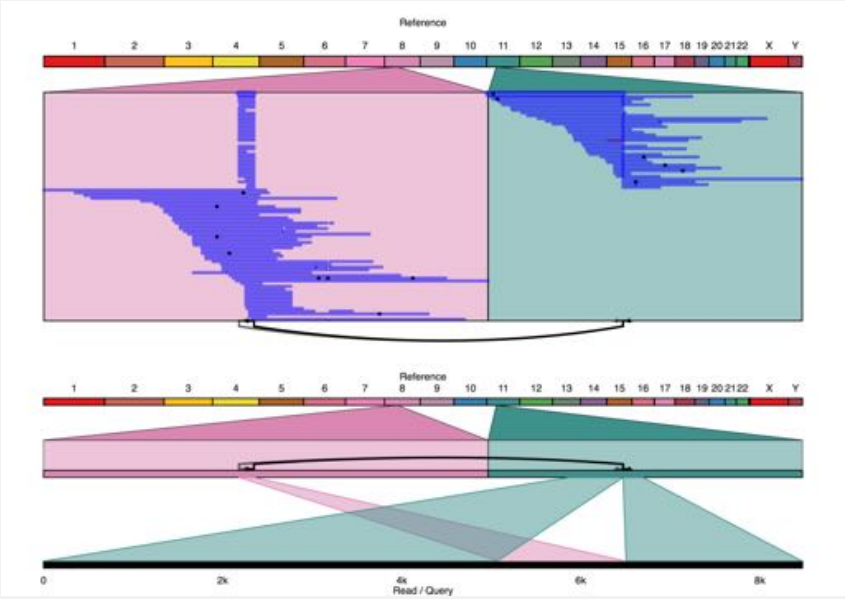
2. Sniffles: Integrates evidence from split-reads, alignment fidelity, breakpoint concordance

Sniffles PacBio Variant Calls

Sniffles calls

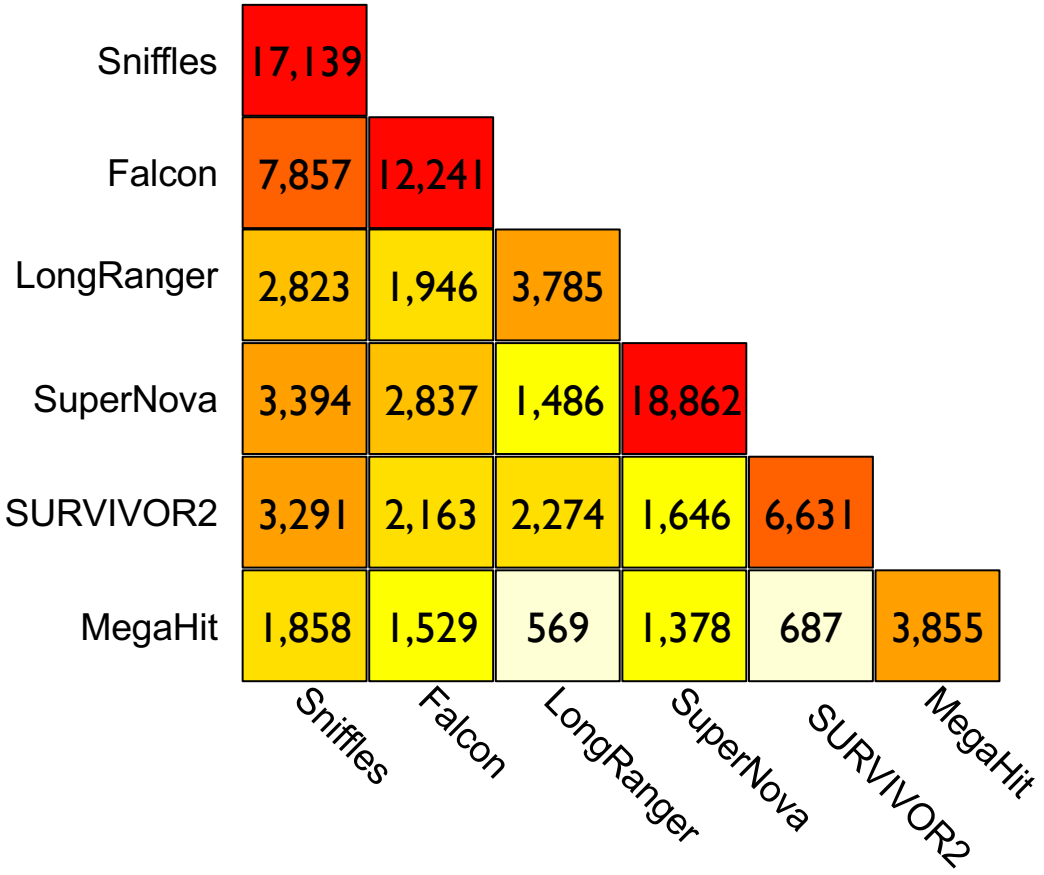
	All SVs (50bp+)	Large SVs (10kbp+)
Deletions	7,389	164
Duplications	1,284	139
Insertions	8,382	4
Inversions	229	116
Translocations	170	170
All	***17,454***	***593***

Translocation in Ribbon

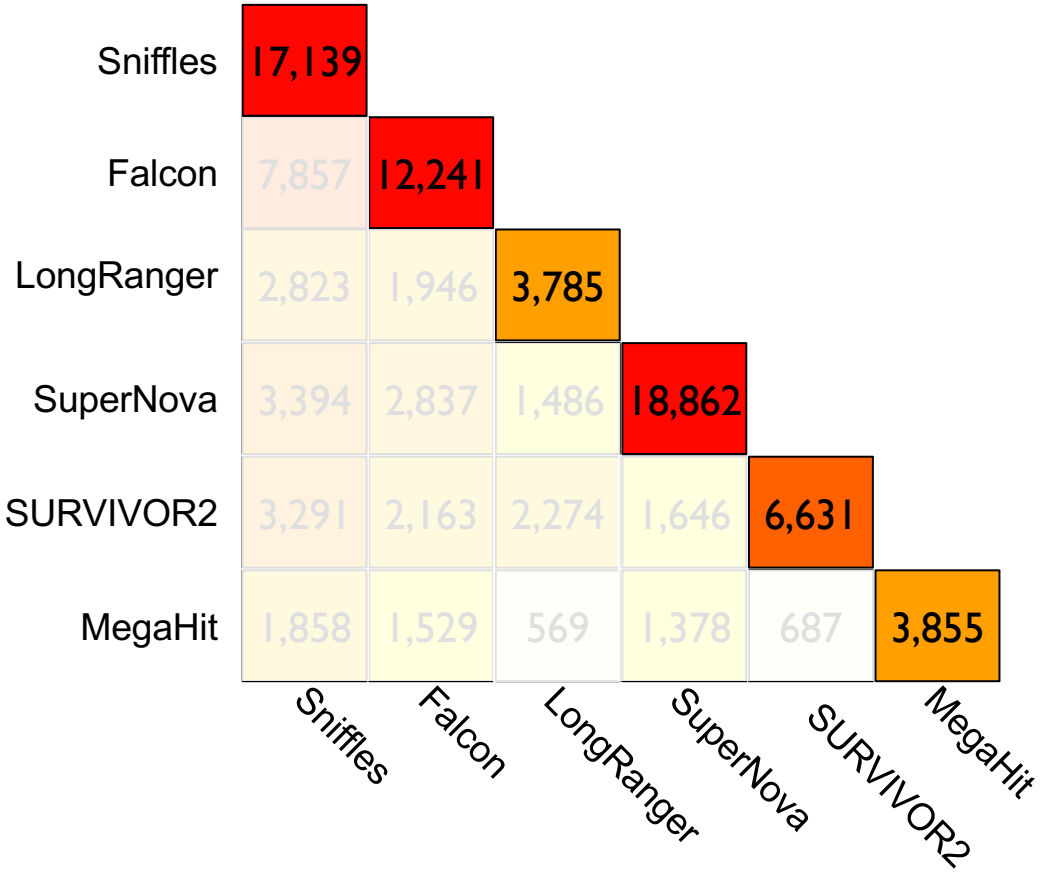


Ribbon: Visualizing complex genome alignments and structural variation
Nattestad et al. (2016) *bioRxiv* doi: http://dx.doi.org/10.1101/082123

Structural Variations Concordance



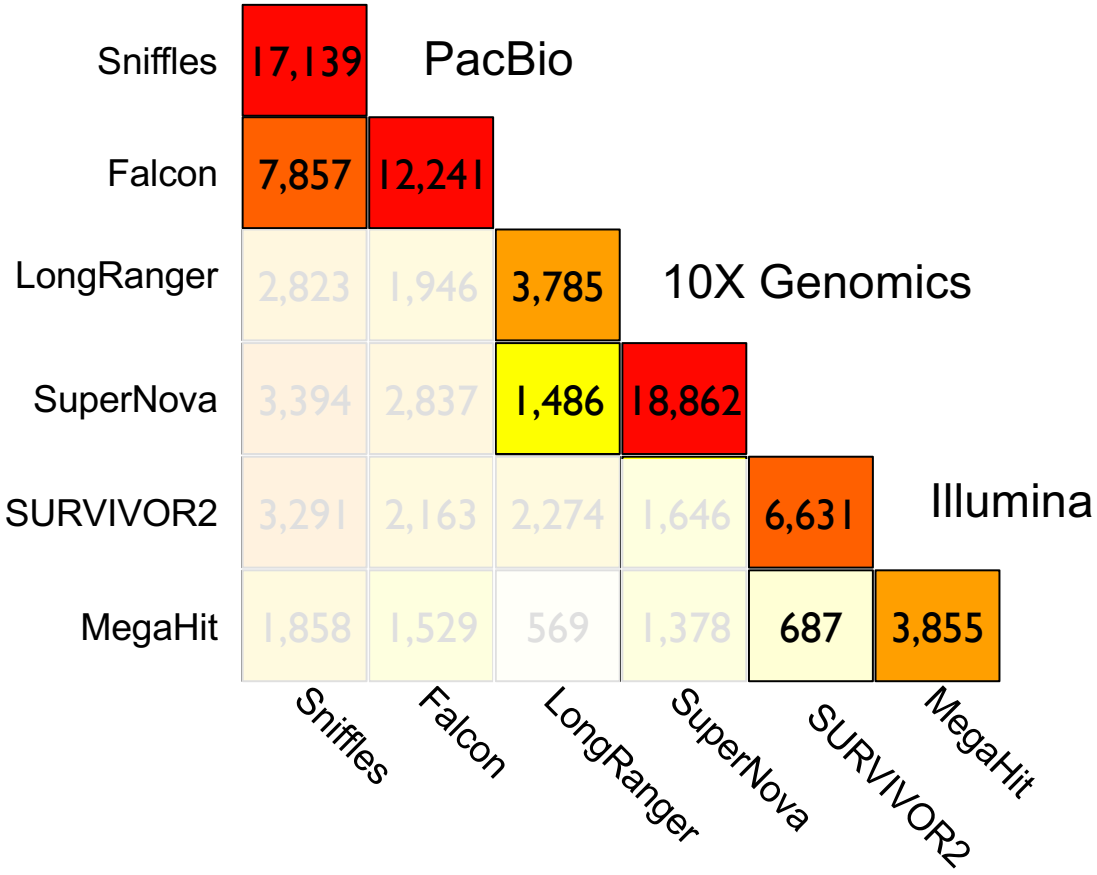
Structural Variations Concordance



Main Diagonal

- Calls per tool

Structural Variations Concordance



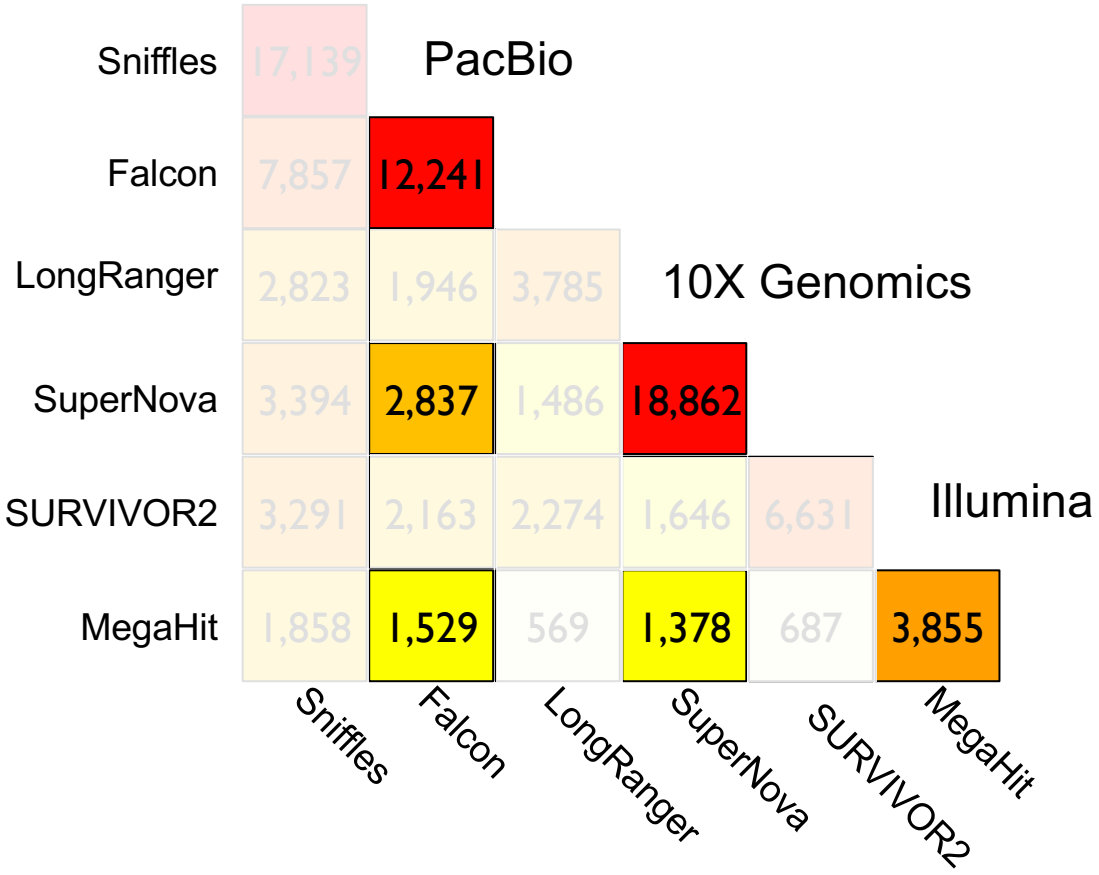
Main Diagonal

- Calls per tool

Outer triplets

- Concordance by Technology

Structural Variations Concordance



Main Diagonal

- Calls per tool

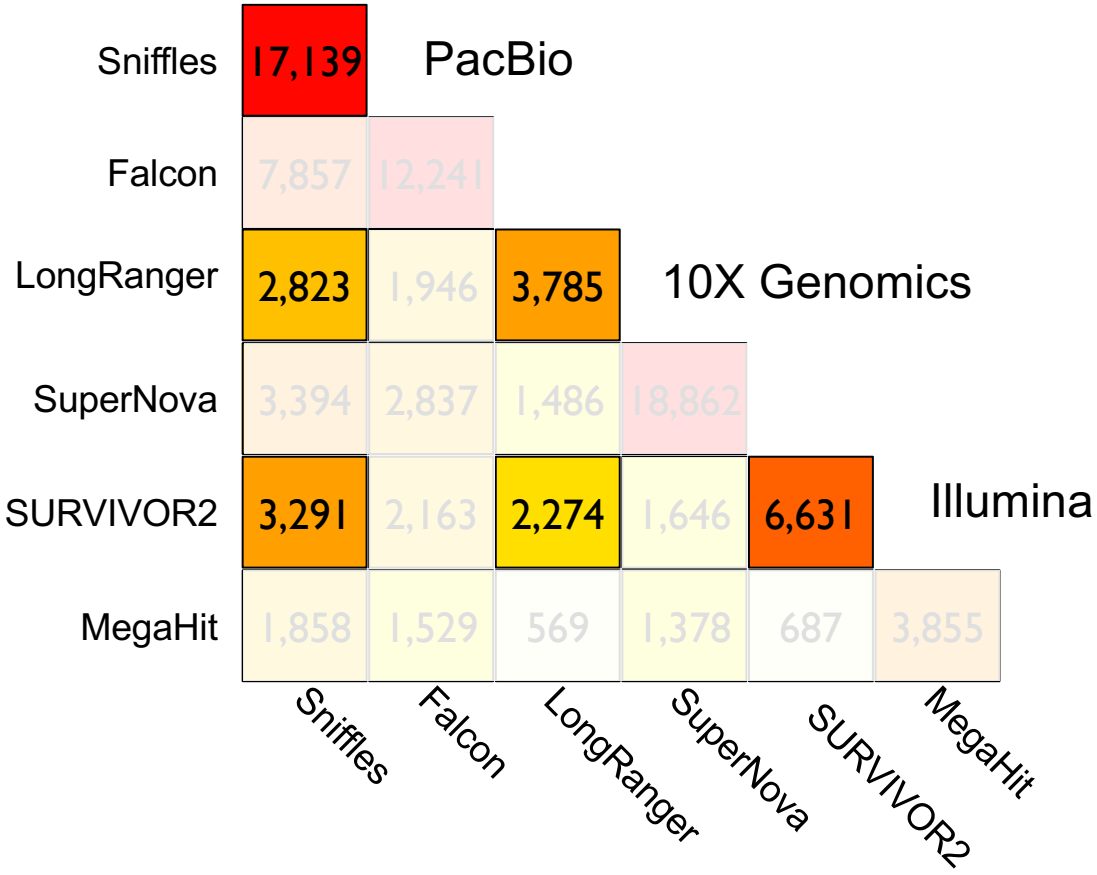
Outer triplets

- Concordance by Technology

Inner triplets

- Concordance by Assembly

Structural Variations Concordance



Main Diagonal

- Calls per tool

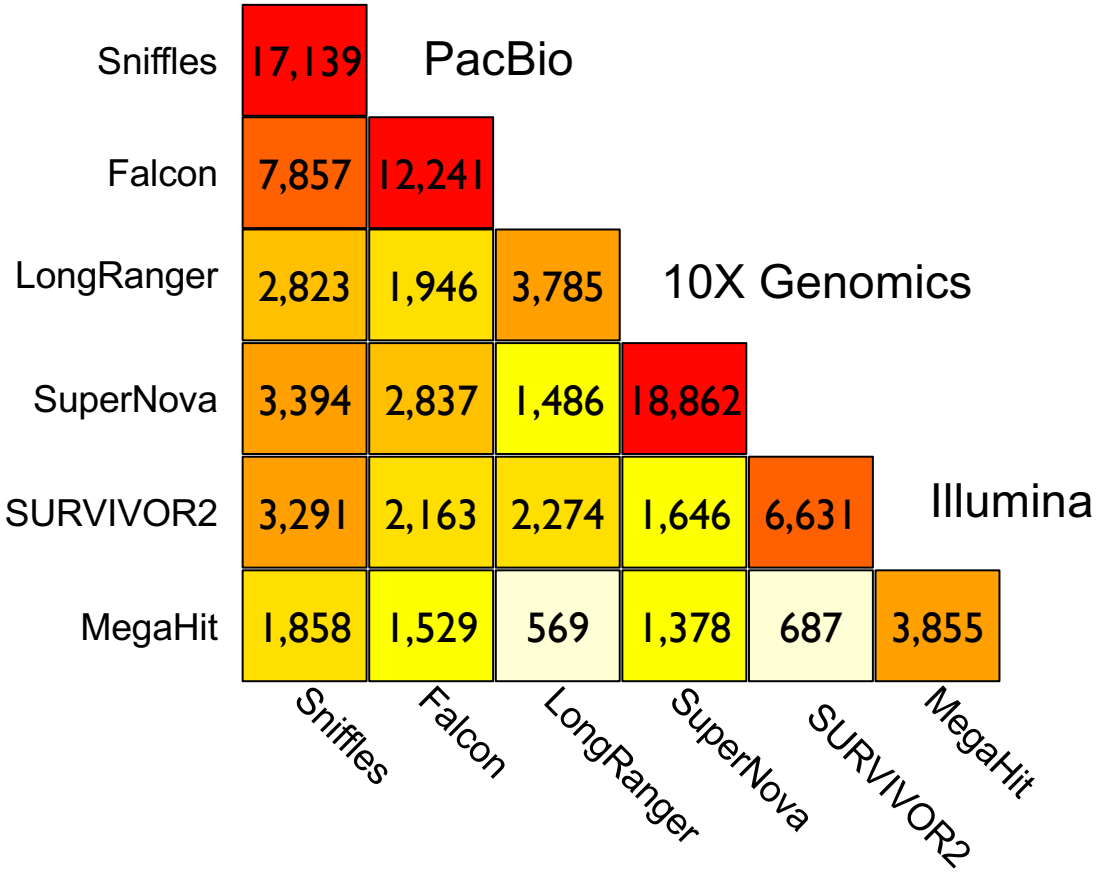
Outer triplets

- Concordance by Technology

Inner triplets

- Concordance by Assembly
- Concordance by Mappers

Structural Variations Concordance



Main Diagonal

- Calls per tool

Outer triplets

- Concordance by Technology

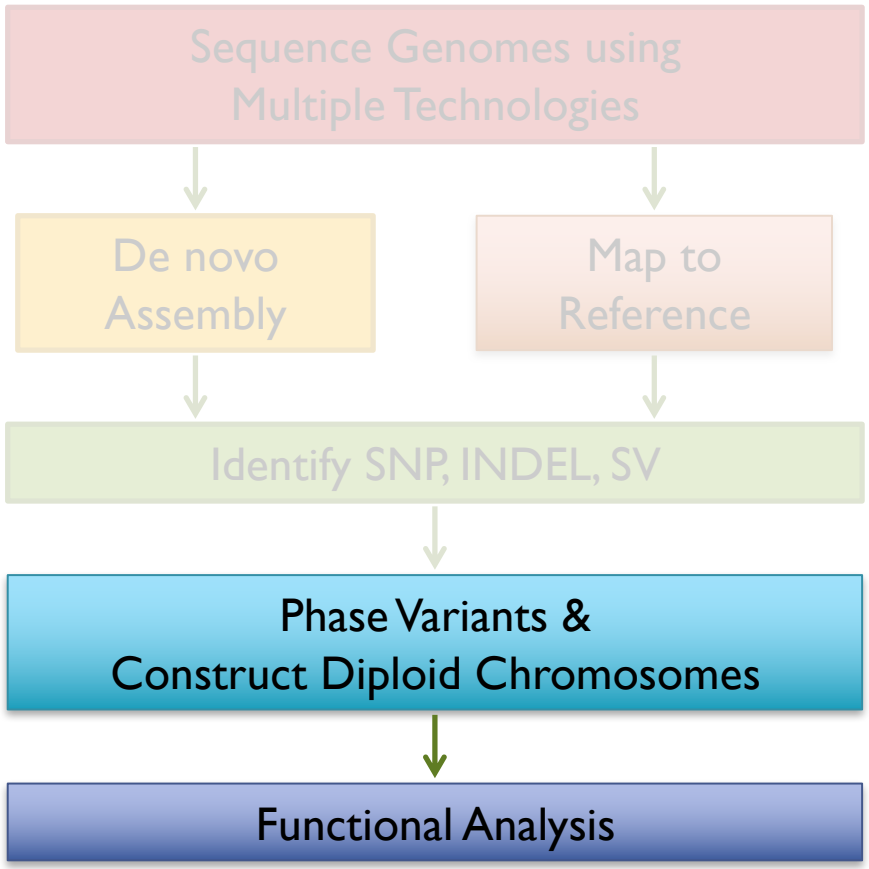
Inner triplets

- Concordance by Assembly
- Concordance by Mappers

Overall:

- We need multiple technologies and approaches

Assembling and Analyzing Personal Genomes

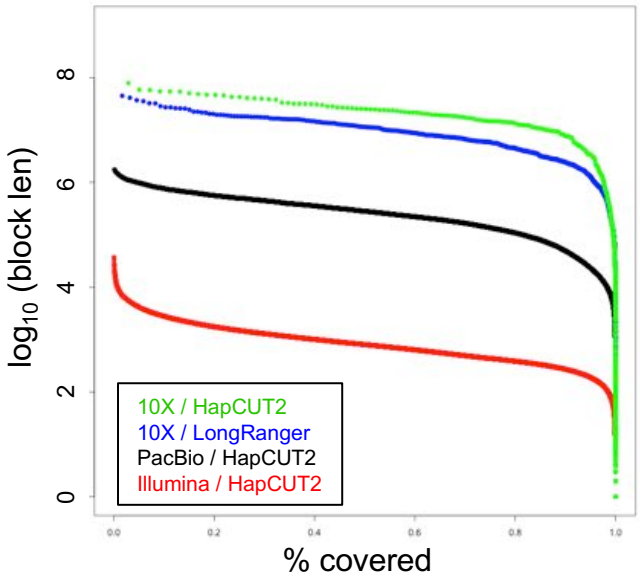


Goals

1. What are the most effective biotechnologies for sequencing?
2. What do we learn from a personalized genome instead of the reference?
3. Can we use the genomic variants as natural perturbations of the encyclopedia elements?

Phasing Results

- Phasing attempts to link together variants that came from the same molecule
- Long reads & fragments are needed to link distant heterozygous sites



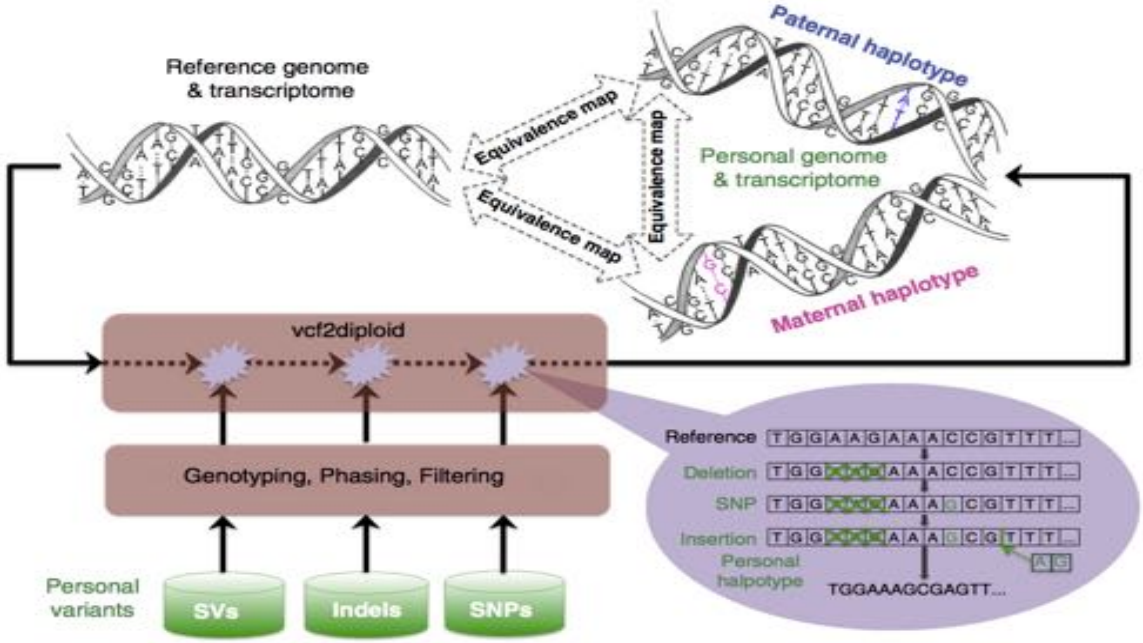
10X + HapCUT2 ***N50: 25.1 Mbp***

10X + LongRanger ***N50: 11.3 Mbp***

PacBio + HapCUT2 ***N50: 283kbp***

HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies
Edge, P, Bafna, V, Bansal, V (2016) *Genome Research.* doi: 10.1101/gr.213462.116

AlleleSeq: Constructing the Personal Genomes

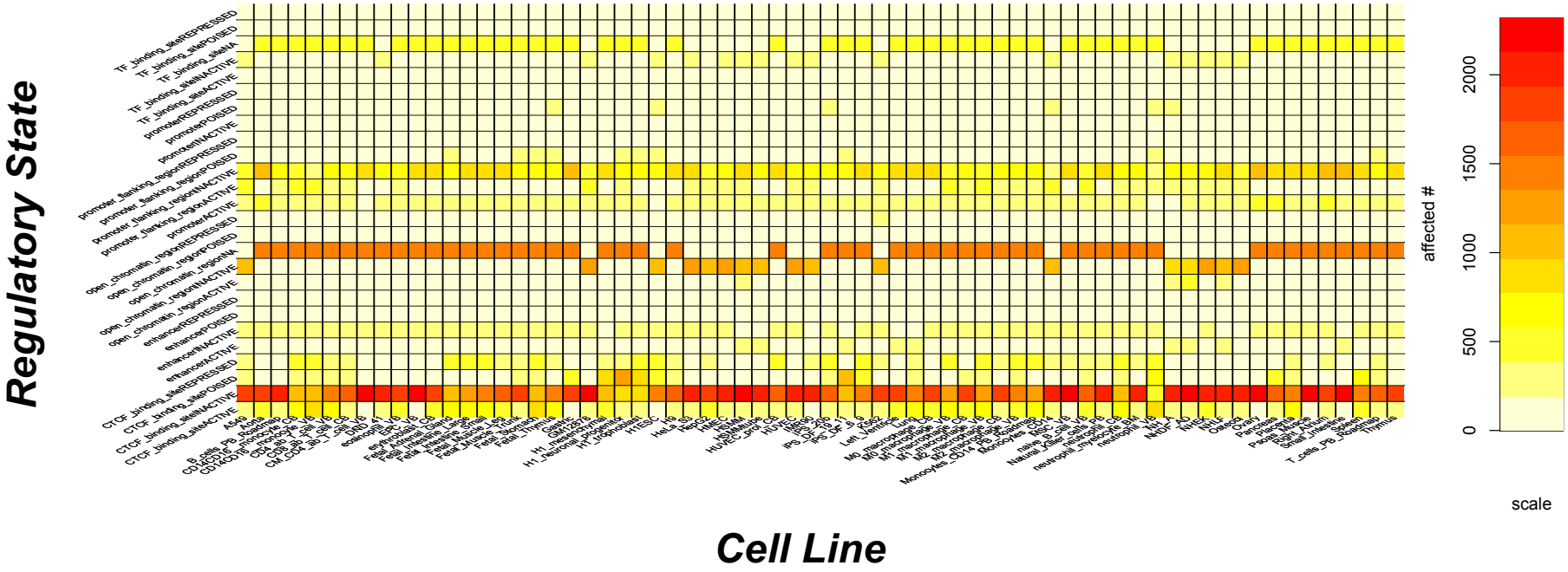


(J Rozowsky et al, 2011)

AlleleSeq/vcf2diploid inserts phased variants from a VCF file into the reference genome to create a pair of phased chromosome fasta files

Structural Variations Across the Genome

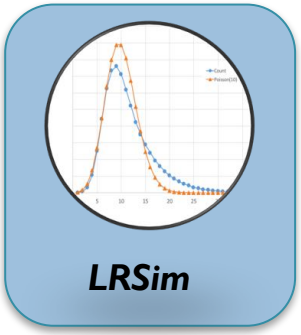
45,741 total candidate SVs
2,133 genes interrupted in coding sequence
6,363 genes interrupted in coding sequences and introns
Thousands of regulatory regions impacted



Personalized Phased Diploid Genomes

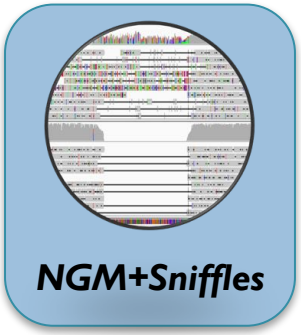
- ***Multiple sequencing technologies & approaches needed***
  - *PacBio*: Best Resolution of SVs
  - *10X*: Best Resolution of Phasing
  - *De novo*: Best Resolution of smaller events
  - *Mapping*: Best resolution of larger events

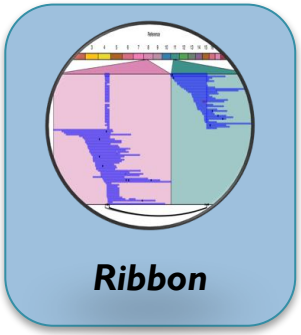
- ***We have just begun to explore the universe of variants that can be detected***
  - Tens of thousands of SVs per person`
  - Thousands of genes, thousands of regulatory elements impacted per person

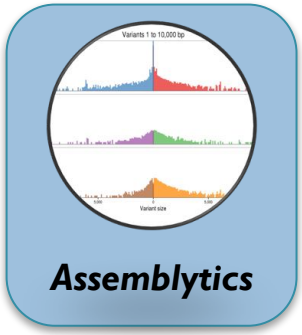

LRSim


TopSorter

SURVIVOR


NGM+Sniffles


Ribbon


Assemblytics

http://schatz-lab.org

Acknowledgements

Schatz Lab
Charlotte Darby
Han Fang
Sam Kovaka
Ruibang Luo
Maria Nattestad
Srividya Ramakrishnan
Philipp Rescheneder
Fritz Sedlazeck

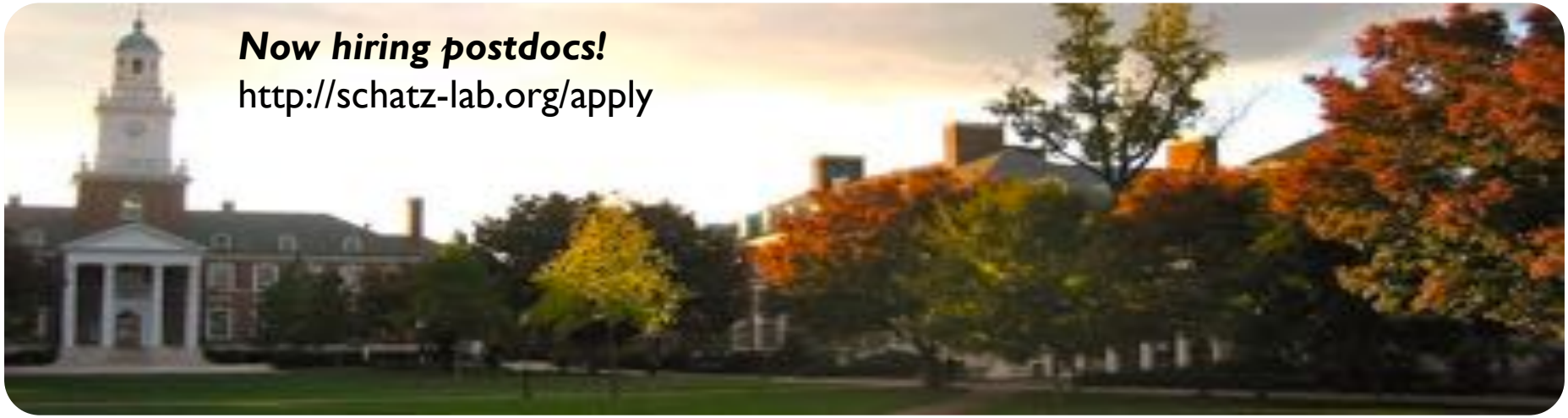
Gingeras Lab
Carrie Davis
Alex Dobin
Ashwin Prakash

McCombie Lab
Sara Goodwin

Guigo Lab
Alessandra Breschi
Anna Vlasova

ENCODE Partners
Berstein Lab
Gerstein Lab
Myers Lab
Ren Lab
Snyder Lab
Stam Lab
Wold Lab
+ All ENCODE Members

National Human Genome Research Institute
NSF
ALFRED P. SLOAN FOUNDATION
PACIFIC BIOSCIENCES®
10X GENOMICS®



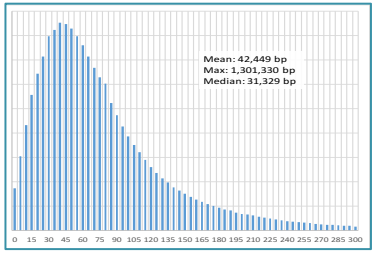
Thank you

http://schatz-lab.org

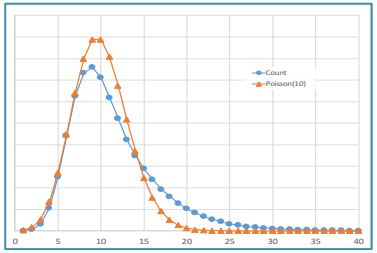
@mike_schatz

LRSim: Linked Read Simulator

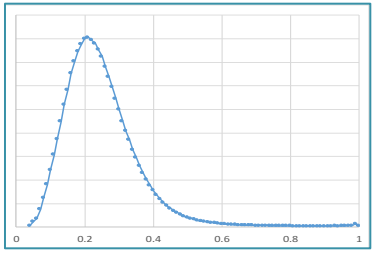
https://github.com/aquaskyline/LRSIM



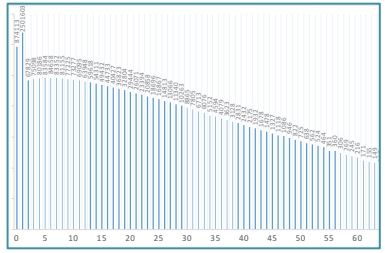
Molecule size (f)



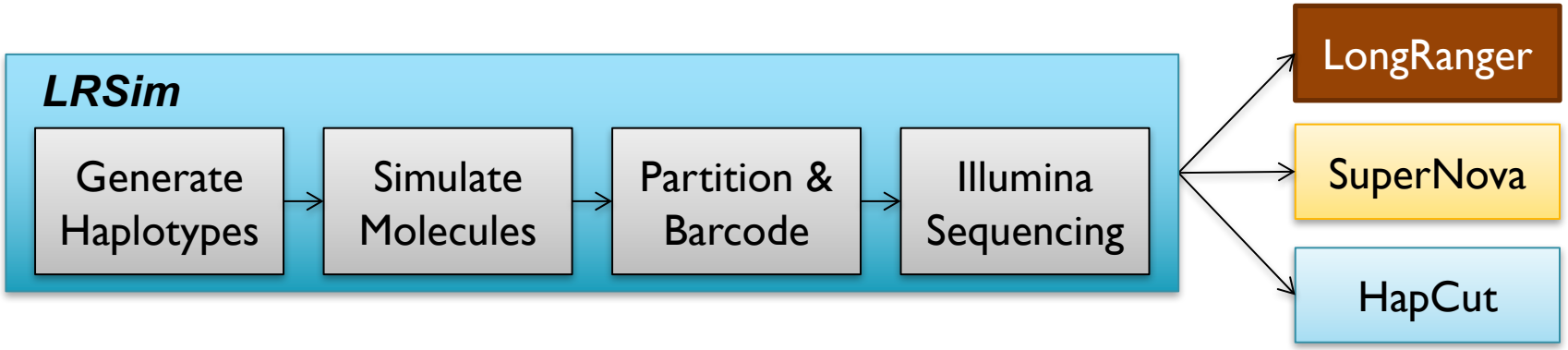
# of molecules per partition (m)



Coverage per molecule (c)



# of partitions (t)



LRSim: a Linked Reads Simulator generating insights for better genome partitioning

Luo, R et al MC (2017) bioRxiv doi: https://doi.org/10.1101/103549