Personal genome construction with SVs

M Gerstein (Yale)

What we have contributed to the SV Group

Deletion Calling
Breakpoint Assembly & Characterization
  Mechanism assignment
Functional Impact of SVs
Retroduplication Calling
SVs & personalized genome construction

...

Future directions

Personal Diploid Genome and Effects on SVs

Functional Impact of Various SVs

SV Formation Mechanism Annotation Using Long Reads

Loss-of-function Annotation

SVs and Disease Associated lncRNAs

Breakpoint Identification

Mobile Elements Using PacBio and 10x Data

CNVnator Calls on the Trios

Future directions

Personal Diploid Genome and Effects on SVs

-- one of the main purposes of having a good SV call set is to be able to build a genome for each person

-- can we move beyond a call-set-centric mindset to a personal-genome-centric mindset?

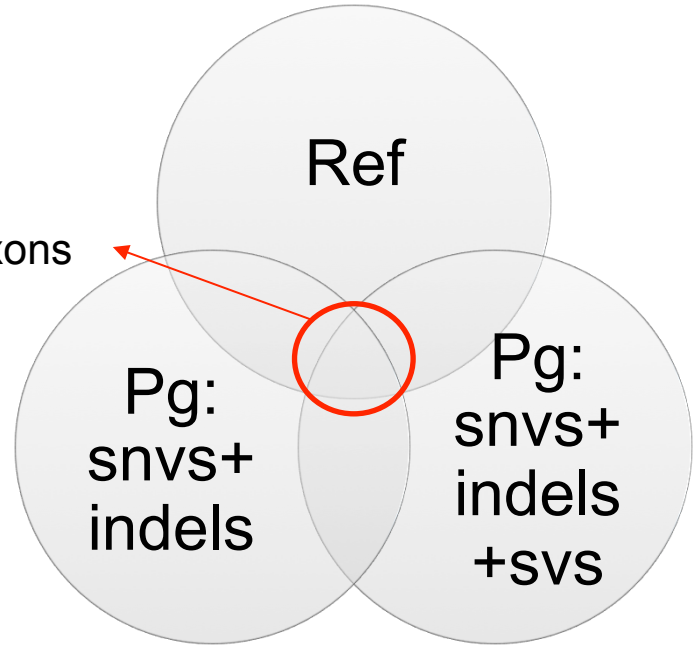
Personalized genomes analyses in Sudmant *et al., Nature* (2015)

Constructed 2 personal genomes of NA12878 based on GRCh37 reference genome

1. 1000 GP P3 **SNVs and indels** integrated call set (low coverage)
2. 1000 GP P3 **SNVs, indels and SVs** with breakpoint information from SVG

Personalized genomes analyses in Sudmant *et al.*, *Nature* (2015)

Number of exons



280,123 GENCODE consensus exons

-- this is a conservative set
-- we are also interested in SVs that completely knock genes out in the personal genomes (these would not have been included)

Oliver Stegle

Utility of SVs:
Exons with direct SV overlap

Comparing between Pgenome-SVs and Pgenome-snpsIndels

-- 18 exons with a direct SV overlap

-- 6/18 exons were expressed (>10 reads)

-- 4/6 showed substantial changes in expression (≥ 2-fold change)

[Sudmant *et al., Nature* (2015)]

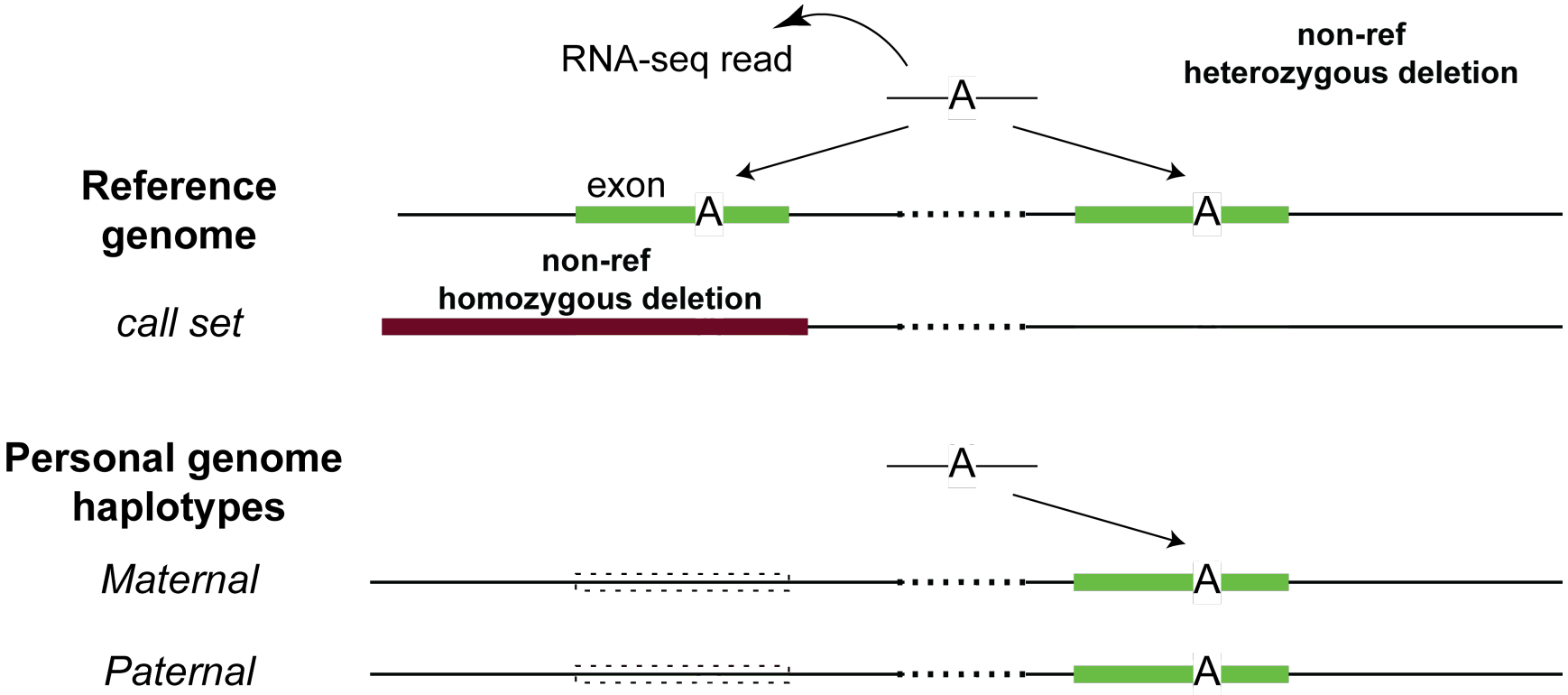
Metrics for Success

How can see how much better a personal genome is than the reference?

How can we compare one personal genome to another?

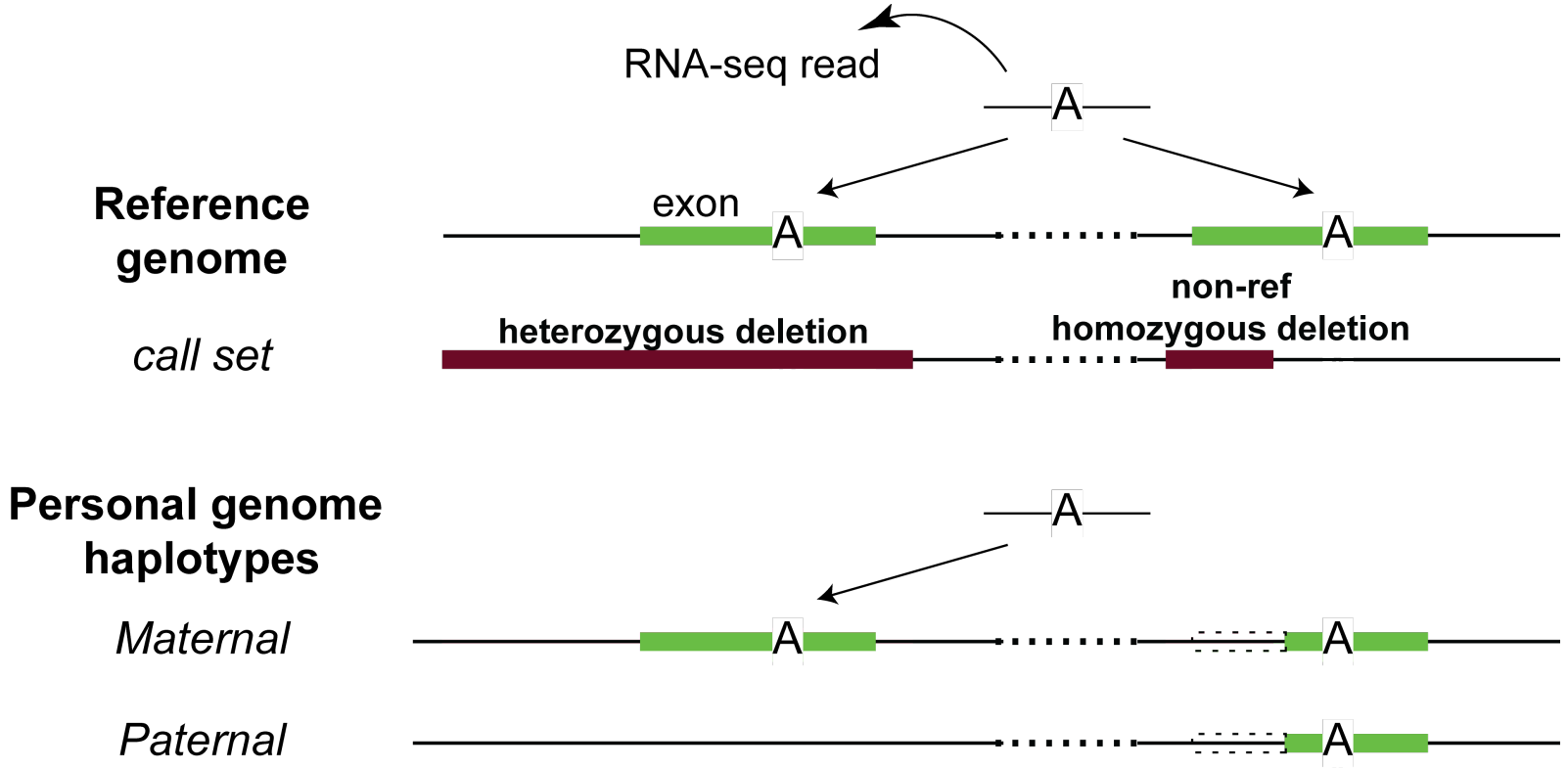
Formal Arguments
Quantities like N50, Ts/Tv

Personal genome is a better reference for alignment: a simple example



[Chen et al. *Nat. Comm.* (in rev.)]

Personal genome is a better reference for alignment : more complex



[Chen et al. *Nat. Comm.* (in rev.)]

Alignment gets better as variant sets are more complete: NA12878 Pol2 ChIP-seq

	Ref genome	Pgenome: snvs only	Pgenome: snvs + indels only	Pgenome: snvs + indels + SVs
Reads processed	208,051,087	208,051,087	208,051,087	208,051,087
# reads uniquely aligned	171,944,588 (82.65%)	172,591,380 (82.96%) M: 171,965,218 (82.66%) P: 171,969,566 (82.66%)	172,738,321 (83.03%) M: 171,982,014 (82.66%) P: 171,982,614 (82.66%)	172,743,175 (83.03%) M: 171,977,765 (82.66%) P: 171,978,147 (82.66%)
# reads that failed to align	18,279,824 (8.79%)	M: 18,290,611 (8.79%) P: 18,276,409 (8.78%)	M: 18,286,906 (8.79%) P: 18,270,944 (8.78%)	M: 18,293,522 (8.79%) P: 18,277,990 (8.79%)
# reads that multimap	17,826,675 (8.57%)	M: 17,795,258 (8.55%) P: 17,805,112 (8.56%)	M: 17,782,167 (8.55%) P: 17,797,529 (8.55%)	M: 17,779,800 (8.55%) P: 17,794,950 (8.55%)

Almost 1M increase in reads

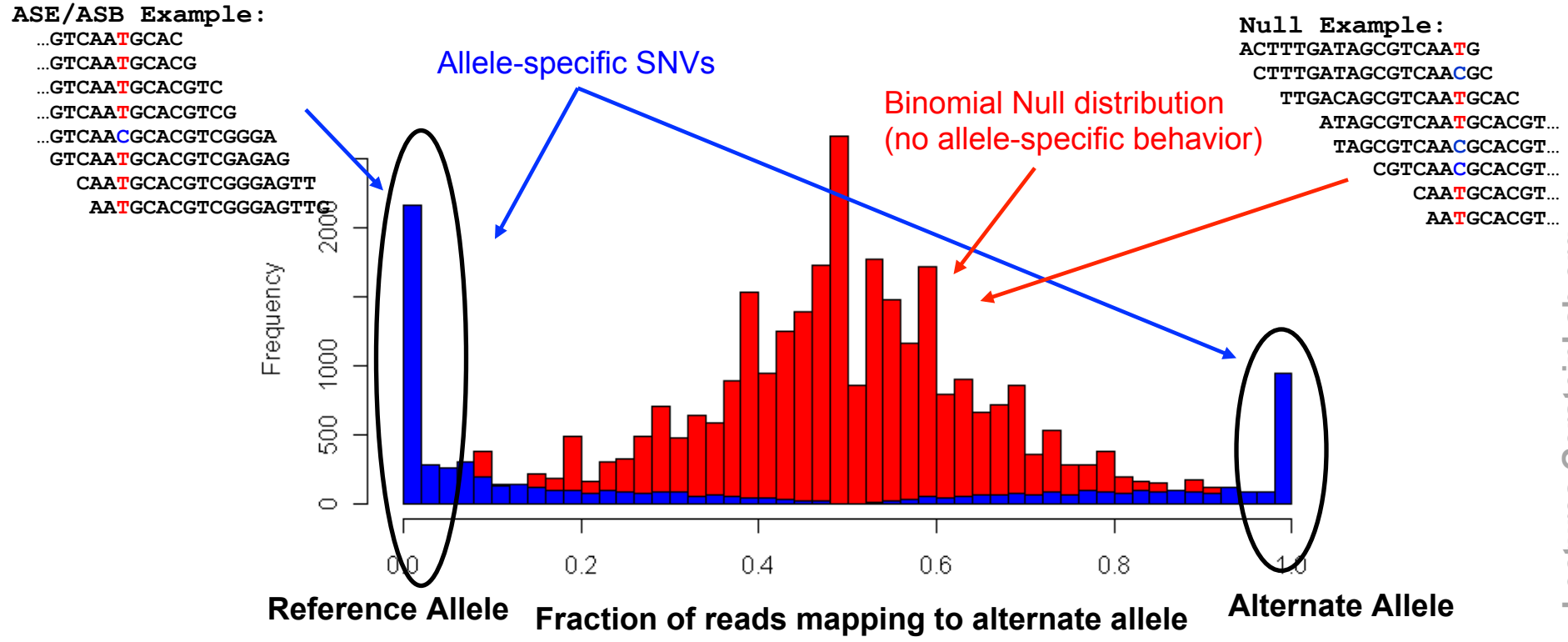
[Chen et al. *Nat. Comm.* (in rev.)]

Alignment gets better as variant sets are more complete: NA12878 RNA-seq (Kilpinen *et al.* 2013)

	Ref genome	Pgenome: snvs only	Pgenome: snvs + indels only	Pgenome: snvs + indels + SVs
Reads processed	37,558,398	37,558,398	37,558,398	37,558,398
# reads uniquely aligned	25,303,498 (67.37%)	25,486,837 (67.86%) M: 25,345,119 (67.48%) P: 25,352,964 (67.50%)	25,538,449 (68.00%) M: 25,371,892 (67.55%) P: 25,383,016 (67.58%)	25,568,042 (68.08%) M: 25,394,098 (67.61%) P: 25,412,184 (67.66%)
# reads that failed to align	8,213,405 (21.87%)	M: 8,195,227 (21.82%) P: 8,195,017 (21.82%)	M: 8,174,209 (21.76%) P: 8,172,957 (21.76%)	M: 8,181,317 (21.78%) P: 8,173,224 (21.76%)
# reads that multimap	4,041,495 (10.76%)	M: 4,018,052 (10.70%) P: 4,010,417 (10.68%)	M: 4,012,297 (10.68%) P: 4,002,425 (10.66%)	M: 3,982,983 (10.60%) P: 3,972,990 (10.58%)

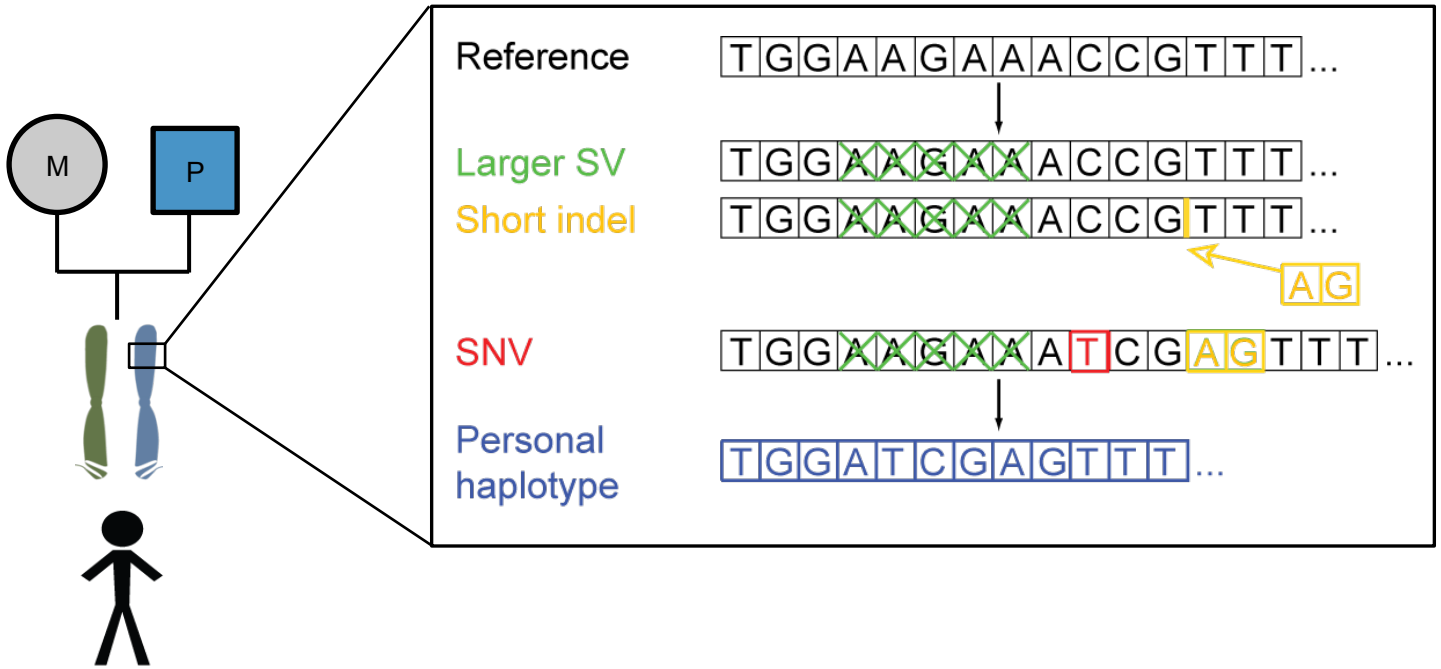
[Chen et al. *Nat. Comm.* (in rev.)]

Reference Bias in functional read mapping (naïve alignment against reference)



[Rozowsky *et al. MSB* ('11)]

Personal genome construction



[Rozowsky *et al. MSB* ('11)]

Construction considerations

1. **Choice of call set(s)**

-- e.g. different versions of 1000GP call sets

2. **Choice of reference**

-- choose the reference genome in which the call set is derived from

3. **Choice of variants**

-- e.g. SVs or indels or SNVs only

4. **Order of incorporation**

-- SVs > indels > SNVs

-- Increased SNV density around SV breakpoints (Abyzov *et al.* 2015)

Many choices of NA12878 call sets: Which one do we use?

1. Genome in a bottle
2. Complete Genomics
3. Illumina Platinum Genomes
4. Broad's GATK Best Practices bundles
5. 1000 Genomes Project
   -- low/high coverage
   -- long reads: PacBio, Moleculo

Assessing quality of call set: Mendelian inconsistency (e.g. GATK HC PCR-free CEU trio)

NA12891 Father	NA12892 Mother	NA12878			total	%Err
		RR	RA	AA		
RR	RR	0	6072	311	---	---
RR	RA	518631	505499	1215	1025345	0.12
RR	AA	1659	194589	1806	198054	1.75
RA	RR	507750	506699	1110	1015559	0.11
RA	RA	194409	397233	195245	786887	---
RA	AA	742	194722	206720	402184	0.18
AA	RR	1485	193636	1551	196672	1.54
AA	RA	653	198416	202366	401435	0.16
AA	AA	113	1316	816825	818254	0.17

*Autosomes only

Jieming Chen

Future development of personal genome construction

1. Iterative personal genome construction

iteration 1: construct pgenome per before

-- map DNA reads to pgenome

-- refine variants on pgenome

iteration 2: rebuild pgenome

2. What makes a personal genome ‘better’ than the other one

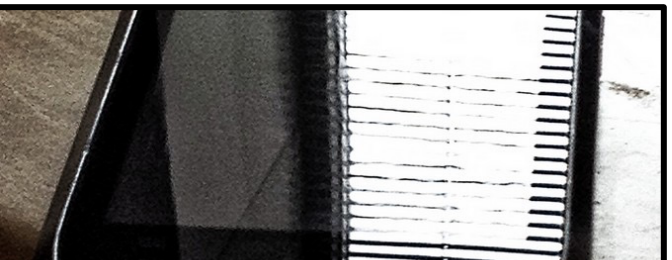
-- metrics to describe a ‘good’ personal genome

Current and future efforts in personal genome construction

- 382 personal genomes from 1000 Genomes Project with RNA-seq and/or ChIP-seq sets -- with only SNVs and indels
- Other high coverage trios in 1000GP and Svtrios

SV Group

Alexej Abyzov
Sushant Kumar
Shantao Li
Fabio Navarro
Yan Zhang



Acknowledgements

Personal genomes

Jieming Chen
Joel Rozowsky
Rob Kitchen
Timur Galeev
Alexej Abyzov
Arif Harmanci