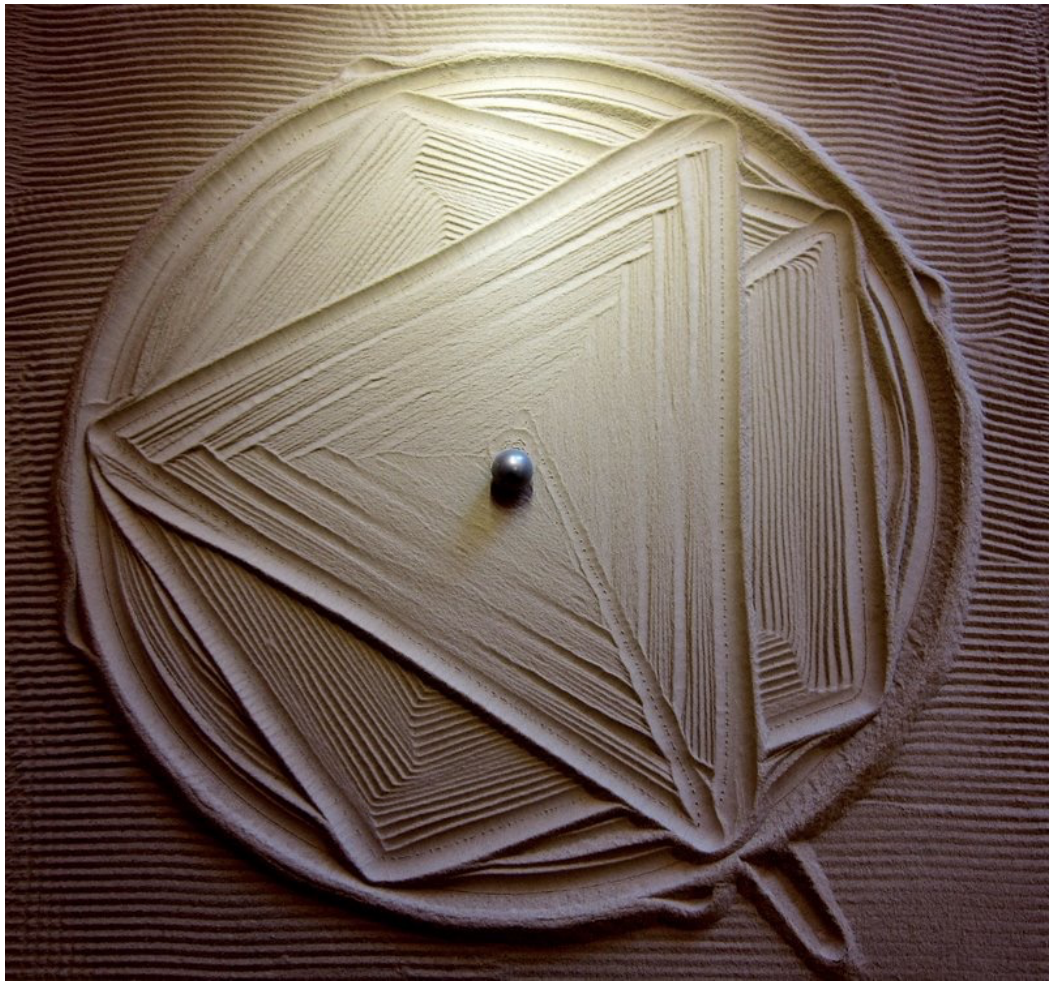
SV Call Sets & Personal Genomes:

new retroduplication calls & building a personal genome with PacBio SVs



Mark Gerstein, Yale

See last slide for more info.

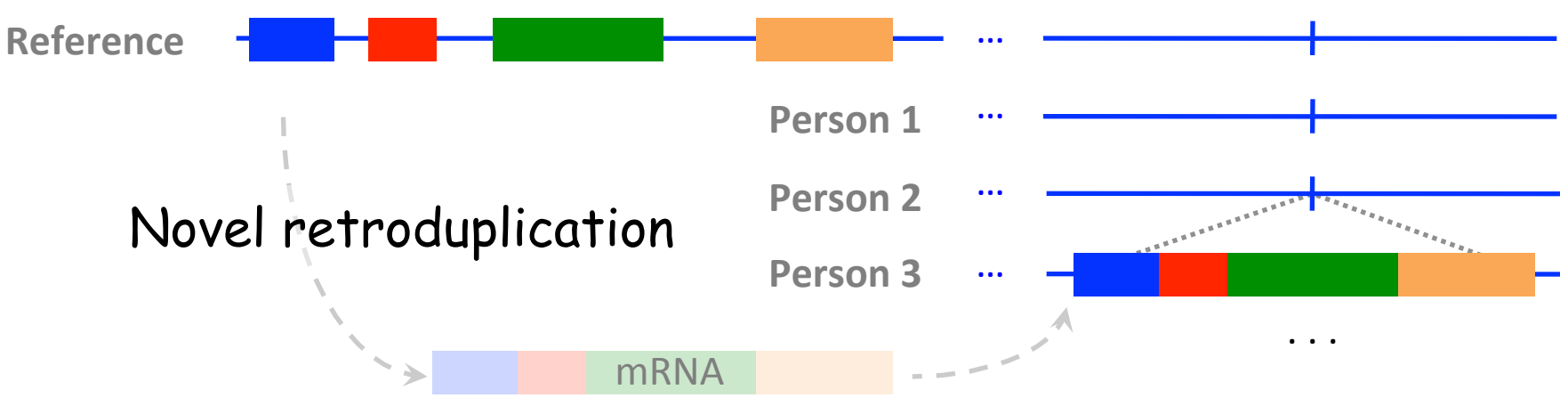
SV Call Sets & Personal Genomes

- Retroduplication SV calls
  - New call set, now for Trio data
- Personal Genomes from SV calls
  - Trying to incorporate a PacBio SV call set
  - Trying to demonstrate QC metrics on genome quality
  - Scaling up

Retroduplication variation (RDV)

- RetroCNVs are duplications of messenger RNAs (mRNAs) mediated by L1 retrotransposons
  - Create intronless copies of protein coding genes with polyA & direct repeats flanking the insertion.
  - Some of these duplications are unfixed in human populations
- Previous callsets based on Low-cov. Illumina WGS. Ongoing working on high-coverage WXS & WGS and also using PacBio

Schrider, D. R., Navarro, F. C. P., Galante, P. A. F., Parmigiani, R. B., Camargo, A. A., Hahn, M. W., & de Souza, S. J. (2013). Gene copy-number polymorphism caused by retrotransposition in humans. PLoS Genetics,
Abyzov, A., Iskow, R., Gokcumen, O., Radke, D. W., Balasubramanian, S., Pei, B., et al. (2013). Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. Genome Research.



New call set based on PCR-free high-cov trio data

CHS

11 — HG00512

ANPEP(HET)
BAZ2A(HET)
CBX3(HOP)
GCSH(HOP)
PCMTD1(HOP)
SET(HOP)
SKA3(HOP)
TDG(HOP)
RBMX (HET)
RALBP1 (HET)
P4HA1 (HET)
FTH1(HET)

10 — HG00513

ANPEP(HET)
CBX3(HOP)
GCSH(HOP)
PCMTD1(HOP)
TDG(HET)
BAZ2A(HET)
RBMX (HET)
RALBP1 (HET)
P4HA1 (HET)
FTH1(HET)

12 — HG00514

ANPEP(HET)
BAZ2A(HET)
CBX3(HOP)
GCSH(HOP)
PCMTD1(HOP)
SET(HET)
SKA3(HET)
TDG(HET)
RBMX (HET)
RALBP1 (HET)
P4HA1 (HET)
FTH1(HET)

PUR

12 — HG00731

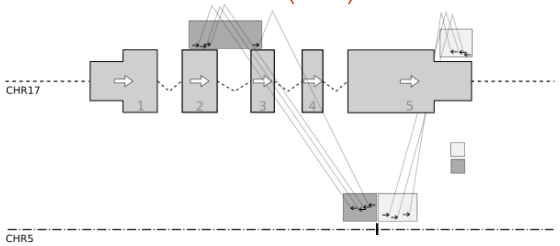
ANPEP(HET)
BAZ2A(HET)
GCSH(HOP)
PCMTD1(HOP)
RPL36(HET)
SET(HET)
TDG(HOP)
SET(HET)
RBMX (HET)
RALBP1 (HET)
P4HA1 (HET)
FTH1(HET)

13 — HG00732

BAZ2A(HET)
CBX3(HET)
GCSH(HOP)
HCAR2(HET)
PCMTD1(HOP)
SKA3(HET)
TDG(HOP)
TYRO3(HET)
SET(HET)
RBMX (HET)
RALBP1 (HET)
P4HA1 (HET)
FTH1(HET)

9 — HG00733

BAZ2A(HET)
CBX3(HET)
GCSH(HOP)
PCMTD1(HOP)
SKA3(HET)
TDG(HOP)
RALBP1 (HET)
P4HA1 (HET)
FTH1(HET)



YRI

14 — NA19239

TDG(HOP)
BAZ2A(HET)
ATP9B(HET)
CBX3(HET)
GCSH(HOP)
MFF(HET)
PCMTD1(HOP)
RPLP0(HET)
SET(HET)
TMEM5(HET)
RBMX (HET)
RALBP1 (HET)
P4HA1 (HET)
FTH1(HET)

11 — NA19240

TDG(HOP)
CBX3(HET)
GCSH(HOP)
PCMTD1(HOP)
RPLP0(HET)
SET(HET)
BAZ2A(HET)
FTH1(HET)
RALBP1 (HET)
RBMX (HET)
P4HA1 (HET)

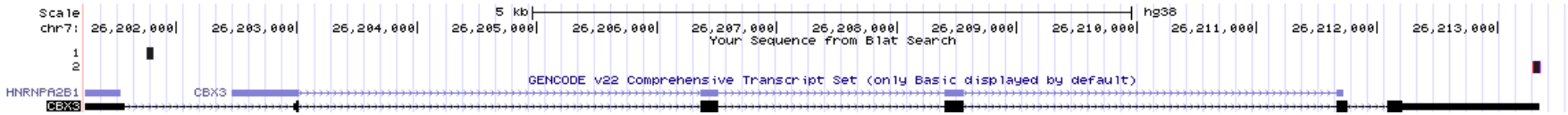
14 — NA19238

ATP9B(HET)
BAZ2A(HET)
CBX3(HET)
GCSH(HOP)
MFF(HET)
PCMTD1(HOP)
RPLP0(HET)
SET(HET)
TDG(HOP)
TMEM5(HET)
RBMX (HET)
RALBP1 (HET)
P4HA1 (HET)
FTH1(HET)

Breakpoint analysis from Illumina High Coverage Call Set

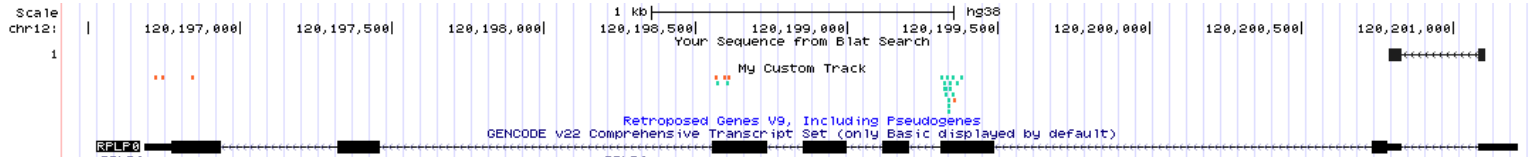
CBX3 inserted into chr15

chr15 40561981 0 313 CTTCGGATGTGGCTTGAGCTGTAGGCGCGGAGGGCCGGAGACGCTGCAGACCCGCGACCCGGAG
chr15 40561992 1 46 ATTTTTTTTTTTAAAGAAATATATAACTATTTATTAACCACTGTTCAGTATTTACAATAAAGTAAAC



RPLP0 inserted into chr11

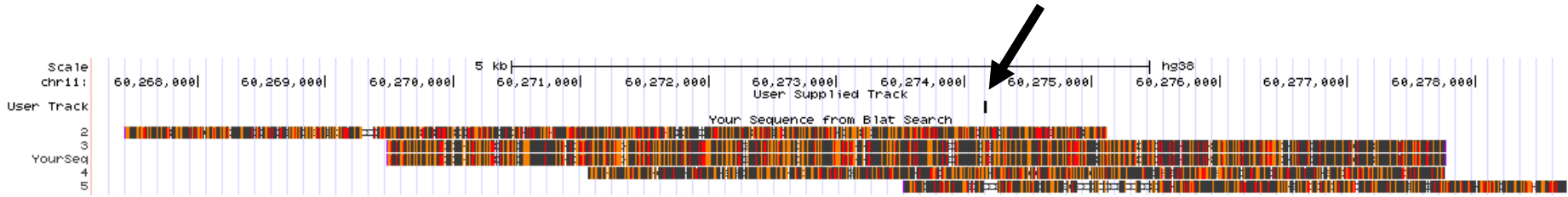
chr11 60274156 0 6 TTTTTTTTTTTTTTTTTAAGAATTAAGCCTTTTTTTCTTTTTTTTTTAATTAAATCTGGCATAGTTGGTTATTTTTTGTGT
chr11 60274167 0 55 CTCTCGCCAGGCGTCCTCGTGGAAGTGACATCGTCTTTAAACCCTGCGTGGCAATCCCTGACGCACCGCC



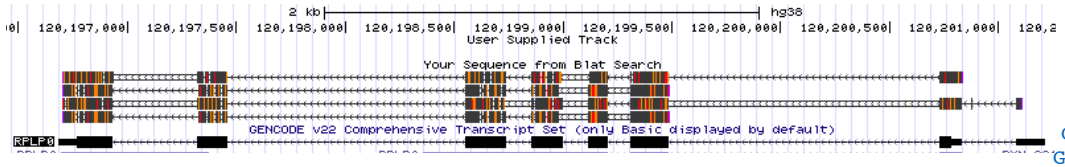
Breakpoints analysis is able to identify all insertions extremities, as well as Target Site Duplications and poly(A)s

Validation of a Single Event using PacBio data
RPLP0 (chr11:60274156-60274179)

Insertion Point



Parental Gene

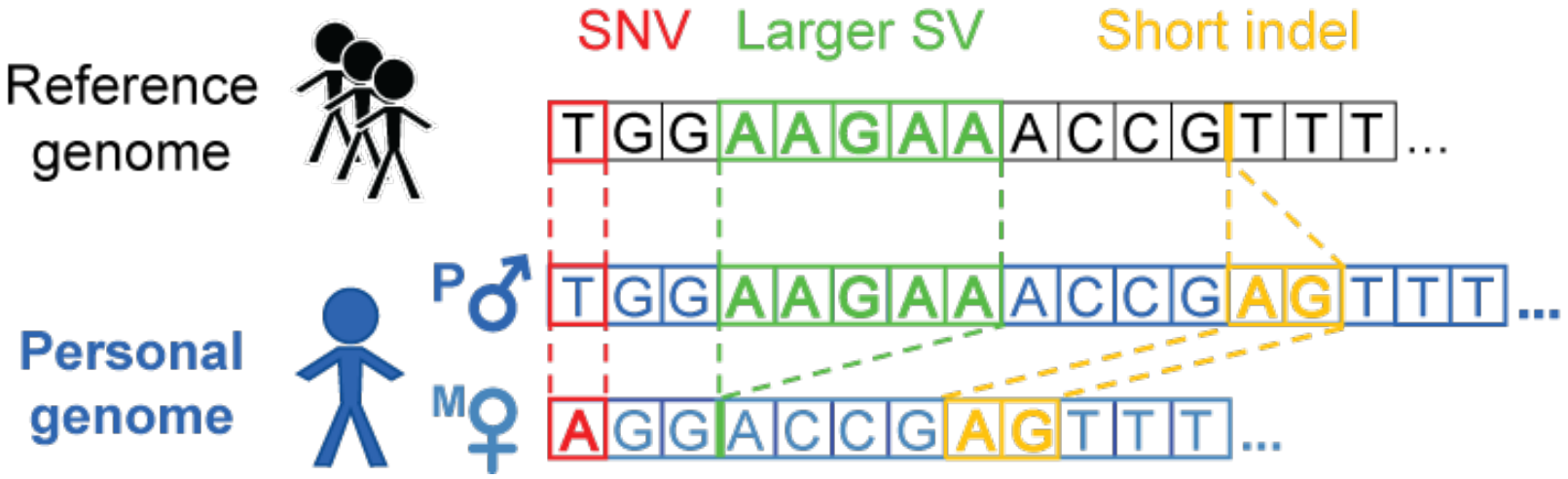



```
CTCTCGCCA GGggCgTCCT CaGTGGAAGT GACATCGcTC TTTAAACCCT  5050
GCGTGCAATC CCTGaACGCA CCGCCGTGAT GCCCAGGGAA GACAaGGGCG  5100
ACCTGGAAGT CCAACTACTT CTTAAGaATC ATCCAACTAa TGGATGATTA  5150
TCCgAAATGt TTTCATtGTG GGAGCAGACA ATGTGGGCTC CAAGCAGATG  5200
gCAGCAGATC CGaATGTCCC tTTCGCGGGA AGGctGTGGT GCTGATGGGC  5250
AAGAACACCA TGATGCGCAA GGCCATCCGA GGCACCTGGA AAACAACCCA  5300
GCTCTGagaA AACTGCTGCC TCgttTATCc GGGGGAATGT GGGCTTTGTG  5350
TTCACCAAGG AGGACCgtTC ACTGAaGATC AGGacCATGT TGCctggcCC  5400
                                   …
ATgcagaCCC ATTCTAaTCA TCAACaGGGT AgCAAACGAG TCCTCCTTGT  5850
CTGTGGAGAC GGATTACACC TTCCAaTTGC TGAAgGTCAA GGCCTTCTTG  5900
GCTaCCATCT gGCCTTTGTG GCTGCcTGCC CCTGTGGCTG CTGCACCACA  5950
GCTCTCCaTG CTGCTGCaTG CAGcCCCCAG CTAAGGTTGA AGCCAAGGAA  6000
GAGTaGGAGG AGTCGGcacg aAGGATATGG GATTTGTCTC TTGACTAAat  6050
ACCAAAgcaa AcCCAACTct agGCCAGgtT TTATTTGCAA ACAAGAAATA  6100
AAGGCTTACT TCTTTAAAAA Aaaaaaaaaa aaaaaacttc cgaaactgaa  6150
gagcaaaagg gaaaaaaatg gaaaaaaaga agcagaacaa cccaaagtgc  6200
```

SV Call Sets & Personal Genomes

- Retroduplication SV calls
  - New call set, now for Trio data
- Personal Genomes from SV calls
  - Trying to incorporate a PacBio SV call set
  - Trying to demonstrate QC metrics on genome quality
  - Scaling up

How we build a personal genome



reference, .fasta

+

personal variants, .vcf

→

personal diploid genome, .fasta

+

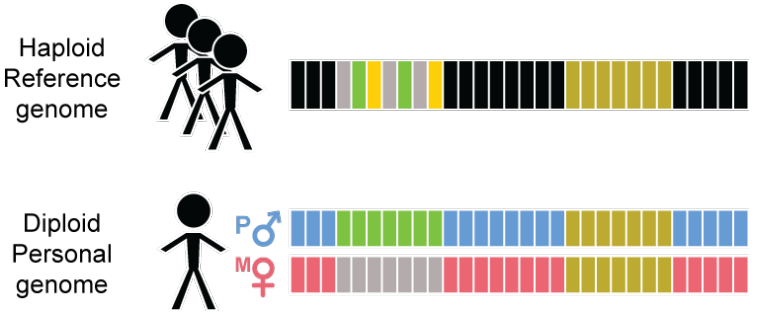
coordinate offset files, .chain

alleleseq.gersteinlab.org

Rozowsky *et al. Mol Syst Biol* (2011)

Why the personal genome (PG) should be a platform for functional genomics

1. **Diploid**
   --Ability to incorporate diverse variants of any size
   --exhibit phase information
2. **Scales easily with more samples & improve with development of sequencing technologies: longer reads and more accurate phase information**
3. **Demonstrably useful in functional genomic assay analyses**
   a) read alignment
   b) RNA-seq quantification
   c) allele-specific analyses



Evolution of NA12878 family of Personal Genomes

	Source	RefGen	Depth	Variants
1	1000 Genomes Project (1000GP) pilot (used for Rozowsky et al., ('11), alleleseq.gersteinlab.org)	hg18	60x	SNVs, indels, deletions (including 33 from fosmid sequencing)
2	GATK Best Practices v3 (UnifiedGenotype)	hg19	64x	SNVs, indels
3	GATK Best Practices v4 (HaplotypeCaller, PCR-free)	hg19	64x	SNVs, indels
4	1000GP Phase 3 SNVs-only	hg19	7.4x	SNVs
5	1000GP Phase 3 SNVs-indels	hg19	7.4x	SNVs, indels
6	1000GP Phase 3 SNVs-indels-SVs	hg19	7.4x	SNVs, indels, SVs
7	**1000GP Phase 3 SNVs-indels-SVs**	hg19	7.4x	SNVs, indels, SVs
8	**GIAB NA12878 pilot genome**	hg19	12x-190x	SNVs, indels, SVs

[7] Updated version of PG used in Sudmant et al, (Nature'15) [#6],
now with added complex SVs Pindel calls

[8] Incl. PacBio-based SV call set from GIAB
SNVs and Indels: High-confidence call set based on 11 WGS & 3 ES datasets (Zook et al, Nat Biotech '14);
SVs: Preliminary PacBio-based call set from
ftp://ftp- trace.ncbi.nih.gov/giab/ftp/data/NA12878/analysis/BCM_PacBio_PBHoney_15.8.24_09012015/

Functional genomics assay read alignment *slightly* improves as variant sets get more complete

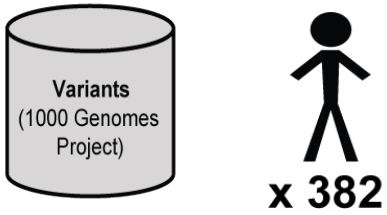
	Reference Genome	1KGP-SV based PG			GIAB based PG		
		SNVs	SNVs & Indels	SNVs, Indels & SVs	SNVs	SNVs & Indels	SNVs, Indels & SVs
# reads uniquely mapped	14,685,701 (**78.20%**)	14,796,823 (**78.79%**)	14,838,547 (79.02%)	14,840,308 (**79.03%**)	14,724,469 (**78.41%**)	14,749,285 (78.54%)	14,754,951 (**78.57%**)
# reads that multimap	671,519 (3.58%)	664,706 (3.54%)	664,876 (3.54%)	663,211 (3.53%)	671,488 (3.58%)	671,695 (3.58%)	670,074 (3.57%)

~18.8 M **Illumina HiSeq 2000 50bp PE RNA-Seq** (Kilpinen et al., *Science*, 2013) reads mapped with STAR: stringent alignment parameters (< 3 mismatches, no short gaps/deletions or soft-clipping permitted)

	Reference Genome	1KGP-SV based PG			GIAB based PG		
		SNVs	SNVs & Indels	SNVs, Indels & SVs	SNVs	SNVs & Indels	SNVs, Indels & SVs
# reads uniquely mapped	554,236 (77.42%)	559,541 (78.16%)	559,569 (78.16%)	559,584 (78.16%)	555,570 (77.60%)	555,706 (77.62%)	555,822 (77.64%)
# reads that multimap	2,812 (0.39%)	2,810 (0.39%)	3,081 (0.43%)	3,083 (0.43%)	2,871 (0.40%)	2,904 (0.41%)	2,938 (0.41%)

~716 K **PacBio RNA-Seq** (Sharon et al., *Nat. Biotechnol*, 2013) reads mapped with STAR
(https://github.com/PacificBiosciences/cDNA_primer/wiki/Bioinfx-study:-Optimizing-STAR-aligner-for-Iso-Seq-data)

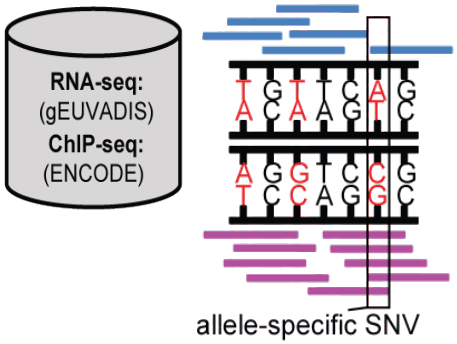
Scaling Personal Genome Construction to 382 1000G individuals



- Construction of PGs is scalable:

We built PGs of 382 individuals using 1000G & trio project variants

Deposited them into the 1000G Project DCC



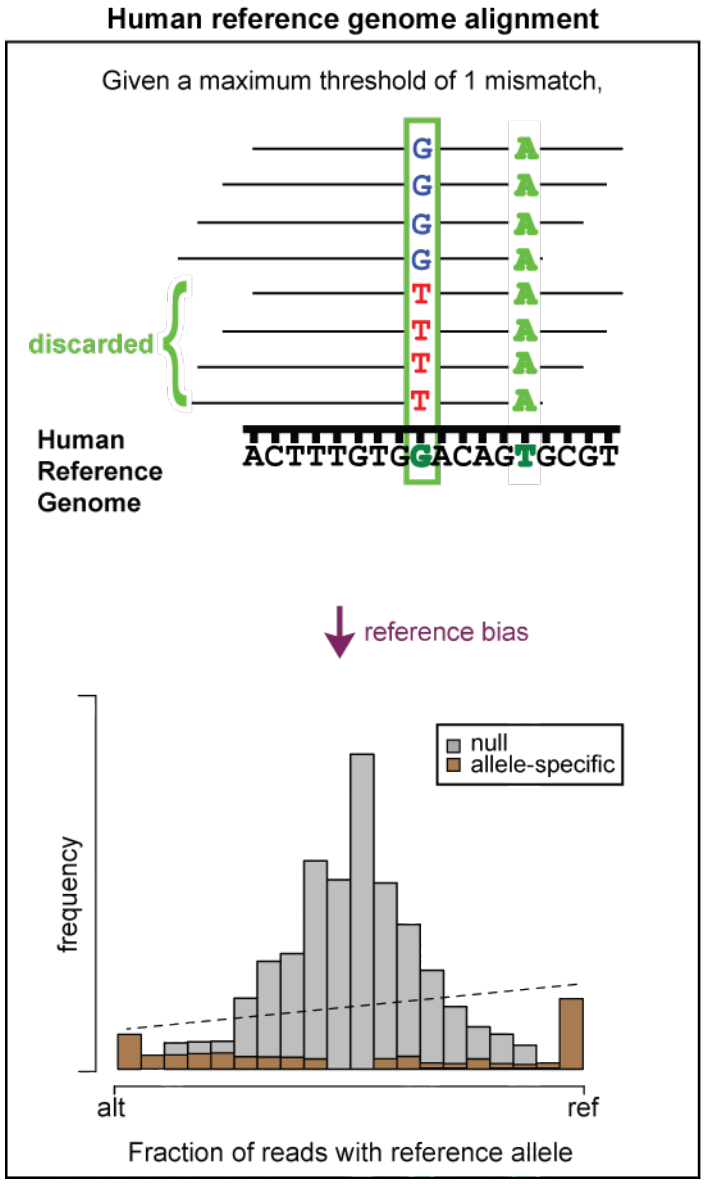
- Analyses of functional datasets based on allelic read counts are most sensitive to mapping biases:

We have developed approaches to account for reference bias, ambiguous mapping bias and read over-dispersion within the PG framework

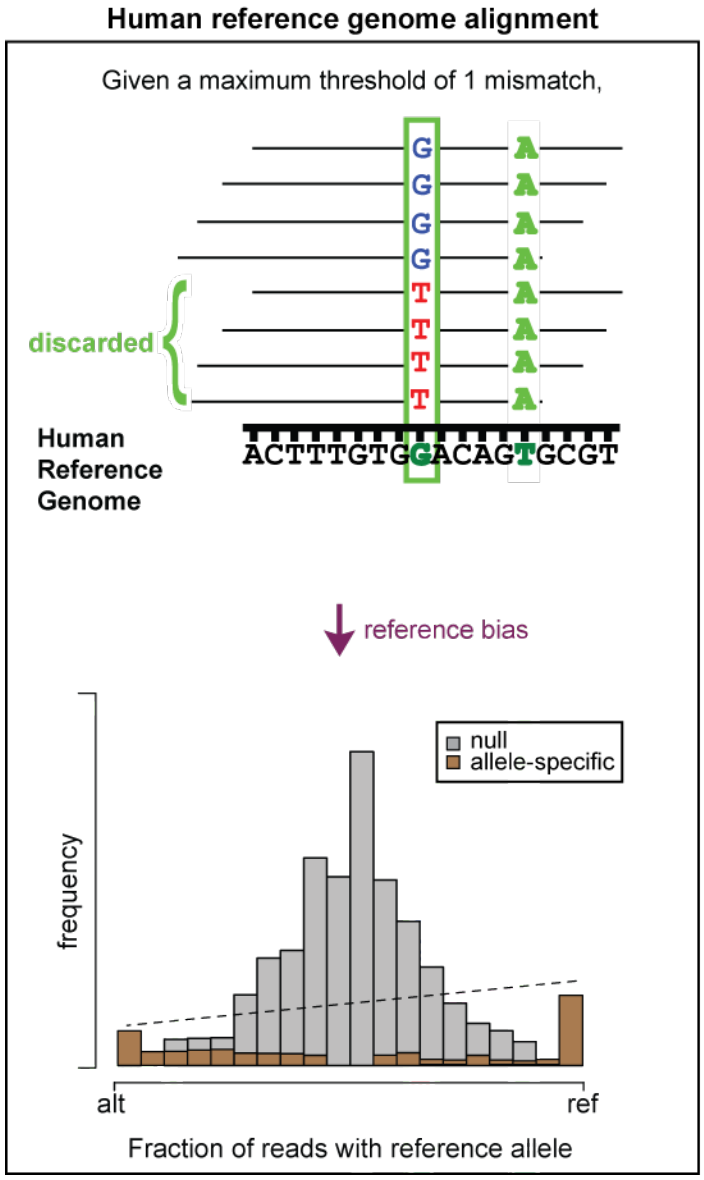
- The PGs and allele-specific annotation of their variants are available from *alleledb.gersteinlab.org*

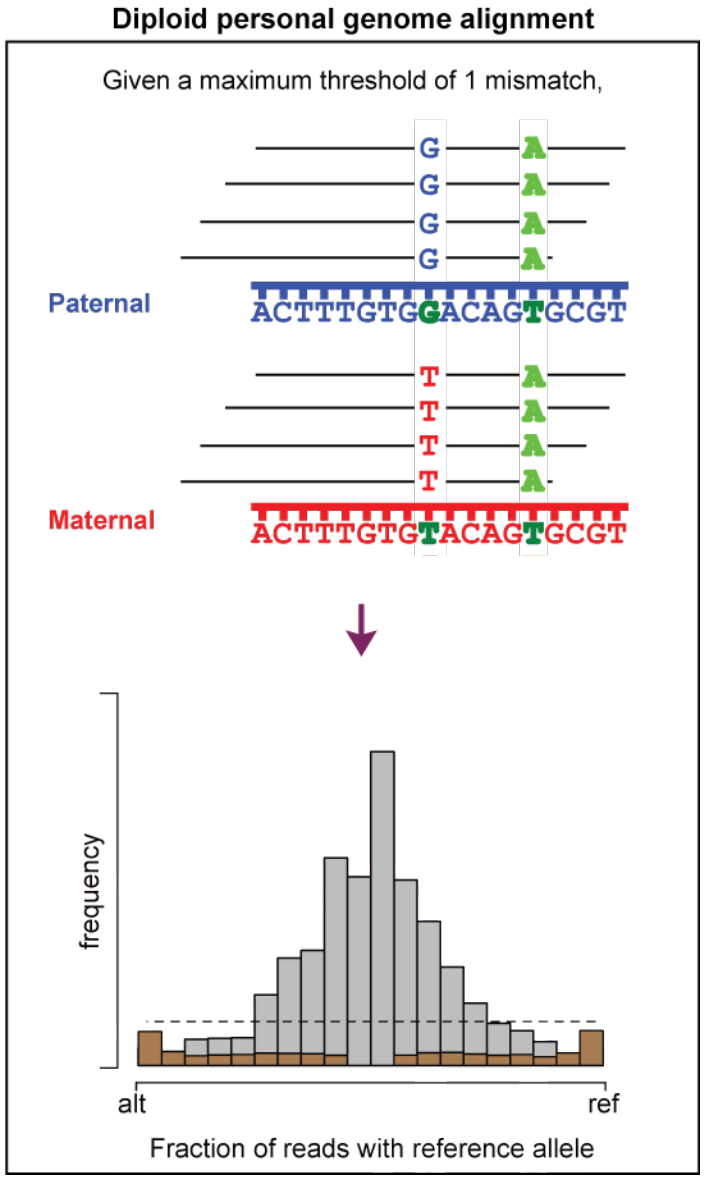
[Chen *et al. Nature Comm.* '16]

PG alleviates reference bias in alignment



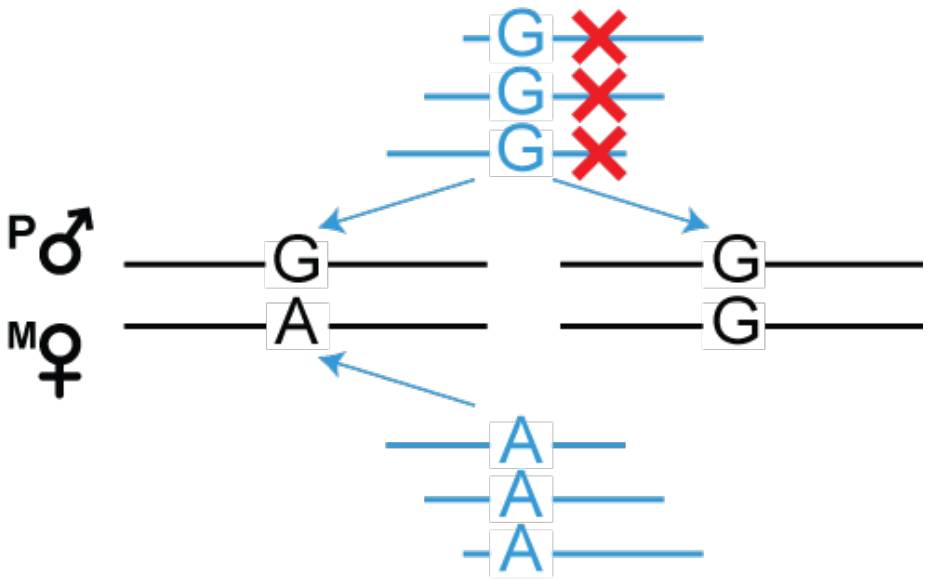
PG alleviates reference bias in alignment





Ambiguous mapping bias

Ambiguous mapping bias (AMB): simple removal of multi-mapping reads may lead to false AS signal



[Chen *et al. Nature Comm.* '16]

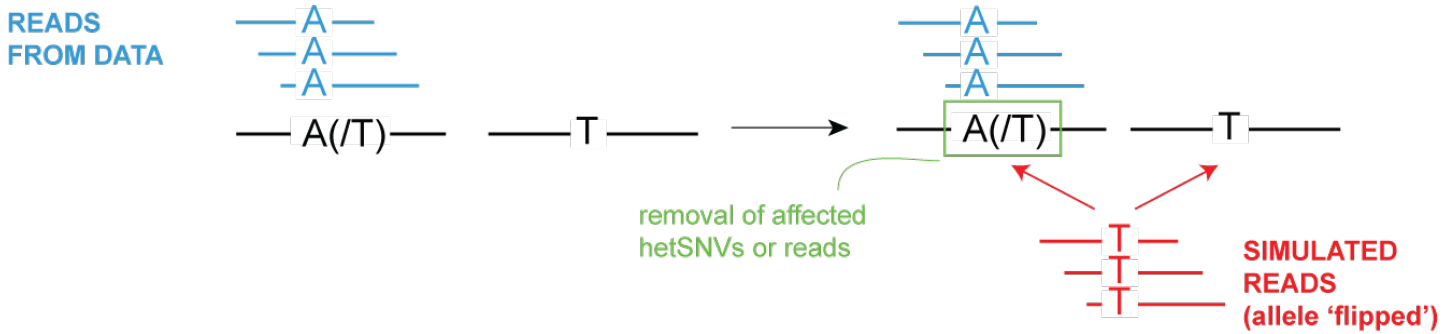
Account for ambiguous mapping bias: reference genome

Current approaches to deal with the reference and ambiguous biases commonly involve

- Filtering sites with potential bias based on mapping of simulated reads generated from genomic sequence;

  or

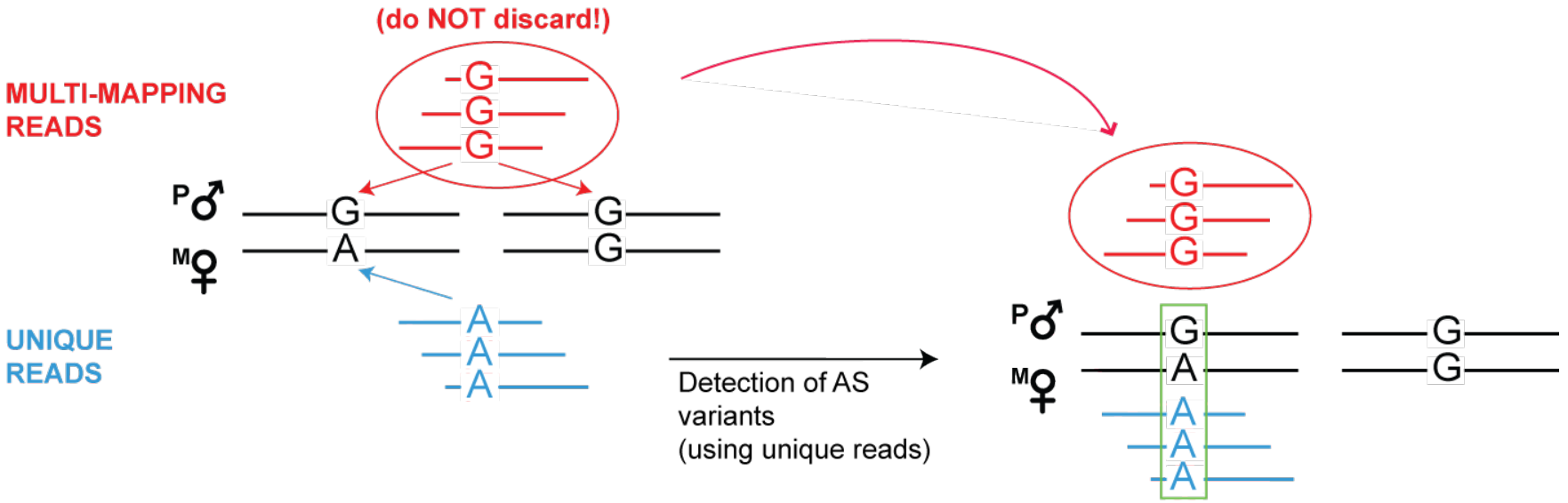
- Using simulated reads obtained by flipping the alleles of original reads at hetSNV positions:



Lappalainen *et al*. (2013)
Panousis *et al.* (2014)
Van de Geijn *et al.* (2015)

Account for ambiguous mapping bias: personal genome

- Using the personal genome, we do not need to simulate reads.
- We can directly test affected sites using multi-mapping read pile



[Chen *et al. Nature Comm.* '16]

SV Call Sets & Personal Genomes

- Retroduplication SV calls
  - New call set, now for Trio data
- Personal Genomes from SV calls
  - Trying to incorporate a PacBio SV call set
  - Trying to demonstrate QC metrics on genome quality
  - Scaling up

SV Call Sets & Personal Genomes

- Retroduplication SV calls
  - New call set, now for Trio data
- Personal Genomes from SV calls
  - Trying to incorporate a PacBio SV call set
  - Trying to demonstrate QC metrics on genome quality
  - Scaling up
- **Qs**
  - **Where can we get more PacBio call sets?**
  - **Thoughts on QC metrics on whether call set improved genome**

Personal genomes

Timur Galeev

Acknowledgments

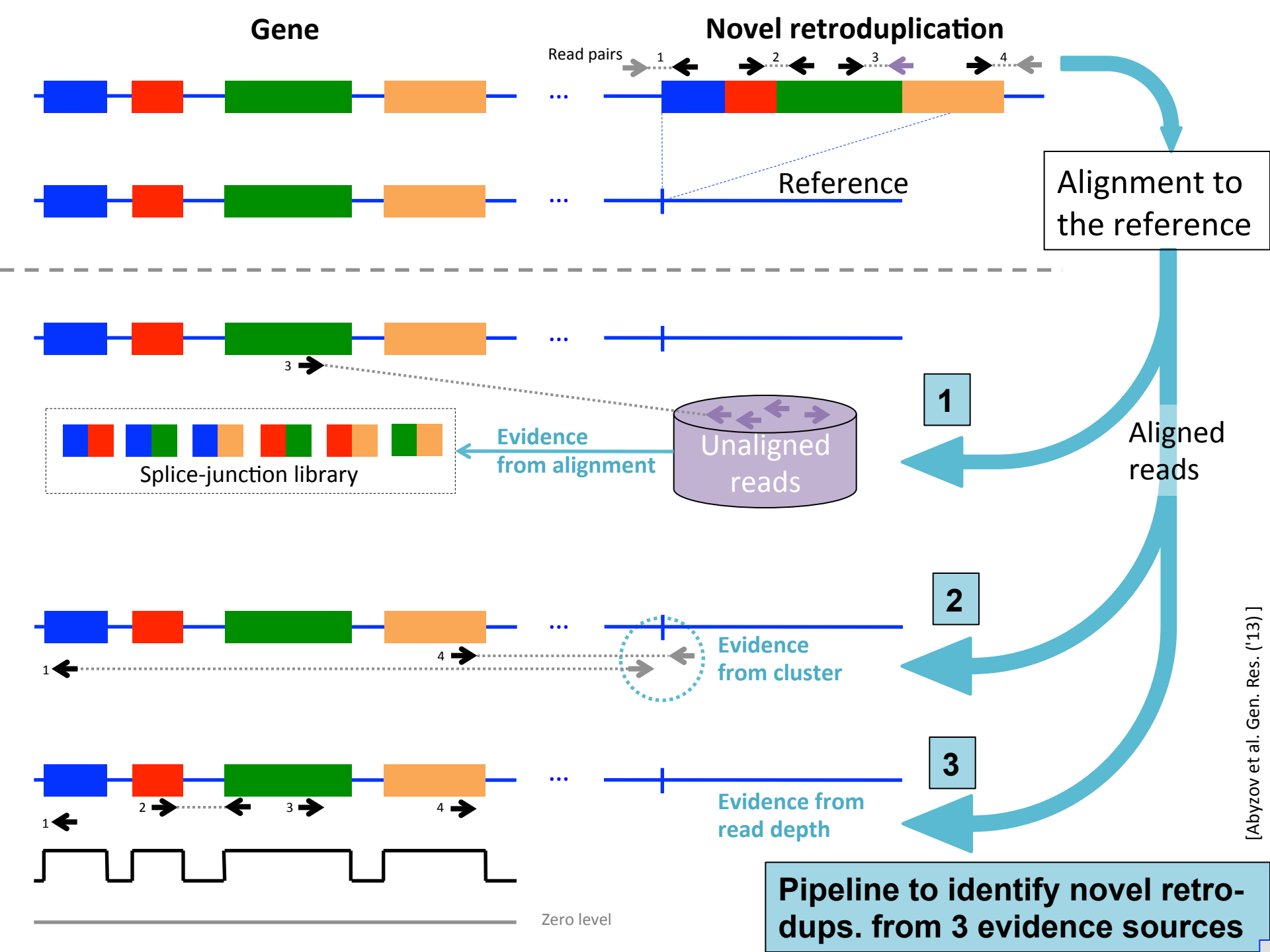
J Chen, J Rozowsky, A Harmanci, R Kitchen,
J Bedford, A Abyzov, Y Kong, L Regan

Retrodups

Fabio Navarro

Alexej Abyzov, Yan Zhang, Shantao Li

Extra


Gene
Novel retroduplication
Read pairs
1
2
3
4
Reference
Alignment to the reference
Aligned reads
Splice-junction library
Evidence from alignment
Unaligned reads
1
2
Evidence from cluster
3
Evidence from read depth
Zero level
[Abyzov et al. Gen. Res. ('13)]
Pipeline to identify novel retro-dups. from 3 evidence sources

Alignment gets better as variant sets get more complete: NA12878 Pol2 ChIP-seq (ENCODE)

	Ref genome	Pgenome: SNVs only	Pgenome: SNVs + indels only	Pgenome: SNVs + indels + SVs
Reads processed	208,051,087			
# reads uniquely aligned	171,944,588 (82.65%)	172,591,380 (82.96%)	172,738,321 (83.03%)	172,743,175 (83.03%)
# reads that multimap	17,826,675 (8.57%)	17,795,258 (8.55%)	17,782,167 (8.55%)	17,779,800 (8.55%)

Almost 1M increase in reads

[Chen *et al. Nature Comm.* '16]

Alignment gets better as variant sets get 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			
# reads uniquely aligned	25,303,498 (67.37%)	25,486,837 (67.86%)	25,538,449 (68.00%)	25,568,042 (68.08%)
# reads that multimap	4,041,495 (10.76%)	4,010,417 (10.68%)	4,012,297 (10.68%)	3,972,990 (10.58%)

[Chen *et al. Nature Comm.* '16]

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University,'16.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers, link to gersteinlab.org or using @markgerstein on twitter).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.
- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt