GIABTR

Adam English
HGSC@BCM

August 9, 2022

Agenda

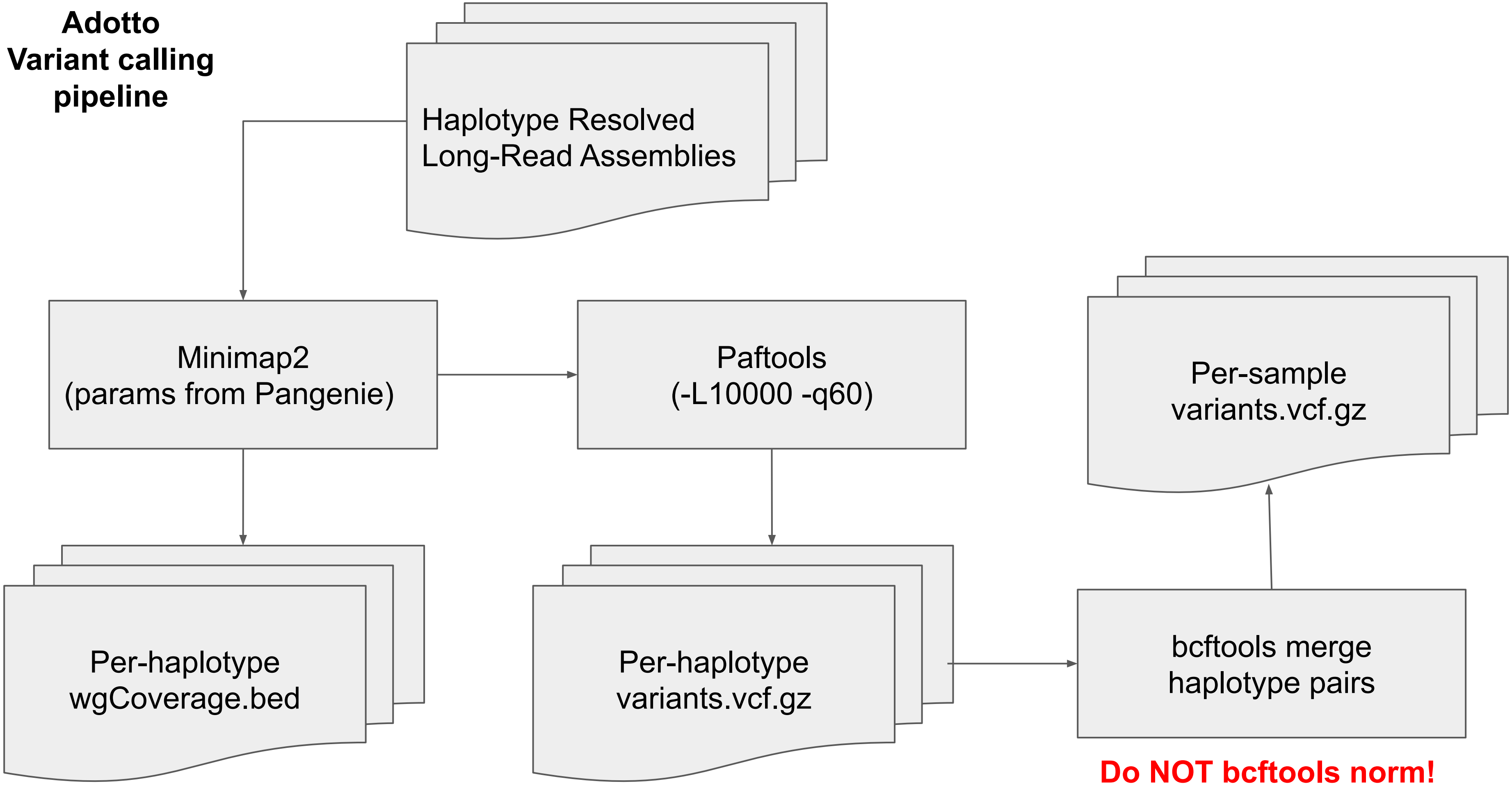
- Last meeting we made v0.1 of the TR Regions/Annotations.
- Describe v0.1 of the pVCF Variants
- Intersecting TR Regions and Variants

Why build a new variant calling pipeline?

- I had most of the parts built already, so why not?
- More control over the pipeline than using an existing caller e.g. dipcall

- By creating a pVCF from multiple assemblies, we may be able to better annotate tandem repeats
  - TandemRepeatFinder is an algorithmic view of TRs
  - Empirically observing copy-number changes of a motif is more definitive TR evidence

Adotto Variant calling pipeline



Haplotype Resolved Long-Read Assemblies

Minimap2 (params from Pangenie)

Paftools (-L10000 -q60)

Per-haplotype wgCoverage.bed

Per-haplotype variants.vcf.gz

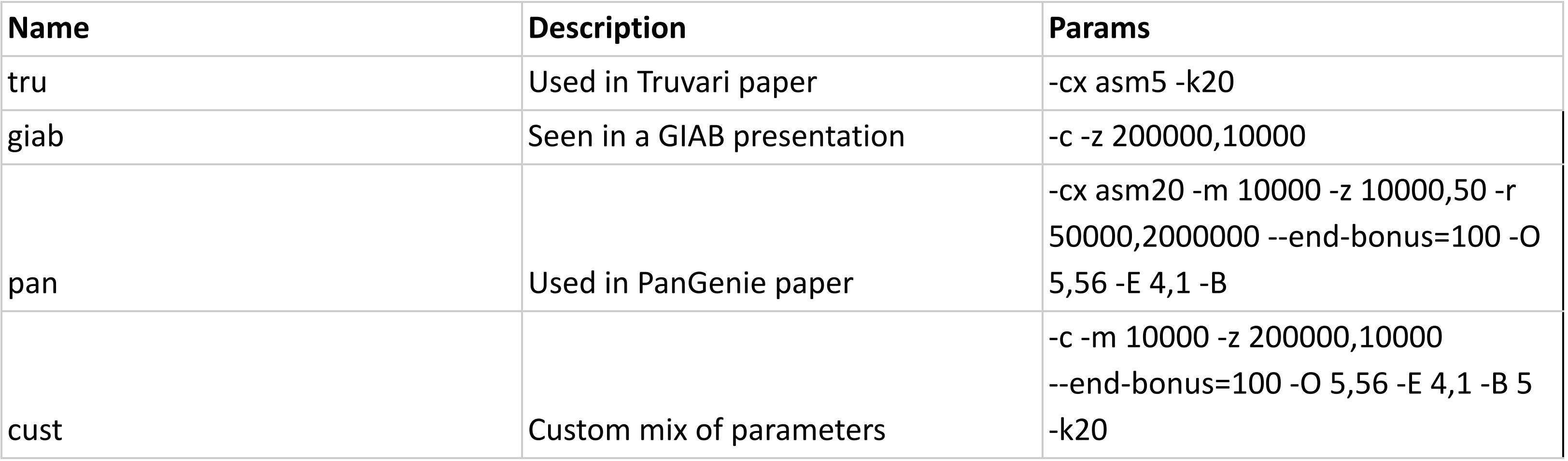
bcftools merge haplotype pairs

Per-sample variants.vcf.gz

Do NOT bcftools norm!

minimap2 parameters

- Previous analysis was performed with unrefined assembly mapping parameters.
- Explore improving calls with different minimap2 parameters
- Map haplotypes individually to hg19
- Annotate PASS as single-contig coverage
- Compare to GIAB SV v0.6



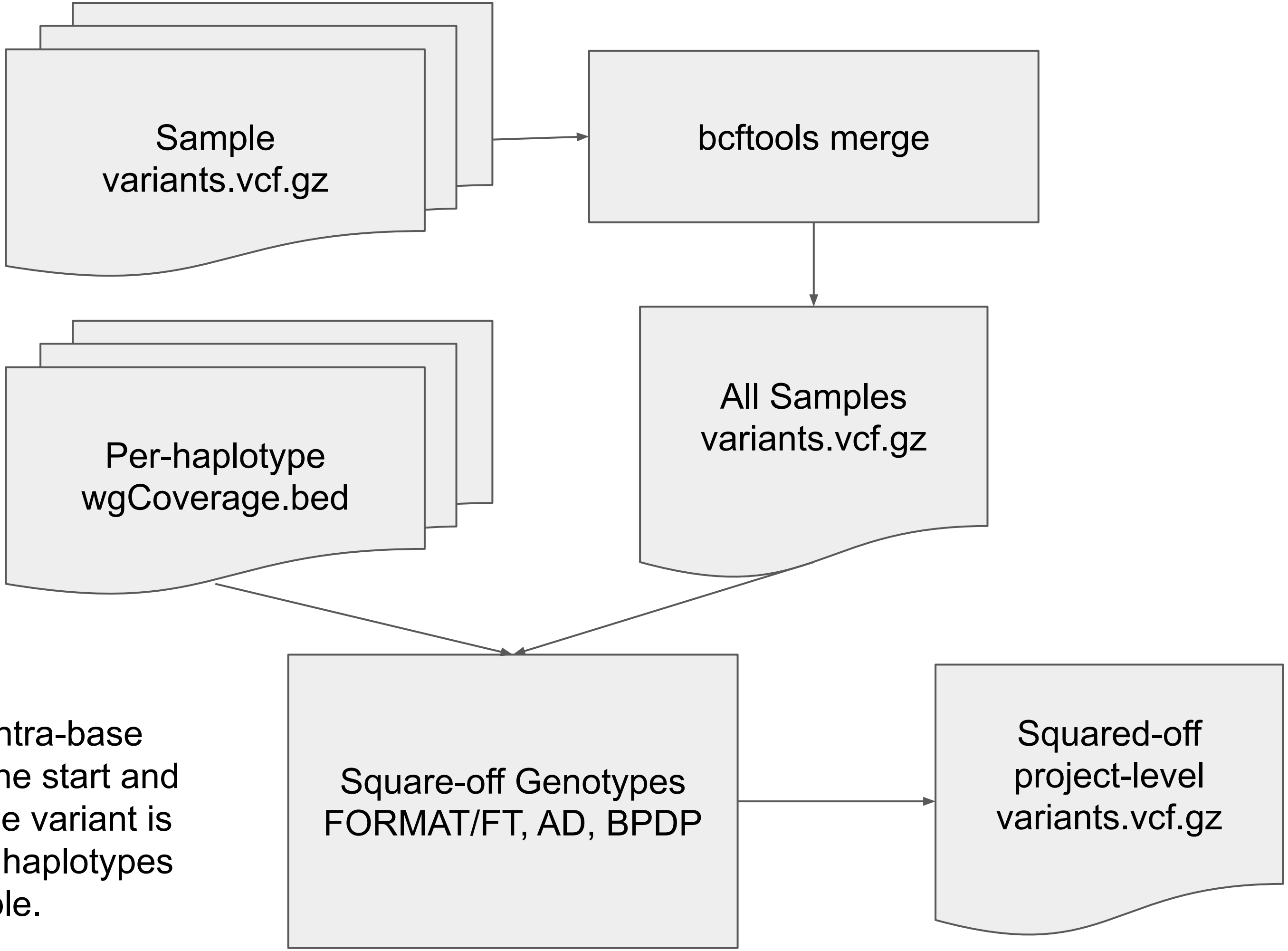
Name	Description	Params
tru	Used in Truvari paper	-cx asm5 -k20
giab	Seen in a GIAB presentation	-c -z 200000,10000
pan	Used in PanGenie paper	-cx asm20 -m 10000 -z 10000,50 -r 50000,2000000 --end-bonus=100 -O 5,56 -E 4,1 -B
cust	Custom mix of parameters	-c -m 10000 -z 200000,10000 --end-bonus=100 -O 5,56 -E 4,1 -B 5 -k20

Parameter Performance GIAB HG002 SV v0.6 (hg19)

Project	Params	True-pos baseline	True-pos call	False-pos	False-neg	Precision	Sensitivity	F-measure
li	**giab**	9,273	10,516	1,093	368	0.906	0.962	0.933
li	**tru**	9,251	10,477	945	390	0.917	0.960	0.938
li	**cust**	9,338	10,595	890	303	0.923	0.969	0.945
li	**pan**	9,335	10,647	712	306	**0.937**	**0.968**	**0.953**
eich	**giab**	9,241	10,448	1,053	400	0.908	0.959	0.933
eich	**tru**	9,217	10,403	935	424	0.918	0.956	0.936
eich	**pan**	9,316	10,590	700	325	**0.938**	**0.966**	**0.952**

The pangenie parameters perform best with f1 of 0.95

Adotto 'Squaring-off' pipeline



Sample variants.vcf.gz

bcftools merge

All Samples variants.vcf.gz

Per-haplotype wgCoverage.bed

Square-off Genotypes FORMAT/FT, AD, BPDP

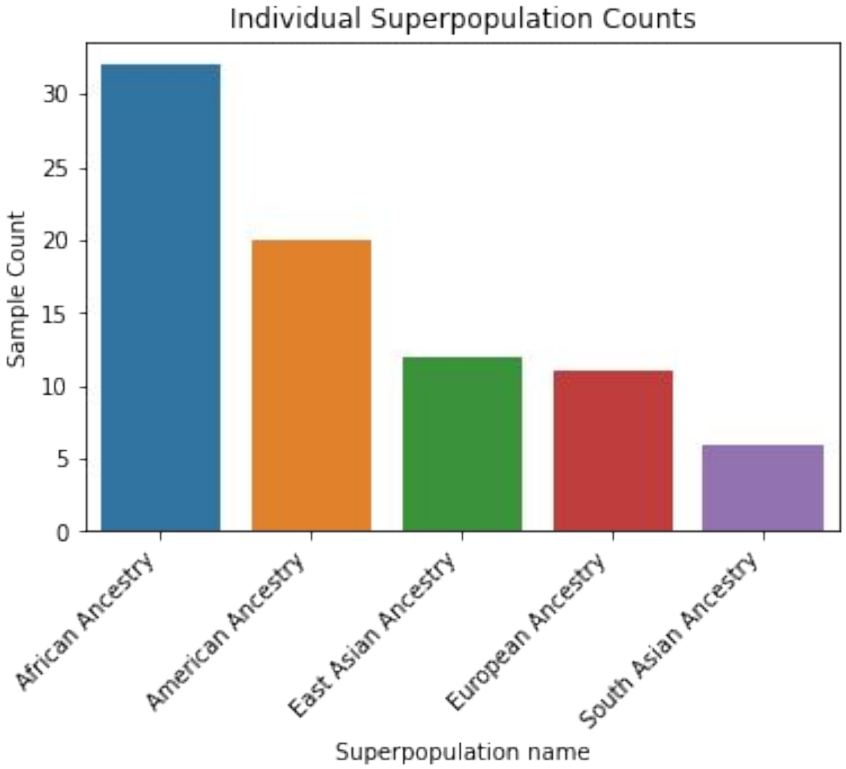
Squared-off project-level variants.vcf.gz

FT == PASS if intra-base coverage before the start and after the end of the variant is exactly 1 for both haplotypes per-sample.

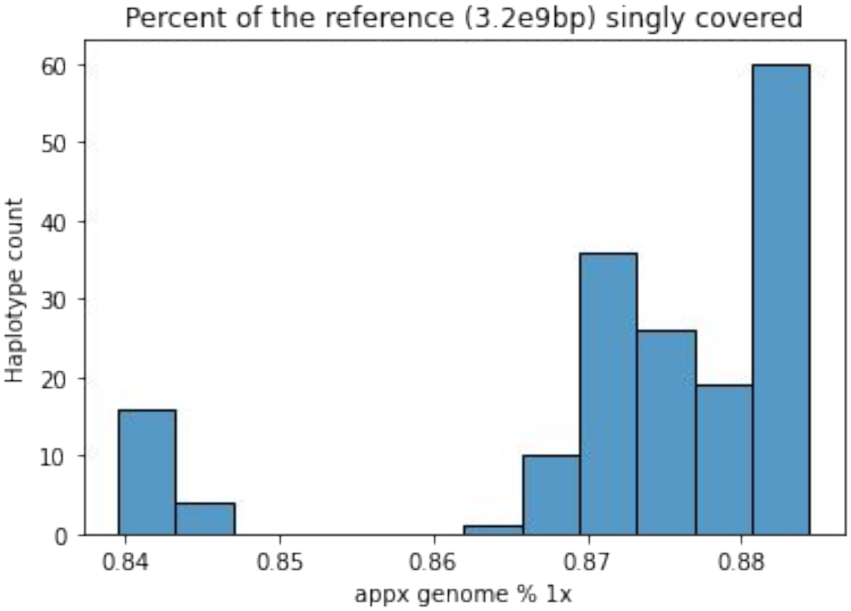
Sample Data

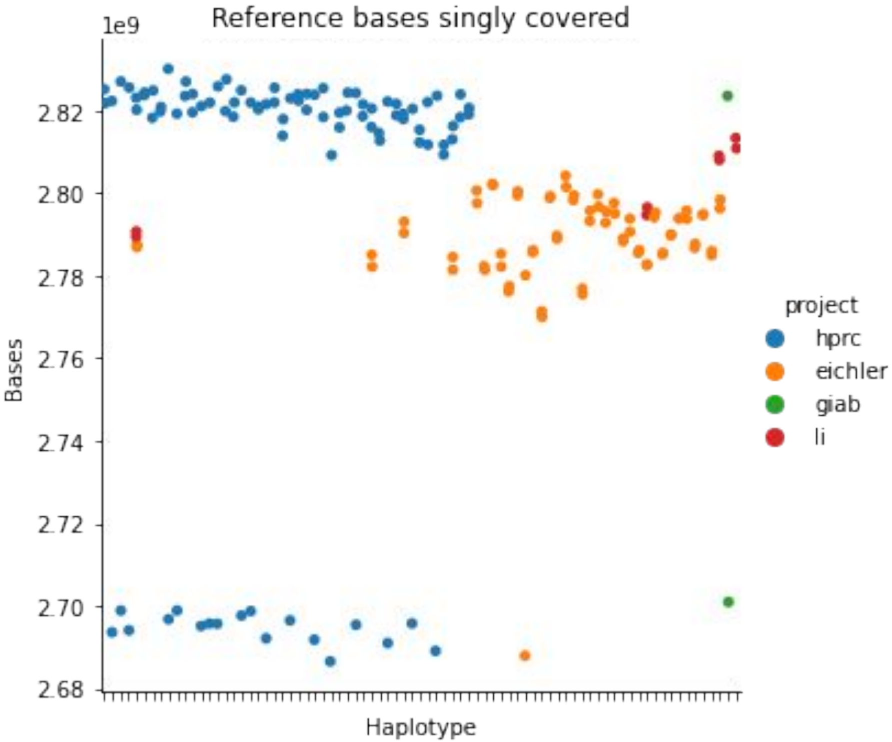
- 3 Projects
  - HPRC (47)
  - Eichler (34)
  - Li (4)
- 172 haplotypes
- 86 samples
- 78 individuals

Replicates	
HG00733	3
NA19240	2
NA24385	3
HG03486	2
HG02818	2
NA12878	2



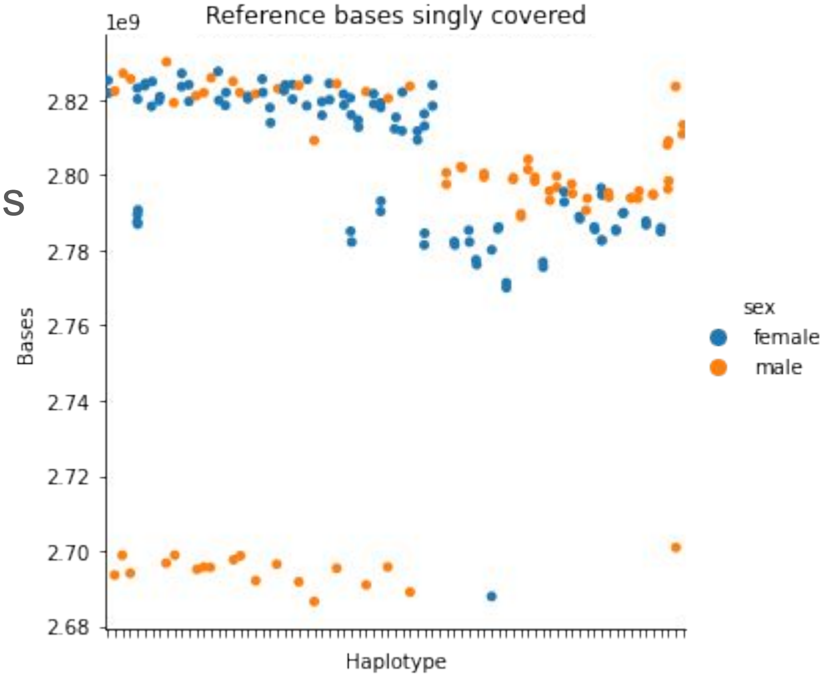
Haplotype Coverage

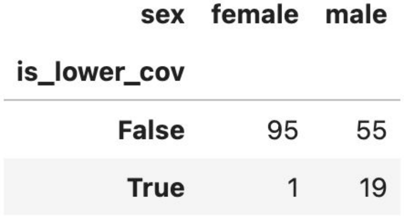




Haplotype Coverage

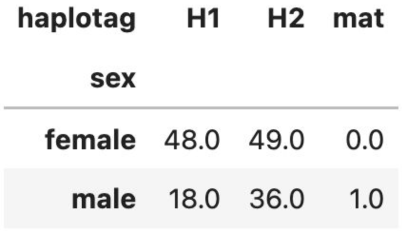
- The 20 ‘low coverage’ haplotypes are almost exclusively male samples
- The 1 female is from Eichler (HG00732)
- All the lower_cov haplotypes are from the paternal assembly





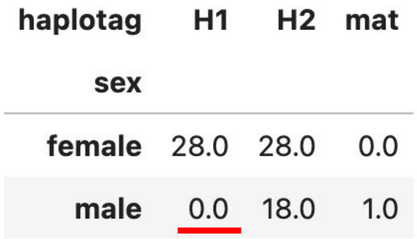
sex	female	male
is_lower_cov		
False	95	55
True	1	19

Not Lower Cov



haplotag	H1	H2	mat
sex			
female	48.0	49.0	0.0
male	18.0	36.0	1.0

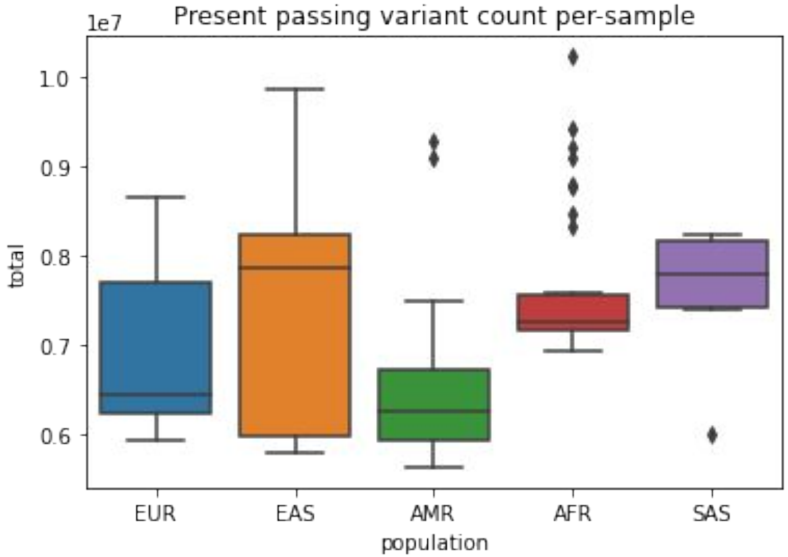
Not Lower Cov (HPRC)

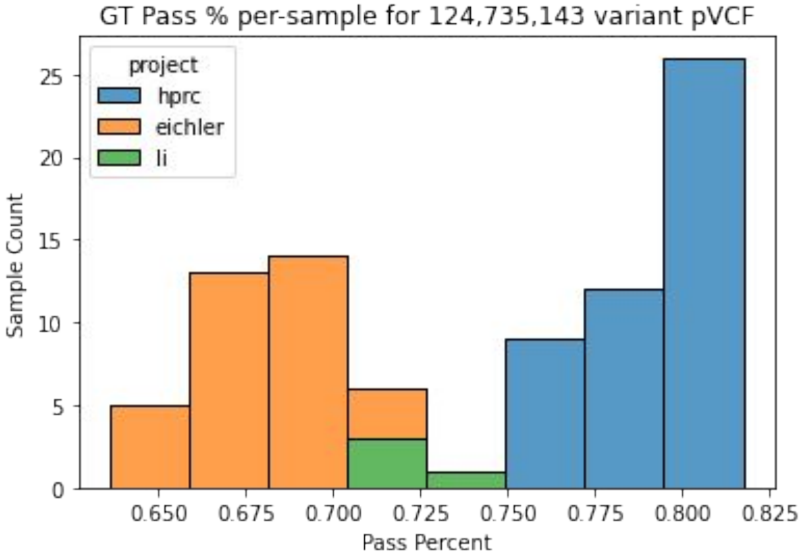


haplotag	H1	H2	mat
sex			
female	28.0	28.0	0.0
male	0.0	18.0	1.0

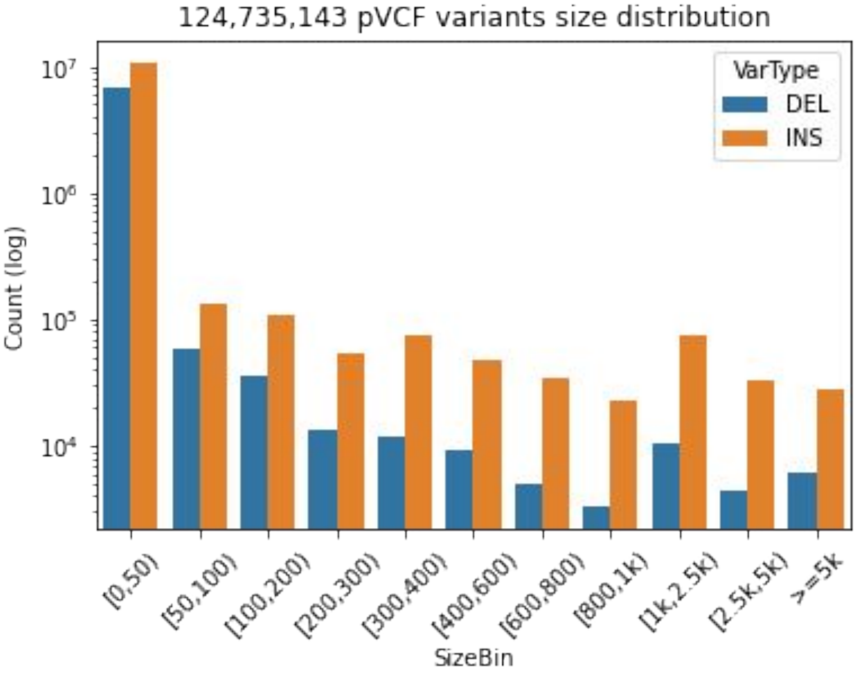
Variant Stats

- Total of 124M variants in pVCF
- Mean of 7,259,633 non-reference-homozygous variants per-sample
- After square-off, ~75% of variants FT==PASS per-sample

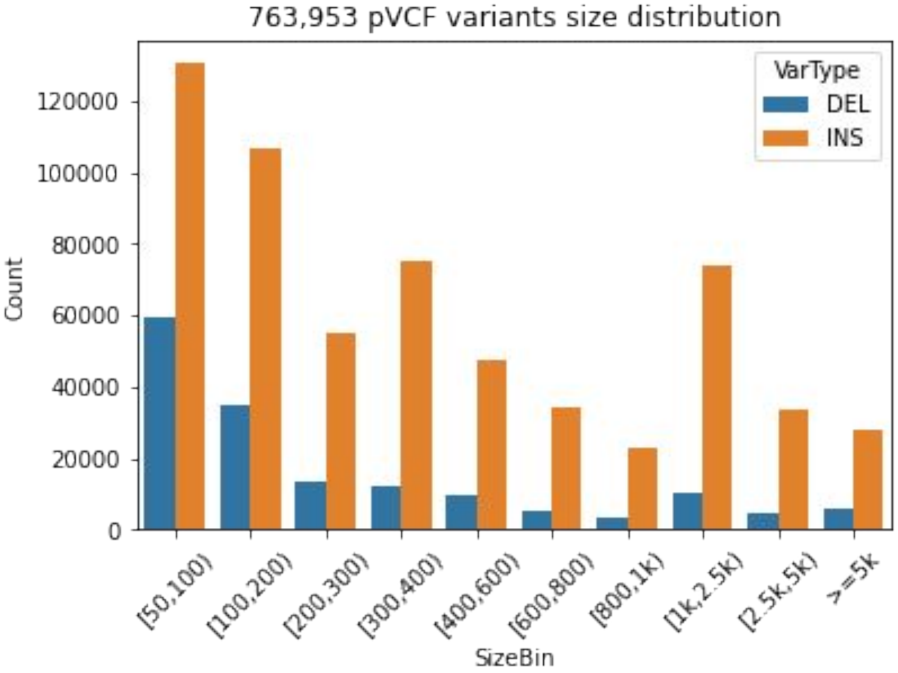




Variant Size Distribution



124,735,143 pVCF variants size distribution



763,953 pVCF variants size distribution

Benchmarking

- Three replicates of HG002/NA24385
- Benchmark against CMRG and GIAB’s TrioHifiAsm with RTG and Truvari

RTG + CMRG smallvar

Replicate	**True-pos baseline**	**True-pos call**	**False-pos**	**False-neg**	**Precision**	**Sensitivity**	**F-measure**
eichler	20,271	22,103	9,599	960	0.697	0.955	0.806
hprc	21,131	22,986	479	100	**0.980**	**0.995**	**0.987**
li	20,288	22,117	8,221	943	0.729	0.956	0.827

Truvari + CMRG SV

eichler	209	209	11	7	0.950	0.968	0.959
hprc	213	213	7	3	**0.968**	**0.986**	**0.977**
li	210	210	17	6	0.925	0.972	0.948

GIAB TrioHifiAsm Benchmarking

Program	Comp	True pos baseline	True pos-call	False-pos	False-neg	Precision	Sensitivity	F-measure
RTG	**eichler**	4,474,711	4,644,812	454,835	187,930	0.9108	0.9597	0.9346
Truvari	**eichler**	21,891	21,891	3,153	8,033	0.874	0.731	0.796
RTG	**li**	4,470,804	4,642,165	600,025	191,837	0.8855	0.9589	0.9207
Truvari	**li**	21948	21,948	3,261	7,976	0.870	0.733	0.796
RTG	**hprc**	4,476,465	4,658,725	119,799	186,176	0.9749	0.9601	0.9674
Truvari	**hprc**	22,384	22,384	2,768	7,540	0.889	0.748	0.812

Why do the SVs have lower Sensitivity?

High consistency of FNs

89% of the 8,345 SV FNs are missed by all replicates

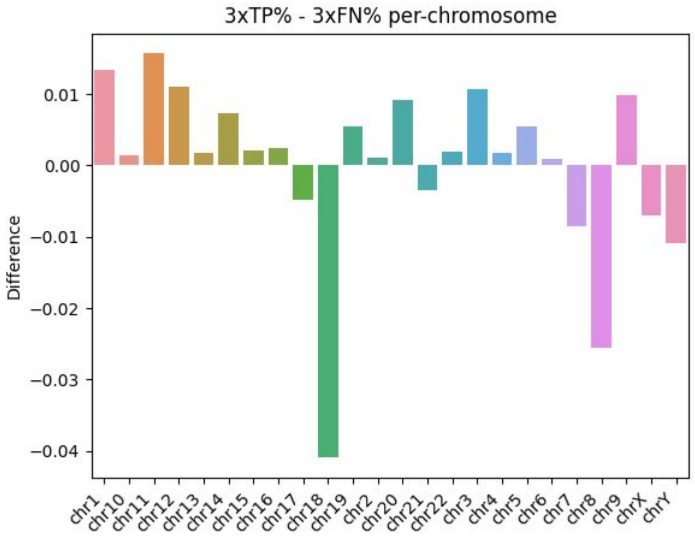
```
#
# Total 8345 calls across 3 VCFs
#
#File   NumCalls
truvari_thfa_eichler/fn.vcf 8033
truvari_thfa_hprc/fn.vcf    7540
truvari_thfa_li/fn.vcf      7976
#
# Summary of consistency
#
#VCFs   Calls   Pct
3       7427    89.00%
2       350     4.19%
1       568     6.81%
```

```
#
# Breakdown of VCFs' consistency
#
#Group  Total   TotalPct PctOfFileCalls
111     7427    89.00%  92.46% 98.50% 93.12%
100     298     3.57%   3.71% 0% 0%
001     258     3.09%   0% 0% 3.23%
101     249     2.98%   3.10% 0% 3.12%
110     59      0.71%   0.73% 0.78% 0%
011     42      0.50%   0% 0.56% 0.53%
010     12      0.14%   0% 0.16% 0%
```

Investigating FNs

- Some patterns may partially describe FNs

		3x FN	3x TP
SVTYPE	**DEL**	3,224	7,793
	INS	4,203	13,786
SVLEN	**Mean**	335	612
	Median	119	185



- More FNs relative to TPs on **chr18, chr8, chrY**, chr7, chrX, chr17, chr21
- 1,647 of the 3x FN have no call within 1kbp.
- Only 794 variants are explained by no-coverage from any HG002 haplotype.
    - 698 have no-coverage from any haplotype.
- 4,525 don’t match due to no multimatching. 4,720 would fail to match even with multimatching

Next Steps

- pVCF v0.1 is available. Zenodo links on github.com/ACEnglish/adotto
- Can pass 3xFN/TP set to GIAB for analysis
  - Are these confident calls?
- Curating tr_regions/pVCF intersection

Variant Intersection with Tandem Repeats

Count non-SNP variants in pVCF within Tandem Repeats regions/annotations

	All TR_Regions		Annotated regions		Unannotated regions	
metric	**count**	**percent**	**count**	**percent**	**count**	**percent**
total regions	2,232,565	100.0%	1,793,027	100.0%	439,538	100.0%
no variant	448,124	20.1%	320,001	17.8%	128,123	29.2%
only SNPs	846,353	37.9%	617,746	34.5%	228,607	52.0%
remaining	938,088	42.0%	855,280	47.7%	82,808	18.8%

Version v0.2 - Available now

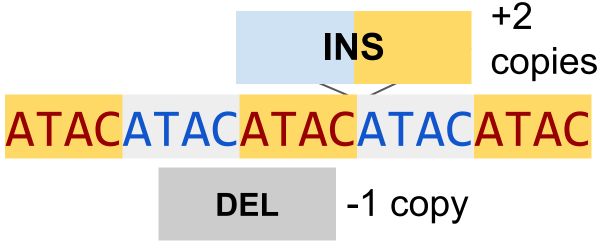
How many variants are TR expansions/contractions?

Truvari anno trf

Goal is to assign known reference TR annotations to VCF entries and calculate the copy-number difference of the variant.

This problem can be tricky because a TR repeat region may have multiple possible TR motifs

If a variant within a TR region cannot be *easily* assigned a TR motif, TandemRepeatFinder is run and new motifs not in the reference may be reported.



INS

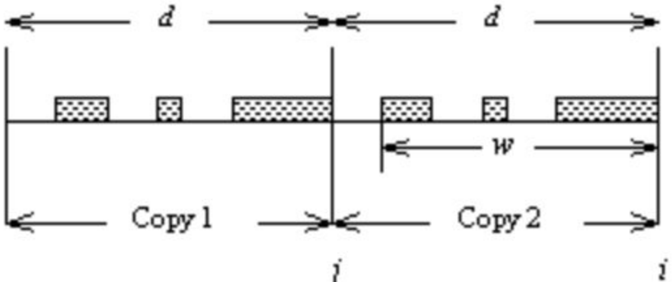
+2 copies

ATACATACATACATACATAC

DEL

-1 copy

```
>chr1:72120-72164
 ATATATATATACACACATATATACATACATACATACATACATAT
 ATATATAcATACACACATATATACATACAcACATAtATACATA
     ATAcATACAtACATATATACATACATAtATACATACATAT
         ATACAtACATAcATACATACATACATACATACATA
```

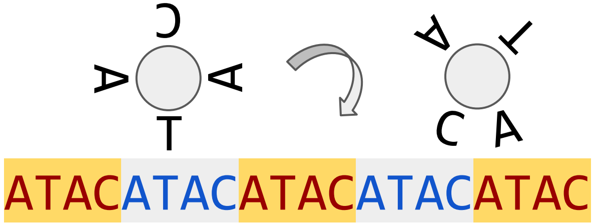


Source: https://tandem.bu.edu/trf/trfdesc.html

‘Unrolling’ Tandem Repeats

- We have a tandem repeat motif **M** of length **N**.
- This motif is repeated **C** times which creates a sequence **S** of length **L = C * N**
- A subsequence $\mathbf{B=S_{p:p+N}}$ for any position $\mathbf{p \in \{0:L\text{-}N\}}$ holds a 'rolled' representation of **M**.
- We can 'unroll' **B** such that **uB == M** with the operation:

```
B = S[p:p+N]
f = p % N
uB = B[-f:] + B[:-f]
```



‘Unrolling’ TR Motifs - Example

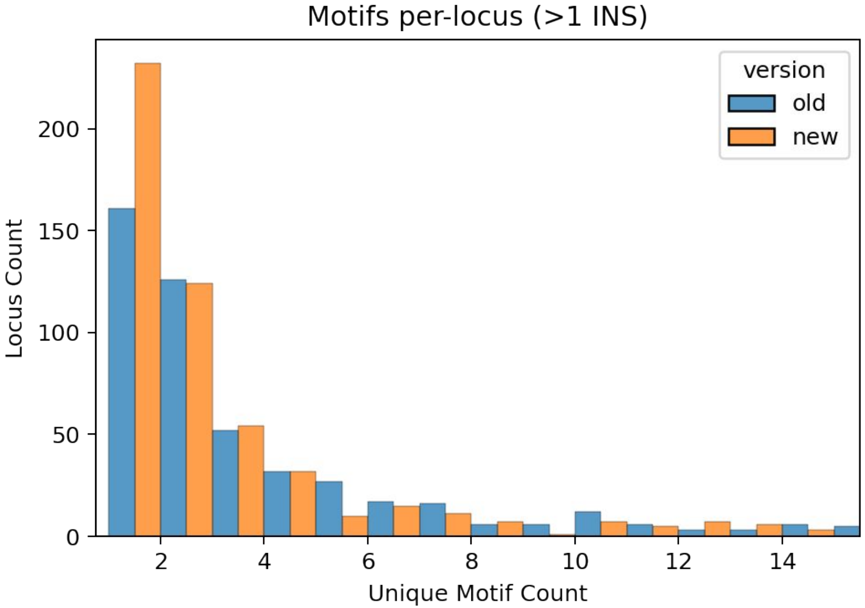
```
Reference motif: 34bp @ chr22:10577401
Alternate motif: 32bp @ chr22:10577405

before unrolling
motif similarity: 0.879
   refTATATGTATGTATACAATACACACACATATAAC-A-
      |----||||||||||||||||-|||||||||||-|--
   altT----GTATGTATACAATACA-ACACATATAACTATA

after unrolling
motif similarity: 0.970
   refTATATGTATGTATACAATACACACACATATAACA
      |||||||||||||||||||||-|||||||||||-
   altTATATGTATGTATACAATACA-ACACATATAAC-
```

Truvari anno trf - Test

- Old approach 785 loci have ~6 motifs per-locus (min 2 insertions)
- New approach 840 loci have ~4 motifs per-locus



Annotated motifs from the new version more frequently match the reference tr_annotation

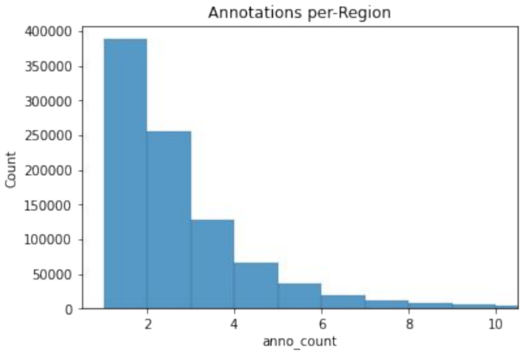
	TRF INS >=50bp	Motif Matches tr_annotation	Percent Matching
Old	10,385	3,285	31.6%
New	10,686	7,152	66.9%

Whole genome compute
Old: ~7,000 hours. New: 100 hours
Analysis on chr22 only with test version the tool

Truvari anno trf Stats

Annotations

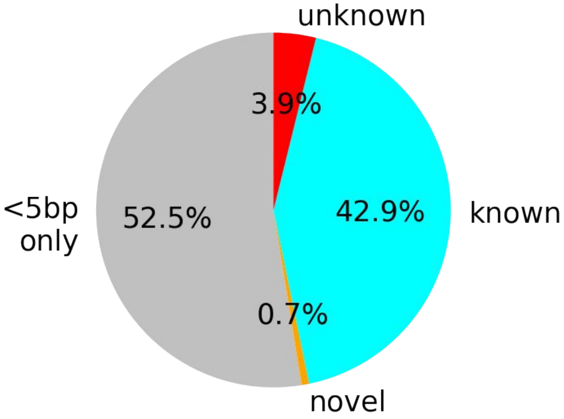
- TR Regions (v0.2):
  - 938,088
  - Spans 121,788,538bp
  - 3.8% of grch38
- TR Annotations
  - 2,337,945
  - Spans 130,866,860bp
  - ~2 annos per-region



Annotations x Variants

- Regions w/ at least one >=5bp variant
  - 445,173 (47.4%)
- Regions w/ >=1 annotated variant
  - 409,010 (91.8%)
- Regions w/ >=1 known TR variant
  - 402,538 (90.4%)

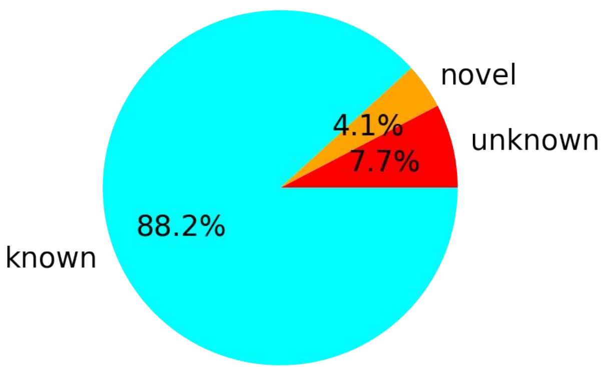
TR Region Breakdown



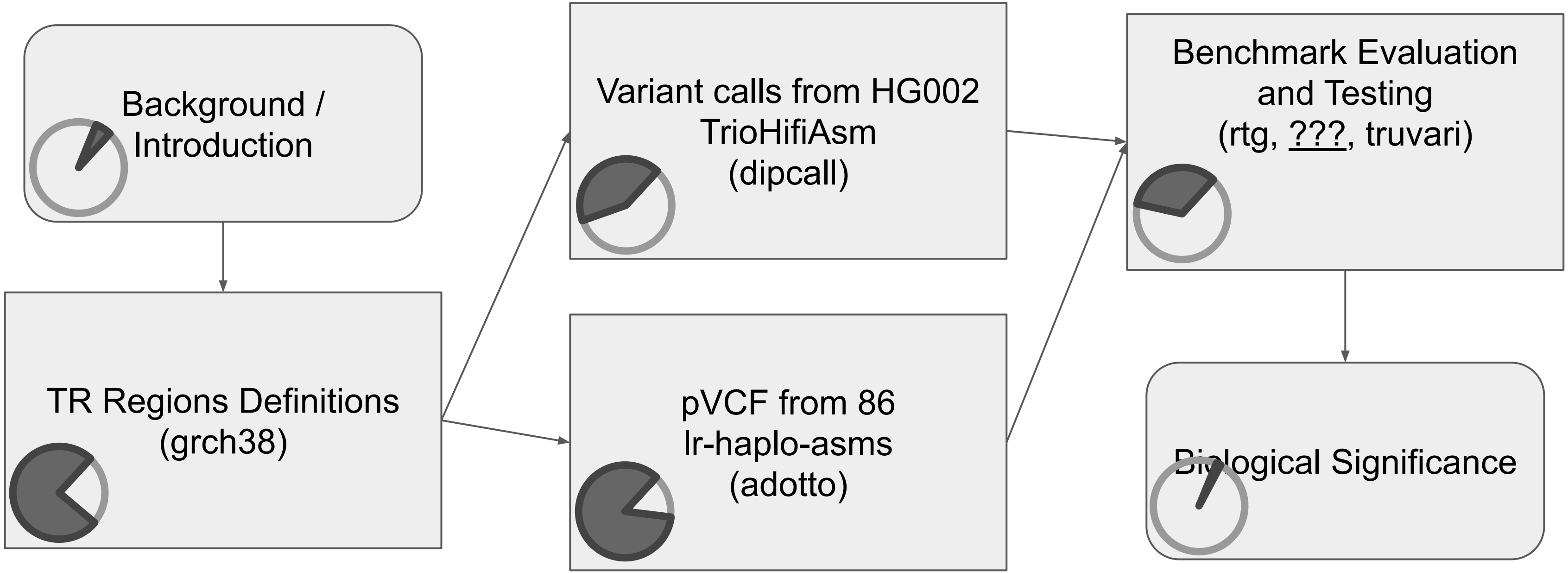
Variants

- Variants in TR regions (>=5bp)
  - 3,278,848
- Annotated variants
  - 3,027,762 (92.3%)
- Annotations matching TR annos
  - 2,892,229 (88.2%)

Variant Breakdown



Revisiting the Roadmap



Background / Introduction

TR Regions Definitions (grch38)

Variant calls from HG002 TrioHifiAsm (dipcall)

pVCF from 86 lr-haplo-asms (adotto)

Benchmark Evaluation and Testing (rtg, ???, truvari)

Biological Significance

Digression: Variant Enrichment in Tandem Repeats

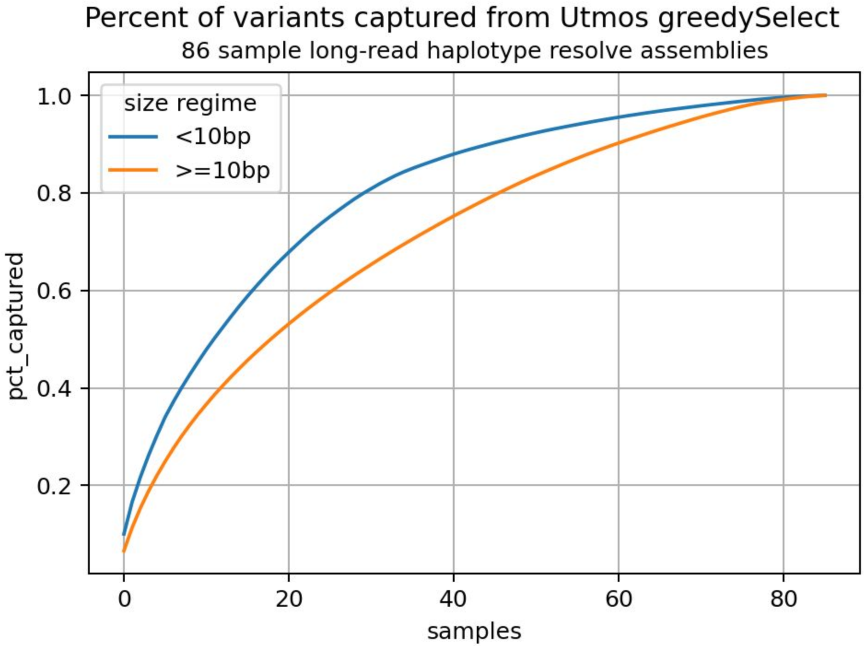
Looking at the variants by count and bases effected, we see most variation occurs in tandem repeat regions.

- v0.2 TR regions genome coverage
    - 121,788,538bp **~3.81%**

var	count	bases
All	17.5%	45.2%
SVs (>=50)	74.5%	47.0%

Digression: Measuring variant diversity

Utmos is a program to perform a greedy approximation of the maximum-coverage problem on genomic variants. We can use it to rank/sort samples by the amount of observed variation each contains and test if there’s more ‘diversity’ (i.e. less variant sharing) in larger events (>=10bp) vs smaller events (<10bp).



- Smaller events taper off more quickly than larger events, suggesting they’re more likely to be shared
    - <10bp AUC*: 68.0
    - >=10bp AUC*: 60.6
- Caveats:
    - Samples vs individuals - there are replicates
    - Alleles vs variants - a single large allele could have multiple variant representations

* Area under curve with composite trapezodial rule using dx=1