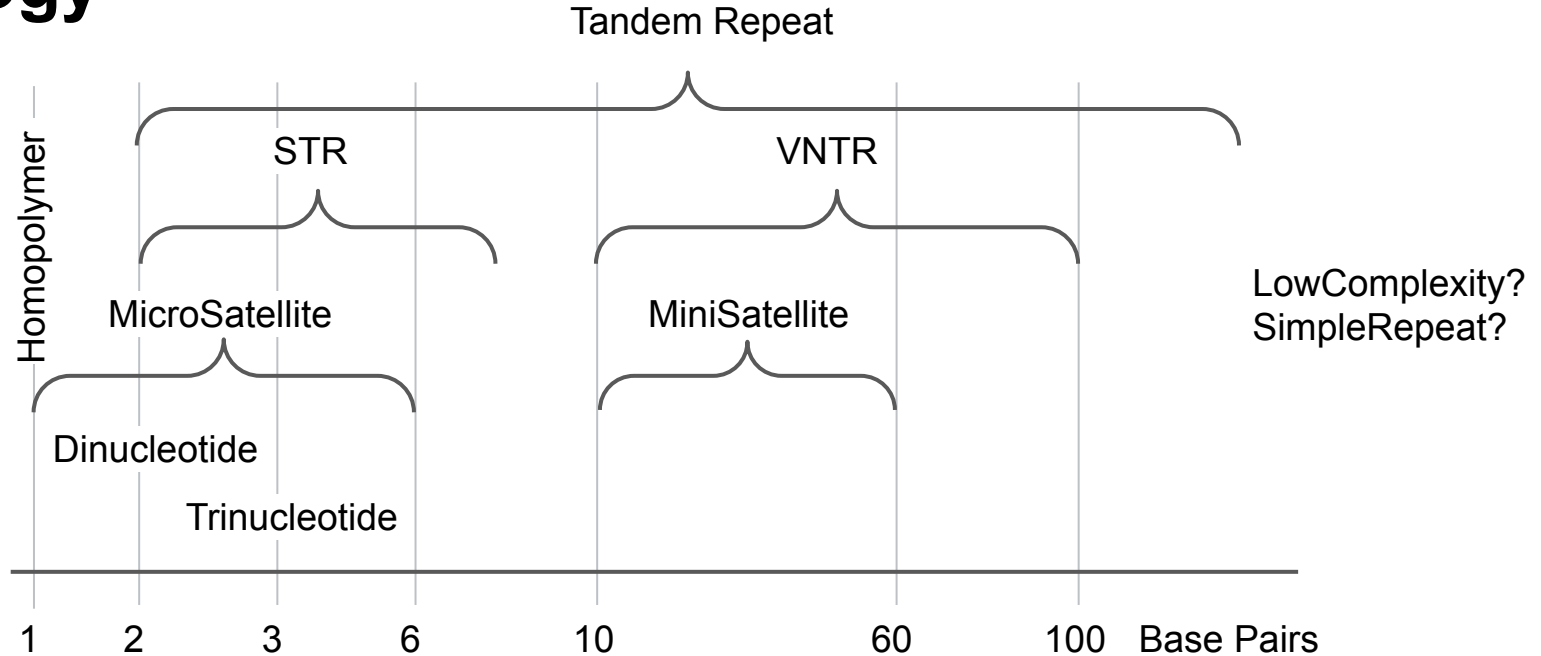



# GIAB TR

Part I

# Ontology



 = 30bp repeat

60bp  
Insertion →  +2 copies



 Repeat Sequence  Repeat Unit/Motif  Variant Sequence

an **insertion/deletion** comprises the **expansion/contraction** of a repeat

# Tandem Repeat Loci Beds

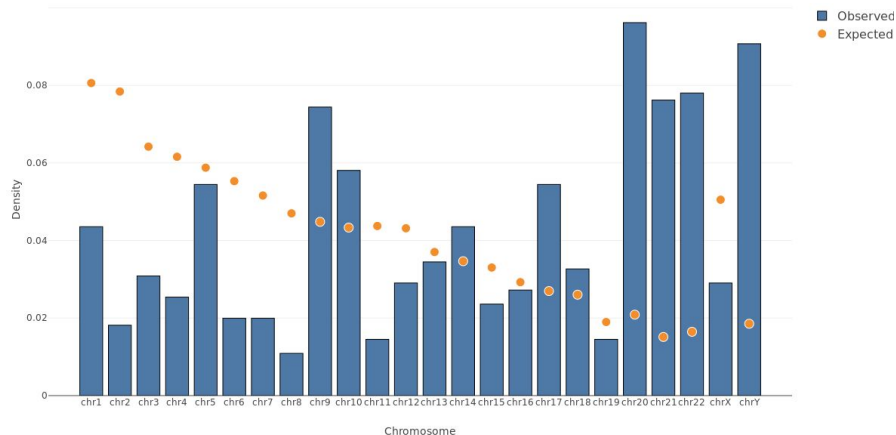
- Beds provided from:
  - GIAB (Jennifer McDaniel)
  - PacBio (Egor Dolzhenko)
  - Baylor (Adam English)
  - UCSD (Helyaneh Ziaei Jam)
  - UCSD2 (Jonghun Park)
- Start with a summary
- Explore intersections
- Goal:
  - Define GRCh38 tandem repeat regions and annotate each region's repeat

# GIAB Beds

23 bed files of repeat annotations

Concat/merged cover ~30% of genome

91% of genome annotated



<https://quinlan-lab.github.io/bedqc/>

## Parent Directory

[GRCh38-LowComplexity-README.md](#)

[GRCh38\\_AllHomopolymers\\_gt6bp\\_imperfectgt10bp\\_slop5.bed.gz](#)

[GRCh38\\_AllTandemRepeats\\_201to10000bp\\_slop5.bed.gz](#)

[GRCh38\\_AllTandemRepeats\\_51to200bp\\_slop5.bed.gz](#)

[GRCh38\\_AllTandemRepeats\\_gt10000bp\\_slop5.bed.gz](#)

[GRCh38\\_AllTandemRepeats\\_gt100bp\\_slop5.bed.gz](#)

[GRCh38\\_AllTandemRepeats\\_lt51bp\\_slop5.bed.gz](#)

[GRCh38\\_AllTandemRepeatsandHomopolymers\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_diTR\\_11to50\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_diTR\\_51to200\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_diTR\\_gt200\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_homopolymer\\_4to6\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_homopolymer\\_7to11\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_homopolymer\\_gt11\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_homopolymer\\_gt20\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_imperfecthomopolgt10\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_imperfecthomopolgt20\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_quadTR\\_20to50\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_quadTR\\_51to200\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_quadTR\\_gt200\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_triTR\\_15to50\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_triTR\\_51to200\\_slop5.bed.gz](#)

[GRCh38\\_SimpleRepeat\\_triTR\\_gt200\\_slop5.bed.gz](#)

[GRCh38\\_notinAllHomopolymers\\_gt6bp\\_imperfectgt10bp\\_slop5.bed.gz](#)

[GRCh38\\_notinAllTandemRepeatsandHomopolymers\\_slop5.bed.gz](#)

# Bed Summary

Source	Total Entries	Uniq Loci	Uniq Span	Genome Span	GIAB Intersection	Percent Intersect	SV Count	SV Percent
GIAB	1,400,092	1,400,092	165,371,166	5.17%	-	-	157,784	80.61%
Baylor	1,031,708	692,882	148,907,732	4.65%	541,252	78.12%	155,472	79.43%
PacBio	171,146	171,145	4,538,738	0.14%	147,970	86.46%	1,783	0.91%
UCSD	1,776,010	1,739,897	35,311,586	1.10%	565,995	32.53%	15,006	7.67%
UCSD2	10,264	10,259	612,921	0.02%	7,535	73.45%	337	0.17%

# Overlapping Repeat Motifs

chr1	72120	72163	16	2.7	59	1.47	ATATATACATACACAC
chr1	72124	72164	12	3.3	62	1.49	ATATATACATAC
chr1	72128	72163	4	8.8	52	1.48	ATAC

>chr1:72120-72164

ATATATATATACACACATATATACATACATACATACATAT

ATATATAcATACACACATATATACATACAcACATAtATACATA

ATAcATACAtACATATATACATACATAtATACATACATAT

ATACAtACATAcATACATACATACATACATACATACATA

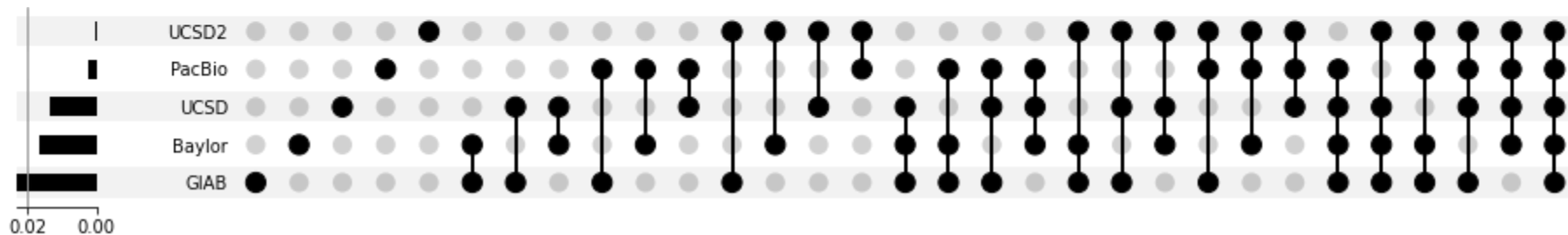
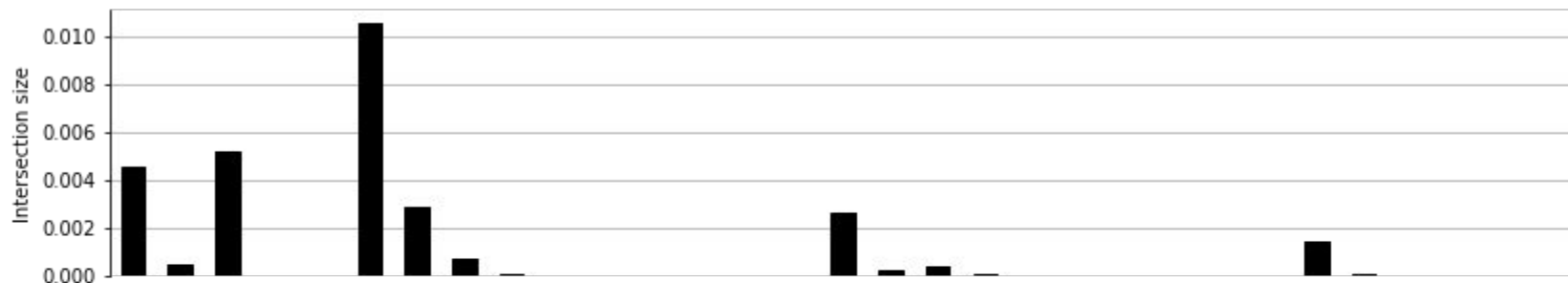
# Intersecting Beds

Intersect the unique/merged regions per-source.

Calculate percent of genome that's covered and number of regions

- Resolved: Each source hits  $\leq 1x$ 
  - 2.98% - 2,423,073 regions
- CPX: At least one source hits  $> 1x$ 
  - 2.98% - 77,337 regions

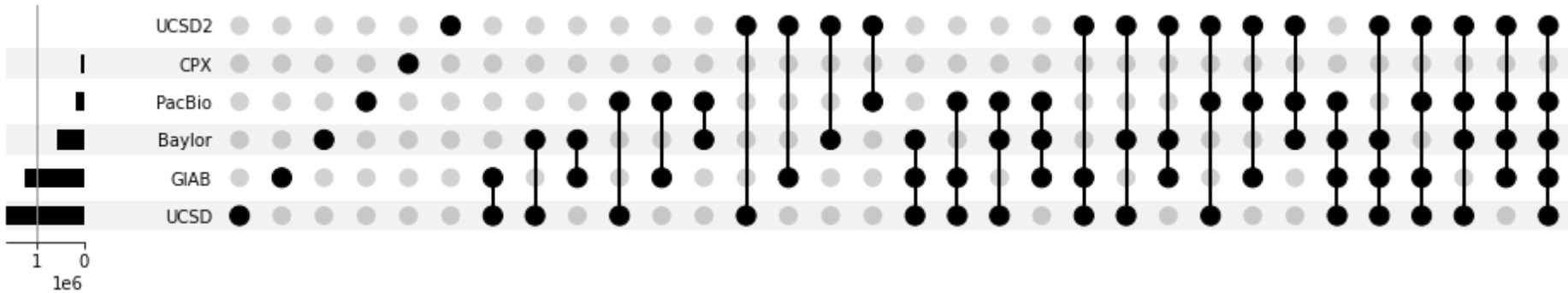
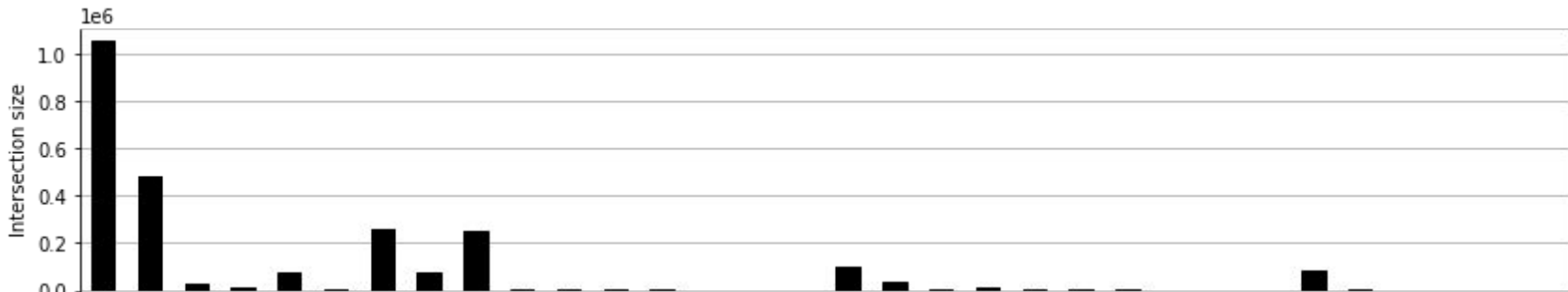
# Intersecting Beds - Genome Coverage Non-Complex



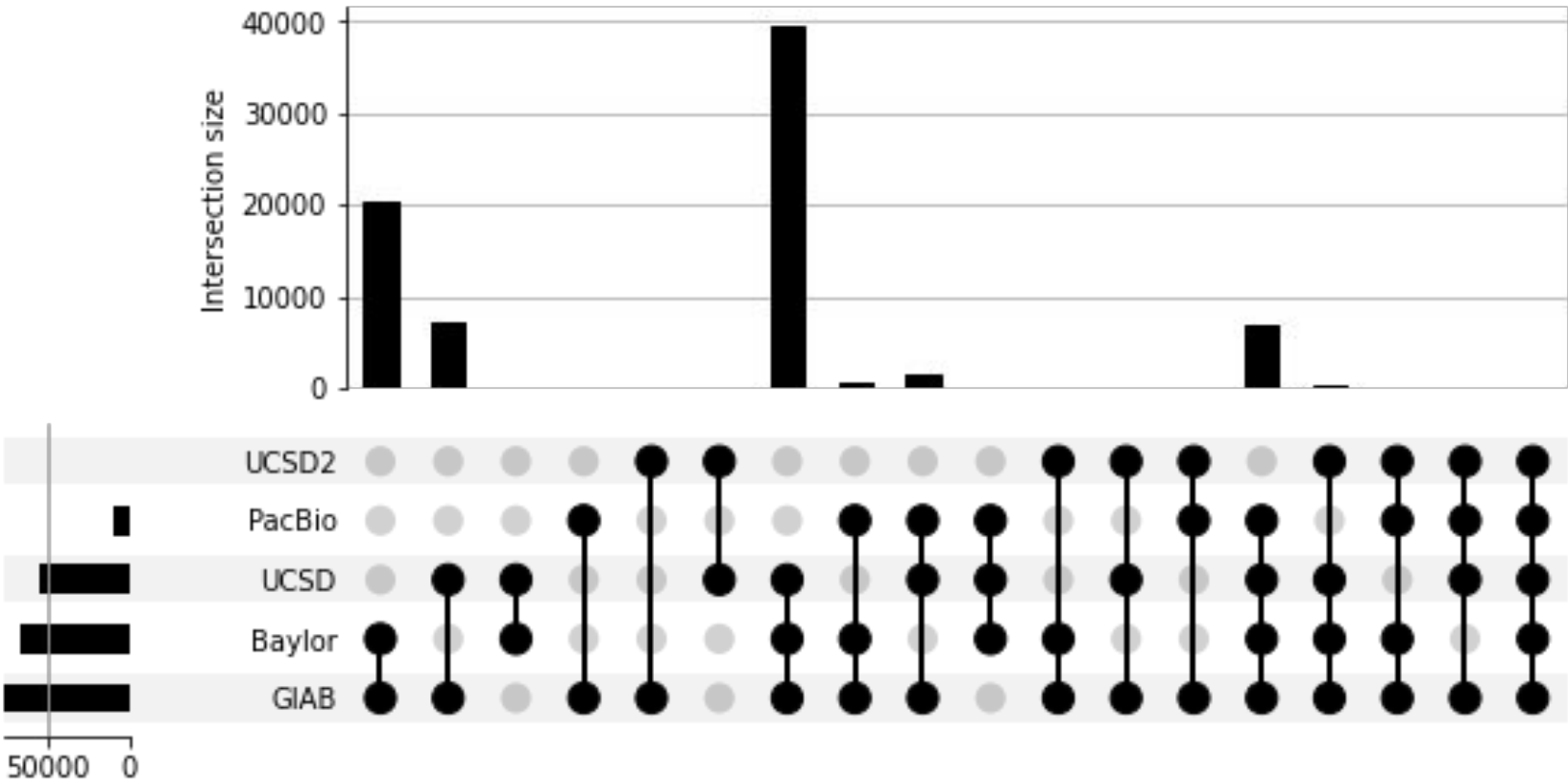
All Agree: 187 regions over 14,579 bp



## Intersecting Beds - Region Counts



# Intersecting Beds - CPX Region Counts



# Conclusions

- Proposed standard terms/definitions for Tandem Repeats
- Found ~3% of genome where 4 sources semi-agreed on Tandem Repeats
  - Motifs aren't resolved, though
- Need to solve overlapping Motifs and sources
  - Is the assumption of disjointed regions/single motif valid?
  - Why are there so many solo UCSD regions?
  - PacBio is the least self-overlapping. Are they the most resolved motifs?
  - UCSC SimpleRepeats are in Baylor and GIAB. Why is Baylor not a subset of GIAB?

# GIAB TR

Part II

# 'Strawman' Truth-Set

- Using Garg (Heng Li) assembly of HG002, call variants with minimap2 and merge haplotypes using Truvari.
  - Minimap ``-cx asm5 -t8 -k20 --secondary=no --cs``
  - Paftools ``-L10000``
- Use ``truvari anno trf`` to annotate variants  $\geq 10\text{bp}$
- Subset to variants within the merged tandem-repeat bed regions.
- Attempt to intersect with gangSTR / hipSTR

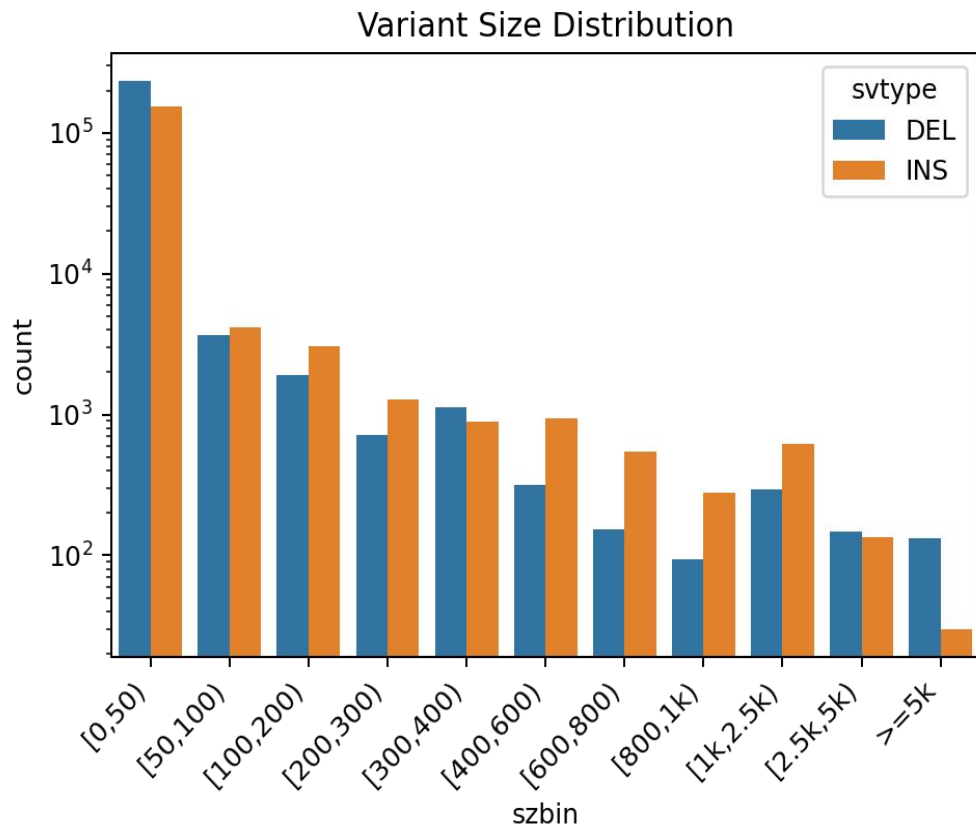
# ‘Strawman’ Assumptions

- 1) The variants called from assemblies are perfect
- 2) Any variant  $\geq 2$ bp inside a TR Region *could* be a tandem repeat
- 3) Truvari anno trf has 100% specificity

## Tandem Repeat Merged Bed

- Merged bed of TR regions from 5 sources
- Total of 2,542,375 regions
- Covers 201,389,980 bp (~6.3% of GRCh38)
- Baylor “SimpleRepeats” from UCSC Track:
  - 692,882 regions, covers 148,907,732 bp

# Long-Read Assemblies



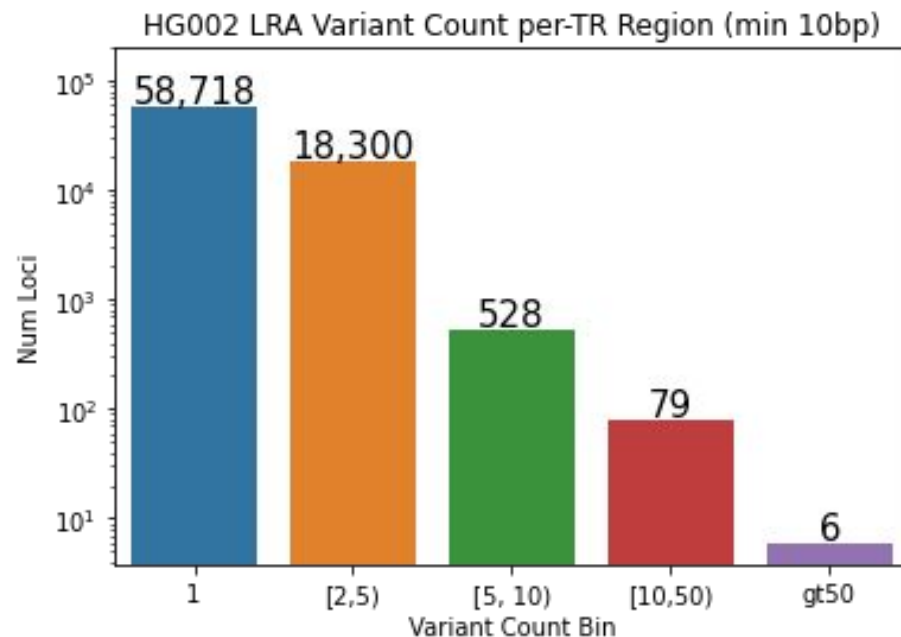
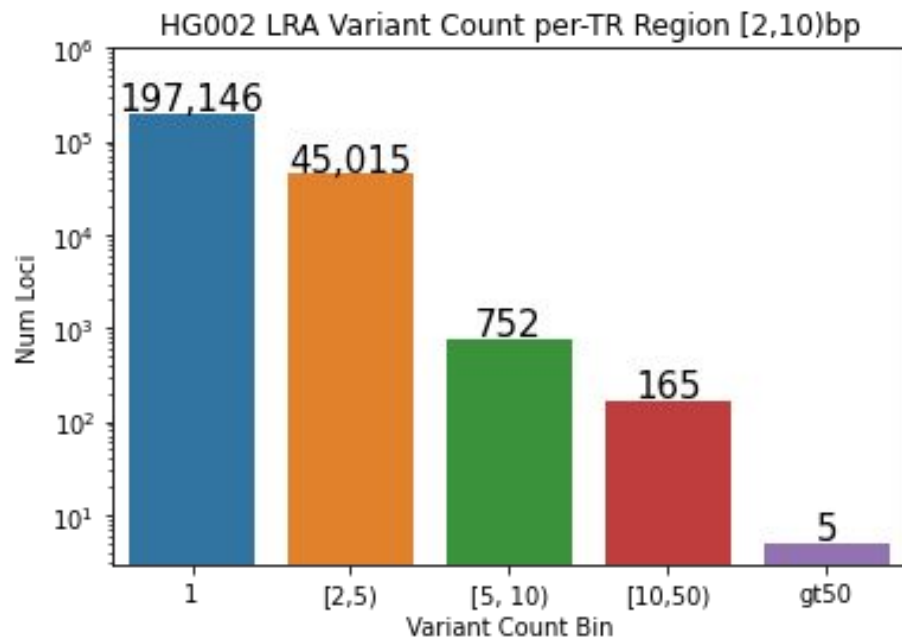
	[2,10)	>=10
Total	480702	303534
inTR	135648	106157

Fisher Exact:

- 1.24 OR
- $P < 0.01$

Enrichment of  $\geq 10$ bp  
variants in TR Regions

# Finding TR from Long-Read Assemblies






# Truvari anno trf

- For all SVs within the UCSC SimpleRepeats track:
  - Incorporate the alternate allele into the SimpleRepeats region
  - Run TandemRepeatFinder on the altered region
  - Match TRF repeats to the **longest repeat motif** from overlapping SimpleRepeat annotations
  - Report motif sequence, copy-number difference, etc
- Only run on variants  $\geq 10\text{bp}$
- SimpleRepeats are  $\sim 73.9\%$  (by bp) of the TR-Regions

# Truvari anno trf - QC

- 48.7% of SVs within TR annotated as tandem repeats
  - 51,740 of the 106,157  $\geq 10$ bp inTR variants
-  9.8% not in TR annotated as tandem repeats.
  - 2,901 of the 29,491  $\geq 10$ bp and not inTR
  - Off-by-one errors somewhere...
- $\geq 10$ bp variants inside TR Regions = 106,157
- 81.3% inside SimpleRepeats
  - 86,339 of the 106,157  $\geq 10$ bp inTR variants
- **48.4%** of candidate variants annotated by Truvari as Tandem Repeats
  - Total of **41,783** SVs

# GangSTR/HipSTR

Annotate GangSTR/HipSTR variants with TR Regions and 'shared' long-read calls

All VCF Entries

	Num vars	Inside Bed	With Var	With 10bp Var	isSR	Has TRF	Percent in TR	Percent w/Var	% w/ ≥10bp	% isSR	% wTR
<b>GangSTR</b>	888,561	748,948	147,797	45,102	35,445	21,027	84.29%	19.73%	30.52%	78%	59%
<b>HipSTR</b>	1,690,933	1,598,028	363,485	77,190	60,379	35,937	94.51%	22.75%	21.24%	78%	59%

HG002 VCF Entries

<b>GangSTR</b>	60,279	58,475	51,154	8,966	6,990	4,222	97.01%	<b>87.48%</b>	17.53%	77%	60%
<b>HipSTR</b>	366,673	348,502	135,348	18,744	15,203	7,867	95.04%	<b>38.84%</b>	13.85%	81%	51%

***\*Rough\**** Specificity Estimate

Assuming Long-Read and Short-Read Variants match



# GangSTR/HipSTR

Annotate GangSTR/HipSTR variants with TR Regions and 'shared' long-read calls

All VCF Entries

	Num vars	Inside Bed	With Var	With 10bp Var	isSR	Has TRF	Percent in TR	Percent w/Var	% w/ ≥10bp	% isSR	% wTR
<b>GangSTR</b>	888,561	748,948	147,797	45,102	35,445	21,027	84.29%	19.73%	30.52%	78%	59%
<b>HipSTR</b>	1,690,933	1,598,028	363,485	77,190	60,379	35,937	94.51%	22.75%	21.24%	78%	59%

HG002 VCF Entries

<b>GangSTR</b>	60,279	58,475	51,154	8,966	6,990	4,222	97.01%	87.48%	<b>17.53%</b>	77%	<b>60%</b>
<b>HipSTR</b>	366,673	348,502	135,348	18,744	15,203	7,867	95.04%	38.84%	<b>13.85%</b>	81%	<b>51%</b>

What we can evaluate.

Because GangSTR/HipSTR report copy-numbers of motifs

# GangSTR

- 3,006 of 4,222 (71%) of the GangSTR RepeatUnits match TRF annotation
- TRFcopies == REPCN Totals:
  - False 2,533
  - True 473
  - = 15% matching.

	REF	HG002_REPCN	TRFDiff	TRFcopies	TRFref
TRBED					
chr7:106984	16	(7, 16)	-7.5	9.0	16.5
chr7:69602	24	(11, 11)	-11.0	13.0	24.0
chr5:55275	27	(20, 20)	-20.0	7.0	27.0
chr12:25359	13	(13, 18)	18.0	31.0	13.0
chr11:97807	14	(14, 22)	6.0	20.5	14.5
chr6:60372	23	(15, 15)	-15.0	8.0	23.0
chr2:135387	7	(5, 5)	-5.3	2.0	7.3
chr4:99608	6	(6, 7)	-394.5	49.0	443.5
chr1:65767	13	(13, 15)	16.0	29.200001	13.200001
chrX:61407	21	(12, 12)	-12.0	9.0	21.0
chr2:119530	20	(25, 25)	5.0	25.5	20.5
chrX:45814	26	(17, 17)	-17.5	9.0	26.5
chr2:129237	12	(12, 18)	18.0	30.5	12.5
chr9:67187	14	(6, 6)	-6.5	8.0	14.5

# TRGT

- Reports SVs that are annotated with TR Information. (a.k.a. sequence resolved)
- VERY easy to compare using Truvari
- Large difference in precision between pVCF and HG002 only
- No bed-file used.
- $\geq 10\text{bp}$
- GTs aren't matching up great for multiple reasons

	0,1	1,0	1,1
1,0	4,768	4,638	226
0,1	5,139	5,144	4,446

	HG002	pVCF
TP-Base	24,361	28,874
TP-Call	24,361	28,874
FP	6,652	2,139
FN	111,287	1,355,554
precision	0.786	0.931
recall	0.180	0.021
f1	0.292	0.041
Base cnt	135,648	1,384,428
Call cnt	31,013	31,013

# Next Steps

- Repeat matching
  - Motif matching
  - Copy-number matching
- Improve Truvari anno trf:
  - Work with all TR Regions (not just SimpleRepeats)
  - Better annotation picking
- Need specific aims:
  - Is using the assemblies and trying to annotate them the right approach?
    - More annotations?
      - Discovered by SRS/LRS
      - Neighborhood Variant Density
  - What separates 'easy' from 'difficult' TRs?
  - Formalize what Truvari's 'sequence resolved' expectations (format standard-ish)
- Dynamic Matching (<50bp)
- **Need to make better variant calls**

# Remaking Variants

- 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



# Maternal Haplotype Performance GIAB SV v0.6

asm	param	precision	TP-base	TP-call	FP
eich	tru	0.925	6,744	6,746	550
	giab	0.918	6,763	6,763	608
	<b>pan</b>	<b>0.947</b>	<b>6,829</b>	<b>6,829</b>	<b>384</b>
	cust	0.935	6,825	6,826	473
li	tru	0.928	6,755	6,757	527
	giab	0.913	6,778	6,780	648
	<b>pan</b>	<b>0.945</b>	<b>6,826</b>	<b>6,827</b>	<b>395</b>
	cust	0.931	6,833	6,835	505

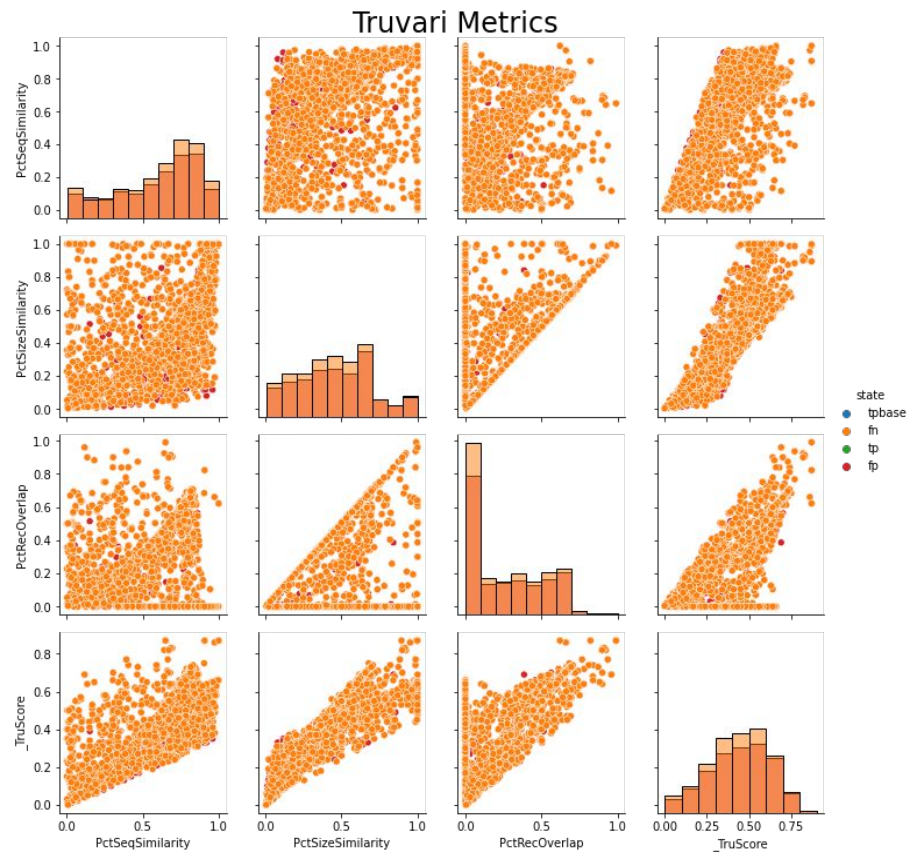
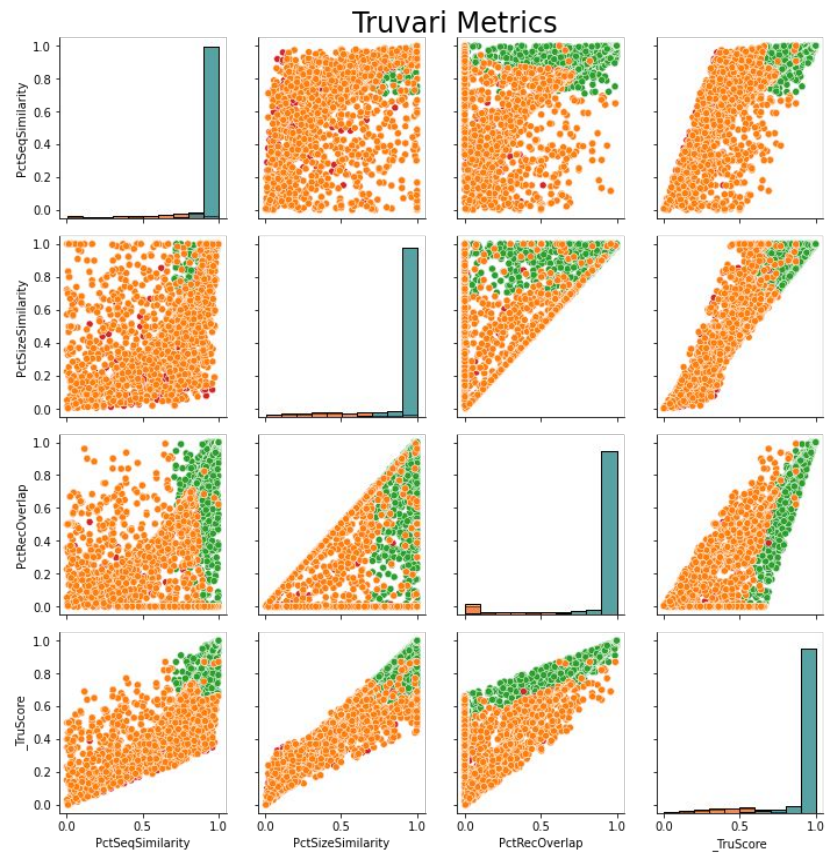
~87.8% of genome covered by single contig per-assembly

# Maternal Haplotype Performance - Consistency

	GIAB	'70%' Similarity				'0%/1000bp'		
param	Base Shared	ASM Shared	ASM1 Unique	ASM2 Unique	% Shared	ASM1 Unique	ASM2 Unique	% Shared
tru	0.652	11,678	<b>4,018</b>	4,509	<b>0.742</b>	<b>2,603</b>	2,568	0.895
giab	0.651	<b>12,197</b>	4,624	4,700	0.723	2,796	2,646	0.885
pan	0.654	11,372	4,129	4,194	0.732	2,690	2,504	<b>0.899</b>
cust	<b>0.655</b>	12,095	4,434	<b>4,117</b>	0.730	2,718	<b>2,457</b>	0.897

~86.2% of genome covered by single contig in both assemblies

# '0%/1000bp' – How Similar?

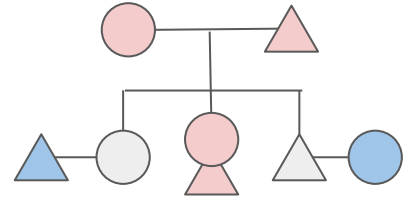
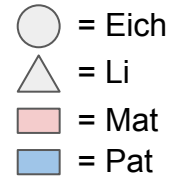


# Sex Check

- Benchmark the eich/li assembly intersection results against GIAB.
- Check the TPs' parent's genotypes
- Assume 0/0 in parent is 'mendelian error'

Intersecting - 5,604/11,372 (49.3%)					
HG004_GT	./.	0/0	0/1	1/1	
HG003_GT					
./.	86	4	136	240	MendErr Tot
0/0	41	1	431	143	77
0/1	81	54	586	515	Pct MendErr
1/1	216	18	397	2416	1.44%
Eich Unique- 1,554/4,194 (37.1%)					
HG004_GT	./.	0/0	0/1	1/1	
HG003_GT					
./.	39	14	61	19	MendErr Tot
0/0	33	3	372	144	216
0/1	58	143	221	152	Pct MendErr
1/1	10	56	79	8	15.30%
Li Unique - 1,560/4,129 (37.8%)					
HG004_GT	./.	0/0	0/1	1/1	
HG003_GT					
./.	30	21	53	10	MendErr Tot
0/0	13	3	155	59	533
0/1	79	381	221	66	Pct MendErr
1/1	23	128	152	10	37.96%

# Intersection with Paternal assembly



Compare the maternal allele SVs intersection sets against the paternal allele SVs (70% similarity)

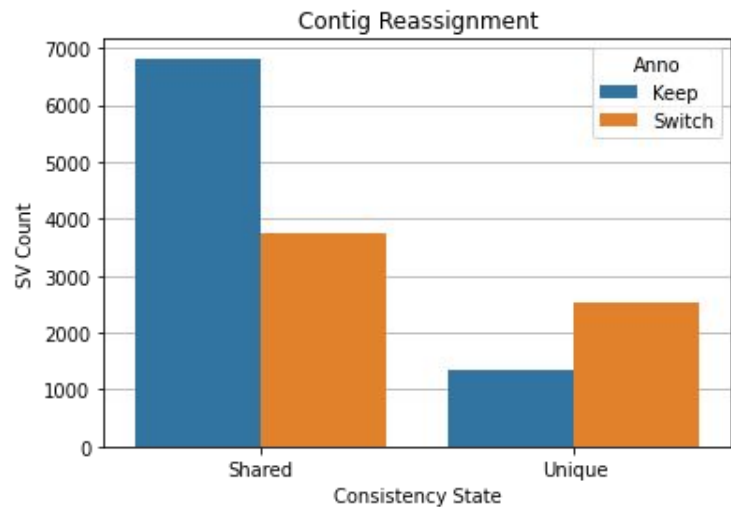
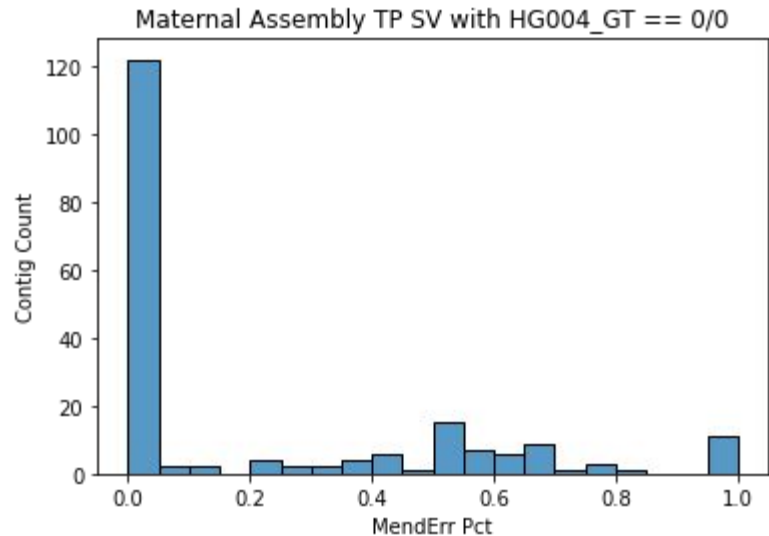
Assembly	Maternal Intersection	TP	FP	Precision
li	Shared	8,237	3,100	<b>72.66%</b>
li	Unique	3,648	481	<b>88.35%</b>
eich	Shared	8,243	3,129	<b>72.49%</b>
eich	Unique	3,850	344	<b>91.80%</b>

Maternal SVs not shared between assemblies are found in the complementary Paternal assemblies

# Can we reassign contigs?

- Asm<->Asm shared state can't be used to determine parental allele
- Procedure:
  - For every TP SV in an assembly that isn't HG002\_GT == 1/1 in GIAB
  - Count how many are MendErr per-contig (Parent\_GT == 0/0).
  - If MendErr >= 15%, annotate that contig as needing a parental switch.

Assembly	Parent	Num Contigs	Contigs w/ TP SV	Contigs Reassigned
li	maternal	5,912	265	72
li	paternal	5,735	281	80
eich	maternal	7,639	249	49
eich	paternal	8,003	239	52



# Reassigned Assembly Results

Original								
		GIAB				Other Asm		
Assembly	Parent	TP	FP	Precision	MendErr %	Shared	Unique	Consistency
li	maternal	6,826	395	94.5%	9.00%	11,372	4,129	73.4%
li	paternal	6,827	393	94.6%	10.21%	11,332	4,212	72.9%
eich	maternal	6,829	384	94.7%	4.29%	11,372	4,194	73.1%
eich	paternal	6,790	395	94.5%	5.44%	11,332	4,001	73.9%
Reassigned								
		GIAB				Other Asm		
Assembly	Parent	TP	FP	Precision	MendErr %	Shared	Unique	Consistency
li	maternal	6,641	391	94.4%	0.99%	13,811	640	95.6%
li	paternal	6,598	359	94.8%	0.20%	13,664	599	95.8%
eich	maternal	6,576	369	94.7%	0.30%	13,811	581	96.0%
eich	paternal	6,525	356	94.8%	0.93%	13,664	572	96.0%

# Contig Reassignment with SNP+INDEL+SV

- Annotate contigs' variants against GIAB v4.2.1 snp/indels.
  - Proband HET variants that are mendelian consistent and present in only one parent.
  - Suggest the contig came from the parent with more shared variants.
- Unite the SNP+INDEL with SV reassignment annotations
  - In cases where the two annotations disagree, no reassignment is performed.

Reassigned (SNP+INDEL+SV)								
Assembly	Parent	GIAB				Other Asm		
		TP	FP	Precision	MendErr %	Shared	Unique	Consistency
li	maternal	6,716	401	94.4%	0.24%	14,308	606	95.9%
li	paternal	6,670	367	94.8%	1.00%	14,183	600	95.9%
eich	maternal	6,709	394	94.5%	0.30%	14,308	556	96.3%
eich	paternal	6,670	378	94.6%	0.90%	14,183	561	96.2%



# Conclusions

- Found a better set of parameters to remake variants
  - Along with a pipeline which annotates things like coverage (dipcall-esque)
- Found methods to increase consistency
  - Leveraged truth-set variants to reassign parental allele of contigs
- Found properties:
  - Assuming 2:1 het/hom - should be upto ~28K high confidence, single-contig covered SVs, which is in-line with expectations
  - ~82% of genome is single covered by both assemblies per-parent. GIAB SV v0.6 Tier1 regions span ~83%.
- Found more to do:
  - Make diploid proband (truvari collapse)
  - Annotate tandem repeats (truvari anno trf)

# GIAB TR

Part III



# Project Key Points

1. Tandem repeats hold important but hard to resolve variations
2. Standard list(s) of tandem repeat regions? (Tiers?)
3. New benchmark improves characterisation and resolution for small variants and structural variants simultaneously
4. New benchmark tools enable accurate comparison of different representations of variants in tandem repeats.

‘Standard list(s) of tandem repeat regions? (Tiers?)’

Step 1 - try to make non-complex annotations

- a. Using the BedFiles collected from the 5 sources
  - b. Merge
  - c. Run TRF on each region's reference, reporting all hits
    - a. Essentially making the UCSC 'simple repeats' track
2. Compare those annotations to the long-read assembly VCFs
- a. Filter/Subset/Pick the annotation that best describes the population

New benchmark improves characterisation and resolution for small variants and structural variants simultaneously

- TrioHifiAsm is really good
  - Possibly areas for improvement, but it is really hard to say for sure if they're better
  - This is the PCTHOM stuff
- I've been wanting to make the pVCF for some time, so I'll describe that process.

# Tandem Repeats and MSA

If we just use a single sample, it'll hard to say “This is the repeat” because we won't know the e.g. motif sequence / step size. There are multiple possible annotations. But presumable over multiple individuals, we can figure out the motif that ‘best’ captures how the genome changes...





# Variant Regularization

- Can 'regularized' variants be more easily compared?
  - This is (slightly) different from 'normalization'
- Hypothesis:
  - Global realignment of haplotypes creates more consistent SVs
- Input:
  - HG002 long-read haplotype resolved assemblies from Garg et.al and Ebert et.al.
- Pipeline:
  - Run minimap2/paftools
  - Bcftools consensus to create full chromosome sequences
  - Remapping/calling the full chromosomes
  - Use `truvari bench` to measure SV ( $\geq 50\text{bp}$ ) similarity of original and 'regularized' VCFs
    - --included regions covered by exactly one contig in each ASM

Reference and alternative alleles of a CA short tandem repeat (STR)	REF ALT	GGGCACACACAGGG GGGCACACAGGG			
			← CA deletion from the reference		
Genome Reference		Variant Call Format			
GGGCACACACAGGG		POS	REF	ALT	
REF	CA	8	CA	.	Not left aligned and alternate allele is empty
ALT	.				
REF	CAC	6	CAC	C	Not left aligned but parsimonious
ALT	C				
REF	GCACA	3	GCACA	GCA	Not right trimmed
ALT	GCA				
REF	GGCA	2	GGCA	GG	Not left trimmed
ALT	GG				
REF	GCA	3	GCA	G	Normalized (left aligned & parsimonious)
ALT	G				
Alleles represented against the human genome reference. Allele pairs are colored the same, all are representations of the same variant.		Alleles represented in Variant Call Format, all are representations of the same variant.			

Source: [https://genome.sph.umich.edu/wiki/Variant\\_Normalization](https://genome.sph.umich.edu/wiki/Variant_Normalization)

# Maternal haplotype comparison

Variant Count

	Original	Regular
Matching	12,326	12,335
Unique (ASM1)	4,571	4,542
Unique (ASM2)	4,770	4,750
Precision	72.9%	73.1%
Recall	72.1%	72.2%
F1	72.5%	72.6%

Altered Bases

Assembly	State	Original	Regular	R - O
1	Matching	6,675,826	6,674,855	-971
2	Matching	6,673,678	6,672,574	-1,104
1	Unmatched	2,386,915	2,385,351	-1,564
2	Unmatched	2,195,803	2,179,734	-16,069

Comparison Metrics of Variants with similarity between [70,100)

Metric	count	mean	std	min	10%	50%	90%	max
Start								
Distance	198	11	184	-760	-165	0	178	990
SizeDiff	198	-9	71	-815	-45	0	30	210

# Intersection between haplotypes from Garg assembly

SV benchmarking/merging use the same fundamental comparison approach.

What is regularization's effect on variants across haplotypes?

Overall, this effort does show a little promise. However, there's still many unknowns.

# Bases by state

State	Original	Regular	R - O
tpbase	5,506,887	5,512,740	5,853
fn	4,208,477	4,195,451	-13,026
tp	5,507,836	5,510,918	3,082
fp	4,348,460	4,338,815	-9,645