GIABTR

Adam English
HGSC @ BCM

July 26, 2022

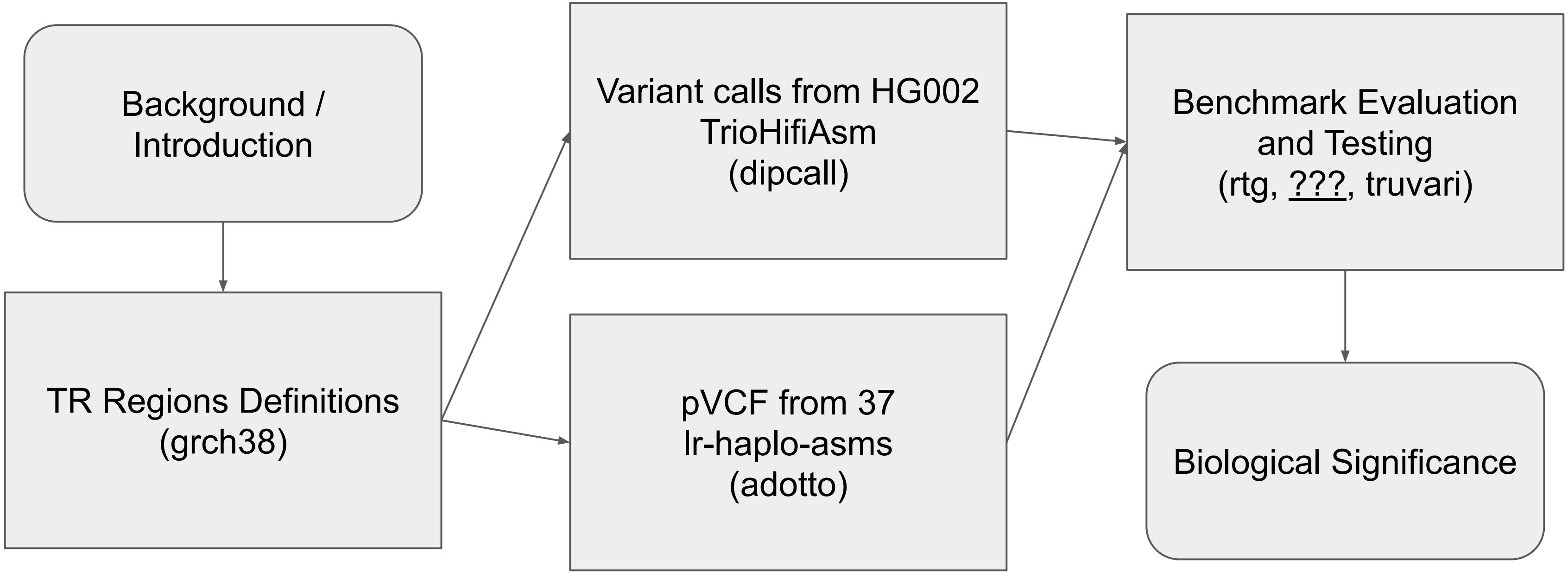
Project Outline

- Fritz has started a project google document to begin writing the manuscript

Key message:

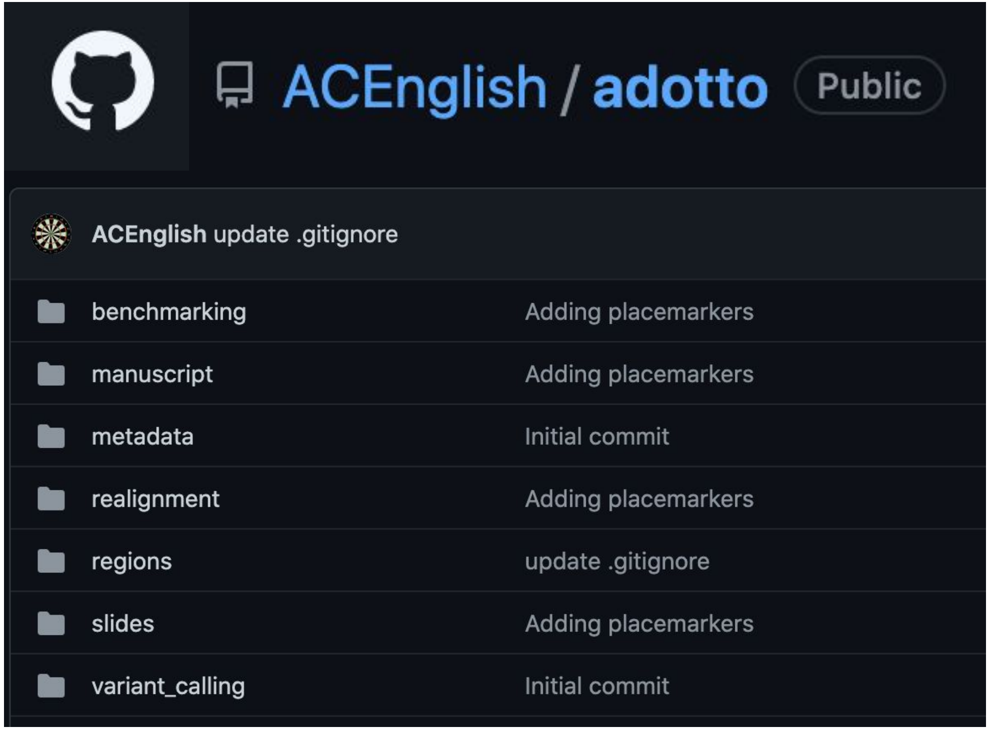
1. Tandem repeats hold important but hard to resolve variations
2. Standard list(s) of tandem repeat regions? (Tiers?)
3. New benchmark improves characterisation and resolution for small variants and structural variants simultaneously
4. New benchmark tools enable accurate comparison of different representations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

Project Roadmap


Background / Introduction
TR Regions Definitions (grch38)
Variant calls from HG002 TrioHifiAsm (dipcall)
pVCF from 37 lr-haplo-asms (adotto)
Benchmark Evaluation and Testing (rtg, ???, truvari)
Biological Significance

Project Management

- Bi-monthly meetings
- Google Doc for manuscript
- Github for analysis and summary tracking
    - Sub-directories and detailed READMEs to organize documentation
    - Would like to use the github's Issues/Discussions pages to delegate/track work



ACEnglish / adotto

ACEnglish update .gitignore

benchmarking	Adding placemarkers
manuscript	Adding placemarkers
metadata	Initial commit
realignment	Adding placemarkers
regions	update .gitignore
slides	Adding placemarkers
variant_calling	Initial commit

TR Regions Definitions

- Collected Tandem Repeat regions bed files from 5 sources
- Attempt to create a grand unified tandem repeat catalogue for grch38
- Strawman tr-regions process:
  - Stats on number/span of regions per-source
  - `bedtools merge` source bed file independently
  - `bedtools merge` between source bed files
  - Run TRF on the regions
  - Add/refine/remove regions from there
- Filtering/Manipulations:
  - Remove source regions:
    - Span < 10bp
    - Span > 50kbp
    - Within 5kbp of grch38 gap annotations
  - Add 25bp slop
- Details in `adotto/regions/README.md`

TR bed input→merged summary stats

	Input		Filtered / Merged		
source	**count**	**span**	**count**	**span**	**~genome %**
baylor	965,511	319,296,434	652,137	74,389,216	2.32%
giab	1,852,129	294,579,269	1,269,585	129,339,212	4.04%
pacbio	171,146	4,538,741	163,355	4,481,815	0.14%
ucsd1	1,776,010	35,852,305	1,738,251	35,300,821	1.10%
ucsd2	10,264	613,138	10,259	612,921	0.02%
TR Regions	3,833,587	244,123,985	2,232,565	238,052,458	**7.44%**

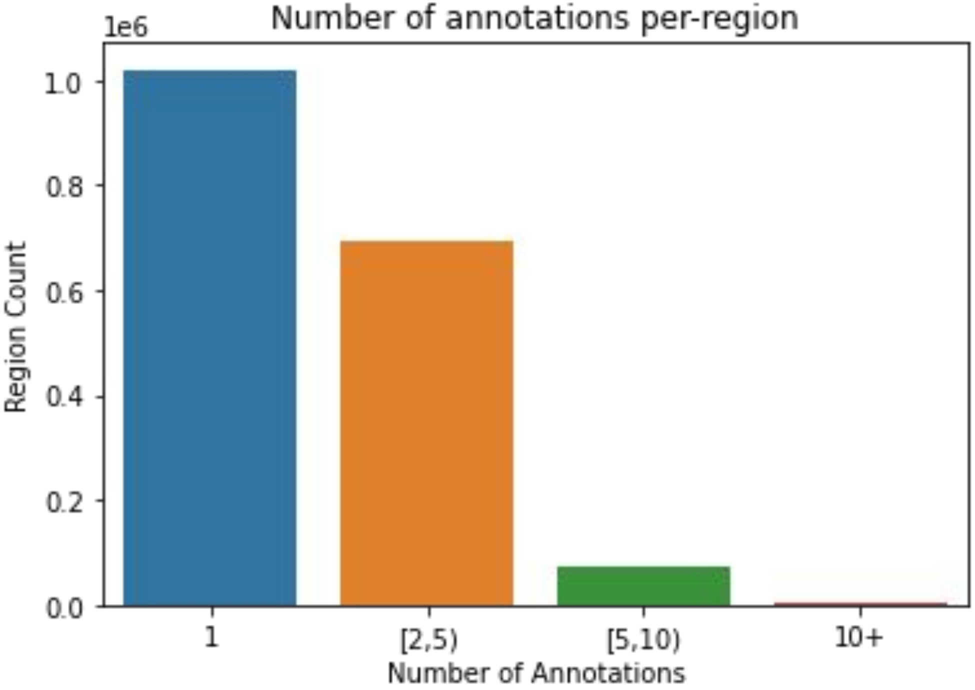
Generating Tandem Repeat Annotations

- Using the set of TR regions, let’s see how many can generate TRF annotations
  - Extract the reference sequence from the region
  - Run TRF
- Table of the expected annotations →
  - Underscored = Used by `truvari anno trf`
- Parameters
  - Match = 3
  - Mismatch = 7
  - Delta = 7
  - PM = 80
  - PI = 5
  - MinScore = 40
  - MaxPeriod = 500

Column	Definition
chrom	chromosome
start	start position of the repeat
end	end position of the repeat
period	period size of the repeat
copies	number of copies of the repeat in the reference
score	alignment score
entropy	entropy measure based on percent composition
repeat	motif of the repeat

QC of TRF Annotations

- Total of **2,232,565** regions
- TRF annotations in 1,793,027 regions (**80.31%**)
- Total of 3,298,925 annotations
- Annotations span 191,486,115 bp
- Merged, the annotations span 110,004,610 bp (**~3.4% of grch38**)



Source intersection with TR annotated

source	Merged Input Count	In TR Annotations	% In TR Annotations	In TR (50% RO)	% In TR (50% RO)
baylor	652,137	648,439	**99.43%**	628,584	**96.39%**
giab	1,269,585	1,072,188	**84.45%**	841,403	**66.27%**
pacbio	163,355	159,535	**97.66%**	130,816	**80.08%**
ucsd1	1,738,251	1,386,733	**79.78%**	1,225,317	**70.49%**
ucsd2	10,259	10,173	**99.16%**	9,811	**95.63%**

Manual Inspection Examples

Next Steps

- Need input on the tr_regions.bed and tr_annotated.bed
  - Spanning the correct places in the tr_regions.bed?
  - tr_annotated.bed entries correct?
  - Can (should) we find tandem-repeat motifs in the 20% of tr_regions.bed without entries in tr_annotated.bed?
    - Different TRF parameters? (RepeatMasker, https://github.com/gregorykucherov/mreps)
- Intersection with HG002 and pVCF
  - Can make a subset of regions/annotations based on their intersection with TrioHifiAsm
  - How many tr_regions/annotated have variants within?
  - Do we count non-SNPs?
  - Would `truvari anno trf` be sufficient?