Evaluation of v3.1 stratifications with Benchmarking

J.McDaniel (June – July 2022)

**Benchmarking Validation Process:**

The defrabb v0.008 pipeline was used to generated draft benchmark from HG2-HPRC-curr.20211005 and perform benchmarking. Since we did not have a benchmark for CHM13 we decided to prepare new draft benchmarks for all references from HG2-HPRC-curr.20211005 using the new v3.1 stratifications as exclusions for ones that have been updated.

AllTandemRepeatsandHomopolymers\_slop5

segdups

AllTandemRepeats\_gt10000bp\_slop5

chainSelf

satellites

rDNA (CHM13 only)

|  |  |  |
| --- | --- | --- |
| **Alignment Reference** | **defrabb run  (on BBD\_Human\_Genomics team Google Drive)** | **HiFi-DV query callsets for benchmarking with hap.py** |
| GRCh37 | 20220705\_v0.008\_HG002-HPRC-GRCh37 | deepvariant\_HG002\_GRCh37\_15kb\_20kb\_52X\_SequelII.vcf.gz |
| GRCh38 | 20220705\_v0.008\_HG002-HPRC-GRCh38 | deepvariant\_HG002\_GRCh38\_15kb\_20kb\_53X\_SequelII.vcf.gz |
| CHM13v2.0 | 20220705\_v0.008\_HG002-HPRC-CHM13v2.0 | HG002.pacbio.chm13.vcf.gz (from A.Carroll) |

**Evaluation:**

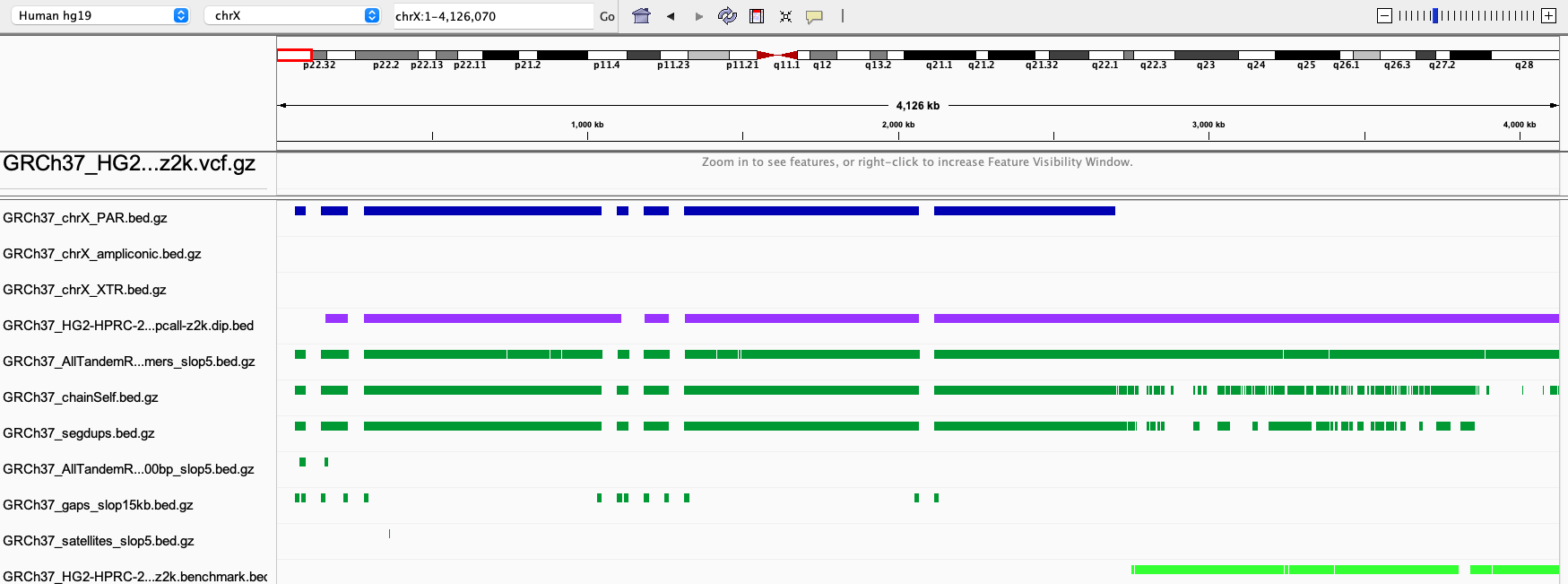
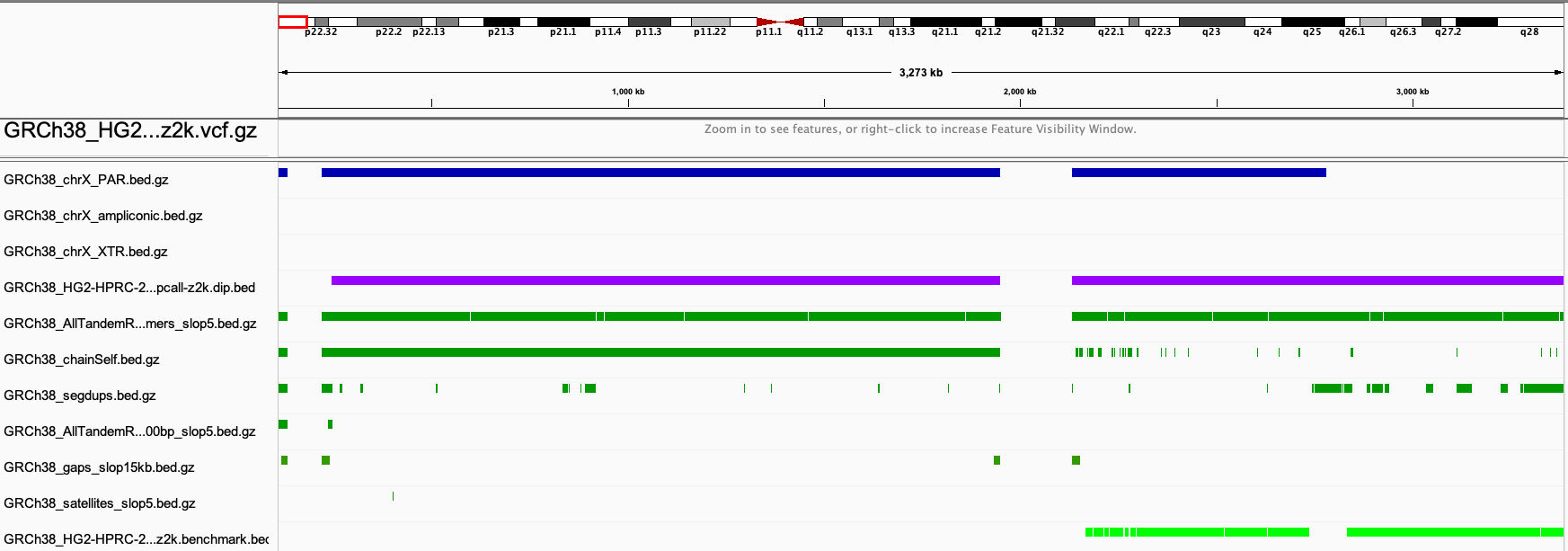
Summary statistics (v3.1-strat-eval\_benchmarking-summary-stats.csv) from the hap.py extended.csvs for all references was generated in R withv3.1\_benchmarking\_summary\_metrics\_for\_all\_stratification\_evaluation.R

Plots of summary metrics were generated in

metrics.html 🡪 contains all summary metrics and stratificaitons JZ was interested in

note: log10 values for TP/.het/.homalt were plotted

**Summary of Issues and Resolutions:**





Upon inspection of benchmarking for validation if v3.1 stratifications using the following defrabb runs

|  |
| --- |
| 20220610\_v0.008\_HG002-HPRC-GRCh38 |
| 20220616\_v0.008\_HG002-HPRC-CHM13v2.0 |
| 20220616\_v0.008\_HG002-HPRC-GRCh37 |

**Issue #1:**

It was found that GRCh37 PARX stratifications did not have any metrics. In looking at the regions (above) JZ noted For GRCh37, the PAR is incorrectly considered a segdup and a self-chain, probably because they didn't mask the Y PAR when creating these files. For GRCh38, this is fixed for segdups, but not for the self-chain, though the self-chain only covers the first ~2Mbp of the PAR. Because we exclude self chains and segdups that have any breaks in the dip.bed inside them or within 15kb of them, the entire PAR gets excluded for 37 and the first 2Mbp gets excluded for 38.

**Resolution:**

Perform the following subtractions using PARX from Heng Li

GRCh38 subtract PAR X from

* chainself\_gt10kb.bed
* chainself.bed

GRCh37 subtract PAR X from

* chainself\_gt10kb.bed
* chainself.bed
* segdups.bed
* segdups\_gt10kb.bed

CHM13

no subtraction as segdups strat is properly annotated

**Issue #2:**

Over time we've used PAR regions from three different sources for strats and dipcall.  I wanted to confirm the regions were the same and found they are not

**Resolution:**

update PARX and Y to use heng (X) and UCSC (Y)

|  |  |  |
| --- | --- | --- |
| **GRCh38** | **PAR- X** | **PAR-Y** |
| Melissa – used for XY strat | 10001 2781479  155701383 156030895 | 10001 2781479  56987321 57217415 |
| Heng – used by dipcall | 0 2781479  155701383 156030895 |  |
| UCSC – used for subtraction in post processing |  | 10000 2781479  56887902 57217415 |
|  |  |  |
| **GRCh37** | **PAR- X** | **PAR-Y** |
| Melissa – used for XY strat | 60001 2699520  154931044 155260560 | 0 2749806  59133470 59373566 |
| Heng – used by dipcall | 0 2699520  154931043 155260560 |  |
| UCSC – used for subtraction in post processing |  | 10001 2649520  59034050 59363566 |

**Issue #3:**

CHM13 chrX\_XTR and chrX\_PAR stratification names are switched, and chrY\_XTR and chrY\_PAR are also switched due to error in stratification tsv with file locations used by hap.py.

**resolution:**

fix fields in tsv

**Issue #4:**

JZ second-guessing decision to make the "nonPAR-nonXTR-nonampliconic" beds.

**Resolution:**

Change these to be just "nonPAR" for 37, 38, and CHM13 (i.e., don't exclude XTR and ampliconic, but just exclude PAR) in the next revision?

**Other Notes:**

JZ noted that it looks like the higher indel recall and precision on CHM13 is almost entirely due to homopolymers longer than 10bp. In contrast, the lower snp recall and precision on CHM13 seem related to tandem repeats and segdups (and snps in homopolymers >10bp are actually better in CHM13)

**Issue #5:**

Following revision of PAR regions and correction of tsv we ran benchmarking again. JZ noted that something was weird with CHM13 PAR-Y and he suspected HiFi-DV query callset didn’t mask chrY PAR when aligning. He asked that I pull up HifiDV vcf in IGV (see below) and look at how many variants in PAR-Y.

**Resolution:**

Looking at query callsets for all reference it does appear PAR-Y was not masked in the CHM13v2.0 query callset as JZ suspected. This is not a problem with our stratifications rather and issue with the contributed callset. PAR-Y was properly masked in GRCh37/38 callsets.

