Vcf Concordance Testing

Pipeline Testing for NA12878 –

1. Creating a manifest file –

python manifest\_to\_map.py /rbdata/binfo/Illumina\_BeadChips/SanfordChipV1/Sanford\_Chip\_v1\_20012521\_A1.bpm /rbdata/binfo/reference/Homo\_sapiens.GRCh37.dna.toplevel.fa

* Location of manifest file –

/rbdata/binfo/Illumina\_BeadChips/SanfordChipV1/Sanford\_Chip\_v1\_20012521\_A1.bpm.map

1. Converting the gtc files to vcf files –

python ./gtc\_vcf.py /rbdata/binfo/Illumina\_BeadChips/SanfordChipV1/Sanford\_Chip\_v1\_20012521\_A1.map 201009170067\_R01C02.gtc

* Filtering out the non-calls in the array vcf file –

grep -v '0/0' test.vcf > new.test.vcf

grep -v '\./\.' new.test.vcf > new.test1.vcf

grep -v 'XY' new.test1.vcf > filtered\_test.vcf

1. Sorting, bgzip and tabix the vcf file –

vcf-sort filtered\_test.vcf > sorted.test.vcf

bgzip -c sorted.test.vcf > sorted.test.vcf.gz

tabix -p vcf sorted.test.vcf.gz

1. Creating the high confidence Snp s file –

bedtools intersect -a NIST\_RTG\_PlatGen\_merged\_highconfidence\_v0.2.primitives.vcf.gz -b union13callableMQonlymerged\_addcert\_nouncert\_excludesimplerep\_excludesegdups\_excludedecoy\_excludeRepSeqSTRs\_noCNVs\_v2.19\_2mindatasets\_5minYesNoRatio\_AddRTGPlatGenConf\_filtNISTclustergt9\_RemNISTfilt\_RemPartComp\_RemRep\_RemPartComp\_v0.2.bed.gz -wb -header > giab.vcf

bedtools intersect -a giab.vcf -b sorted.test.vcf.gz -header > giab\_chip\_intersection.vcf

vcf-sort giab\_chip\_intersection.vcf > sorted.giab\_chip\_intersection.vcf

bgzip -c sorted.giab\_chip\_intersection.vcf > sorted.giab\_chip\_intersection.vcf.gz

tabix -p vcf sorted.giab\_chip\_intersection.vcf.gz

1. Vcftools –

vcf-compare sorted.test.vcf.gz sorted.giab\_chip\_intersection.vcf.gz

# This file was generated by vcf-compare.

# The command line was: vcf-compare(v0.1.14-12-gcdb80b8) sorted.test.vcf.gz sorted.giab\_chip\_intersection.vcf.gz

#

#VN 'Venn-Diagram Numbers'. Use `grep ^VN | cut -f 2-` to extract this part.

#VN The columns are:

#VN        1  .. number of sites unique to this particular combination of files

#VN        2- .. combination of files and space-separated number, a fraction of sites in the file

VN 855 sorted.giab\_chip\_intersection.vcf.gz (0.6%)

VN 48745 sorted.test.vcf.gz (25.9%)

VN 139557 sorted.giab\_chip\_intersection.vcf.gz (99.4%) sorted.test.vcf.gz (74.1%)

#SN Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.

SN Number of REF matches: 139428

SN Number of ALT matches: 120669

SN Number of REF mismatches: 129

SN Number of ALT mismatches: 18759

SN Number of samples in GT comparison: 0

# Number of sites lost due to grouping (e.g. duplicate sites): lost, %lost, read, reported, file

SN Number of lost sites: 2 0.0% 140414 140412 sorted.giab\_chip\_intersection.vcf.gz

zcat sorted.test.vcf.gz | vcf-convert -r /rbdata/binfo/reference/Homo\_sapiens.GRCh37.dna.toplevel.fa > new.sorted.test.vcf