HG002 Integration

The following datasets were used for HG002 integration v.3.2:

Illumina 150bp PE 300X Novoaligned mapped data

SOLiD 5500W 75bp SE LifeScope mapped data

Complete Genomics

Ion Exome

Preparation of the data for integration included variant calling, conversion of gvcf to vcf and determining callable/not-callable regions. All pre-processing steps and integration was performed separately for each chromosome. Unless otherwise noted all processing was performed on DNAnexus.

**Integration Pre-processing**

Illumina

Variant calling was performed for all chromosomes using GATKHC v3.5 and FreeBayes. Following variant calling the the gvcf files were converted to vcf using the Genotype-GVCFs v3.5 applet. Callable loci was then run for each chromosome using the same parameters as NA12878. In hindsight I realized I should have looked at the depth of coverage and adjusted the maxDepth option in the callable loci command to be twice the depth of coverage. Looking at the vcf for chromosome 20 I found that the median depth of coverage for this chromosome was 263. The maxDepth used for callable loci was 560 and in retrospect should have adjusted to 526. The vcf from GATK and FreeBayes files were then combined vcf-combineall chrom applet and the bed files from callable loci were combined using bed-combineall applet. A new “callableonly” bed file was generated from the combined callable loci output by subsetting for “callable” regions. The combined files were used for comparison and manual curation steps following integration.

SOLiD 5500W

HG002 data was mapped to hg19 by John Bodeau at LifeTechnologies using LifeScope v2.5.1. Picard-tools-1.141 AddReadGroups was used to add read groups to the bam files. Samtools-1.2 merge was used to merge bams for all HG002 sequencing runs. Merged bams were then indexed and reheadered using Samtools index and reheader respectively.

Variant calling was performed for all chromosomes with exception of MT as no bam was generated for MT. Variant calling was performed using GATKHC v3.5. Following variant calling the the gvcf files were converted to vcf using the Genotype-GVCFs v3.5 applet. Prior to running callable loci the depth of coverage was estimated using the vcf from chromosome 20. The median depth of coverage for chromosome 20 was found to be 42. For the callable loci command the maxDepth was adjusted to 84, twice the depth of coverage. Callable loci was run for all chromosomes, with exception of MT. Vcf and bed file for SOLiD were not combined as no comparisons were made with SOLiD variant calls.

Complete Genomics

To generate vcf and bed files for Complete Genomics data the integration-prepare-cg applet was run for all chromosomes with exception of MT which failed in integration-prep. A separate command for each chromosome was run to split by chromosome. This applet was updated during the processing of HG002 to also output a “callable” loci bed, in addition to “notcallable”, to be used in comparison following integration. Using the “callable” bed files the bed-combineallchrom applet was to generated combined “callable” region bed file for use in comparison following integration. The combined files were used for comparison and manual curation steps following integration.

Ion Exome

To generate bed files callable loci was run for all chromosomes. MT failed callable loci run and was not further processed. To generate vcf for Ion Exome data the integration-prepare-cg applet was run for all chromosomes with exception of MT. A separate command for each chromosome was run to split by chromosome. Chromosome Y failed during integration-prep likely due to no variants being called.

**Integration (5/20/16)**

Nist-integration-v3.2 was run by chromosome for chromosomes 1-22. Following integration, the following integration by chromosome output files were combine separately: highconf.vcf.gz, highconf.bed, callablemultinter\_gt0.bed, filteredsites.bed, and all.vcf.gz. The high confidence call file was then compared with the following callsets: NA12878 integration v3.2, HG002 Illumina GATKHC and HG002 Complete Genomics. As previously mentioned “callable” bed files needed to be generated for Illumina and Complete Genomics, as output from callable loci generates bed files with callable status of a region and is not limited to callable regions.