Exome Capture Metrics GVA2020

> ⚠ **Data used in this tutorial**
>
> Recommended but not required that you first complete the trios tutorial and use the data you generated here. Alternatively, canned data provided.

Evaluating capture metrics

There are many ways to measure sequence capture. You might care more about minimizing off-target capture, to make your sequencing dollars go as far as possible. Or you might care more about maximizing on-target capture, to make sure you get data from every region of interest. These two are usually negatively correlated.

Using Picard's "CollectHsMetrics" function to evaluate capture

Here is a link to the full picard documentation and here is a link to the CollectHsMetrics tool

To run CollectHsMetrics on Lonestar, there are three prerequisites: 1) A bam file and 2) a list of the genomic intervals that were to be captured and 3) the reference (.fa). As you would guess, the BAM and interval list both have to be based on exactly the same genomic reference file.

For our tutorial, the bam files are one of these:

BAM files for exome capture evaluation tutorial

```
/corral-repl/utexas/BioITeam/ngs_course/human_variation/NA12878.chrom20.ILLUMINA.bwa.CEU.exome.20111114.bam
/corral-repl/utexas/BioITeam/ngs_course/human_variation/NA12892.chrom20.ILLUMINA.bwa.CEU.exome.20111114.bam
/corral-repl/utexas/BioITeam/ngs_course/human_variation/NA12891.chrom20.ILLUMINA.bwa.CEU.exome.20111114.bam
```

I've started with one of Illumina's target capture definitions (the vendor of your capture kit will provide this) but since the bam files only represent chr21 data I've created a target definitions file from chr21 only as well. Here they are:

Two relevant target list definitions

```
/corral-repl/utexas/BioITeam/ngs_course/human_variation/target_intervals.chr20.reduced.withhead.intervallist
/corral-repl/utexas/BioITeam/ngs_course/human_variation/target_intervals.reduced.withhead.intervallist
```

And the relevant reference is:

Reference for exome metrics

```
/corral-repl/utexas/BioITeam/ngs_course/human_variation/ref/hs37d5.fa
/corral-repl/utexas/BioITeam/ngs_course/human_variation/ref/hs37d5.fa.fai
```

This block will work on data you generated in the human trios analysis

```
mkdir $SCRATCH/GVA_Exome_Capture
cd $SCRATCH/GVA_Exome_Capture
cp $SCRATCH/GVA_Human_trios/raw_files/NA12878.chrom20.ILLUMINA.bwa.CEU.exome.20111114.bam .
cp $SCRATCH/GVA_Human_trios/raw_files/target_intervals.chr20.reduced.withhead.intervallist .
cp $SCRATCH/GVA_Human_trios/raw_files/ref/hs37d5.fa .
cp $SCRATCH/GVA_Human_trios/raw_files/ref/hs37d5.fa.fai .
```

This block will work if you have not completed the human trios tutorial

```
mkdir $SCRATCH/GVA_Exome_Capture
cd $SCRATCH/GVA_Exome_Capture
cp /corral-repl/utexas/BioITeam/ngs_course/human_variation/NA12878.chrom20.ILLUMINA.bwa.CEU.exome.20111114.bam .
cp /corral-repl/utexas/BioITeam/ngs_course/human_variation/NA12892.chrom20.ILLUMINA.bwa.CEU.exome.20111114.bam .
cp /corral-repl/utexas/BioITeam/ngs_course/human_variation/NA12891.chrom20.ILLUMINA.bwa.CEU.exome.20111114.bam .
cp /corral-repl/utexas/BioITeam/ngs_course/human_variation/target_intervals.chr20.reduced.withhead.intervallist
.
cp /corral-repl/utexas/BioITeam/ngs_course/human_variation/target_intervals.reduced.withhead.intervallist .
cp /corral-repl/utexas/BioITeam/ngs_course/human_variation/ref/hs37d5.fa .
cp /corral-repl/utexas/BioITeam/ngs_course/human_variation/ref/hs37d5.fa.fai .
```

The run command looks long but isn't that complicated (like most java programs):

How to run exactly these files on Lonestar

```
java -Xmx4g -Djava.io.tmpdir=/tmp -jar /corral-repl/utexas/BioITeam/bin/picard.jar CollectHsMetrics
BAIT_INTERVALS=target_intervals.chr20.reduced.withhead.intervallist TARGET_INTERVALS=target_intervals.chr20.
reduced.withhead.intervallist INPUT=NA12878.chrom20.ILLUMINA.bwa.CEU.exome.20111114.bam
REFERENCE_SEQUENCE=hs37d5.fa  OUTPUT=exome.picard.stats PER_TARGET_COVERAGE=exome.pertarget.stats
```

You may notice that the picard tool is found in the BioITeam directory and it is called using the full path to the .jar file. In tomorrows closing tutorial, you'll see two different options to create a small bash script to avoid the java invocation or at least avoid having to remember where picard.jar is stored as even though it is in our path, jar files are not found with the which command.

The aggregate capture data is in exome.picard.stats, but it's format isn't very nice; here's a linux one-liner to reformat the two useful lines (one is the header, the other is the data) into columns, along with the result:

```
grep -A 1 '^BAIT' exome.picard.stats | awk 'BEGIN {FS="\t"} {for (i=1;i<=NF;i++) {a[NR"_"i]=$i}} END {for (i=1;
i<=NF;i++) {print a[1"_"i]"\t"a[2"_"i]}}'
```

Here is the output of the above command, **DO NOT!** paste this into the command line.

```
BAIT_SET        target_intervals
GENOME_SIZE        3137454505
BAIT_TERRITORY        1843371
TARGET_TERRITORY        1843371
BAIT_DESIGN_EFFICIENCY        1
TOTAL_READS        4579959
PF_READS        4579959
PF_UNIQUE_READS        4208881
PCT_PF_READS        1
PCT_PF_UQ_READS        0.918978
PF_UQ_READS_ALIGNED        4114249
PCT_PF_UQ_READS_ALIGNED        0.977516
PF_UQ_BASES_ALIGNED        283708397
ON_BAIT_BASES        85464280
NEAR_BAIT_BASES        49788346
OFF_BAIT_BASES        148455771
ON_TARGET_BASES        85464280
PCT_SELECTED_BASES        0.476731
PCT_OFF_BAIT        0.523269
ON_BAIT_VS_SELECTED        0.631886
MEAN_BAIT_COVERAGE        46.363038
MEAN_TARGET_COVERAGE        46.76568
PCT_USABLE_BASES_ON_BAIT        0.245533
PCT_USABLE_BASES_ON_TARGET        0.245533
FOLD_ENRICHMENT        512.716312
ZERO_CVG_TARGETS_PCT        0.009438
FOLD_80_BASE_PENALTY        23.38284
PCT_TARGET_BASES_2X        0.849372
PCT_TARGET_BASES_10X        0.484824
PCT_TARGET_BASES_20X        0.435911
PCT_TARGET_BASES_30X        0.401622
PCT_TARGET_BASES_40X        0.36876
PCT_TARGET_BASES_50X        0.335459
PCT_TARGET_BASES_100X        0.173683
HS_LIBRARY_SIZE        5325189
HS_PENALTY_10X        232.05224
HS_PENALTY_20X        -1
HS_PENALTY_30X        -1
HS_PENALTY_40X        -1
HS_PENALTY_50X        -1
HS_PENALTY_100X        -1
AT_DROPOUT        2.143632
GC_DROPOUT        10.000011
SAMPLE
LIBRARY
READ_GROUP
```

Taking the output even further

It is rare that you ever want to work with a single sample. While this format is nice for a single sample, comparing the same data across multiple samples would not be the easiest to do with this format. Instead, putting this information to a file, then using grep and awk you could make a small table of the specific information you want.

Since I don't actually know what capture kit was used to produce these libraries, these may or may not accurately reflect how well the library prep went, but generally speaking having >40x average coverage on your baits (the target regions) is good, as is over 500 fold enrichment. While it may be tempting to consider 52% of reads being 'off bait' as a bad thing, instead consider that ~48% of reads mapped to just ~0.06% of the genome.

Additional Exercises:

These results were based on sample NA12878. How do the other 2 samples (NA12891, and NA12892) from the trios tutorial compare for their enrichment?

Return to GVA2020