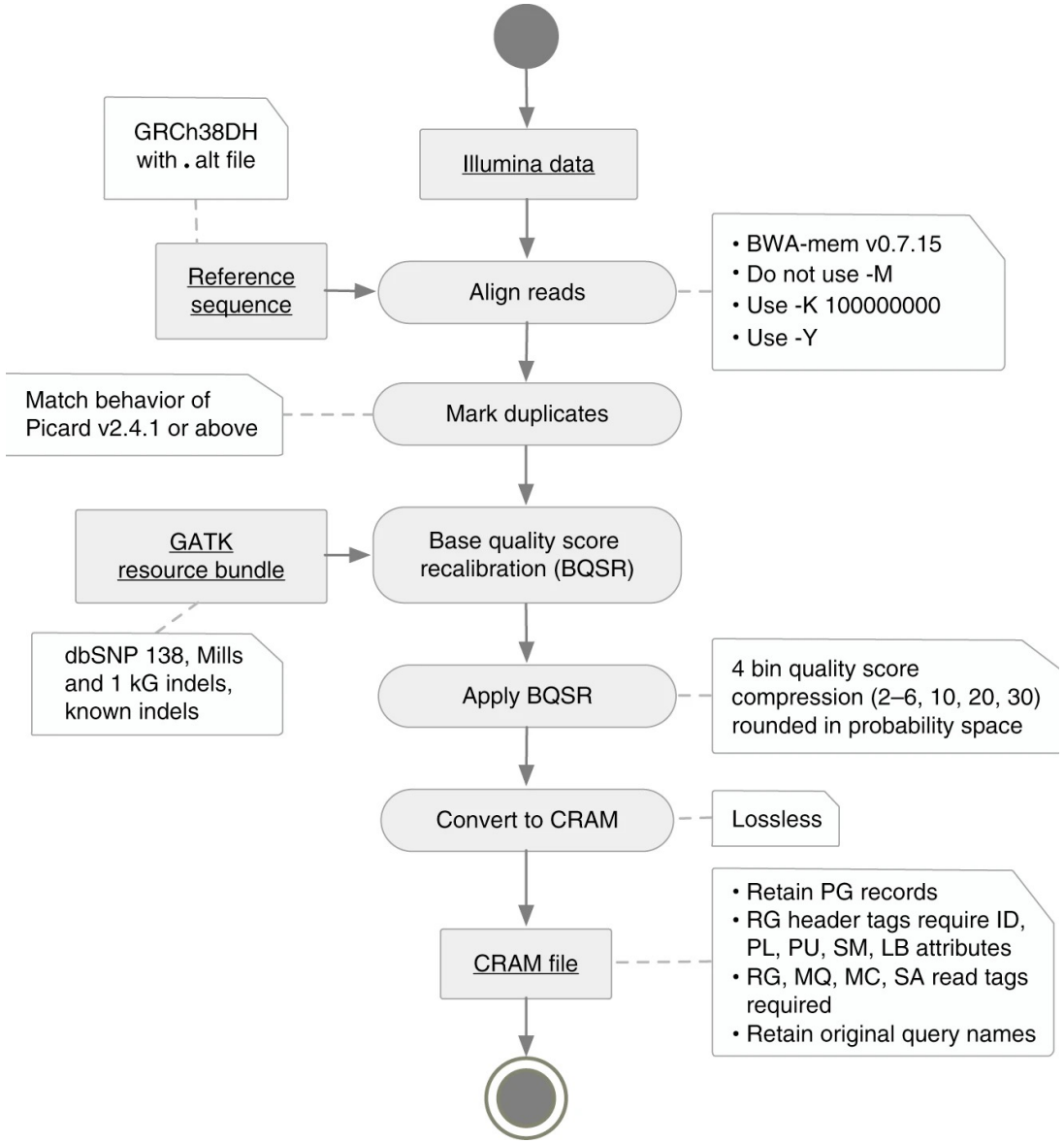
DNA Alignment Continued…

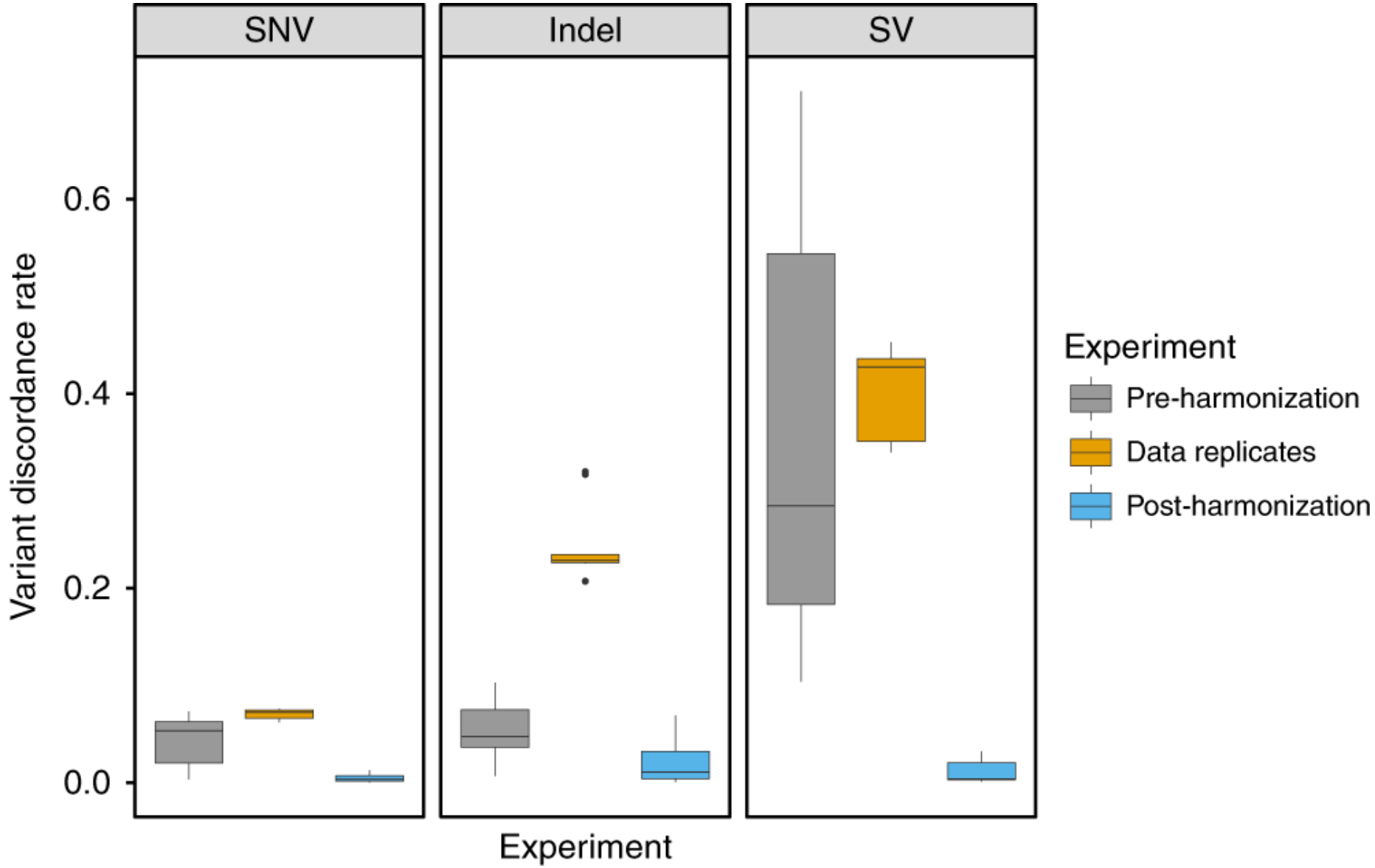
BFX Workshop

10/17/22

Functional Equivalence (FE)

- Regier, A.A., Farjoun, Y., Larson, D.E. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun* **9,** 4038 (2018). https://doi.org/10.1038/s41467-018-06159-4




SNV
Indel
SV
Variant discordance rate
0.6
0.4
0.2
0.0
Experiment
Experiment
Pre-harmonization
Data replicates
Post-harmonization

FE Post-harmonization @ MGI

- Align each read group separately to the GRCh38 reference
  - bwa-mem (v0.7.15-r1140) with the parameters “-K 100000000 -p -Y”.
- MC and MQ tags are added using samblaster (v0.1.24) with the parameters “-a --addMateTags”.
- Read group BAM files are merged together with “samtools merge” (v1.3.1-2).
- The resulting file is name-sorted with “sambamba sort -n” (v0.6.4).
- Duplicates are marked using Picard MarkDuplicates (v2.4.1) with the parameter “ASSUME_SORT_ORDER = queryname”, then the results are coordinate sorted using “sambamba sort”.
- A base quality recalibration table is generated using GATK BaseRecalibrator (v3.6) with knownSites files (dbSNP138, Mills and 1 kg indels, and known indels) from the GATK resource bundle (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0)
  - parameters “--preserve_qscores_less_than 6 -dfrac .1 -nct 4 -L chr1 -L chr2 -L chr3 -L chr4 -L chr5 -L chr6 -L chr7 -L chr8 -L chr9 -L chr10 -L chr11 -L chr12 -L chr13 -L chr14 -L chr15 -L chr16 -L chr17 -L chr18 -L chr19 -L chr20 -L chr21 -L chr22”.
- The base recalibration table is applied using GATK PrintReads with the parameters “-preserveQ 6 -BQSR “${bqsrt}” -SQQ 10 -SQQ 20 -SQQ 30 --disable_indel_quals”.
- Finally, the output is converted to CRAM using “samtools view”.

Standard SAM Tags

Samblaster adds MC and MQ tags to alignments.

Tag	Type	Description
AM	i	The smallest template-independent mapping quality in the template
AS	i	Alignment score generated by aligner
BC	Z	Barcode sequence identifying the sample
BQ	Z	Offset to base alignment quality (BAQ)
BZ	Z	Phred quality of the unique molecular barcode bases in the OX tag
CB	Z	Cell identifier
CC	Z	Reference name of the next hit
CG	B,I	BAM only: CIGAR in BAM's binary encoding if (and only if) it consists of >65535 operators
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CO	Z	Free-text comments
CP	i	Leftmost coordinate of the next hit
CQ	Z	Color read base qualities
CR	Z	Cellular barcode sequence bases (uncorrected)
CS	Z	Color read sequence
CT	Z	Complete read annotation tag, used for consensus annotation dummy features
CY	Z	Phred quality of the cellular barcode sequence in the CR tag
E2	Z	The 2nd most likely base calls
FI	i	The index of segment in the template
FS	Z	Segment suffix
FZ	B,S	Flow signal intensities
GC	?	Reserved for backwards compatibility reasons
GQ	?	Reserved for backwards compatibility reasons
GS	?	Reserved for backwards compatibility reasons
H0	i	Number of perfect hits
H1	i	Number of 1-difference hits (see also NM)
H2	i	Number of 2-difference hits
HI	i	Query hit index
IH	i	Query hit total count
LB	Z	Library
MC	Z	CIGAR string for mate/next segment
MD	Z	String encoding mismatched and deleted reference bases
MF	?	Reserved for backwards compatibility reasons
MI	Z	Molecular identifier; a string that uniquely identifies the molecule from which the record was derived
MQ	i	Mapping quality of the mate/next segment
NH	i	Number of reported alignments that contain the query in the current record
NM	i	Edit distance to the reference
OA	Z	Original alignment
OC	Z	Original CIGAR (deprecated; use OA instead)
OP	i	Original mapping position (deprecated; use OA instead)
OQ	Z	Original base quality
OX	Z	Original unique molecular barcode bases
PG	Z	Program
PQ	i	Phred likelihood of the template
PT	Z	Read annotations for parts of the padded read sequence
PU	Z	Platform unit
Q2	Z	Phred quality of the mate/next segment sequence in the R2 tag
QT	Z	Phred quality of the sample barcode sequence in the BC tag
QX	Z	Quality score of the unique molecular identifier in the RX tag
R2	Z	Sequence of the mate/next segment in the template
RG	Z	Read group
RT	?	Reserved for backwards compatibility reasons
RX	Z	Sequence bases of the (possibly corrected) unique molecular identifier
S2	?	Reserved for backwards compatibility reasons
SA	Z	Other canonical alignments in a chimeric alignment
SM	i	Template-independent mapping quality
SQ	?	Reserved for backwards compatibility reasons
TC	i	The number of segments in the template
TS	A	Transcript strand
U2	Z	Phred probability of the 2nd call being wrong conditional on the best being wrong
UQ	i	Phred likelihood of the segment, conditional on the mapping being correct
X?	?	Reserved for end users
Y?	?	Reserved for end users
Z?	?	Reserved for end users

Picard MarkDuplicates – Duplication Metrics

Metrics that are calculated during the process of marking duplicates within a stream of SAMRecords.

Field	Description
LIBRARY	The library on which the duplicate marking was performed.
UNPAIRED_READS_EXAMINED	The number of mapped reads examined which did not have a mapped mate pair, either because the read is unpaired, or the read is paired to an unmapped mate.
READ_PAIRS_EXAMINED	The number of mapped read pairs examined. (Primary, non-supplemental)
SECONDARY_OR_SUPPLEMENTARY_RDS	The number of reads that were either secondary or supplementary
UNMAPPED_READS	The total number of unmapped reads examined. (Primary, non-supplemental)
UNPAIRED_READ_DUPLICATES	The number of fragments that were marked as duplicates.
READ_PAIR_DUPLICATES	The number of read pairs that were marked as duplicates.
READ_PAIR_OPTICAL_DUPLICATES	The number of read pairs duplicates that were caused by optical duplication. Value is always < READ_PAIR_DUPLICATES, which counts all duplicates regardless of source.
PERCENT_DUPLICATION	The fraction of mapped sequence that is marked as duplicate.
ESTIMATED_LIBRARY_SIZE	The estimated number of unique molecules in the library based on PE duplication.

https://gatk.broadinstitute.org/hc/en-us/articles/360036350292-MarkDuplicates-Picard-
http://broadinstitute.github.io/picard/picard-metric-definitions.html#DuplicationMetrics

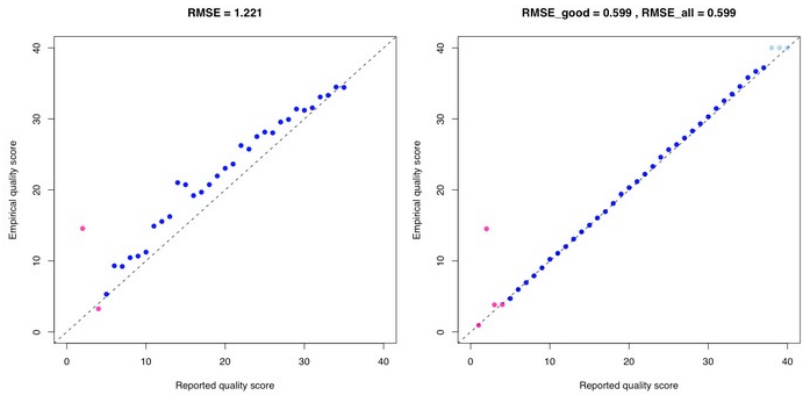
BaseRecalibrator builds the model

- Features Collected
  - read group the read belongs to
  - quality score reported by the machine
  - machine cycle producing this base (Nth cycle = Nth base from the start of the read)
  - current base + previous base (dinucleotide)
- “For each bin, we count the number of bases within the bin and how often such bases mismatch the reference base…”
  - Exclude dbSNP aka. Known Variants
- Apply ”yates” correction for sparse quality bins
- Output to a recalibration file in GATKReport format

ApplyBQSR, aka. PrintReads, adjusts the scores

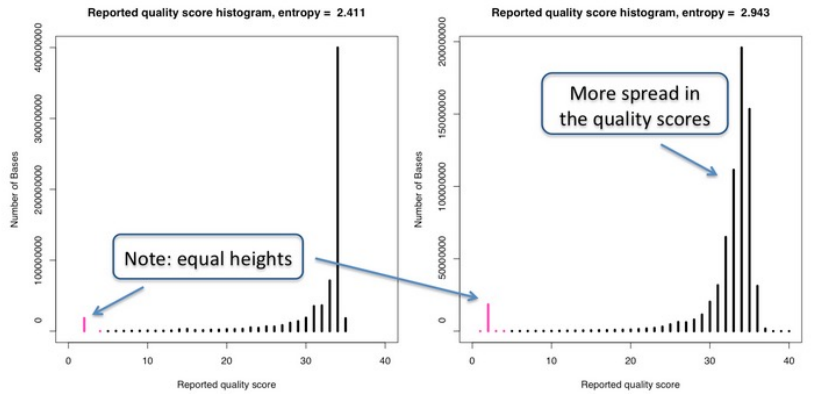
- Using the recalibration file from BaseRecalibrator, adjust each base's score based on which bins it falls in.
- New quality score is:
  - the sum of the global difference between reported quality scores and the empirical quality
  - plus the quality bin specific shift
  - plus the cycle x qual and dinucleotide x qual effect
- Following recalibration, the read quality scores are much closer to their empirical scores than before.

Reported Quality vs. Empirical Quality



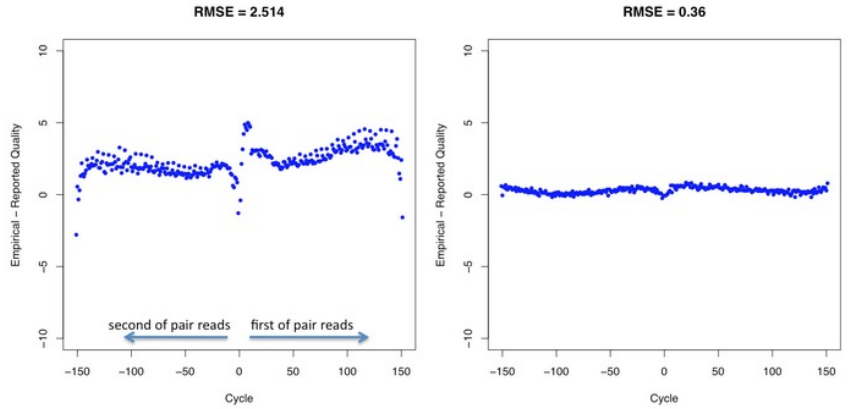
Original Data | After GATK Recalibration

Distribution of Quality Scores



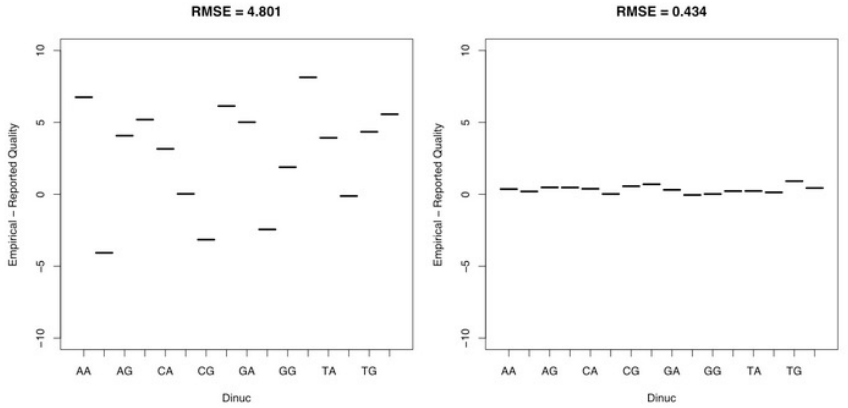
Original Data | After GATK Recalibration

Residual Error by Machine Cycle



Original Data | After GATK Recalibration

Residual Error by Dinucleotide



Original Data | After GATK Recalibration

https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR-

InsertSizeMetrics

Metrics about the insert size distribution of a paired-end library, created by the CollectInsertSizeMetrics program and usually written to a file with the extension ".insert_size_metrics". In addition the insert size distribution is plotted to a file with the extension ".insert_size_Histogram.pdf".

Field	Description
MEDIAN_INSERT_SIZE	The MEDIAN insert size of all paired end reads where both ends mapped to the same chromosome.
MEDIAN_ABSOLUTE_DEVIATION	The median absolute deviation of the distribution. If the distribution is essentially normal then the standard deviation can be estimated as ~1.4826 * MAD.
MIN_INSERT_SIZE	The minimum measured insert size. This is usually 1 and not very useful as it is likely artifactual.
MAX_INSERT_SIZE	The maximum measure insert size by alignment. This is usually very high representing either an artifact or possibly the presence of a structural re-arrangement.
MEAN_INSERT_SIZE	The mean insert size of the "core" of the distribution. Artefactual outliers in the distribution often cause calculation of nonsensical mean and stdev values. To avoid this the distribution is first trimmed to a "core" distribution of +/- N median absolute deviations around the median insert size. By default N=10, but this is configurable.
STANDARD_DEVIATION	Standard deviation of insert sizes over the "core" of the distribution.
READ_PAIRS	The total number of read pairs that were examined in the entire distribution.
PAIR_ORIENTATION	The pair orientation of the reads in this data category.
WIDTH_OF_10_PERCENT	The "width" of the bins, centered around the median, that encompass 10% of all read pairs.
WIDTH_OF_20_PERCENT	The "width" of the bins, centered around the median, that encompass 20% of all read pairs.
WIDTH_OF_30_PERCENT	The "width" of the bins, centered around the median, that encompass 30% of all read pairs.
WIDTH_OF_40_PERCENT	The "width" of the bins, centered around the median, that encompass 40% of all read pairs.
WIDTH_OF_50_PERCENT	The "width" of the bins, centered around the median, that encompass 50% of all read pairs.
WIDTH_OF_60_PERCENT	The "width" of the bins, centered around the median, that encompass 60% of all read pairs.
WIDTH_OF_70_PERCENT	The "width" of the bins, centered around the median, that encompass 70% of all read pairs. This metric divided by 2 should approximate the standard deviation when the insert size distribution is a normal distribution.
WIDTH_OF_80_PERCENT	The "width" of the bins, centered around the median, that encompass 80% of all read pairs.
WIDTH_OF_90_PERCENT	The "width" of the bins, centered around the median, that encompass 90% of all read pairs.
WIDTH_OF_99_PERCENT	The "width" of the bins, centered around the median, that encompass 100% of all read pairs.

AlignmentSummaryMetrics

High level metrics about the alignment of reads within a SAM file, produced by the CollectAlignmentSummaryMetrics program and usually stored in a file with the extension ".alignment_summary_metrics".

CATEGORY	One of either UNPAIRED (for a fragment run), FIRST_OF_PAIR when metrics are for only the first read in a paired run, SECOND_OF_PAIR when the metrics are for only the second read in a paired run or PAIR when the metrics are aggregated for both first and second reads in a pair.
TOTAL_READS	The total number of reads including all PF and non-PF reads. When CATEGORY equals PAIR this value will be 2x the number of clusters.
PF_READS	The number of PF reads where PF is defined as passing Illumina's filter.
PCT_PF_READS	The fraction of reads that are PF (PF_READS / TOTAL_READS)
PF_NOISE_READS	The number of PF reads that are marked as noise reads. A noise read is one which is composed entirely of A bases and/or N bases. These reads are marked as they are usually artifactual and are of no use in downstream analysis.
PF_READS_ALIGNED	The number of PF reads that were aligned to the reference sequence. This includes reads that aligned with low quality (i.e. their alignments are ambiguous).
PCT_PF_READS_ALIGNED	The percentage of PF reads that aligned to the reference sequence. PF_READS_ALIGNED / PF_READS
PF_ALIGNED_BASES	The total number of aligned bases, in all mapped PF reads, that are aligned to the reference sequence.
PF_HQ_ALIGNED_READS	The number of PF reads that were aligned to the reference sequence with a mapping quality of Q20 or higher signifying that the aligner estimates a 1/100 (or smaller) chance that the alignment is wrong.
PF_HQ_ALIGNED_BASES	The number of bases aligned to the reference sequence in reads that were mapped at high quality. Will usually approximate PF_HQ_ALIGNED_READS * READ_LENGTH but may differ when either mixed read lengths are present or many reads are aligned with gaps.
PF_HQ_ALIGNED_Q20_BASES	The subset of PF_HQ_ALIGNED_BASES where the base call quality was Q20 or higher.
PF_HQ_MEDIAN_MISMATCHES	The median number of mismatches versus the reference sequence in reads that were aligned to the reference at high quality (i.e. PF_HQ_ALIGNED READS).
PF_MISMATCH_RATE	The rate of bases mismatching the reference for all bases aligned to the reference sequence.
PF_HQ_ERROR_RATE	The fraction of bases that mismatch the reference in PF HQ aligned reads.
PF_INDEL_RATE	The number of insertion and deletion events per 100 aligned bases. Uses the number of events as the numerator, not the number of inserted or deleted bases.
MEAN_READ_LENGTH	The mean read length of the set of reads examined. When looking at the data for a single lane with equal length reads this number is just the read length. When looking at data for merged lanes with differing read lengths this is the mean read length of all reads.
READS_ALIGNED_IN_PAIRS	The number of aligned reads whose mate pair was also aligned to the reference.
PCT_READS_ALIGNED_IN_PAIRS	The fraction of reads whose mate pair was also aligned to the reference. READS_ALIGNED_IN_PAIRS / PF_READS_ALIGNED
PF_READS_IMPROPER_PAIRS	The number of (primary) aligned reads that are **not** "properly" aligned in pairs (as per SAM flag 0x2).
PCT_PF_READS_IMPROPER_PAIRS	The fraction of (primary) reads that are *not* "properly" aligned in pairs (as per SAM flag 0x2). PF_READS_IMPROPER_PAIRS / PF_READS_ALIGNED
BAD_CYCLES	The number of instrument cycles in which 80% or more of base calls were no-calls.
STRAND_BALANCE	The number of PF reads aligned to the positive strand of the genome divided by the number of PF reads aligned to the genome.
PCT_CHIMERAS	The fraction of reads that map outside of a maximum insert size (usually 100kb) or that have the two ends mapping to different chromosomes.
PCT_ADAPTER	The fraction of PF reads that are unaligned and match to a known adapter sequence right from the start of the read.

HsMetrics

http://broadinstitute.github.io/picard/picard-metric-definitions.html#HsMetrics

- Metrics generated by CollectHsMetrics for the analysis of target-capture sequencing experiments. The metrics in this class fall broadly into three categories:
  - Basic sequencing metrics that are either generated as a baseline against which to evaluate other metrics or because they are used in the calculation of other metrics.
    - This includes things like the genome size, the number of reads, the number of aligned reads etc.
  - Metrics that are intended for evaluating the performance of the wet-lab assay that generated the data.
    - This group includes metrics like the number of bases mapping on/off/near baits, %selected, fold 80 base penalty, hs library size and the hs penalty metrics. These metrics are calculated prior to some of the filters are applied (e.g. low mapping quality reads, low base quality bases and bases overlapping in the middle of paired-end reads are all counted).
  - Metrics for assessing target coverage as a proxy for how well the data is likely to perform in downstream applications like variant calling.
    - This group includes metrics like mean target coverage, the percentage of bases reaching various coverage levels, and the percentage of bases excluded by various filters. These metrics are computed using the strictest subset of the data, after all filters have been applied.

Field	Description
BAIT_SET	The name of the bait set used in the hybrid selection.
GENOME_SIZE	The number of bases in the reference genome used for alignment.
BAIT_TERRITORY	The number of bases which are localized to one or more baits.
TARGET_TERRITORY	The unique number of target bases in the experiment, where the target sequence is usually exons etc.
BAIT_DESIGN_EFFICIENCY	The ratio of TARGET_TERRITORY/BAIT_TERRITORY. A value of 1 indicates a perfect design efficiency, while a valud of 0.5 indicates that half of bases within the bait region are not within the target region.
TOTAL_READS	The total number of reads in the SAM or BAM file examined.
PF_READS	The total number of reads that pass the vendor's filter.
PF_UNIQUE_READS	The number of PF reads that are not marked as duplicates.
PCT_PF_READS	The fraction of reads passing the vendor's filter, PF_READS/TOTAL_READS.
PCT_PF_UQ_READS	The fraction of PF_UNIQUE_READS from the TOTAL_READS, PF_UNIQUE_READS/TOTAL_READS.
PF_UQ_READS_ALIGNED	The number of PF_UNIQUE_READS that aligned to the reference genome with a mapping score > 0.
PCT_PF_UQ_READS_ALIGNED	The fraction of PF_UQ_READS_ALIGNED from the total number of PF reads.
PF_BASES_ALIGNED	The number of PF unique bases that are aligned to the reference genome with mapping scores > 0.
PF_UQ_BASES_ALIGNED	The number of bases in the PF_UQ_READS_ALIGNED reads. Accounts for clipping and gaps.
ON_BAIT_BASES	The number of PF_BASES_ALIGNED that are mapped to the baited regions of the genome.
NEAR_BAIT_BASES	The number of PF_BASES_ALIGNED that are mapped to within a fixed interval containing a baited region, but not within the baited section per se.
OFF_BAIT_BASES	The number of PF_BASES_ALIGNED that are mapped away from any baited region.
ON_TARGET_BASES	The number of PF_BASES_ALIGNED that are mapped to a targeted region of the genome.
PCT_SELECTED_BASES	The fraction of PF_BASES_ALIGNED located on or near a baited region (ON_BAIT_BASES + NEAR_BAIT_BASES)/PF_BASES_ALIGNED.
PCT_OFF_BAIT	The fraction of PF_BASES_ALIGNED that are mapped away from any baited region, OFF_BAIT_BASES/PF_BASES_ALIGNED.
ON_BAIT_VS_SELECTED	The fraction of bases on or near baits that are covered by baits, ON_BAIT_BASES/(ON_BAIT_BASES + NEAR_BAIT_BASES).
MEAN_BAIT_COVERAGE	The mean coverage of all baits in the experiment.
MEAN_TARGET_COVERAGE	The mean coverage of a target region.
MEDIAN_TARGET_COVERAGE	The median coverage of a target region.
MAX_TARGET_COVERAGE	The maximum coverage of reads that mapped to target regions of an experiment.
PCT_USABLE_BASES_ON_BAIT	The number of aligned, de-duped, on-bait bases out of the PF bases available.
PCT_USABLE_BASES_ON_TARGET	The number of aligned, de-duped, on-target bases out of all of the PF bases available.
FOLD_ENRICHMENT	The fold by which the baited region has been amplified above genomic background.
ZERO_CVG_TARGETS_PCT	The fraction of targets that did not reach coverage=1 over any base.
PCT_EXC_DUPE	The fraction of aligned bases that were filtered out because they were in reads marked as duplicates.
PCT_EXC_MAPQ	The fraction of aligned bases that were filtered out because they were in reads with low mapping quality.
PCT_EXC_BASEQ	The fraction of aligned bases that were filtered out because they were of low base quality.
PCT_EXC_OVERLAP	The fraction of aligned bases that were filtered out because they were the second observation from an insert with overlapping reads.
PCT_EXC_OFF_TARGET	The fraction of aligned bases that were filtered out because they did not align over a target base.
FOLD_80_BASE_PENALTY	The fold over-coverage necessary to raise 80% of bases in "non-zero-cvg" targets to the mean coverage level in those targets.
PCT_TARGET_BASES_1X	The fraction of all target bases achieving 1X or greater coverage.
PCT_TARGET_BASES_2X	The fraction of all target bases achieving 2X or greater coverage.
PCT_TARGET_BASES_10X	The fraction of all target bases achieving 10X or greater coverage.
PCT_TARGET_BASES_20X	The fraction of all target bases achieving 20X or greater coverage.
PCT_TARGET_BASES_30X	The fraction of all target bases achieving 30X or greater coverage.
PCT_TARGET_BASES_40X	The fraction of all target bases achieving 40X or greater coverage.
PCT_TARGET_BASES_50X	The fraction of all target bases achieving 50X or greater coverage.
PCT_TARGET_BASES_100X	The fraction of all target bases achieving 100X or greater coverage.
HS_LIBRARY_SIZE	The estimated number of unique molecules in the selected part of the library.
HS_PENALTY_10X	The "hybrid selection penalty" incurred to get 80% of target bases to 10X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 10X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 10 * HS_PENALTY_10X.
HS_PENALTY_20X	The "hybrid selection penalty" incurred to get 80% of target bases to 20X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 20X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 20 * HS_PENALTY_20X.
HS_PENALTY_30X	The "hybrid selection penalty" incurred to get 80% of target bases to 30X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 30X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 30 * HS_PENALTY_30X.
HS_PENALTY_40X	The "hybrid selection penalty" incurred to get 80% of target bases to 40X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 40X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 40 * HS_PENALTY_40X.
HS_PENALTY_50X	The "hybrid selection penalty" incurred to get 80% of target bases to 50X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 50X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 50 * HS_PENALTY_50X.
HS_PENALTY_100X	The "hybrid selection penalty" incurred to get 80% of target bases to 100X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 100X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 100 * HS_PENALTY_100X.
AT_DROPOUT	A measure of how undercovered <= 50% GC regions are relative to the mean. For each GC bin [0..50] we calculate a = % of target territory, and b = % of aligned reads aligned to these targets. AT DROPOUT is then abs(sum(a-b when a-b < 0)). E.g. if the value is 5% this implies that 5% of total reads that should have mapped to GC<=50% regions mapped elsewhere.
GC_DROPOUT	A measure of how undercovered >= 50% GC regions are relative to the mean. For each GC bin [50..100] we calculate a = % of target territory, and b = % of aligned reads aligned to these targets. GC DROPOUT is then abs(sum(a-b when a-b < 0)). E.g. if the value is 5% this implies that 5% of total reads that should have mapped to GC>=50% regions mapped elsewhere.
HET_SNP_SENSITIVITY	The theoretical HET SNP sensitivity.
HET_SNP_Q	The Phred Scaled Q Score of the theoretical HET SNP sensitivity.