Standard Operating Procedures for Data Quality Control (QC)

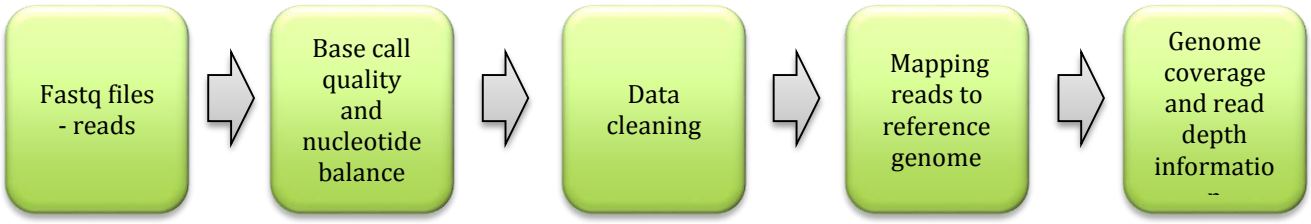
In this document we provide some recommended guidelines for the quality control of genomic data. Full Standard Operating Procedure (SOP) Guidelines can be found at: http://www.h3abionet.org/tools-and-resources.

Standard Operating Procedures for NGS QC: Targeted and Whole Genome Re-sequencing

The following describes steps and general considerations included in QC analysis of NGS data before any particular application (such as variant calling). These are not platform specific, although quality value and read length thresholds mentioned are for Illumina.

Goal: To ensure that there are enough unique mappable reads with high-quality base calls to cover genomic regions of interest (whole genome or in case of targeted sequencing e.g. exome). Depending on application, targeted average coverage can be 30-50X per individual, or in low coverage sequencing with many individuals, as low as 4X.

Workflow: Targeted/whole genome re-sequencing
Input: FASTQ data after demultiplexing (reads separated according to bar codes)
Output: 1) Trimmed reads in FASTQ format, 2) Read mapping in BAM format, 3) Genome coverage and read depth information, and 4) Summary of each QC step in **QC report**.
Data acceptance criterion: Is coverage/read depth sufficient in genomic regions of interest



Workflow	Process	Quality Gate	Quality Metric	Software
targeted/whole genome re-sequencing	Base call quality and nucleotide balance	Base quality Nucleotide balance	average Phred > 30 during first 50bp -	FASTQC or FASTX FASTQC or FASTX
	Data cleaning	Adapter removal Quality trimming Removal of short reads Removal of duplicates	- Phred<3 hard trimming from both 5' and 3' ends + sliding window keeping longest segment where average Phred>25 in all 4-base windows read length > 50bp -	Trimmomatic, FASTX, CLCbio Trimmomatic, FASTX, CLCbio Trimmomatic, FASTX, CLCbio FASTX, CLCbio
	Mapping reads to reference genome			BWA or Bowtie2
	Removal of duplicates (if not done before mapping)		-	Picard or Samtools
	Genome coverage and read depth information	Mapping quality Mode of read depth histogram	>20 >30X	in-house Perl script utilizing Samtools and R or GATK CallableLoci tool
	Visualization			Samtools, IGV or Gbrowse

Note: save all summary files in each step.

1. Procedure for checking base call quality and nucleotide balance

Histogram of base call quality values along read length reveals quickly evident sequencing run failures. Look quality value distribution as a function of read base position. If average quality drops early (below Phred score 30 during the first 50 bp of reads), it indicates problems with data/sequencing. Software FASTQC gives also amount of duplicated sequences and list of most abundant k-mers, the latter can be used to detect the presence of adapter sequences. FASTQC gives also nucleotide composition, A/C/G/T balance per read position, which ideally should be constant over entire read length. Fluctuations indicate bias, either problems in sequencing process or sequencing library. **Include in QC report**: FASTQC output files.

Note: even if average quality does not meet quality criteria, there can still be useful reads in the data set which can be combined with new data. Therefore, in order to find mappable reads, the remaining steps of the workflow should be still performed.

2. Procedure for data cleaning

i) adapter trimming

If DNA fragment is shorter read length, read will contain (beginning or full) adapter sequence in its 3' end. This might still be good-quality sequence, so quality based trimming does not remove it and if not removed, it will affect negatively to read mappability.

ii) low-quality base trimming

Low-quality bases are likely erroneous so trimming them improves mappability of reads. Usually read quality decreases towards the end of read. The task is to clip high-quality part of the read. Individual low-quality base calls can be accepted in the read in order not to clip reads too short.

Note1: it is better to trim each read individually based on individual base quality values rather than remove the same number of bases from all reads based on average quality of all reads.
Note2: there can be Ns (indicated by Phred QV 2 in Illumina data) or otherwise low-quality bases also in the beginning of read, which. should be trimmed

iii) minimum read length filtering

If read is too short after trimming, it might not map uniquely. Because of their limited use, short reads can be completely dropped from further analysis.

Software: steps i-iii can be performed in one command line using Trimmomatic or in multiple steps using (commercial) CLCbio software. Both of them allow quality based read trimming from both 5' and 3' ends of read. For paired-end read data, output files include paired-end reads where both reads passed filtering and separate file(s) where only one read of the pair passed filtering. Output files are in FASTQ format. **Include in QC report**: summary of data cleaning.

iv) duplicate read removing

Ideally there should be only one read from each original molecule in sequencing library, but due to PCR there can be several copies. Presence of large amount of duplicates indicates PCR bias or problems with library preparation (low quality/concentration of DNA). Removing duplicates can be done in two different ways: before mapping (FASTX tool fastx_collapser or CLCbio software) or after mapping reads onto genome (Samtools rmdup or Picard MarkDuplicates). **Include in QC report:** amount of duplicates.

3. Procedure for genome coverage and read depth checking

i) map (align) trimmed reads/pairs onto reference genome, current standard is hg19

Software e.g. BWA or Bowtie2, mapping output file is in BAM format. **Include in QC report**: number of mappable and unmappable reads.

ii) remove duplicate reads/pairs (if not done before read mapping)

Software based on read mapping (BAM file) e.g. Samtools rmdup, Picard MarkDuplicates. **Include in QC report:** amount of duplicates.

iii) calculate coverage histogram based on mapped nonduplicate reads

Software e.g. in-house Perl script utilizing Samtools mpileup and a visualization tool such as R. Read BAM file and based on read depths per position from targeted regions (exome or whole genome) create histogram, check its shape and whether the mode (not plain average) is where expected (e.g. 30X). **Include in QC report:** coverage histogram and read depth information.

Note1: include only uniquely mapped reads/pairs (mapping quality > threshold depending on the mapping software)
Note2: take into account both mapping quality and base call quality. See e.g. GATK CallableLoci tool.
Note3: histogram should include all targeted regions, not only those with mapped reads

Quality gate check point: data ok, if coverage criterion is met. Since coverage is measured with good-quality data without duplicates and it takes into account genomic regions of interest (e.g. exome), it indicates usefulness of data. If genomic coverage or read depth not sufficient, more data should be sequenced.

Note: if the amount of duplicates is large, it indicates that number of unique reads has already saturated and new sequencing library should be prepared.

If specific predefined SNPs/other genomic regions are of interest, coverage of those regions can be checked individually with Samtools or genome browser (e.g. IGV).

4. Further additional checks based on mapped data

i) check unmapped reads

Compare the number of mapped reads against number of unmapped reads. Unmapped reads can be due to true insertions or sequence artifacts. If proportion of unmapped reads is high, check if adapter trimming was successful. E.g. if remains of adapter in the read are so short that adapter trimming cannot detect them, it is possible to trim fixed number of bases (estimated length of remaining adapter sequence in the reads) from all unmapped reads.

ii) check insertsize distribution (if paired-end reads)

Calculate insertsize distribution based on mapped read pairs and compare against expected insertsize. Width of insertsize distribution reflects the specificity of size selection (ideally narrow peak). If empirical average insertsize is smaller than targeted, library preparation may require an additional size selection to remove short fragments (short fragments tend to become overemphasized in most sequencing platforms).

5. QC report

QC report should contain information of all steps along workflow:

- Description of data set
- Names and versions of software used and parameter settings
- Summaries of QC outputs in each step

6. Software

- Bowtie2 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- BWA http://bio-bwa.sourceforge.net/
- CLCbio http://www.clcbio.com
- FASTQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- FASTX http://hannonlab.cshl.edu/fastx_toolkit/
- GATK http://www.broadinstitute.org/gatk/
- IGV http://www.broadinstitute.org/igv/

- Perl http://www.perl.org/
- Picard http://picard.sourceforge.net/
- R http://www.r-project.org/
- Samtools http://samtools.sourceforge.net/
- Trimmomatic http://www.usadellab.org/cms/?page=trimmomatic

Quality control checks on Affymetrix (<= 6.0), Axiom and DMET Genome-Wide Human SNP Arrays

1. Introduction

Although SNPs data set from old genotyping arrays are useful, the most complete genotyping array to screen functional variants for disease relevance is currently Axiom Genome-Wide Human Arrays Plates, designed:

- To maximize genomic coverage of common, rare SNP, novel SNP and indels in Caucasian, Asian, and African populations (**Genome-Wide Population-Optimized Human Arrays**).

- To screen for putative-functional variants in protein coding regions of the human genome. When combined with genotype data from genome-wide arrays, this panel of rare non-synonymous coding SNPs, indel mutations, and other key GWAS markers is a powerful tool to identify causal variants in complex diseases (**Exome Genotyping Arrays**).

- To include known and predicted SNPs and indels within promoters, miRNA seed sites, precursor miRNA stem-loop regions, mRNA target binding sites, and miRNA processing proteins. Over 80% of the array content is not found on any other genotyping arrays, making it ideal for supplementing GWAS or gene regulation studies (**miRNA Target Site Genotyping Arrays**).

2. Purpose

To provide quality control guidelines to ensure that Affymetrix (<=6.0), Axiom and DMET Genome-Wide Human SNP Arrays data are fit for further analysis.

3. Consideration and special notes

SNP call-rate is the percentage of SNP markers for which a genotype call has successfully been made. Genotype calls can be AA, BB, or AB. Failure to make a call ('no call') is denoted by NC. QC call-rate is a measure of the precision between probe replicates for a given marker.

Median of the absolute values of all Pairwise Differences (MAPD) is an overall measure of the variability between paired log2 ratios for a given chip/sample. Markers and their respective log2 ratios are paired based on their genomic location.

Contrast QC (CQC) measures the ability of a given chip/sample to resolve SNP genotype calls into three clusters – i.e. AA, BB, or AB – based on the performance of 10,000 random SNP markers.

Dish QC is a metric that evaluates the overlap between the two homozygous peaks (AT versus GC) using normalized intensities of control non-polymorphic probes from both channels.

4. Inputs

Affymetrix <= 6.0, Axiom and DMET Plus Array: **CEL files (1 per chip/sample).**

5. Output

Affymetrix <= 6.0, Axiom and DMET Plus Array**: genotypes data set**

6. Software

BRLMM in Affymetrix Power Tools (APT) for making genotype calls from mapping arrays.
The Affymetrix Genotyping Console software (GTC) creates genotype calls for collections of CEL files.

7. Details of genotyping and SNP QC procedures

CEL files	Normalization of raw data	Identify novel varients between control/case	Any Applications eg. GWAS	Replication
	Annotate SNPs/marker	Call Rate cutoff		Fine mapping Variant discovery

1. Low-intensity markers: removed markers with the mean normalized intensity (R) for the heterozygote cluster(AB) <= 0.25.
2. Heterozygote clusters shifted too close to a homozygote cluster: removed markers with the mean normalized theta value (T) for the heterozygote cluster (AB T mean) < 0.2 or ≥ 0.8.
3. Visually inspected the plots for markers on the boundaries (0.2 ≤ AB T mean ≤ 0.25; 0.75 ≤ AB T mean < 0.8) and determined appropriate cutoff.
4. Poor separation on theta axis: removed markers with cluster separation < 0.3.
5. Visually reviewed markers with cluster separation between 0.3-0.32; removed these as appropriate.
6. More heterozygotes called than expected under HWE: removed markers with het excess ≥ 0.2.
7. Fewer heterozygotes called than expected under HWE: visually inspected a subset of intensity plots for polymorphic markers with het excess ≤ -0.3.

8. Replication errors: removed any remaining SNPs with >3 replication errors. Call rate: removed all remaining markers with call rate < 98%.

8. Final Protocol

After running QC for a given CEL file, the corresponding default threshold values determine the samples that are in or out of bounds:

1. SNP call rate ≥ 97%
2. QC call rate ≥ 86%
3. Median Absolute Pairwise Difference (*MAPD*) ≤ 0.35
4. Contrast QC (CQC) ≥ 0.4
5. Dish QC (DQC) ≥ 0.82

These are recommended QC metric for the AxiomTM Genome-Wide Human Array.

Where a sample is out of bounds for one or more QC parameters, a consensus decision on whether to include or exclude the sample in end analysis should be made between the bioinformatician performing the analysis and the laboratory personnel who provided the DNA sample. If multiple samples are out of bands for one or more QC parameters, then the QC thresholds can be modified accordingly.