Standard operating procedure for genotype calling of raw Affymetrix SNP 6 microarray data

Introduction

This document describes the background and recommended practises for genotype calling and Quality Control of Affymetrix SNP6 data from the raw CEL file stage. Various packages exist for processing this sort of data, hence the software used here (Affymetrix Power Tools) is only one example of a possible workflow.

Most of the steps were adapted from the following documents:

- APT MANUAL: apt-probeset-genotype (1.14.3)[1]

Experimental design

Confirm that the experimental design included steps to control for effects such as:

- case/control imbalance across plates
- case/controls not done in batches
- known genotypes included if larger sample numbers
- power
- population structure (will be checked post-genotyping anyway)

(see section I-A in whitepaper)

Genotyping

1. Sample Quality Control:
   Run **apt-geno-qc** to get QCR and CQC scores for each CEL file. Remove the cell files with QCR < 0.86 and CQC < 0.4
2. First pass genotyping
   Use **apt-probeset-genotype** to do an initial run of genotype calling
3. Remove samples with call rate < 97%
4. Run **apt-probeset-genotype** again on only these samples

(see section I-B in whitepaper)

Convert to PLINK format

The output of the APT calling procedure is a text file in “calls” format. This consists of a matrix of values (either -9,0,1,2) arranged in markers by row, samples by columns format. It needs to be converted to a format that can be read by PLINK to perform subsequent analysis. This entails:

1. Reading the text file of array annotations provided by Affymetrix

[1] http://www.affymetrix.com/support/developer/powertools/changelog/apt-probeset-genotype.html#quickstartbirdseed

2. Replacing the Affymetrix probe ID with the RSID of the marker
3. Replacing the -9/0/1/2 value with the appropriate genotype for that marker

Care should be taken when converting calls to their genotype values, specifically when the probe is on the “reverse” strand of the genome relative to the strand usually reported. See the README packaged with the Affymetrix CSV annotation file. Converted data should also be checked against public releases of genotype data, using the plink concordance test.

Batch/Plate QC

1. Remove plates abnormally low average DQCs. An example would be to flag plates with upper 25th percentile lower than median of median DQC for all other plates. See Affy document
2. Filter according to post-genotyping performance metrics
    - sample pass rate > 95%
    - average SNP call rate of passing samples > 99%
    - 95% of SNPs should have call rate > 97% (of passing samples)
4. Check concordance of replicates/known genotypes
    - If there are control samples included (eg HapMap samples), check the concordance of the called genotypes with the known genotypes of these samples
    - If there are duplicates, check the agreement between them. Check that gender call matches
5. Check for sample contamination
    - Plot DQC vs Sample Call Rate for each plate
6. Check for plate-wise MAF differences
    - Perform a chi-squared analysis of MAF per SNP between plates. These should not differ systematically

(see section I-C in whitepaper)

Sample QC

1. Check for sample mixup
    - Check against “Signature SNPs”
2. Check for contaminated samples
    - High autosomal heterozygosity
    - DQC vs sample call rate
    - evidence of allelic imbalance
    - Unexpected patterns of relatedness
3. Check called gender against recorded gender

(see section I-D in whitepaper)

SNP QC

1. SNP call rate
    - Remove SNPs with call rate <95% (beware of not filtering Y SNPs)
2. FLD
    - Remove SNPs with FLD < 3.6

3. Heterozygous strength offset
    - Remove SNPs with low HetSO
4. Other filters
    - Hardy Weinberg p-value
    - Mendelian trio errors
    - Reproducibility between replicates
    - Fixed alleles

(see section I-E in whitepaper)