PERLEGEN
SCIENCES

Individual Genotype Dataset Description

- Cleaned data -

Overview

The data set consists of six files, including this document. All data files are plain text with two or more tab separated columns. The first row of each data file consists of column headers.

SNP Information

The file SNP_INFO.TXT was sent separately on 12-19-05, and is not included in this transfer.

SNPs not included in this data transfer

The file SNPS_NOT_GENOTYPED contains 8,781 SNPs that were not transferred with this data set.

Column Name	*Description*
SNP_ID	Perlegen internal SNP identifier.

Reasons why genotypes were not transferred for a given SNP include:

1. The call rate was <80%.
2. QC criteria related to X and Y SNP performance.
3. Monomorphic SNPs.
4. HWE p-value < 1e-15 in either cases or controls.
5. SNPs with HWE p-value between 1e-15 and 1e-4 that were visually inspected and where problems with the clustering were detected.
6. 3 SNPs that cross-hyb with Y and are fixed difference (males P ~ 0.5 and females P ~ 1)

Samples not included in this data transfer

The file SAMPLES_NOT_GENOTYPED contains 79 samples that were not transferred with this data set.

Column Name	*Description*
Sample_ID	Sample Identifier from customer manifest.
Reason	Numeric code for reason

Reasons why samples were not genotyped include:

1. The call rate was <80% (3).
2. The sample identity was questionable (3).
3. Possible contamination with other sample(s) (8).
4. Insufficient DNA received (35).
5. Samples labeled as controls with ftnd > 0 (3).

6. No phenotype information (1).
7. Discrepant phenotype information (1).
8. The worse data for each of the duplicated sample pairs (2).
9. Discordant monozygotic twins (2).
10. Suspect phenotype information (2).
11. Unresolved gender discrepancy, suspect phenotype information (5).
12. Samples with low heterozygocity on X and low call rate on Y; they have autosomal heterozygocity in the normal range. These may be XO females, but represent an unusually high proportion of the subjects, so left out for potential data quality issues (8).
13. Unresolved FTND score discrepancies (6).

Individual Genotype Data

The file GENOTYPE.TXT contains individual genotypes for 35,673 SNPs in a compact representation. Each genotype is represented by two characters, representing nucleotide codes for the alleles in that individual, separated by tabs. The header row is a tab-delimited list of the sample ID's.

Column Name	*Description*
SNP_ID	Perlegen internal SNP identifier.
Genotype	A tab delimited list of 1,929 2 character individual genotypes. The two characters are the nucleotide codes for the two alleles, or 'NN' for genotype not determined. Genotypes for male samples at sex-linked loci are reported with homozygous diploid codes (e.g. A male hemizygous for the 'G' allele of an X-linked SNP has the reported genotype 'GG'), and females are reported as 'NN' on Y-linked SNPs.

Note: No quality score (QS) cutoff was used in reporting the genotypes. A QS cutoff of 7 should be used to eliminate unreliable genotypes.

Quality Score Data

The file QUALITY.TXT contains individual quality scores for 35,673 SNPs in a compact representation. The header row is a tab-delimited list of the sample ID's.

The quality scores are calculated as $[-10 \log_{10} (p)]$, where p is an estimated probability of a discordant genotype with another platform. A value of 20 indicates that about 1 such genotype out of 100 would show a discordance if genotyped on another platform.

Column Name	*Description*
SNP_ID	Perlegen internal SNP identifier.
QS	A tab delimited list of 1,929 2 digit individual genotype quality scores, with "00" for 'NN' genotype calls.

Copyrights

The file COPYRIGHTNOTICE_V3.TXT is a document describing the copyrights covering all of the files in this directory.

Remaining sample phenotype issues:

The following switches have been made to the sample genders:

04NA14755	Switch gender	1
03NA08538	Switch gender	1
All site 4 samples	Switch gender	2

1) Resolved gender discrepancy. The gender will be switched from originally reported to genetically determined gender for the analyses.
2) Systematically change genders for all site 4 samples.