Pooled Genotyping Dataset Description

Overview

The data set consists of seven files. All data files are plain text with tab separated columns. The first row of each data file consists of column headers.

SNP Information

The file SNP_INFO.TXT has the following information for a total of 2,427,354 SNPs.

Column Name	*Description*
SNP_ID	Perlegen internal SNP identifier.
dbSNP_rsID	The dbSNP RefSNP cluster ID (from NCBI) when available. Can be null.
dbSNP_ssID	The dbSNP submission ID for SNPs Perlegen submitted to dbSNP. Can be null.
Allele 1	The nucleotide code for allele 1.
Allele 2	The nucleotide code for allele 2.
Chromosome	Chromosome number of the NCBI Build 35 contig on which the best alignment was found. X is used for the X chromosome, Y for the Y chromosome.
Sex-linked	A, autosomal; P, pseudoautosomal (on X or Y, in the pseudoautosomal region); S, sex linked (on X or Y, not in the pseudoautosomal region); U, unknown (where the mapping is either to a contig without a chromosomal association or to a contig associated with X or Y but whose position on the X or Y assembly is unknown).
Accession ID	The accession number from NCBI Build 35 of the contig to which the SNP aligns.
Contig Position	Nucleotide position in NCBI build 35 contig of the reference base in the alignment; may be null.
Strand	+ or -, based on the strand for the reported alleles on NCBI Build 35.
Assayed sequence	The 29mer assayed for this SNP, with an ambiguity character representing the SNP at the middle base.

Pooled Genotype Data

The file RESULTS.TXT contains the results of Perlegen's SNP assays applied to the pooled samples. Records are included for all assay results. Each SNP has exactly 16 measurements. Total rows: 38,837,665 plus one header row, corresponding to 2,427,354 SNPs.

Column Name	*Description*
SNP_ID	Perlegen internal SNP identifier.
SAMPLE_NAME	The pool designator.
SCAN_CODE	Perlegen's internal scan code, including an 8 character internal identifier and a timestamp.
PHAT	p-hat value or ND in cases where data quality was insufficient.[1]

Notes

[1] SNP measurements are designated ND if the measurement had any of the following:

A. A conformance of <0.9
   a. Could indicate absence of target.
B. Saturated probes.
   a. p-hat is unreliable if some probes are saturated.
C. Signal to Background ratio of <1.5
   a. p-hat is unreliable with low signal to background.

SNPs not included in SNP_SUMMARY.TXT

The file SNPS_NOT_GENOTYPED contains 249,636 SNPs that were not included in SNP_SUMMARY.txt file because the number of sub-pools for either cases or controls that passed our QC criteria was less than 2 (see Pooled Genotyping Results, above for a description of the QC criteria).

Column Name	*Description*
SNP_ID	Perlegen internal SNP identifier.

Summary of pooled information for each SNP.

SNP_SUMMARY.TXT contains the following information for a total of 2,177,718 SNPs for which at least two case and two control pools had passing p-hat values.

Column Name	*Description*
SNP_ID	Perlegen internal SNP identifier. This matches the SNP_ID key in the other files.
NUM_PASSED_CASES	The number of measurements that passed QC criteria (i.e. are not ND in results.txt) for the Case pool.
NUM_PASSED_CONTROLS	The number of measurements that passed QC criteria (i.e. are not ND in results.txt) for the Control pool.
AVG_PHAT_CASES	The average of the PHAT measurements for the Case pool.
AVG_PHAT_CONTROLS	The average of the PHAT measurements for the Control pool.
DELTA_PHAT	The estimated difference in allele frequency between the pools, expressed as (AVG_PHAT_CASES - AVG_PHAT_CONTROLS).
STD_ERROR	The standard error of the estimate of DELTA_PHAT.
STD_ERROR_CORRECTION	The standard error correction used in the modified t-test. The standard error correction was obtained separately for each chip design by minimizing the coefficient of variation standard_deviation(t-test)/mean(t-test), where t-tests were computed for each SNP as DELTA_PHAT/(STD_ERROR+STD_ERROR_CORRECTION).
MODIFIED_T_TEST_P_VALUE	p-value obtained from a t-statistic DELTA_PHAT/(STD_ERROR+STD_ERROR_CORRECTION) with NUM_PASSED_CASES+NUM_PASSED_CONTROLS-2 degrees of freedom.
EMPIRICAL_P_VALUE	Obtained as rank of MODIFIED_T_TEST_P_VALUE on each chip design divided by the total number of passing SNP measurements for each chip design.
LD_BIN_ID	Perlegen internal unique LD bin identifier. Only populated if LD bin has >1 passing SNP. This matches LD_BIN_ID in the file LD_BIN_SUMMARY.txt.
PASSING	1 when the snp passed our internal QC criteria in an independent individual genotyping study (Science paper) using this set of chips, 0 otherwise.

Summary of LD bin information.

LD_BIN_SUMMARY.TXT contains the following information for a total of 139,046 LD bins for which at least two SNPs have passed our QC criteria.

Column Name	*Description*
LD_BIN_ID	Perlegen internal unique LD bin identifier. This matches LD_BIN_ID in the file SNP_SUMMARY.txt.
BEST_SNP_ID	SNP_ID of a SNP in the LD block that received the lowest EMPIRICAL_P_VALUE.
BEST_EMPIRICAL_P_VALUE	EMPIRICAL_P_VALUE of the snp identified by the BEST_SNP_ID.
NUM_SNPS_IN_LD_BIN	Number of snps that passed our QC criteria in the LD bin.
T_TEST_P_VALUE	T-test p-value computed for the LD bin that measures the level of correlation among the DELTA_PHAT values in the LD bin. The t statistic was obtained from passing SNPs in the LD bin as the average(DELTA_PHAT)/SEM, where the standard error of mean SEM was obtained as standard_deviation(DELTA_PHAT)/sqrt(NUM_SNPS_IN_LD_BIN).

Copyrights

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