NIDA - Individual Genotype Dataset Description

Overview

The data set consists of three files. The data file is plain text with tab separated columns. The first row of the data file consists of column headers.

SNP Information

The file SNP_INFO.TXT has the following information for a total of 44,454 SNPs.

Column Name	*Description*
SNP_ID	Perlegen internal SNP identifier.
Refsnp_rsID	The RefSnp rsID (from NCBI) when available. Can be null.
RefSnp_ssID	The RefSnp ssID for SNPs Perlegen submitted to dbSNP. Can be null.
Allele 1	The nucleotide code for Perlegen's reference allele.
Allele 2	The nucleotide code for Perlegen's alternate allele.
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 and U for sequences not assigned to any chromosome in Build 35. May be null if SNP could not be placed on any contig.
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, unassigned (or unknown pseudoautosomal status for X and Y).
Accession ID	The accession number from NCBI Build 35 of the contig to which the SNP aligns; may be null.
Contig Position	Nucleotide position on the NCBI build 35 contig of the 'N' in the assayed sequence; may be null. For SNPs this is always a single position, but in the case of a deletion-insertion polymorphism (DIP), the mapping may be a range. See the table below describing the full range of position notations.
Strand	+ or -, based on the strand for Allele 1 on NCBI Build 35; may be null.
Assayed sequence	The 29mer we assayed for this SNP, with an 'N' representing the SNP or deletion-insertion polymorphism (DIP) at the middle base.

Genes	A comma-separated list of NCBI gene IDs for which the SNP is contained within a footprint extending +/- 10 kb of the largest transcribed sequence.
Selected candidate genes	A comma-separated list of NCBI gene IDs that provided the basis for selecting the SNP. The physical footprint of the gene was extended by use of linkage equilibrium data.
Flags	A comma delimited list indicating selection categories and other information, as described in the table below.

Position notations:

Contig Position	*Definition*
12345	The N in the assayed sequence aligns to the single base 12345 on the genomic contig.
12345^12346	The N in the assayed sequence maps to the genomic contig between bases 12345 and 12346 (i.e. the N is an insertion with respect to the reference assembly)
12345..12346	The N in the assayed sequence maps to the range of genomic bases beginning with 12345 and ending with 12346.

Flags:

Flags	*Definition*
CAND_GENE	The SNP was selected because of its association with a candidate gene.
DIP	The variation is a deletion-insertion polymorphism (DIP) rather than a SNP.
CDS	The SNP falls within +/- 10bp of the CDS portion of an exon in a candidate gene.
LD	The SNP was selected because it was found to be in linkage disequilibrium with SNPs in the footprint of a candidate gene for European Americans.
LD_EXONIC	The SNP was selected because it was found to be in linkage disequilibrium with SNPs in an exon of a candidate gene for European Americans.
NIDA_CHOSEN	The SNP was specifically indicated by NIDA for assay.
POOLED	The SNP was assayed in the previous pooled study.
PRLGN_CHOSEN	The SNP was selected by Perlegen.
PROMOTER	The SNP was selected because it was found in the promoter region defined for a candidate gene.
STRAT	The SNP is a Perlegen population stratification analysis SNP.
UTR	The SNP is within +/- 10bp of the UTR portion of an exon in a candidate gene.

WK	The SNP is known to genotype well on the Perlegen platform and is polymorphic in some tested population.
WK_FREQ	The SNP is known to genotype well on the Perlegen platform and is polymorphic in some tested population with MAF >4% in European Americans.