Mayo - PGRN Gemcitabine Study Data

Genotype Data: Illumina 550K snp data

Genotype data has been provided on all snps (prior to quality control snp removal) and is split up by chromosome formatted as PED and MAP files.

PED files - white-space (space) delimited file: the first six columns are mandatory:

```
     Family ID  (Copied the Individual ID since no Family information)
     Individual ID  (Note: The Individual ID identifies their race)
     Paternal ID  (All zero for these cell lines)
     Maternal ID  (All zero for these cell lines)
     Sex (1=male; 2=female; other=unknown)
     Phenotype  (All -9 for these cell lines)
```

Genotypes (column 7 onwards) are also white-space delimited; they are character (A,C,G,T) except 0 which is, by default, the missing genotype character for PLINK. For Example, here are 2 individuals typed for 3 SNPs (one row/person):

```
        AA01  AA01  0 0  1  -9  A A  G G  A C
        AA02  AA02  0 0  1  -9  A A  A G  0 0
        ...
```

Note that since the first 6 columns are standard in a PED file we include dummy columns in order to follow the format. The value -9 was entered for the phenotype and indicates all missing values since the standard missing value in PLINK is -9 and we do not have a case control phenotype for these cell lines.

MAP files - space delimited file with the following 4 columns:

```
     chromosome (1-22, X, Y or XY)
     rs# or snp identifier
     Genetic distance (All zero for these cell lines)
     Base-pair position (bp units)
```

snps_removed.txt – A text file containing one column of RS# ID's indicating snps that we removed based on our quality control analysis (SNPs included in analysis of data that has been published). Quality control was completed by excluding SNPs with Hardy-Weinberg Equilibrium (HWE) p-values < 0.001 (minimum p-value between exact test for HWE [Guo and Thompson 1992; Wigginton, et al. 2005] and stratified test for HWE [Schaid, et al. 2006]), minor allele frequency < 5%, or call rate < 95% from further analyses.

Phenotype data:

gemc_raw_data.csv – contains the cytotoxicity data for gemcitabine for 8 dosage levels.

```
     Individual ID
     Dose  (Dosage level of cytotoxicity experiment)
     Value
```

gemc_logistic_curvefit_phenos.csv

```
     Individual ID
     AUC  (As calculated by our curvefit algorithm)
     GI50  (AKA IC50, as calculated by our curvefit algorithm)
```

Expression Data: Affymetrix U133 Plus 2.0 54K expression data
LCL_exp_adj.csv – contains expression data that has been processed with GCRMA to obtain normalized probe-level intensity measurements. They expression values have been further adjusted for race, gender, and batch effects.

```
     ID (Expression probe ID)
     Annotation Date
     Representative Public ID
     Genome Version
     Alignments
     Gene Title
     Gene Symbol
     Chromosomal Location
     Entrez Gene
     OMIM
     RefSeq Protein ID
     RefSeq Transcript ID
```

LCL_exp_unadj.csv - contains Affymetrix U133Plus2.0 expression data that has been processed with GCRMA to obtain normalized probe-level intensity measurements. No adjustments for race, gender, or batch effects were made.

```
     ID (Expression probe ID)
     Annotation Date
     Representative Public ID
     Genome Version
     Alignments
     Gene Title
     Gene Symbol
     Chromosomal Location
     Entrez Gene
     OMIM
     RefSeq Protein ID
     RefSeq Transcript ID
```