Genetic and Nongenetic Variation Revealed for the Principal Components of Human Gene Expression

Anita Goldinger,*,†,[1] Anjali K. Henders,‡ Allan F. McRae,*,†,‡ Nicholas G. Martin,‡ Greg Gibson,§ Grant W. Montgomery,‡ Peter M. Visscher,*,† and Joseph E. Powell*,†

*University of Queensland Diamantina Institute, The Translational Research Institute, Brisbane, Queensland 4102, Australia, †The Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia, ‡Queensland Institute of Medical Research, Herston, Brisbane, Queensland 4006, Australia, and §School of Biology and Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, Georgia 30332

ABSTRACT Principal components analysis has been employed in gene expression studies to correct for population substructure and batch and environmental effects. This method typically involves the removal of variation contained in as many as 50 principal components (PCs), which can constitute a large proportion of total variation present in the data. Each PC, however, can detect many sources of variation, including gene expression networks and genetic variation influencing transcript levels. We demonstrate that PCs generated from gene expression data can simultaneously contain both genetic and nongenetic factors. From heritability estimates we show that all PCs contain a considerable portion of genetic variation while nongenetic artifacts such as batch effects were associated to varying degrees with the first 60 PCs. These PCs demonstrate an enrichment of biological pathways, including core immune function and metabolic pathways. The use of PC correction in two independent data sets resulted in a reduction in the number of *cis*- and *trans*-expression QTL detected. Comparisons of PC and linear model correction revealed that PC correction was not as efficient at removing known batch effects and had a higher penalty on genetic variation. Therefore, this study highlights the danger of eliminating biologically relevant data when employing PC correction in gene expression data.

GENE expression profiling has become a very popular technique used to quantify regulatory changes in messenger (m)RNA expression associated with disease and environmental factors. Gene expression acts as an intermediate phenotype between genotypes and complex traits and is known to act as a modifier to disease susceptibility (Nica and Dermitzakis 2008; Li *et al.* 2012). Genetic variation underlying gene expression levels has been well established and reported within the literature, with the transcript levels for the majority of genes being heritable to some degree (Price *et al.* 2011; Grundberg *et al.* 2012; Powell *et al.* 2012b).

Microarray technology can simultaneously capture the expression of thousands of transcripts within an individual. However, these arrays are sensitive to environmental or experimental perturbations, for example due to different laboratory technicians and reagents (Churchill 2002; Irizarry *et al.* 2005), microarray chip and chip position (Luo *et al.* 2010), temperature (Scherer 2009), and even ozone levels (Thomas *et al.* 2003). These effects can constitute a substantial proportion of variance within a data set (Leek *et al.* 2010).

Normalization strategies have become standard in gene expression studies to correct for nonnormal distributions and inconsistencies between arrays (Allison *et al.* 2006). However, normalization techniques do not control for batch effects caused by technical artifacts. These batch effects require additional correction techniques (Scherer 2009) and failure to do so has led to spurious associations (Spielman and Cheung 2007; Baggerly *et al.* 2008).

Many different correction and normalization techniques are currently used in gene expression studies (for review see Chen *et al.* 2011; Qin *et al.* 2012). Principal components analysis (PCA) is one method that has been used for the correction of widespread batch effects (Leek and Storey

doi: 10.1534/genetics.113.153221
Manuscript received June 1, 2013; accepted for publication August 26, 2013
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.153221/-/DC1.
[1]Corresponding author: University of Queensland Diamantina Institute, Translational Research Institute, Brisbane 4102, Australia. Email: a.goldinger@uq.edu.au

2007; Pickrell *et al.* 2010; Fehrmann *et al.* 2011; Qin *et al.* 2012). PCA determines linear combinations of variables and projects them into orthogonal vectors that are ranked on variance explained (Jackson and Wiley 1991). PCA is often used for dimensionality reduction (Holter *et al.* 2000), identifying gene correlations (Fukushima *et al.* 2008); determining coexpressed networks (Hirai *et al.* 2007); classifying gene expression modules (Fukushima *et al.* 2008); analyzing time-series data (Raychaudhuri *et al.* 2000); and determining differentially expressed genes in different tissues (Misra *et al.* 2002), environments, and various conditions (Ma and Kosorok 2009).

Clearly PCA is a powerful tool to analyze and understand high-dimension gene expression data. However, while PCA is commonly used to decompose variation into common axes, the components of variation contributing to each principal component (PC) are often unknown. Information on the components of variation captured by PCs is important if the correct inferences and interpretation of PCA results are to be made. A clear example is in the use of PCs to correct for batch effects, where variation in the first 1–50 PCs is removed from the data set by fitting these PCs as covariates in a linear model and using the residuals as a corrected phenotype (Leek and Storey 2007; Stegle *et al.* 2008; Leek *et al.* 2010; Fehrmann *et al.* 2011; Fu *et al.* 2012; Qin *et al.* 2012; Stranger *et al.* 2012). Although this method has been demonstrated to increase the power to detect eQTL by eliminating confounding batch and environmental effects (Fehrmann *et al.* 2011), there is a risk of the indirect removal of biologically relevant information due to the large amount of variance held in the first few PCs (Qin *et al.* 2012).

In this study we investigated the sources of variation driving PCs generated on a gene expression data set produced from whole blood. This data set is composed of 860 individuals from 314 families from the Brisbane Systems Genetics Study (BSGS) (Powell *et al.* 2012a). We combine the power of pedigree and SNP-based designs to quantify and dissect the total genetic variance of PCs. From these results we demonstrate that the PC correction methods used in published literature (for example, in Fehrmann *et al.* 2011; Fu *et al.* 2012) simultaneously remove both genetic variation and batch effects. We also show that with careful analysis, the information contained within PCs can be leveraged to provide an understanding of gene expression pathways underlying complex disease. We use these results to emphasize the importance of using other correction methods when applicable instead of removing PCs to correct for batch effects.

Materials and Methods

Samples

A total of 335 unrelated individuals were selected from the BSGS, a cohort comprising 860 individuals from 314 families (Powell *et al.* 2012a). Individuals were selected on two criteria: the first was to obtain the maximum number of unrelated individuals by selecting parents ($n = 165$) and a single individual from families with no parents ($n = 170$). Unrelated individuals were selected from this data set to exclude family effects from being present within the PCs. All individuals were genotyped on an Illumina 610-Quad Beadchip (Illumina, San Diego) by the Scientific Services Division at deCODE Genetics, Reykjavik, Iceland. A total of 488,462 SNPs were present after appropriate quality control (see *Genome-wide association study on PCs* below).

RNA was collected from whole-blood samples that were collected in a PAXgene tube (QIAGEN, Valencia, CA), analyzed for purity on an Agilent Bioanalyzer, converted to cDNA, amplified, and purified using the Ambion Illumina TotalPrep RNA Amplification Kit (Ambion). The expression levels were quantified on an Illumina HumanHT-12 v4.0 Beadchip. Samples were randomized across the chip to minimize batch effects due to families, sex, and generation. Quality control methods for selecting highly expressed probes in the samples are described in detail in Powell *et al.* (2012a). After appropriate quality control, the probes were further filtered for expression in 100% of samples. Probe names starting with HS, KIAA, and LOC were removed from the data set, as they did not map to characterized ref-seq genes. After filtering, a total of 9086 probes remained.

Replication sample

The Center for Health Discovery and Well Being (CDHWB) study is a population-based cohort consisting of 139 individuals collected in Atlanta (Nath *et al.* 2012). Gene expression profiles were generated with Illumina HT-12 V3.0 arrays from whole blood collected with Tempus tubes that preserve RNA. Whole-genome genotypes were measured using Illumina Omni Quad arrays.

Normalization and batch effect correction

The gene expression levels of the 9086 probes were normalized using the quantile, log2, and z-score transformation. Correction methods were then applied to create four different data sets that were used for further analysis. These are (a) standard normalization (log, quantile, and z-score transformation), (b) standard normalization with linear model correction for batch effects, (c) standard normalization with correction for 1–25 PCs, and (d) standard normalization with correction for 1–50 PCs. These data sets/correction procedures are referred throughout this article as data sets a–d.

Batch effects arise from various sources during the generation of the data and processing date records can provide a useful estimate of differences between subsets of groups (Qin *et al.* 2012). In previous work (Powell *et al.* 2012a), we showed that chip identification (ID) and chip

position comprehensively account for batch effects. Data set b was created using the residuals from a linear model with the batch effects of chip ID and chip position as well as sex and age as covariates. This model was fitted as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

is a vector of probe values from data set a with $n = 335$ individual values.

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & \cdots & x_{1c} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_{nc} \end{pmatrix}$$

is a matrix with $c = 86$ covariates: chip position (11 levels), chip ID (73 levels), and sex and generation, which were coded as dichotomous variables.

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \beta_1 \\ \vdots \\ \beta_c \end{pmatrix}$$

is a vector of parameters for the model. The parameter μ represents the mean expression level across all the individuals. The vector

$$\mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

holds the model residuals. The parameters in $\boldsymbol{\beta}$ are estimated by

$$\boldsymbol{\beta} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}. \tag{2}$$

Data sets c and d were created from the residuals obtained after correcting for the first 25 and 50 PCs using Equation 1, where $\mathbf{y}$ was the normalized data set (a) and the PC score values (Equation 3, see below) obtained from data set a were fitted as covariates in $\mathbf{X}$. The first 50 PCs were selected to follow the procedure used in previously published gene expression articles (Fehrmann *et al.* 2011; Fu *et al.* 2012).

Principal components

PCs were calculated on data set a to facilitate PC correction, which was used to generate data sets c and d. PCs were also calculated on all four data sets a–d to analyze the extent of batch effect associations. To follow the methods used in previous gene expression studies (Leek and Storey 2007; Fehrmann *et al.* 2011; Fu *et al.* 2012), the principal components used in this study were generated using singular value decomposition (SVD). SVD decomposes the high-dimensional data set

$$\mathbf{M} = \begin{pmatrix} m_{11} & \cdots & m_{1p} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{np} \end{pmatrix},$$

with $n = 335$ (samples) and $p = 9{,}085$ (probes), into a set of uncorrelated, orthogonal vectors

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathbf{T}}, \tag{3}$$

where

$$\mathbf{U} = \begin{pmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nn} \end{pmatrix}$$

is a matrix with score values for each PC (columns), which gives the correlation values between the samples (rows).

$$\boldsymbol{\Sigma} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

is a diagonal matrix containing the eigenvalues, which represent the amount of variance each PC explains.

$$\mathbf{V} = \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{p1} & \cdots & v_{pn} \end{pmatrix}$$

are the eigenvectors that hold the correlation values for each probe (rows) to the PC (columns) (Golub and Reinsch 1970).

Data set a demonstrates a homogenous variance structure within the score plots (Supporting Information, Figure S1) with no clear population stratification or substructure comprising independent clusters of individuals. As this data set contains the same information as b–d, this indicates the uniform nature of the samples within the BSGS data set.

The eigenvector values in $\mathbf{V}$ (Equation 3) represent the correlation between the probes and the principal component. As multiple probes are correlated with each PC, the selection was based on the eigenvector values for each probe. The minimum eigenvector demonstrates the largest negative correction, while the maximum eigenvector demonstrates the largest positive correlation. Across all PCs the maximum eigenvector values ranged between 0.02 and 0.08 and the minimum values ranged between -0.02 and -0.08 (Figure S2A). The PCs that demonstrated the smallest maximum and minimum values were the first 1–20 PCs. As the number of probes selected for each PC can affect later enrichment analysis, we wanted to select a consistent number of probes between all PCs. Therefore a cutoff eigenvector value was used to select probes as this produced the most consistent results and provided a standard approach to use for all PCs. This cutoff value, selected to be < -0.02 or >0.02, incorporated the maximum and minimum values

for all PCs to ensure that only the most highly correlated probes for each PC were selected. This eigenvector cutoff value chose a consistent number of probes (~550) across all the PCs (Figure S2B). The eigenvalues for the first 10 PCs were confirmed via linear regression with the relationship tested using an *F*-test and *P*-values corrected for multiple testing, using a Bonferroni adjustment.

Association with batch effects

Principal variance components analysis (PVCA) was used to quantify the extent of batch, age, and sex effects within data sets a–d. This method has been described fully previously (see Li *et al.* 2009). PCs are generated by an eigenvalue decomposition of a covariance matrix, and the batch effects are quantified with a linear mixed model, using the batch effect terms as covariates. The variance components in each model are estimated using restricted maximum likelihood and are scaled by the eigenvalues obtained for each PC. The variance attributed to each factor is divided by the variance (determined from the eigenvalue) of the corresponding PC and then standardized across all factors.

A similar method was used to quantify the extent of batch effects on the principal components obtained from SVD. Score values obtained from the SVD of data sets a–d were tested for batch effect association, using a multiple linear regression model (Equations 1 and 2), where $\mathbf{y}$ is a vector of score values for one PC, with $n = 335$ individual score values. $\mathbf{X}$ is a matrix with $c = 86$ covariates: chip position (11 levels), chip ID (73 levels), and sex and generation, which were coded as dichotomous variables.

The R^2 values for the linear regression are calculated as the sum of squares explained by the regression (SSL) divided by the total sum of squares (SST), which gives the percentage of variance in each PC that is associated with batch effects,

$$R^2 = \frac{1 - \mathrm{SSL}}{\mathrm{SST}}, \tag{4}$$

and are adjusted for multiple covariates using

$$R^2_{\mathrm{adj}} = 1 - \frac{(1-R^2)(n-1)}{(n-c-1)}. \tag{5}$$

P-values were obtained from an *F*-test of the total variance explained by the model as opposed to total variance. Multiple testing was accounted for by Bonferroni correction.

Estimation of heritability

To assess the extent of genetic variability held within the PCs that are removed for PC correction in data sets c and d, estimates for heritability were calculated for all PCs generated from data set a. To compute heritability for the PCs the score values for related individuals had to be estimated. This was done to minimize the presence of family trends within the PCs and to retain the same PCs between expression (e) QTL and heritability analyses. The estimated scores were calculated by multiplying the probe values for those individuals by the eigenvectors calculated from the PCA decomposition,

$$\mathbf{U_{est}} = \mathbf{XV}, \tag{6}$$

where

$$\mathbf{U_{est}} = \begin{pmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nn} \end{pmatrix}$$

is the resulting matrix of estimated PC score values with $n = 425$ individuals.

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

is a matrix of probe values with $P = 9085$ probes.

$$\mathbf{V} = \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ v_{p1} & \cdots & v_{pn} \end{pmatrix}$$

is a matrix of eigenvectors, previously derived from the analysis of data from the 335 unrelated individuals.

$\mathbf{U_{est}}$ was combined with $\mathbf{U}$ from the original PCA to create an $\mathbb{R}^{860 \times 335}$score matrix containing values of all individuals for the 335 PCs.

Heritabilities were estimated for the PCs generated from data set a, using Quantitative Trait Disequilibrium Test (QTDT), which partitions variance components attributed to additive genetic, environmental, or common family variance (Abecasis *et al.* 2000). This method utilizes the complete pedigree structure within the data. Two models were compared in this analysis: an AE model, which includes an additive genetic component and a unique environment component, and a CE model, which includes common and unique environment components. Additive genetic, common, and unique environment variance estimates were divided by the total phenotypic variance to determine the proportion contributed by each factor.

Heritability estimates for all probes were calculated using QTDT on the full data set of 860 individuals. The full data set was corrected using the same four methods used for data sets a–d. The distributions of probe heritability estimates obtained from the four different correction methods were compared. QTDT heritability estimates are constrained to have a minimum value of 0.

Genome-wide association study on PCs

A genome-wide association analysis for each PC was performed using PLINK software (Purcell *et al.* 2007) on data set a to provide an independent assessment of genetic effects present for PCs. SNPs were filtered based on a minor allele frequency >0.05, missingness >0.10, and Hardy–Weinberg equilibrium *P*-value <1*e*-6. After filtering, 488,462

SNPs remained for analysis. Significance was determined both at a family-wise error rate, using Bonferroni correction at an α-level of 0.05, and with an empirical *P*-value estimation for each PC, using 1000 permutations in PLINK (Purcell *et al.* 2007).

eQTL analysis

To determine the impact of each of the different correction methods employed in data sets a–d, an eQTL analysis was performed for each probe within these four data sets, using the same procedure as described above. To replicate the effect on eQTL detection we performed the sample analysis for an independent sample comprising 139 unrelated individuals (CHDWB). *Cis*- and *trans*-eQTL at multiple FDR levels (Benjamini and Hochberg 1995) were extracted and compared between the data sets. *Cis*-eQTL associations are defined as being within ± 1 Mb of the gene tested.

Biological pathway analysis

The pathway analysis was performed on the online Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resources 6.7 (Da Wei Huang and Lempicki 2008). Probes associated to the PCs of interest were submitted as a list of Illumina probe IDs and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment was performed using the Functional Annotation Chart implemented by DAVID (Sherman *et al.* 2007). Significance of pathway enrichment was determined from a modified Fisher's test, which calculates the chance that a set of genes of related terms is presented at a certain percentage in the list. Multitesting was accounted for, using a Benjamini–Hochberg false discovery rate (FDR) of 0.05 (Benjamini and Hochberg 1995).

Results

Nongenetic contributions to PCs

PVCA was used to estimate the extent of batch, sex, and age effects within the whole data set (Boedigheimer *et al.* 2008; Li *et al.* 2009). This method involves first decomposing the data set into a series of PCs and estimating the effects of each covariate on these components. The total variance of each covariance is then summed across all PCs and scaled by the respective eigenvalue. PVCA was applied to data sets a–d to compare the effectiveness of these methods at removing batch effects from the data.

For data set a, batch effects explained 29.2% of the total variance, with chip ID having the largest effect (Figure 1A). This proportion is far less than the 75.5% cumulative variance explained by the first 50 PCs (Figure S3A), which were removed in previous studies to correct for batch effects (Fehrmann *et al.* 2011; Fu *et al.* 2012). The results from linear model correction (Figure 1B) and PC correction (Figure 1, C and D) show that the majority of batch effects have been removed using these methods.

To analyze how batch effects are distributed across PCs, each of the four data sets a–d was decomposed using SVD into 335 PCs. Multiple regression of each PC on chip ID, chip position, sex, and age was used to analyze the distribution of these effects across all PCs. The majority of variance attributed to batch effects in the normalized data set a was held within the first 58 PCs of the data set (Figure 2A), with the proportion of variance explained by batch effects ranging from 24% to 74%. This indicates that unknown sources of variance, not associated to batch, sex, or age effects, are also present within these PCs. As the initial PCs pick up the majority of variance within the data set (Figure S3A), significant associations to these PCs indicate a much larger impact of batch, sex, and age effects within the data set. Therefore signification association to the later PCs as in Figure 2B demonstrates a negligible effect (as quantified for each data set in Figure 1). Batch effect associations to the first few PCs are expected to occur more frequently than in later PCs, because the initial PCs can capture the majority of correlation structure within the data set (Jackson and Wiley 1991).

Overall it is clear that batch effects have a large impact on gene expression variation. However, fitting these as covariates in a linear model during the correction procedure removes the presence of these batch effects (Figure 1B). This correction procedure (data set b), which fits all batch effects as individual factors in a linear model (see *Materials and Methods*), removes the vast majority of associations with batch effects in the principal components generated on the residuals (Figure 2B). There was only one significant association with PC330 present in this corrected data set and that accounted $\sim$0% of the variance, which explains why the variance attributed to batch effects was calculated to be zero with PVCA (Figure 1B). Batch effects were also efficiently removed in data sets c and d, which were corrected for the first 25 PCs (68.4% of the variance) and 50 PCs (75.5% of the variance) (Figure 1, C and D, respectively). However, there are still 6% and 2% associations to batch effects in data sets c and d, respectively, and analyses of PCs generated on these two data sets show many components capturing these residual batch effects (Figure 2, C and D, respectively). These results indicate that PC correction does not remove known batch effects as effectively as correcting for chip ID, chip position, sex, and generation with linear models.

Traditionally the last few PCs, which contain only a small fraction of the variance, are removed as they are attributed to noise or experimental error (Gauch 1982). However, the correlation structure present in the data set can drastically change which PCs are attributed to noise or batch effects (Peres-Neto *et al.* 2005). In gene expression data, it has been demonstrated that the final PCs usually hold genetically relevant information (Yeung and Ruzzo 2001; Ma and Kosorok 2009). Due to the large amount of variance that has been attributed to batch effects within gene expression data sets (Leek *et al.* 2010), removal of PCs has recently

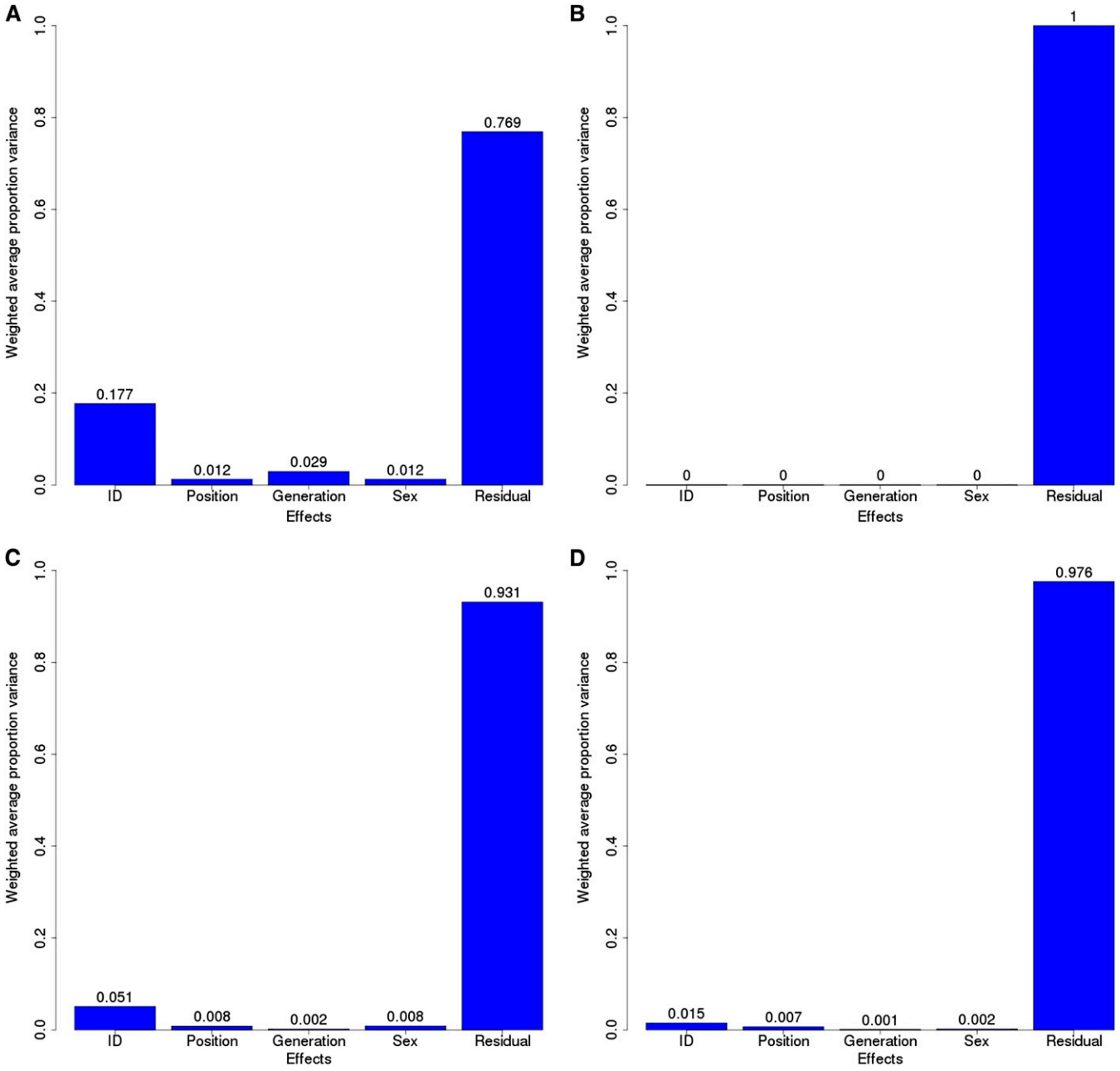


Figure 1 (A–D) Principal variance components analysis (PVCA) of gene expression data sets. PVCA calculates the proportion of variance in the entire data set that is attributed to certain batch, sex, and age effects. Altogether batch, sex, and age effects account for 29% of the variance in the noncorrected data set (A) and are fully removed in the corrected data set (B). Batch effects present in PC corrected dataset are represented in (C) PC25 and (D) PC50. The residuals represent the remaining variance in the data set not attributed to these batch effects.

focused on the initial PCs (up to 50) that explain the majority of the variance within the data sets (Stegle *et al.* 2008; Fehrmann *et al.* 2011; Fu *et al.* 2012; Qin *et al.* 2012). We demonstrated that the initial PCs did have the highest association to these batch effects. However, the variability in which PCs are associated with the batch effects (Figure 2A) makes it difficult to just select an arbitrary number of components to correct for. Removing an uninformed number of PCs could lead to inefficient correction as opposed to using linear models with recorded batch effects. Other factors including biological information could be present alongside these artifacts in these PCs, although we acknowledge that additional batch effects, not recorded, may also be present.

Genetic contribution to PCs

To determine the total genetic component of the PCs that would be removed when using PC correction (as in data sets c and d), heritabilities were estimated for all 335 PCs generated from the normalized data set a (see *Materials and*

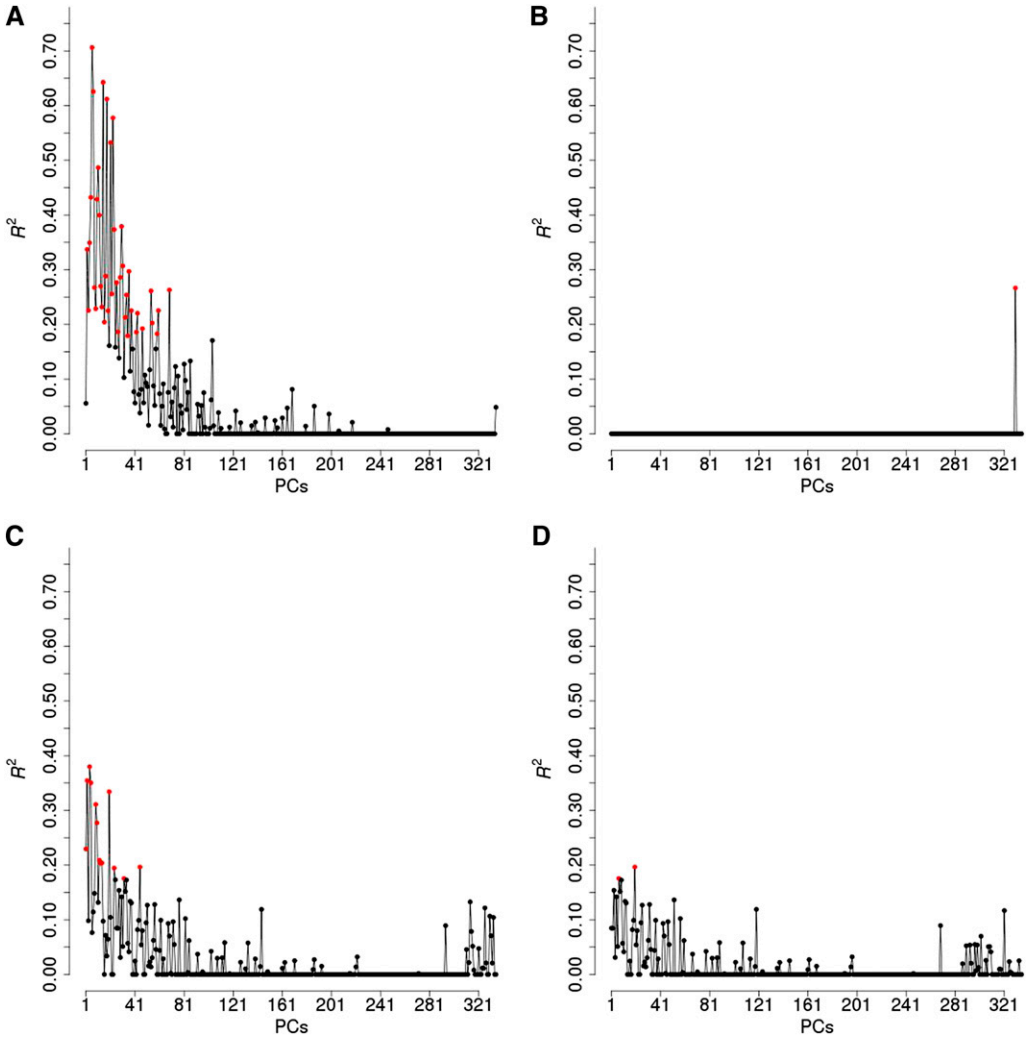


Figure 2 Batch effect associations to PCs. Significant values are highlighted in red and are obtained from an *F*-test on the linear regression models. (A) R^2 from a regression analysis of each PC on batch, sex, and age effects in a data set that is not corrected for any of those factors. (B) R^2 from a regression analysis of each PC on the batch, sex, and age effects in a data set that is corrected for these factors. There is no significant association to batch effects present. (C) Association to batch effects in a data set corrected for the first 25 PCs. (D) Association to batch effects in a data set corrected for the first 50 PCs.

Methods). The heritability estimates showed that there was a considerable genetic component to nearly all 335 PCs generated (Figure 3). The mean of the heritability across all PCs was 0.429 (SD 0.1), with the first 50 PCs having an average heritability of 0.39 (SD 0.13). This demonstrates that a high genetic component is still captured in the first 50 PCs despite the strong association to batch effects. Comparisons between genetic and common family environment models indicate that there is a slight confounding between the two (Figure S4). There is a significant association between the estimates between the two models with an $R^2 = 0.08$ ($P = 8.57e\text{-}08$). This indicates that the heritability estimates could be slightly biased upward due to confounding with common family effects (Lynch and Walsh 1998).

To assess the impact of linear model and PC correction on genetic variation, we estimated the heritabilities of the 9086 probes in the entire data set of 860 individuals, using the same four correction methods a–d (see *Materials and Methods*). Mean heritabilities for the 9086 probes from the four normalization strategies were 0.32, 0.23, 0.21, and 0.18, respectively (Figure 4). The high mean heritability from strategy a is likely due to inflation from batch effects such as the date of RNA extraction, which was performed in family groups and therefore could not be corrected without removing heritable variation. With PC corrections c and d, there was a much higher proportion of zero heritability probes as opposed to those in strategy b. The lower mean estimate and the higher proportion of zero heritability probes when correcting for 1–50 PCs suggest that PC correction has a much higher penalty of genetic variation than linear model correction.

To investigate the impact of PC correction on eQTL detection, we ran a series of eQTL analyses for each probe in the unrelated BSGS data sets a–d. Associated *cis* and *trans* variants were extracted at multiple FDR thresholds (Table 1). The results demonstrated a much higher number of both *cis*- and *trans*-eQTL detected within the data set corrected with linear models (b) as opposed to PC-corrected data sets (c and d). PC correction, however, enhanced eQTL detection when compared to noncorrected data sets (a), likely reflecting a removal of false positives caused by batch effects. The improved removal of batch effects in the PC50 (d) *vs.* PC25 (c) corrected data set (Figure 1 and Figure 2) is also reflected by an increased number of *cis*- and *trans*-eQTL in

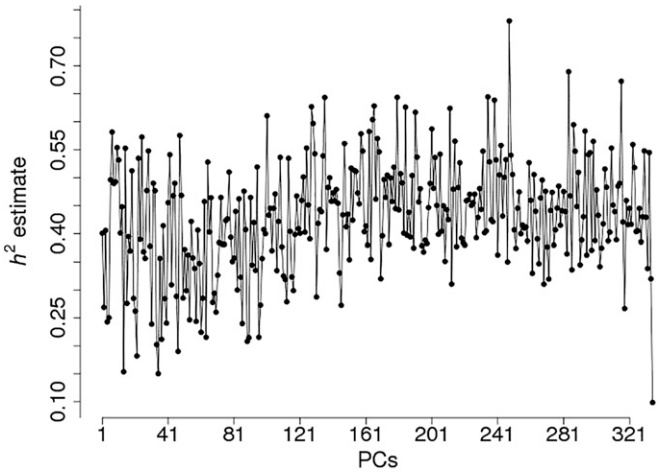


Figure 3 Heritability estimates. Narrow-sense heritability estimates of each PC obtained from QTDT. These results indicate that nearly all PCs hold genetic information.

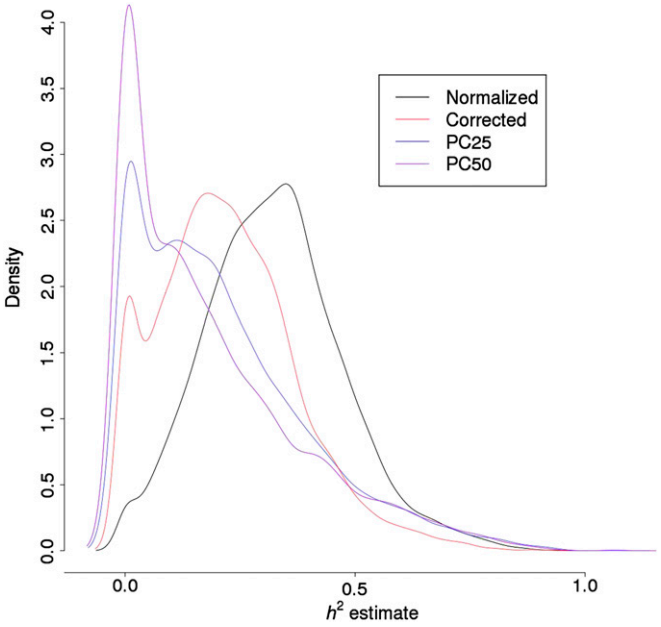


Figure 4 Distribution of heritability estimates for 9086 probes from four different correction methods. Black, standard normalized; red, standard normalization with linear model correction for batch effects; blue, corrected for batch effects using the first 25 PCs; purple, corrected for batch effects using the first 50 PCs. There is a drop in heritability using the correction methods; however this is more pronounced in the PC corrected data set. The heritability estimates are constrained to zero due to the nature of genetic variance component estimating in QTDT.

data set d. These same trends were also observed in our replication sample (CHDWB) (Table 1). The results from these eQTL studies indicate that the PC correction method negatively affects the number of eQTL that can be detected within gene expression data sets.

To investigate loci driving genetic variation captured within PCs generated from data set a, we performed a genome-wide association study for each PC. The PCs were tested for association with 488,462 genotyped SNPs, using linear models implemented in PLINK software (Purcell *et al.* 2007). At a study-wide significance level determined by Bonferroni correction ($0.05/(488{,}462 \text{ SNPs} \times 335 \text{ PCs}) = 3e^{-10}$), no significant SNPs were found. We next examined the top SNPs that were significant after Bonferroni correction on each PC ($0.05/448{,}462 = 1.0e^{-7}$). There were 23 SNPs that were significant at this threshold and these were confirmed with 1000 permutations (Figure S5 and Table S1). The lack of genome-wide significant associations could be attributed to the study not having enough power due to a small sample size (Park *et al.* 2010). Another explanation is that the genetic variance attributed to PCs is highly polygenic. As multiple probes are driving each PC (Figure S2B), the magnitude of different signals can prevent the detection of individual SNP effects.

Biological pathway analysis

We next sought to evaluate whether the first 50 PCs generated from data set a contained linear combinations of genes involved in an expression pathway. Pathway analysis was performed using the DAVID Bioinformatics Resources 6.7, Functional Annotation Tool (Da Wei Huang and Lempicki 2008) on the first 50 PCs (Figure S6 and Table S2). For numerous PCs we are able to demonstrate significant enrichment for multiple different KEGG pathways (FDR = 0.05).

Immune function was the most common process with PC3, PC12, PC24, PC25, and PC26 all showing significant enrichment for immune functional pathways. PC3 showed enrichment for Fc-gamma *R*-mediated phagocytosis ($P = 5.4e^{-3}$). PC12 showed enrichment for hematopoietic cell lineage ($P = 2e^{-3}$), B-cell receptor signaling ($P = 3.8e^{-2}$), graft-*vs.*-host disease ($P = 4.4e^{-2}$), non-small-cell lung cancer ($P = 4e^{-2}$), and asthma ($P = 4.4e^{-2}$). PC24 and PC25 were also enriched for B-cell receptor signaling ($P = 2.7e^{-5}$ and $P = 1.6e^{-2}$, respectively) and PC26 showed enrichment for primary immunodeficiency ($P = 1.2e^{-2}$).

Metabolic processes were enriched in PC8, PC13, and PC32. Enrichment for oxidative phosphorylation was found in PC8, PC13, and PC32 ($P = 4e^{-5}$, $P = 5.6e^{-3}$, and $P = 3.7e^{-2}$). PC8 also showed enrichment for genes known to be involved in susceptibility to Parkinson's ($P = 3e^{-3}$), Alzheimer's ($P = 3e^{-3}$), and Huntington's disease ($P = 9e^{-5}$) and proteasomal components involved in peptide processing ($P = 4e^{-5}$). These brain conditions have been linked to oxidative metabolic dysfunction, due to oxidative damage to neurons (Browne *et al.* 1997; Mecocci *et al.* 2004; Rhein *et al.* 2009) and neuronal energy deficiency (Hoyer 1992) and also improper peptide processing that leads to the buildup of amyloid plaques (Jonsson *et al.* 2012).

Enrichments for ribosomal components were found in PC1, PC13, and PC18 ($P = 4.5e^{-5}$, $P = 2.4e^{-2}$, and $P = 1e^{-4}$). PC13 also showed enrichment for RNA degradation

Table 1 eQTL results for datasets a–d

	BSGS							
	No correction, data set a		PC 25 correction, dataset c		PC 50 correction, dataset d		Linear model correction, dataset b	
FDR	*Cis*	*Trans*	*Cis*	*Trans*	*Cis*	*Trans*	*Cis*	*trans*
0.2	1586	8056	1676	9085	1707	9085	2824	9085
0.1	840	2144	929	2530	1149	2564	1914	3025
0.05	596	170	650	254	806	301	1746	490
0.01	449	99	502	137	737	169	1199	264
0.001	349	75	402	104	640	114	1005	218
0.0001	273	56	316	88	510	91	848	183
					CHDWB			
0.2	743	8102	815	8234	909	8695	1164	8841
0.1	513	1249	597	1471	684	1538	822	2076
0.05	345	90	373	98	457	152	627	388
0.01	189	51	252	46	366	120	594	215
0.001	143	22	207	31	295	86	470	136
0.0001	101	9	168	15	183	27	385	141

Cis regions were defined as ±1 Mb either side of the transcription start site. *Trans* was defined as elsewhere in the genome. The numbers of probes with a *cis* or *trans* association significant at various study-wide FDR thresholds is provided for each of the four correction methods

processes ($P = 1.8e^{-3}$). PC7 was enriched for porphyrin metabolism ($P = 2e^{-2}$) involved in heme biosynthesis.

Of the 11 PCs that showed a significant enrichment (1, 3, 7, 8, 12, 13, 18, 24, 25, 26, and 32), most of them also demonstrated relatively high heritability estimates (0.4, 0.4, 0.58, 0.49, 0.4, 0.45, 0.37, 0.39, 0.57, 0.37, and 0.49, respectively), with an average heritability of 0.45. These results together demonstrate that biologically relevant and interesting probe enrichments are present within the first 50 PCs despite the high association with batch effects within the data set.

Discussion

The removal of genetic variation alongside batch effects when using PC correction has been alluded to but never formally investigated in several articles (Stegle *et al.* 2010; Fehrmann *et al.* 2011; Brown *et al.* 2012; Fu *et al.* 2012). Due to the unique study design of the BSGS, which contains both unrelated and family information, we are able to quantify the genetic components driving each PC with two independent approaches: SNP association analysis and heritability estimates. From this, we have demonstrated that all PCs obtained from the decomposition of a gene expression data set contain relevant genetic information. Most importantly, we show that the first 50 PCs, which have been removed in previously published articles to correct for batch effects (Leek and Storey 2007; Stegle *et al.* 2008; Leek *et al.* 2010; Fehrmann *et al.* 2011; Fu *et al.* 2012; Qin *et al.* 2012; Stranger *et al.* 2012), contain both a considerable proportion of genetic variation influencing gene expression (Figure 3) and an enrichment for biological networks. The considerable genetic variation found within the initial PCs cautions against the removal of such components in the data set due to the potential loss of genetic information.

We also show that batch effects are distributed across the first 59 PCs with varying effect sizes. As these initial PCs contain a combination of both genetic and batch effects, there appears to be a trade-off between removing batch effects and removing biologically relevant data when employing PC correction. Removing a higher number of PCs improves batch effect correction while at the same time increasing the amount of genetic variation that is removed (Figure 4 and Table 1) and the removal of a smaller number of PCs (as in some studies, *e.g.*, Pickrell *et al.* 2010) can lead to the incomplete removal of batch effects. While the exact proportions of genetic and batch effect variance observed in the PCs here are unique to this data set, similar patterns of variance distributions are expected to be present in other high-throughput expression data sets.

PC correction became prevalent in gene expression studies after being used to correct for expression heterogeneity by a method called surrogate variable analysis (SVA) (Leek and Storey 2007). This method corrects for noise in the data set that is not accounted for by the primary variable of interest, which may be different experimental conditions or genes. It has been demonstrated as an effective means of enhancing the genetic signals of interest and minimizing false discovery rates. One key difference between SVA and PCA correction is that the principal components were generated on the residuals of the data that were corrected for the primary variable of interest. These principal components contain residual "noise" within the data set and their subsequent removal enhanced the power to detect signals associated with the primary variable. Later studies removed principal components from the whole data set without former corrections for variables of interest (Stegle *et al.* 2008; Pickrell *et al.* 2010; Fehrmann *et al.* 2011; Fu *et al.* 2012). We have demonstrated here that this can remove genetically driven variation within the gene expression data set that

could be of interest to the researcher, due to the ability of PCs to pick up multiple sources of variation (Stegle *et al.* 2010; Qin *et al.* 2012).

As the PCs pick up linear combinations of factors in the data set, they also have the power to detect gene regulatory networks and coexpressed modules (Holter *et al.* 2000). Gene regulatory networks have a considerable impact on disease as opposed to single-gene changes (Chen *et al.* 2008), as groups of genes are known to interact and respond to environmental perturbations together (Ihmels *et al.* 2003). These networks are made of coexpressed gene modules that are robust to environmental changes and even different microarray platforms (Chaussabel *et al.* 2008). Principal components have been used to quantify regulatory SNPs governing metabolic networks in both yeast (Biswas *et al.* 2008) and human data (Abo *et al.* 2012). We observed in our data set that there was enrichment for biological networks within the initial PCs. These were dominated by immune function processes, which have been detected previously with clustering analysis of gene expression data (Chaussabel *et al.* 2008) as well as of metabolic, ribosomal, protein, and heme biosynthesis pathways. As these PCs also demonstrate a high heritability, this indicates that a considerable proportion of the variance in these PCs is being explained by biological factors. The removal of such PCs to compensate for batch effects would have led to the removal of this interesting information. It has also been demonstrated previously that PC correction also removes a large proportion of covariance in a data set, which could constitute these gene networks and interactions (Qin *et al.* 2012).

Our results show that PC correction negatively affects both average probe heritability (Figure 4) and the number of eQTL hits detected (Table 1). While the removal of a larger number of PCs improved batch effect correction (Figure 1) and increased the number of eQTL detected in the data set (Table 1), it had a much larger impact on mean probe heritability. This genetic variation could be composed of additional factors such as genetic covariance contributing to gene networks that may not necessarily be found in an eQTL study. Linear model correction, on the other hand, demonstrated superior eQTL detection and retained a much higher probe average heritability. Therefore, if recorded processing dates are present, these values can be used to correct for batch effects in the data by incorporating them as covariates in a linear model (Li and Rabinovic 2007). This not only increases eQTL detection by increasing power and reducing false positives (Stegle *et al.* 2008) but also ensures that, unlike PC correction, large amounts genetic variance are not removed (Figure 4). Strong associations with PCs explaining large proportions of variance in a data set can be used to select for appropriate batch effects to model within a data set if a large number of different factors have been recorded (Parts *et al.* 2012). We demonstrate here that linear model correction also has the advantage over PC correction in that it is more effective in correcting for known batch effects (Figure 2).

We demonstrate that the removal of PCs from gene expression data sets to correct for batch and environmental effects should be treated with caution, as many different sources of variation can be present within them. This comes from the ability of PCs to detect many linear combinations of trait values (Yeung and Ruzzo 2001). As it can be difficult to distinguish between which factors are driving the PCs without prior knowledge of the batch and environmental trends in the data set, removal of PCs runs the risk of removing biologically interesting data. However, removal of the first 50 PCs can sometimes provide a good method to account for technical artifacts that can lead to spurious associations when no batch effects have been recorded. Our results here show that it is clearly preferable to record such information during the generation of the data and correct for it using standard linear approaches (Scherer 2009).

The data used in this study are available to potential researchers through the Consortium for Genetic Architecture of Gene Expression (CAGE). Should individuals wish to obtain a copy of the data they can apply with a research proposal for membership to CAGE. Applications can be made by e-mailing Peter Visscher (peter.visscher@uq.edu.au) or Joseph Powell (joseph.powell@uq.edu.au).

Acknowledgments

The authors declare that they have no conflicts of interest. The authors gratefully acknowledge the participation of the individuals sampled in this work. The authors thank Sara Smith and Anthony Caracella for their technical assistance with the microarray hybridizations; Alison Mackenzie, Marlene Grace, and Ann Eldridge for data collection; and Dale Nyholt, David Smyth, and Scott Gordon for data management. This research was supported by Australian National Health and Medical Research Council (NHMRC) grants 389892, 496667, 613601, 1010374, and 1046880 and National Institutes of Health (NIH) grant GM057091. Grant W. Montgomery and Peter M. Visscher are supported by the NHMRC Fellowship Scheme. Joseph E. Powell is supported by the Australia Research Council (ARC) Fellowship Scheme. Anita Goldinger is supported by an Australian Postgraduate Award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Literature Cited

Abecasis, G., L. Cardon, and W. Cookson, 2000 A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet. 66: 279–292.

Abo, R., G. D. Jenkins, L. Wang, and B. L. Fridley, 2012 Identifying the genetic variation of gene expression using gene sets: application of novel gene set eQTL approach to PharmGKB and KEGG. PLoS ONE 7: e43301.

Allison, D. B., X. Cui, G. P. Page, and M. Sabripour, 2006 Microarray data analysis: from disarray to consolidation and consensus. Nat. Rev. Genet. 7: 55–65.

Baggerly, K. A., K. R. Coombes, and E. S. Neeley, 2008 Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. J. Clin. Oncol. 26: 1186–1187.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57: 289–300.

Biswas, S., J. D. Storey, and J. M. Akey, 2008 Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. BMC Bioinformatics 9: 244.

Boedigheimer, M. J., R. D. Wolfinger, M. B. Bass, P. R. Bushel, J. W. Chou *et al.*, 2008 Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. BMC Genomics 9: 285.

Brown, C. D., L. M. Mangravite, and B. E. Engelhardt, 2012 Integrative modeling of eQTLs and cis-regulatory elements suggest mechanisms underlying cell type specificity of eQTLs. arXiv preprint. arXiv:1210.3294.

Browne, S. E., A. C. Bowling, U. Macgarvey, M. J. Baik, S. C. Berger *et al.*, 1997 Oxidative damage and metabolic dysfunction in Huntington's disease: selective vulnerability of the basal ganglia. Ann. Neurol. 41: 646–653.

Chaussabel, D., C. Quinn, J. Shen, P. Patel, C. Glaser *et al.*, 2008 A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity 29: 150–164.

Chen, C., K. Grennan, J. Badner, D. Zhang, E. Gershon *et al.*, 2011 Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS ONE 6: e17238.

Chen, Y., J. Zhu, P. Y. Lum, X. Yang, S. Pinto *et al.*, 2008 Variations in DNA elucidate molecular networks that cause disease. Nature 452: 429–435.

Churchill, G. A., 2002 Fundamentals of experimental design for cDNA microarrays. Nat. Genet. 32: 490–495.

Da Wei Huang, B. T. S., and R. A. Lempicki, 2008 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4: 44–57.

Fehrmann, R. S. N., R. C. Jansen, J. H. Veldink, H. J. Westra, D. Arends *et al.*, 2011 Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 7: e1002197.

Fu, J., M. G. M. Wolfs, P. Deelen, H. J. Westra, R. S. N. Fehrmann *et al.*, 2012 Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 8: e1002431.

Fukushima, A., M. Wada, S. Kanaya, and M. Arita, 2008 SVD-based anatomy of gene expressions for correlation analysis in Arabidopsis thaliana. DNA Res. 15: 367–374.

Gauch, H. G., Jr., 1982 Noise reduction by eigenvector ordinations. Ecology 63: 1643–1649.

Golub, G. H., and C. Reinsch, 1970 Singular value decomposition and least squares solutions. Numerische Mathematik 14: 403–420.

Grundberg, E., K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil *et al.*, 2012 Mapping cis-and trans-regulatory effects across multiple tissues in twins. Nat. Genet. 44: 1084–1089.

Hirai, M. Y., K. Sugiyama, Y. Sawada, T. Tohge, T. Obayashi *et al.*, 2007 Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc. Natl. Acad. Sci. USA 104: 6478–6483.

Holter, N. S., M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar *et al.*, 2000 Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc. Natl. Acad. Sci. USA 97: 8409–8414.

Hoyer, S., 1992 Oxidative energy metabolism in Alzheimer brain. Mol. Chem. Neuropathol. 16: 207–224.

Ihmels, J., R. Levy, and N. Barkai, 2003 Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nat. Biotechnol. 22: 86–92.

Irizarry, R. A., D. Warren, F. Spencer, I. F. Kim, S. Biswal *et al.*, 2005 Multiple-laboratory comparison of microarray platforms. Nat. Methods 2: 345–350.

Jackson, J. E., and J. Wiley, 1991 *A User's Guide to Principal Components*. Wiley Online Library. Wiley & Sons, New York.

Jonsson, T., J. K. Atwal, S. Steinberg, J. Snaedal, P. V. Jonsson *et al.*, 2012 A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488: 96–99.

Leek, J. T., and J. D. Storey, 2007 Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3: e161.

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead *et al.*, 2010 Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11: 733–739.

Li, C., and A. Rabinovic, 2007 Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118–127.

Li, J., P. R. Bushel, T. M. Chu, and R. D. Wolfinger, 2009 Principal variance components analysis: estimating batch effects in microarray gene expression data, pp. 141–154 in *Batch Effects and Noise in Microarray Experiments*: *Sources and Solutions.* Wiley & Sons, Ltd, Chichester, UK.

Li, Y., J. Huang, and C. I. Amos, 2012 Genetic association analysis of complex diseases incorporating intermediate phenotype information. PLoS ONE 7: e46612.

Luo, J., M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi *et al.*, 2010 A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 10: 278–291.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Ma, S., and M. R. Kosorok, 2009 Identification of differential gene pathways with principal component analysis. Bioinformatics 25: 882–889.

Mecocci, P., U. MacGarvey, and M. F. Beal, 2004 Oxidative damage to mitochondrial DNA is increased in Alzheimer's disease. Ann. Neurol. 36: 747–751.

Misra, J., W. Schmitt, D. Hwang, L. L. Hsiao, S. Gullans *et al.*, 2002 Interactive exploration of microarray gene expression patterns in a reduced dimensional space. Genome Res. 12: 1112–1120.

Nath, A. P., D. Arafat, and G. Gibson, 2012 Using blood informative transcripts in geographical genomics: impact of lifestyle on gene expression in Fijians. Front. Genet. 3: 243.

Nica, A. C., and E. T. Dermitzakis, 2008 Using gene expression to investigate the genetic basis of complex disorders. Hum. Mol. Genet. 17: R129–R134.

Park, J. H., S. Wacholder, M. H. Gail, U. Peters, K. B. Jacobs *et al.*, 2010 Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. 42: 570–575.

Parts, L., Å. K. Hedman, S. Keildson, A. J. Knights, C. Abreu-Goodger *et al.*, 2012 Extent, causes, and consequences of small RNA expression variation in human adipose tissue. PLoS Genet. 8: e1002704.

Peres-Neto, P. R., D. A. Jackson, and K. M. Somers, 2005 How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comput. Stat. Data Anal. 49: 974–997.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt *et al.*, 2010 Understanding mechanisms underlying human

gene expression variation with RNA sequencing. Nature 464: 768–772.

Powell, J. E., A. K. Henders, A. F. McRae, A. Caracella, S. Smith *et al.*, 2012a The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. PLoS ONE 7: e35430.

Powell, J. E., A. K. Henders, A. F. McRae, M. J. Wright, N. G. Martin *et al.*, 2012b Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Res. 22: 456–466.

Price, A. L., A. Helgason, G. Thorleifsson, S. A. McCarroll, A. Kong *et al.*, 2011 Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. 7: e1001317.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–575.

Qin, S. P., J. Kim, D. Arafat, and G. Gibson, 2012 Effect of normalization on statistical and biological interpretation of gene expression profiles. Front. Genet. 3: 160.

Raychaudhuri, S., J. M. Stuart, and R. B. Altman, 2000 Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput. 2000: 455–466.

Rhein, V., X. Song, A. Wiesner, L. M. Ittner, G. Baysang *et al.*, 2009 Amyloid-β and tau synergistically impair the oxidative phosphorylation system in triple transgenic Alzheimer's disease mice. Proc. Natl. Acad. Sci. USA 106: 20057–20062.

Scherer, A., 2009 *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley & Sons, New York.

Sherman, B. T., Q. Tan, J. Kir, D. Liu, D. Bryant *et al.*, 2007 DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35: W169–W175.

Spielman, R. S., and V. G. Cheung, 2007 Reply to "On the design and analysis of gene expression studies in human populations". Nat. Genet. 39: 808–809.

Stegle, O., A. Kannan, R. Durbin, and J. Winn, 2008 Accounting for non-genetic factors improves the power of eQTL studies, pp. 411–422 in *Research in Computational Molecular Biology*. Springer, Berlin, Germany/New York.

Stegle, O., L. Parts, R. Durbin, and J. Winn, 2010 A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLOS Comput. Biol. 6: e1000770.

Stranger, B. E., S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle *et al.*, 2012 Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 8: e1002639.

Thomas, L., E. M. Coffey, H. Dai, Y. D. He, D. A. Kessler *et al.*, 2003 Effects of atmospheric ozone on microarray data quality. Anal. Chem. 75: 4672–4675.

Yeung, K. Y., and W. L. Ruzzo, 2001 Principal component analysis for clustering gene expression data. Bioinformatics 17: 763–774.

Communicating editor: E. Petretto

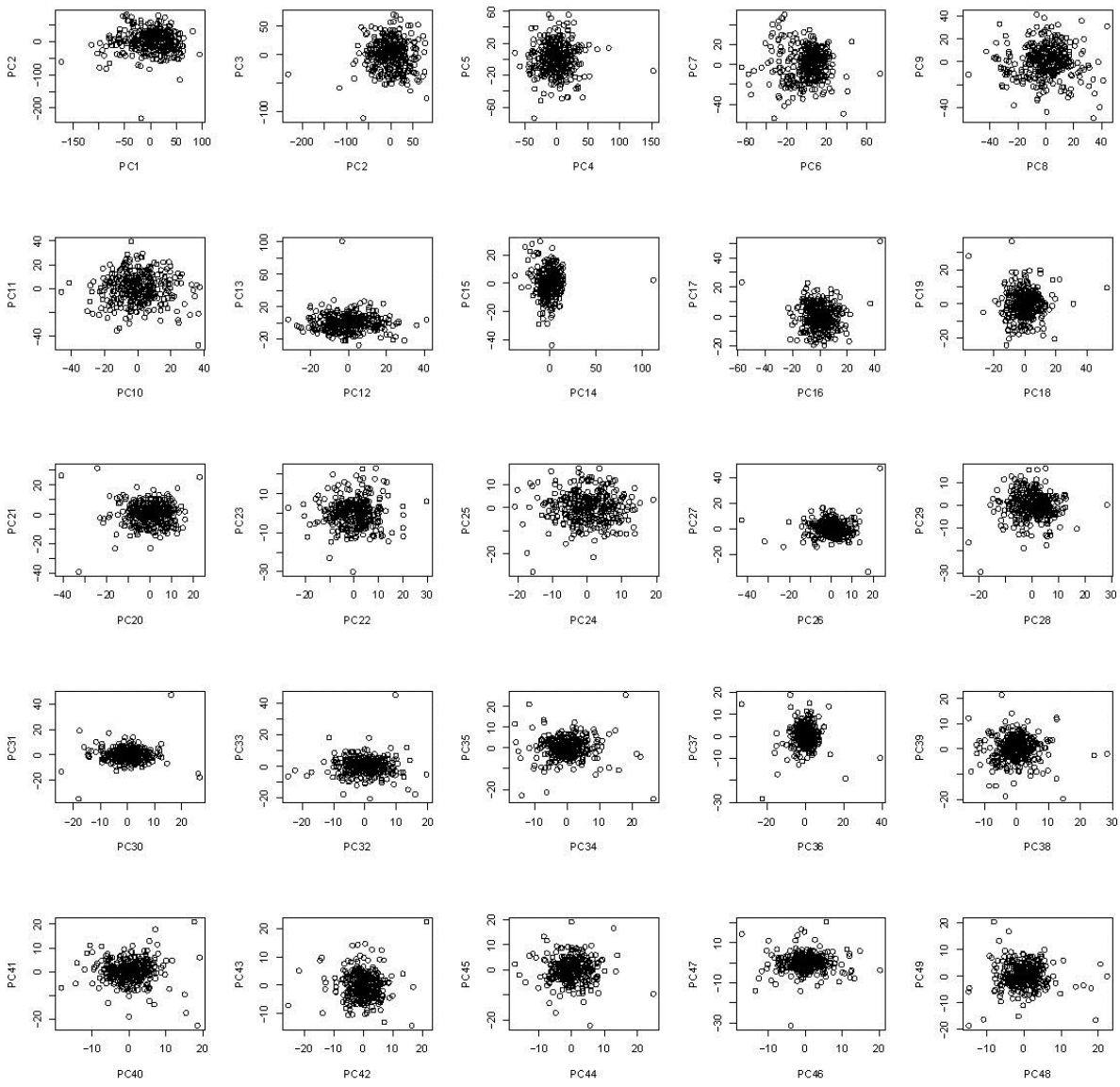
GENETICS

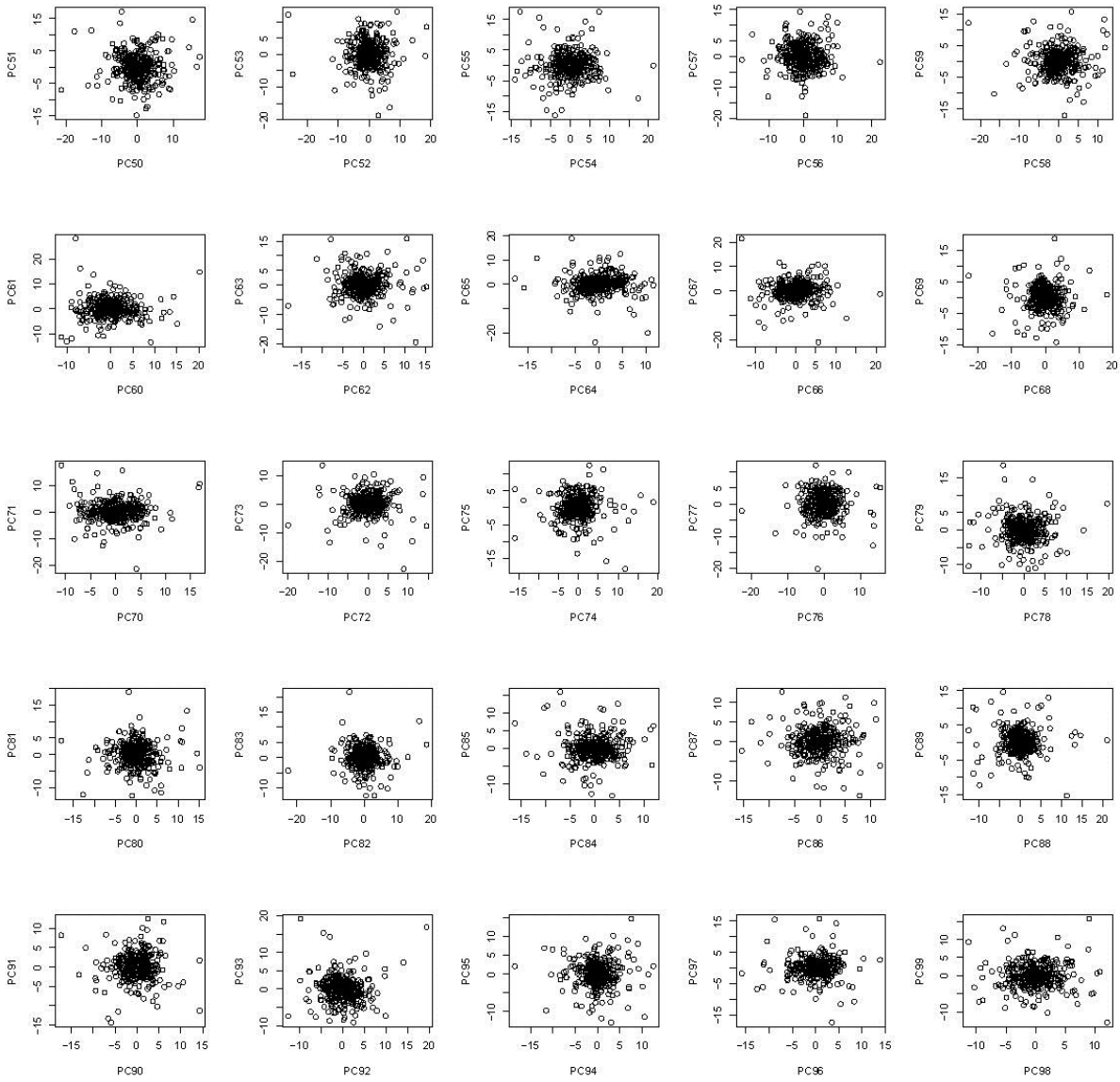
Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.153221/-/DC1

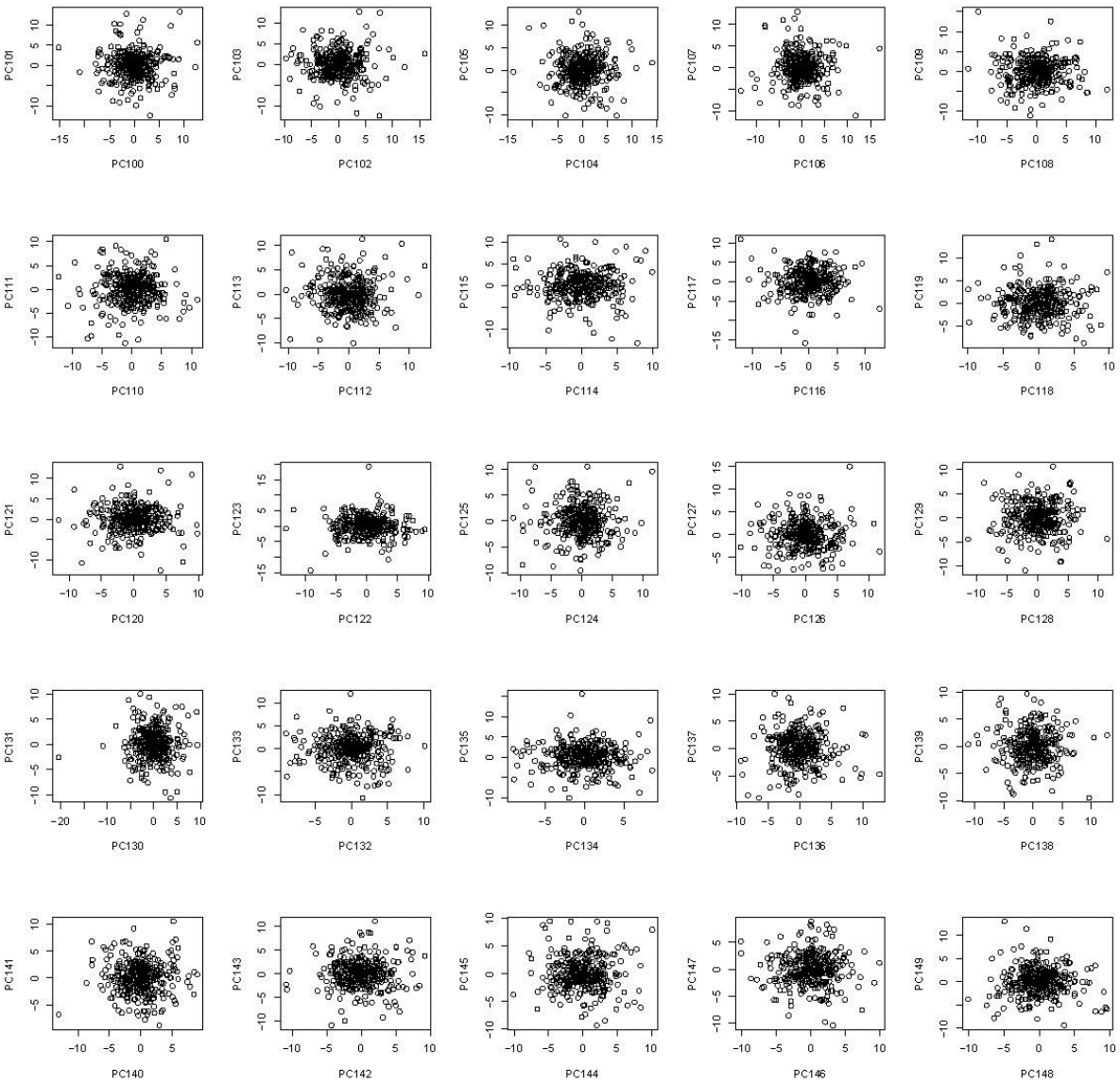
Genetic and Nongenetic Variation Revealed for the Principal Components of Human Gene Expression

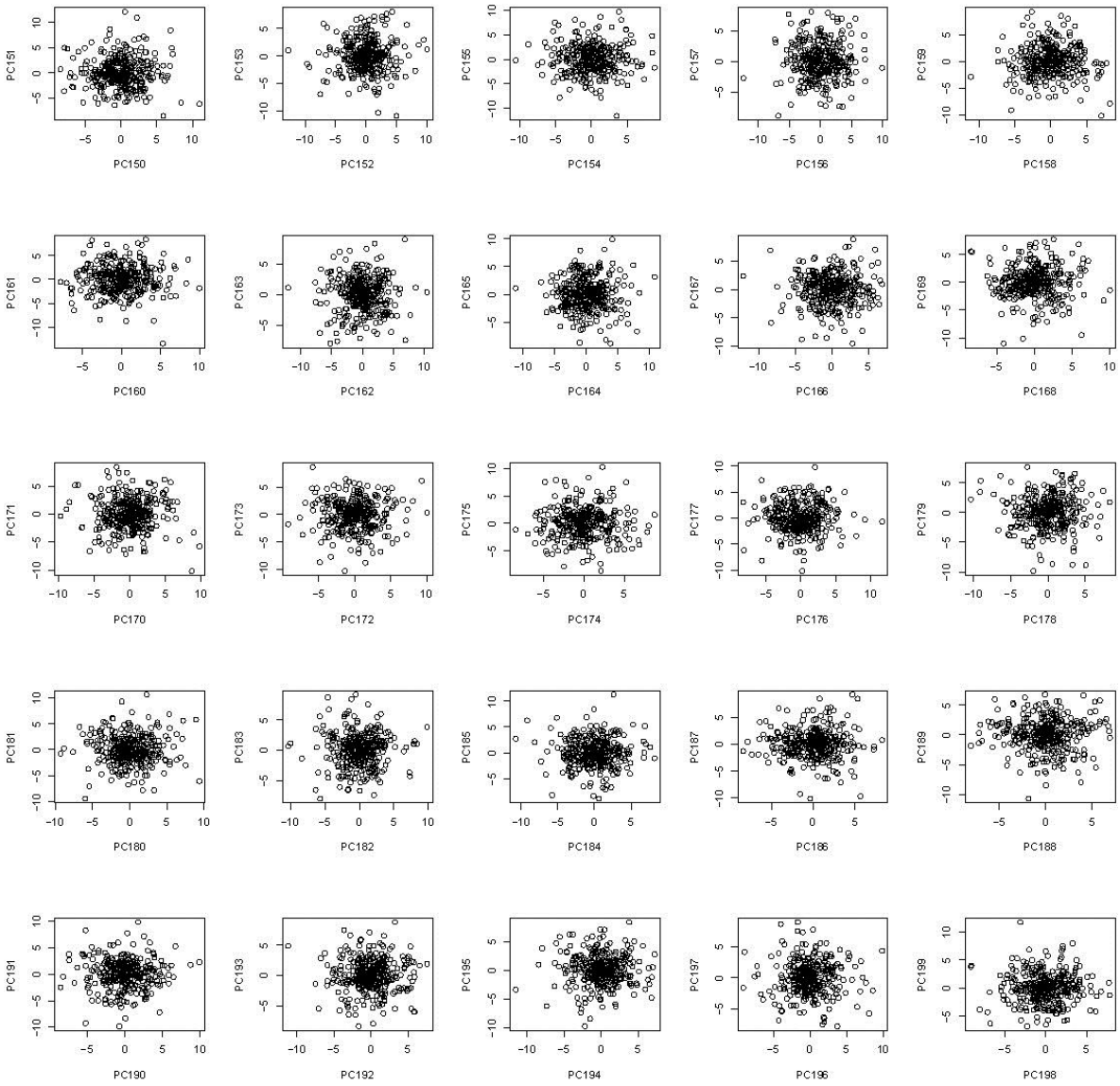
Anita Goldinger, Anjali K. Henders, Allan F. McRae, Nicholas G. Martin, Greg Gibson, Grant W. Montgomery, Peter M. Visscher, and Joseph E. Powell

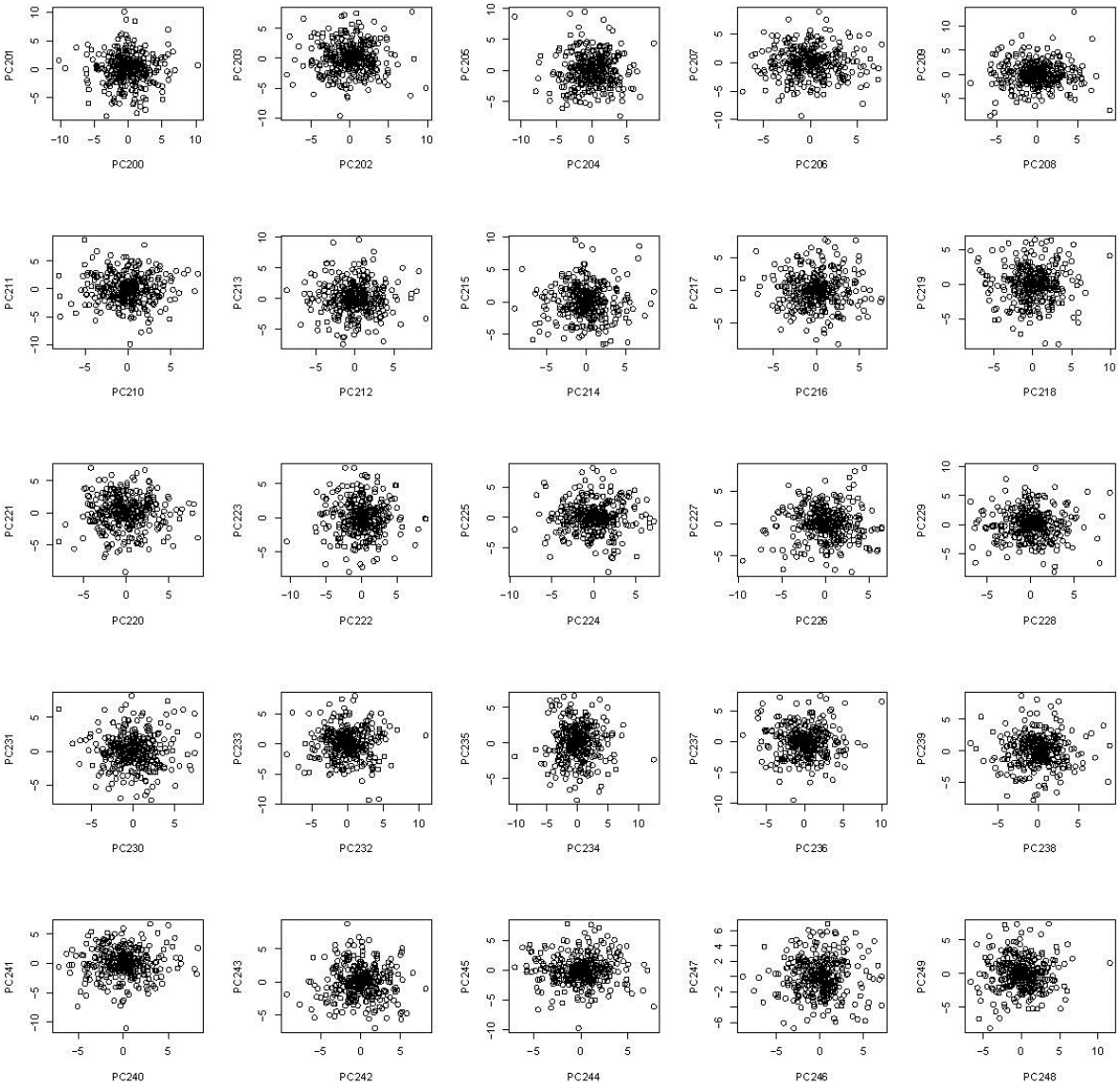
DOI: 10.1534/genetics.113.153221

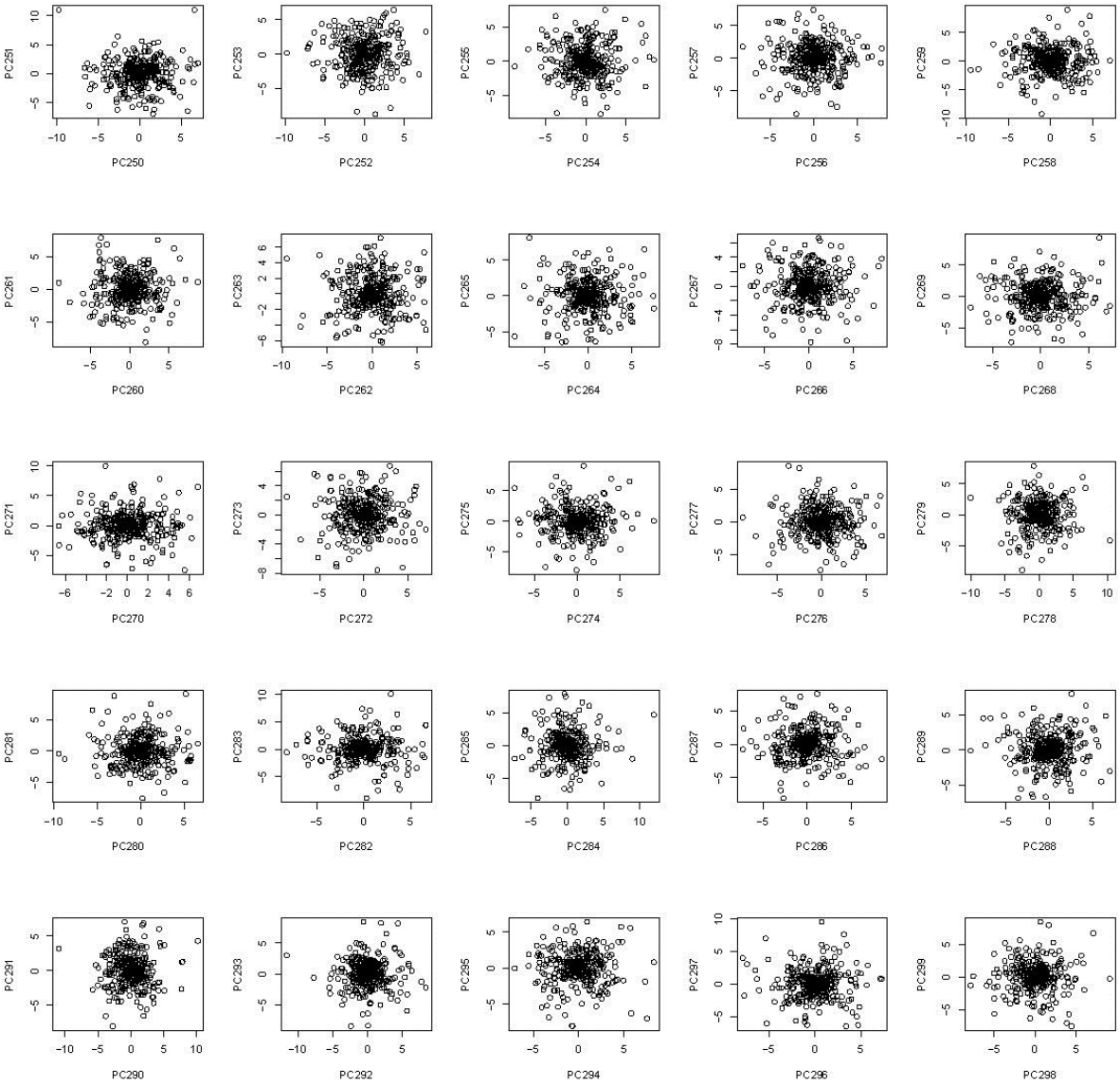

PC2
PC1
PC3
PC2
PC5
PC4
PC7
PC6
PC9
PC8
PC11
PC10
PC13
PC12
PC15
PC14
PC17
PC16
PC19
PC18
PC21
PC20
PC23
PC22
PC25
PC24
PC27
PC26
PC29
PC28
PC31
PC30
PC33
PC32
PC35
PC34
PC37
PC36
PC39
PC38
PC41
PC40
PC43
PC42
PC45
PC44
PC47
PC46
PC49
PC48


PC51 PC50
PC53 PC52
PC55 PC54
PC57 PC56
PC59 PC58
PC61 PC60
PC63 PC62
PC65 PC64
PC67 PC66
PC69 PC68
PC71 PC70
PC73 PC72
PC75 PC74
PC77 PC76
PC79 PC78
PC81 PC80
PC83 PC82
PC85 PC84
PC87 PC86
PC89 PC88
PC91 PC90
PC93 PC92
PC95 PC94
PC97 PC96
PC99 PC98


PC101
PC100
PC103
PC102
PC105
PC104
PC107
PC106
PC109
PC108
PC111
PC110
PC113
PC112
PC115
PC114
PC117
PC116
PC119
PC118
PC121
PC120
PC123
PC122
PC125
PC124
PC127
PC126
PC129
PC128
PC131
PC130
PC133
PC132
PC135
PC134
PC137
PC136
PC139
PC138
PC141
PC140
PC143
PC142
PC145
PC144
PC147
PC146
PC149
PC148


PC151
PC150
PC153
PC152
PC155
PC154
PC157
PC156
PC159
PC158
PC161
PC160
PC163
PC162
PC165
PC164
PC167
PC166
PC169
PC168
PC171
PC170
PC173
PC172
PC175
PC174
PC177
PC176
PC179
PC178
PC181
PC180
PC183
PC182
PC185
PC184
PC187
PC186
PC189
PC188
PC191
PC190
PC193
PC192
PC195
PC194
PC197
PC196
PC199
PC198


PC201
PC200
PC203
PC202
PC205
PC204
PC207
PC206
PC209
PC208
PC211
PC210
PC213
PC212
PC215
PC214
PC217
PC216
PC219
PC218
PC221
PC220
PC223
PC222
PC225
PC224
PC227
PC226
PC229
PC228
PC231
PC230
PC233
PC232
PC235
PC234
PC237
PC236
PC239
PC238
PC241
PC240
PC243
PC242
PC245
PC244
PC247
PC246
PC249
PC248


PC251
PC250
PC253
PC252
PC255
PC254
PC257
PC256
PC259
PC258
PC261
PC260
PC263
PC262
PC265
PC264
PC267
PC266
PC269
PC268
PC271
PC270
PC273
PC272
PC275
PC274
PC277
PC276
PC279
PC278
PC281
PC280
PC283
PC282
PC285
PC284
PC287
PC286
PC289
PC288
PC291
PC290
PC293
PC292
PC295
PC294
PC297
PC296
PC299
PC298

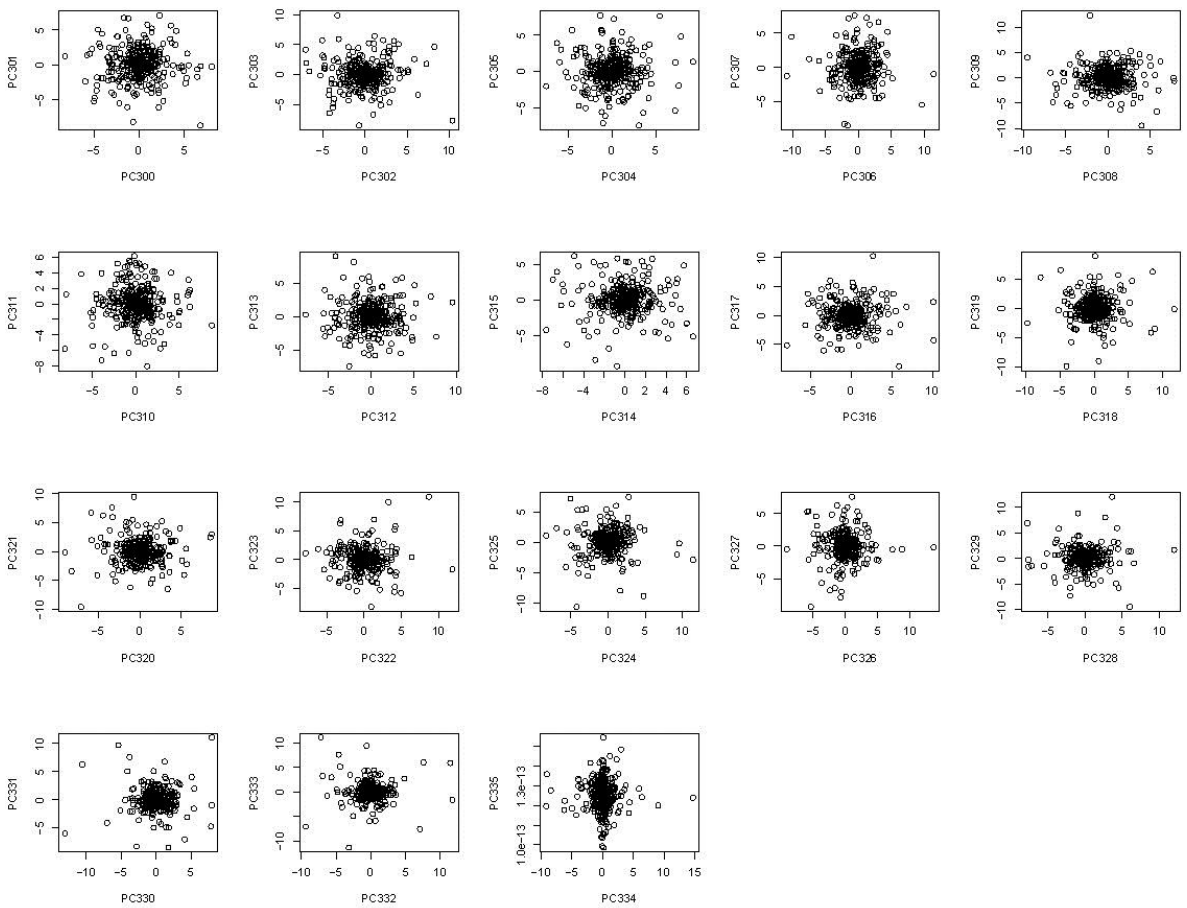


Figure S1 Scores plots of PCs against their adjacent vectors. Demonstrates a homogenous population with no clear substructure or independent clustering of groups of individuals.

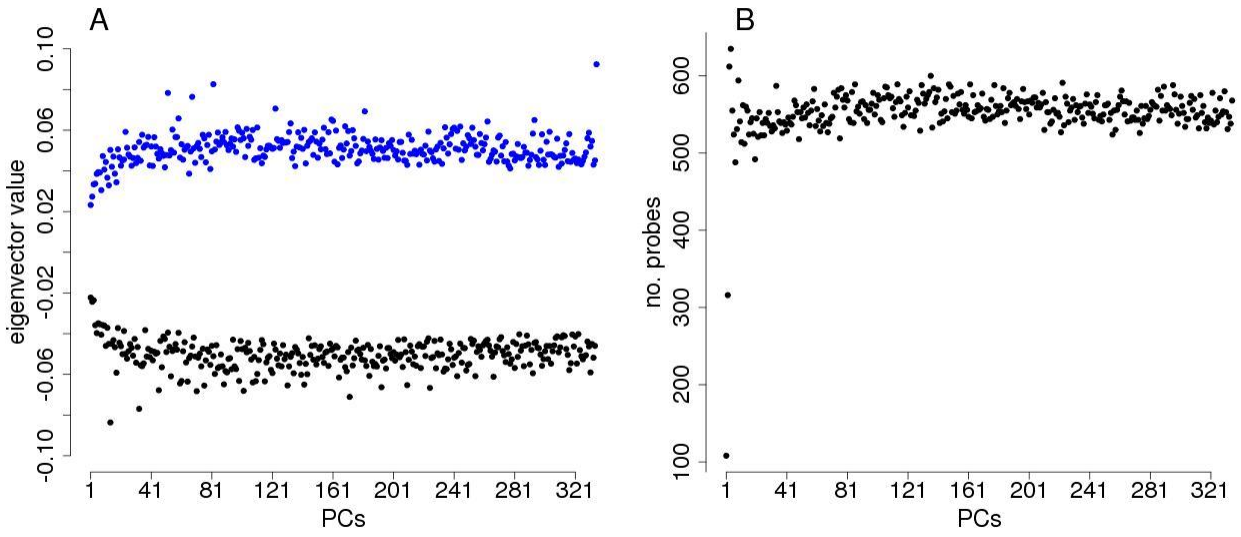


Figure S2 Selection of probes driving PCs**.** (S2A) Maximum (blue) and minimum (black) eigenvector values for each PC. The eigenvector values represent the extent of the correlation between a probe and a PC, with 0 indicating no association. Selection of probes driving each PC is based on the optimal number of probes for each PC that have the same eigenvector value cut off. As multiple probes can contribute a small amount of variance to each PC it is reasonable that a low cut off value can pick up many of the significant probes driving each PC. Probes that have an eigenvector value of greater than 0.02 or less than -0.02 were selected for further biological enrichment analysis as this incorporated all the maximally and minimally expressed probes in this section. (S2B) Selection of probes at this cut off value enabled approximately similar numbers of probes to be selected for each PC. The slightly lower number of probes that are selected in the initial PCs is due to the lower maximum and minimum eigenvector values in these PCs as shown in (S2A).

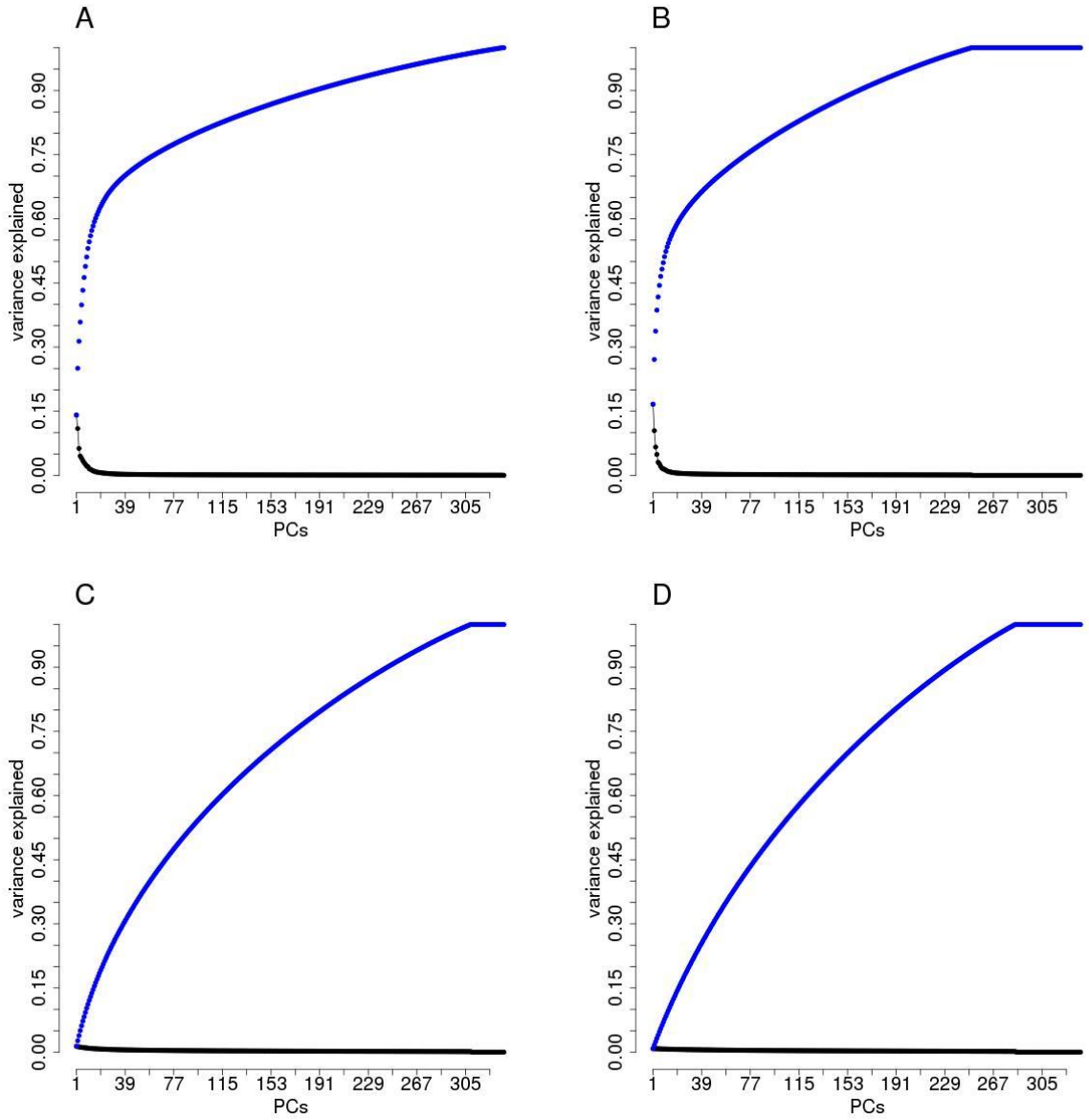


Figure S3 Variance explained by PCs. Calculated from the eigenvalues obtained from the Singular Value Decomposition. Variance explained by each PC is plotted in black and cumulative variance in blue. A) Normalized dataset, B) Corrected with linear models, C) PC25 corrected D) PC50 corrected. All variances add up to 1. Cum

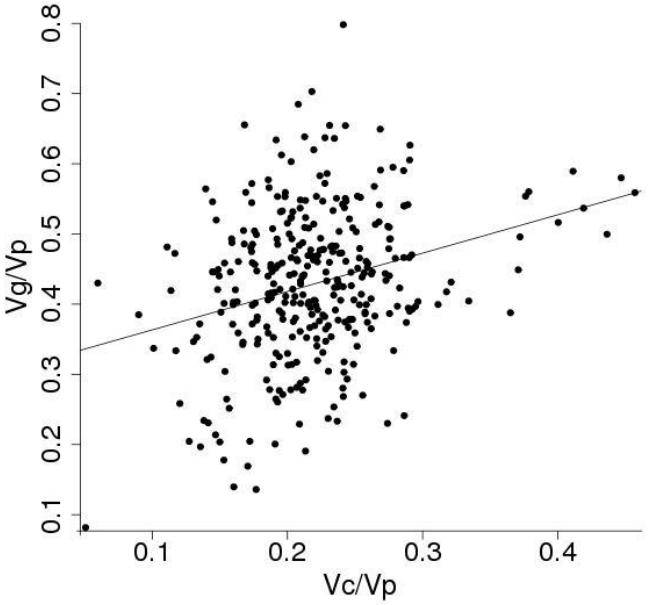
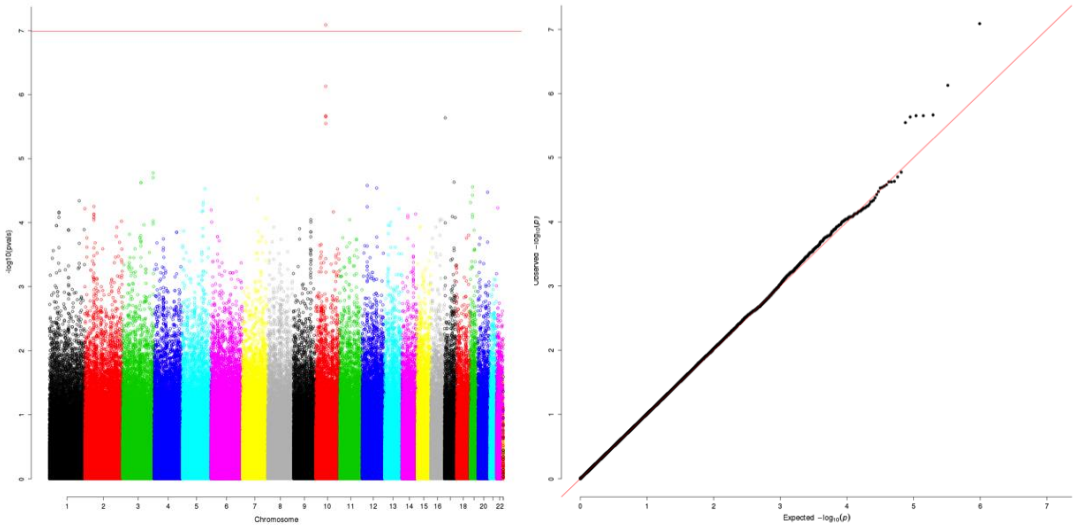
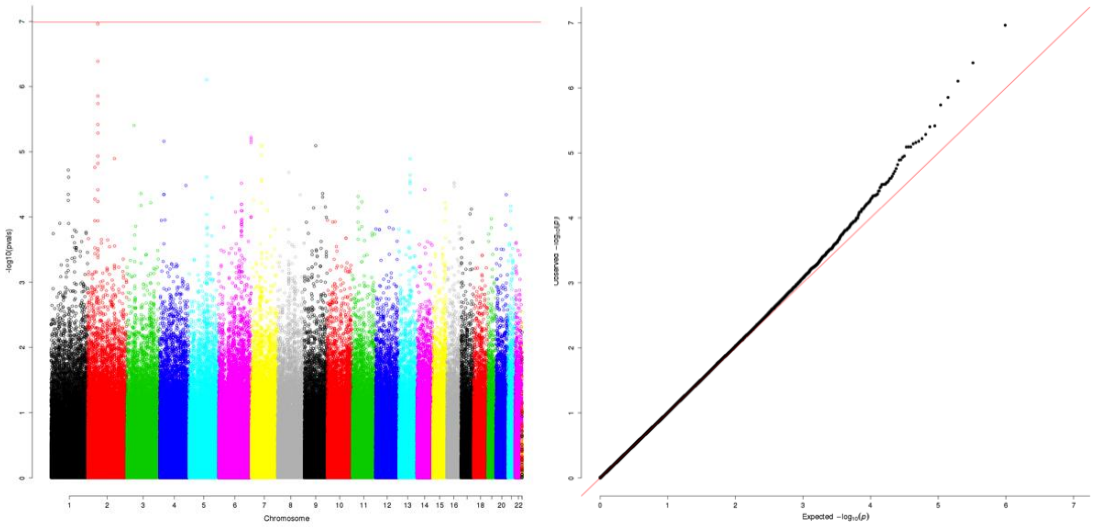


Figure S4 Correlation between additive genetic and common environmental factors. Significant association (p = 8.57e-08 and R2=0.08) between the common environment and genetic components estimated in an AC and AE models. The proportion of common environment variance was calculated by dividing the variance attributed to common environment (Vc) by the total phenotypic variance (Vp). The proportion of genetic variability (heritability) was calculated by dividing the additive genetic component (Vg) by the total phenotypic variance (Vp). This result indicates that the heritability estimates obtained are confounded with common environment variance and therefore inflated upwards by common family effects.

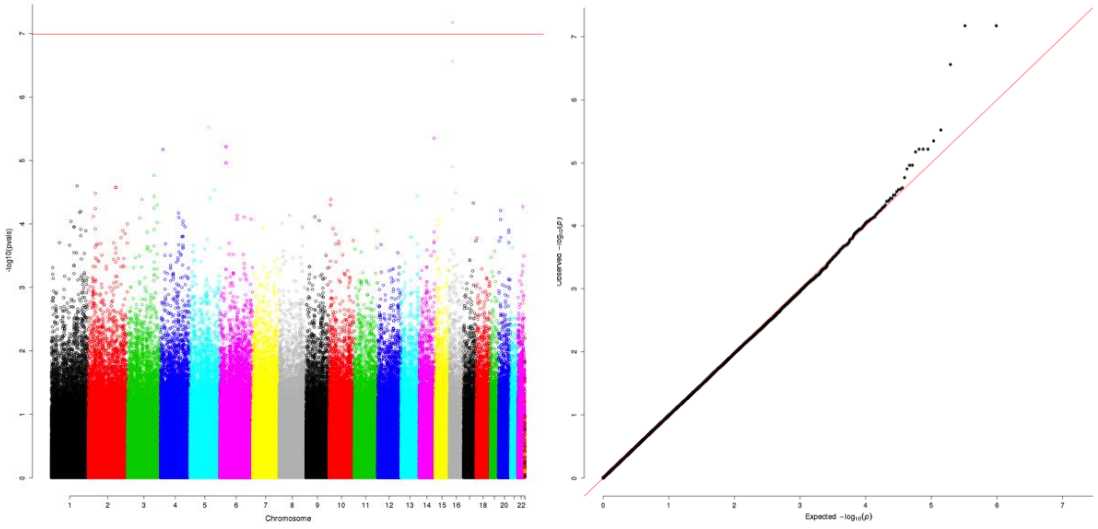
PC22



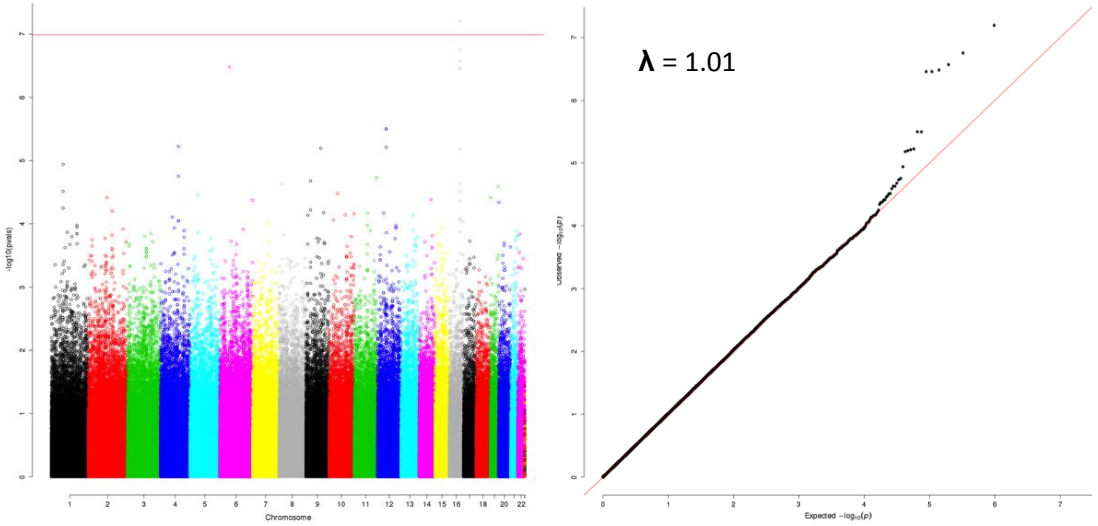
PC35



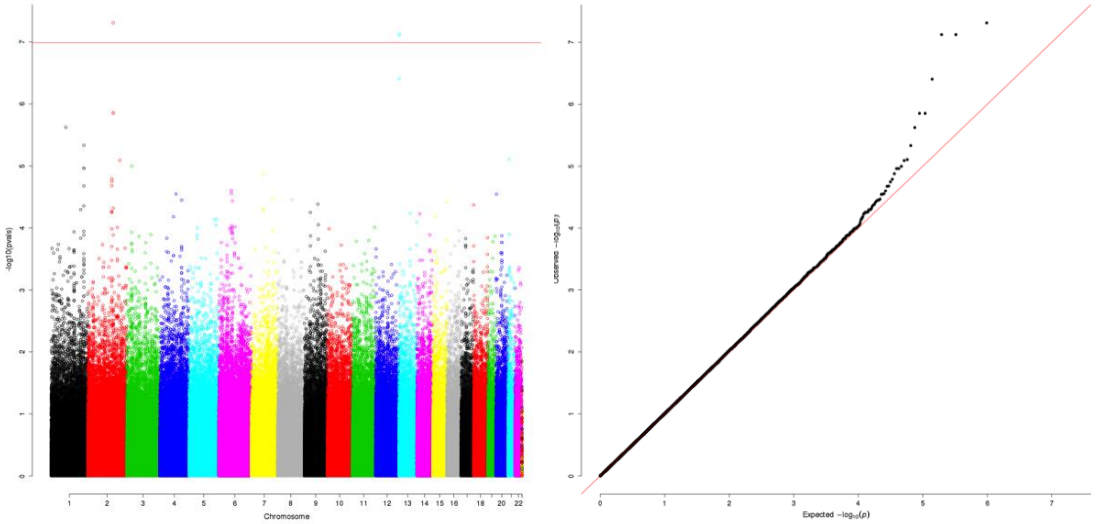
PC81



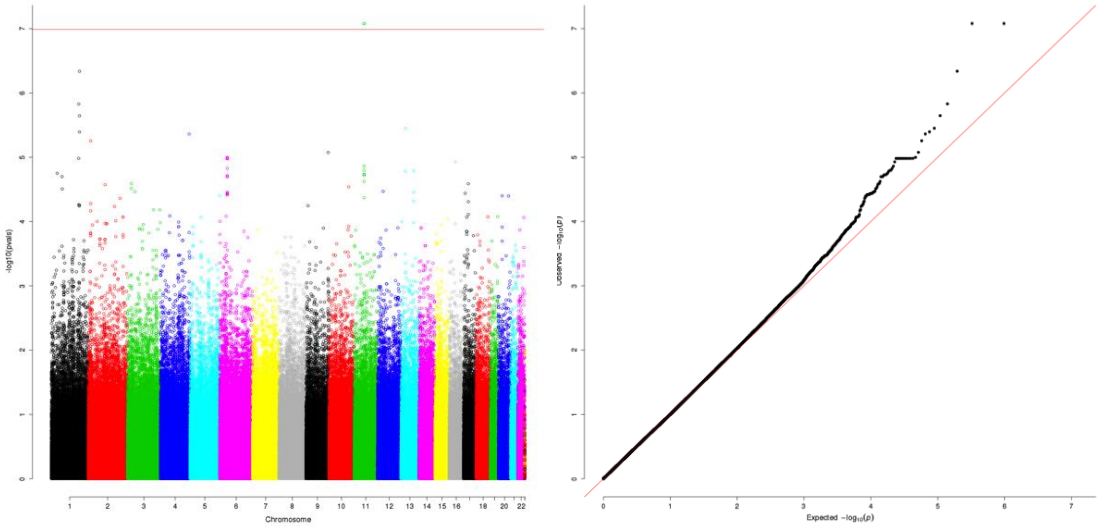
PC100



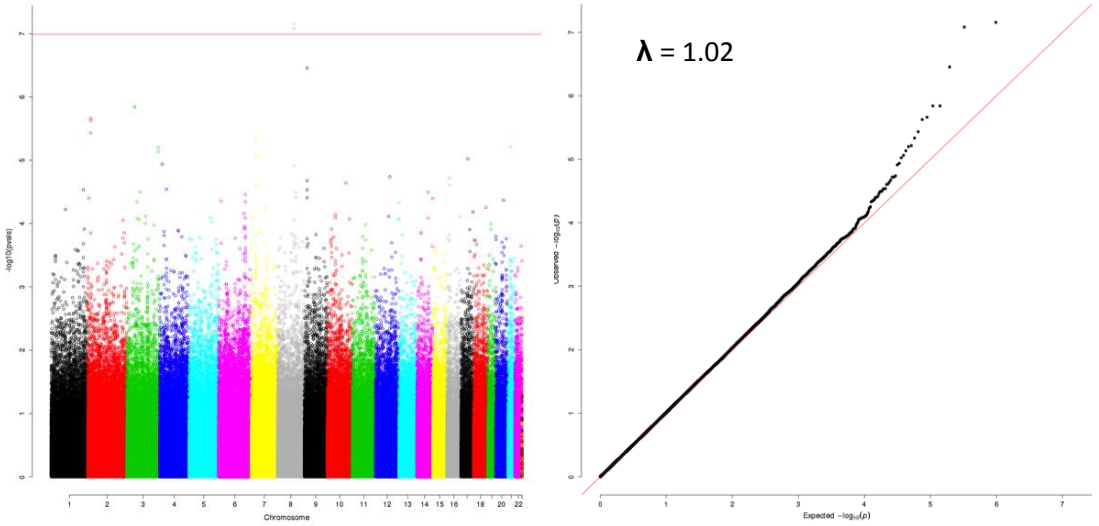
PC106



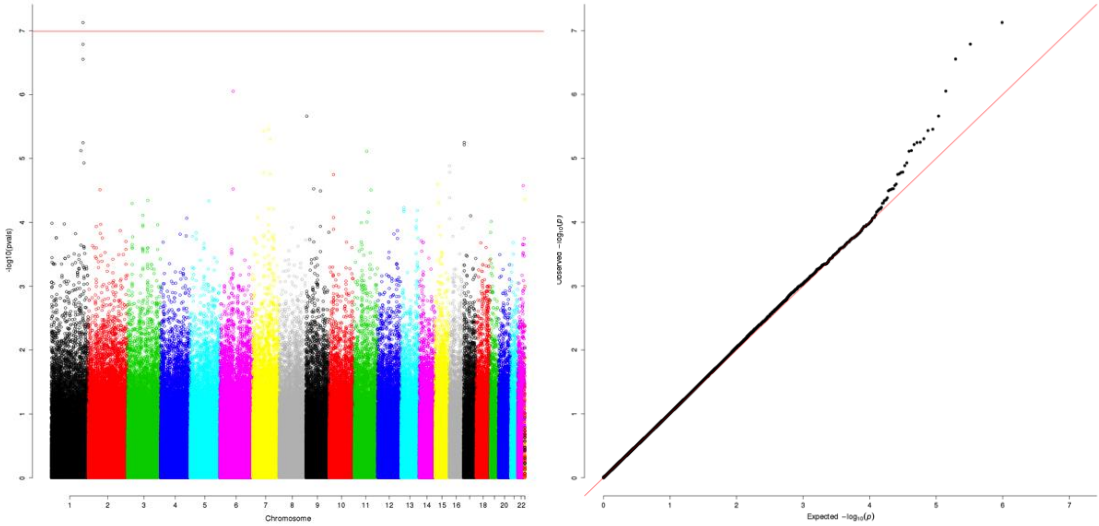
PC110



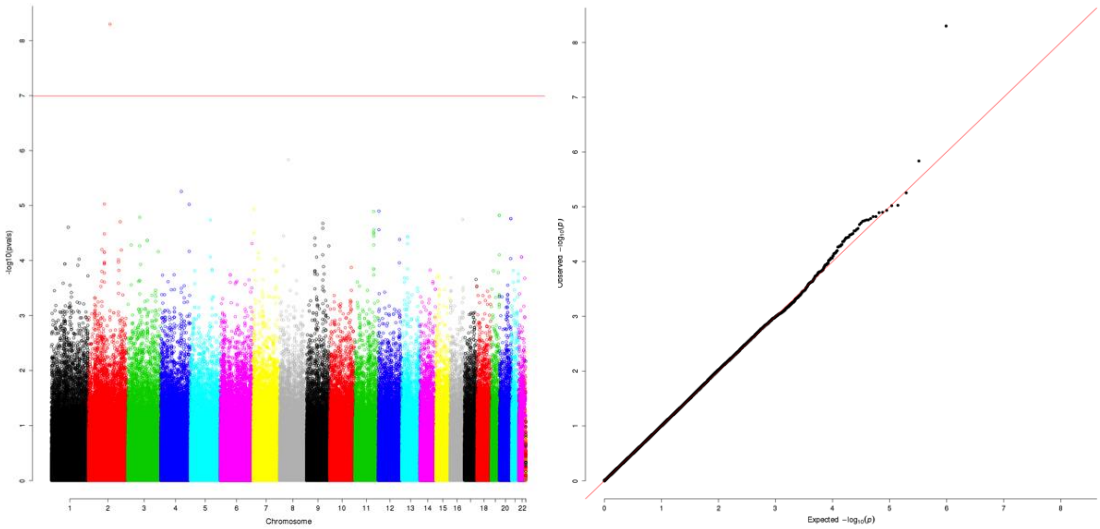
PC119



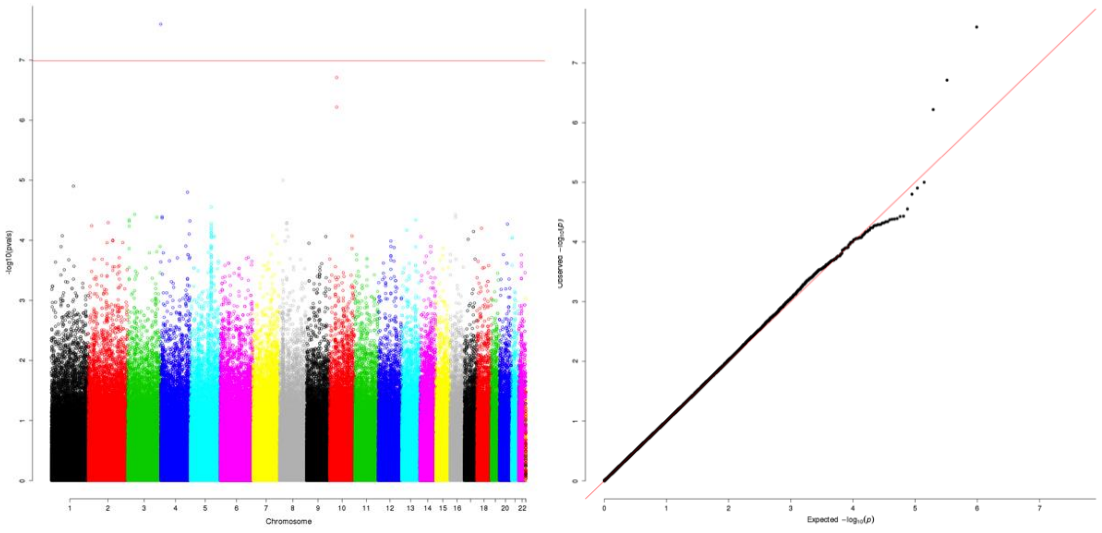
PC156



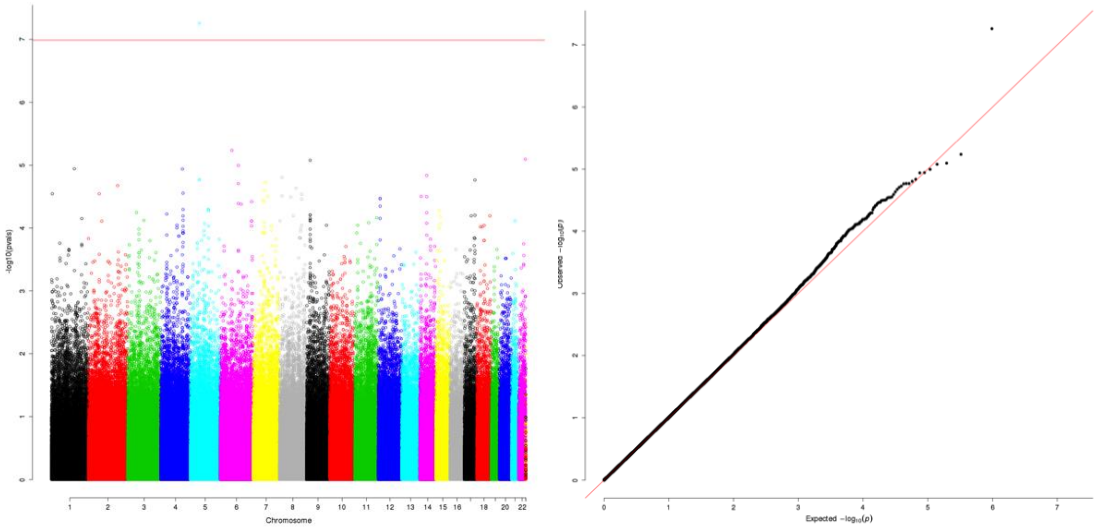
PC157



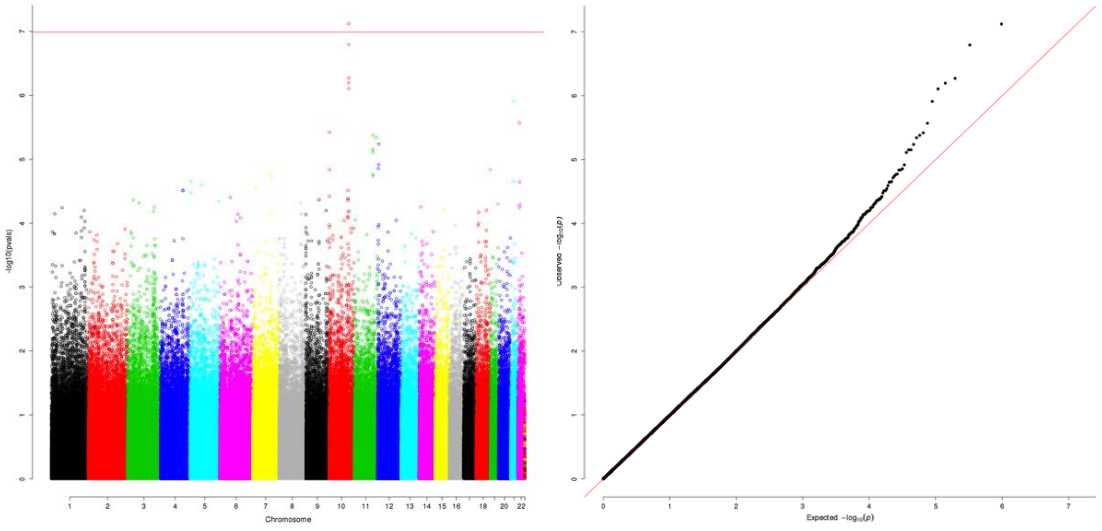
PC175



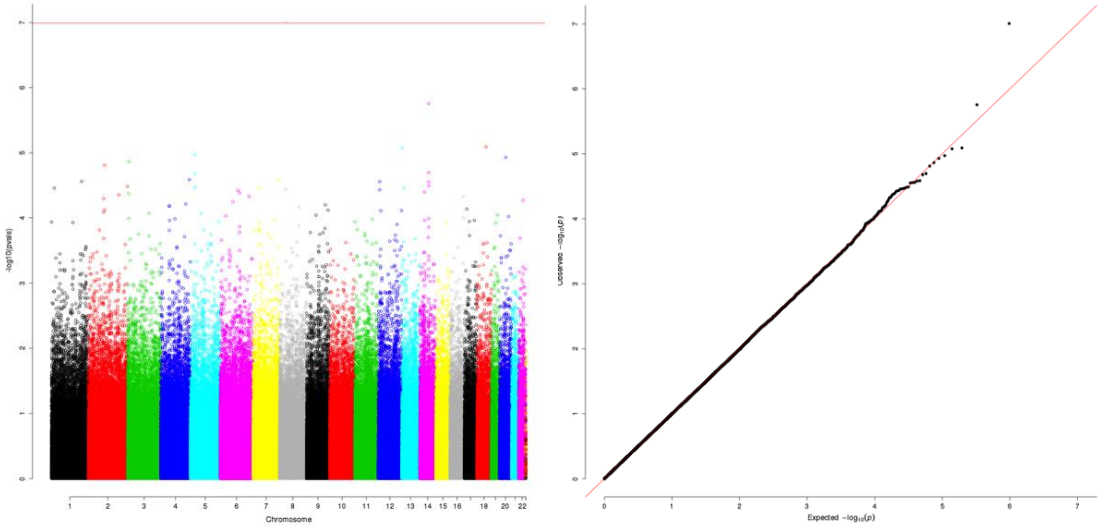
PC214



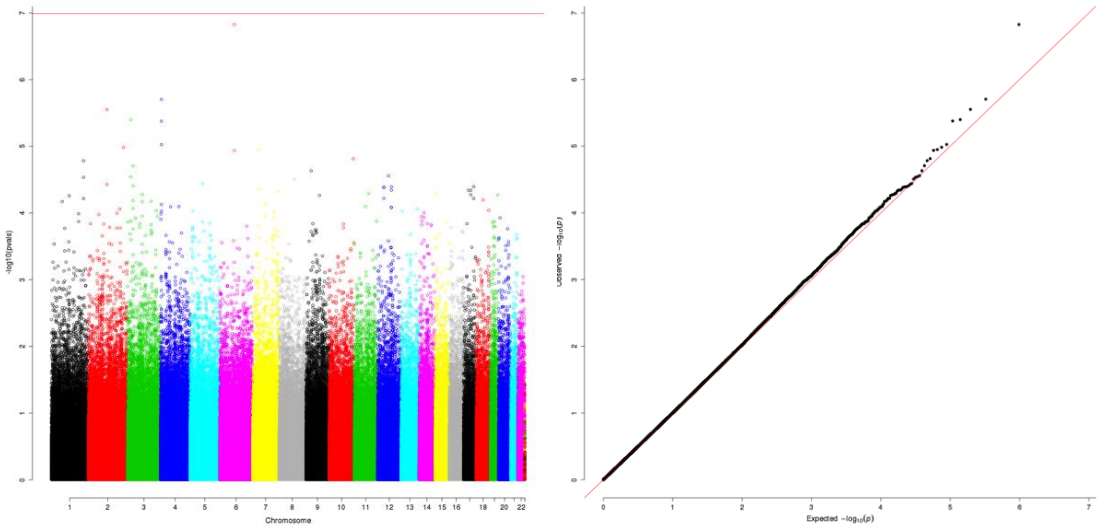
PC225



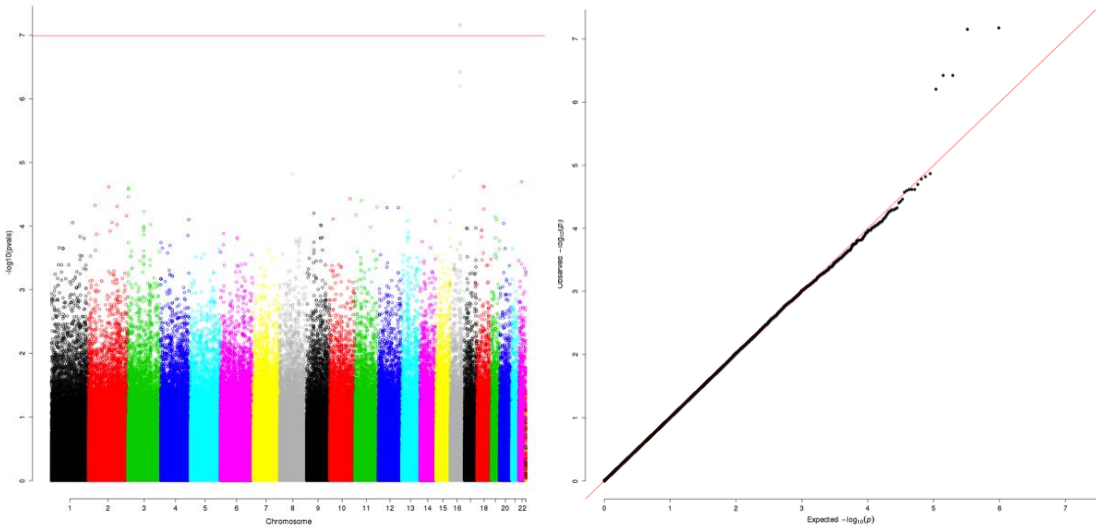
PC245



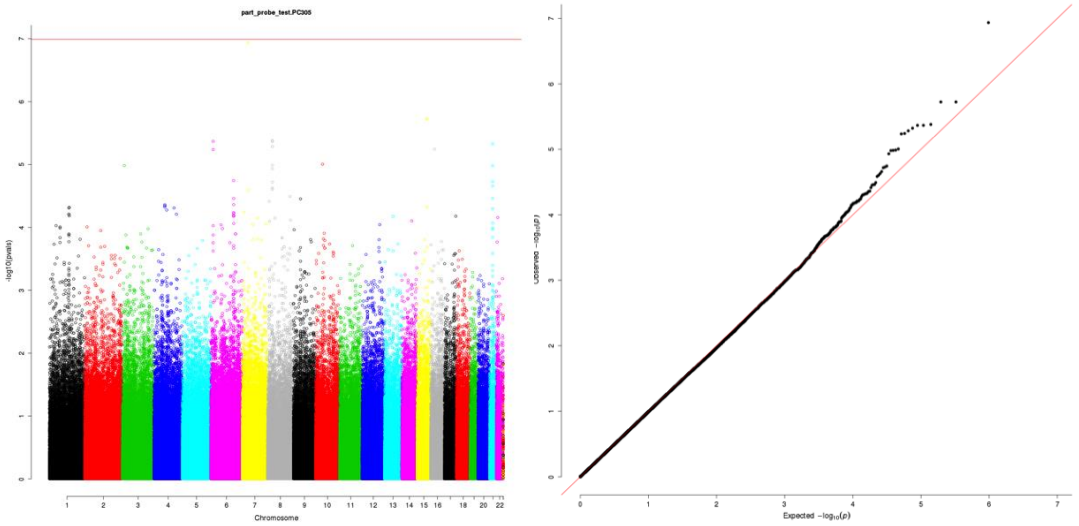
PC264



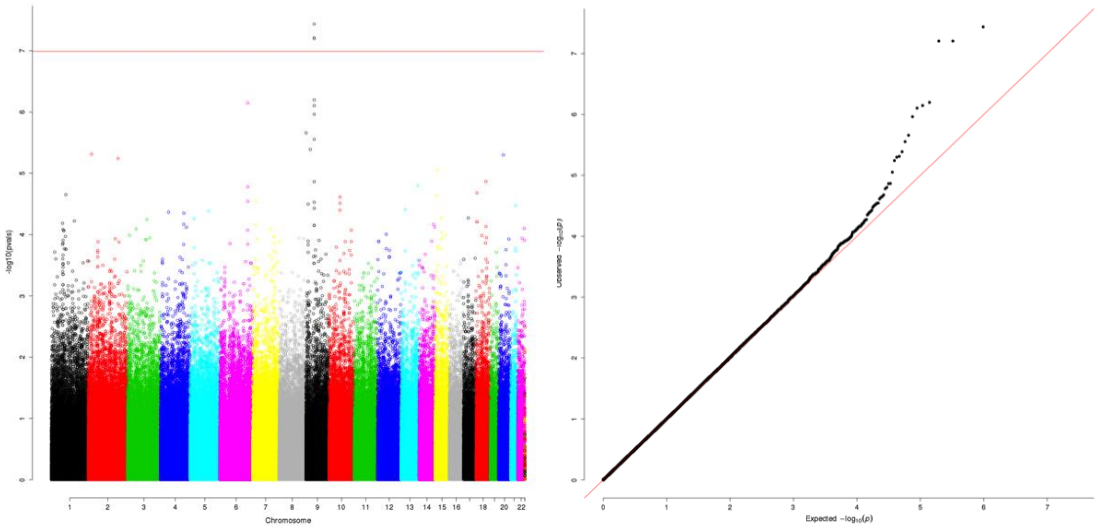
PC274



PC305



PC323



PC324

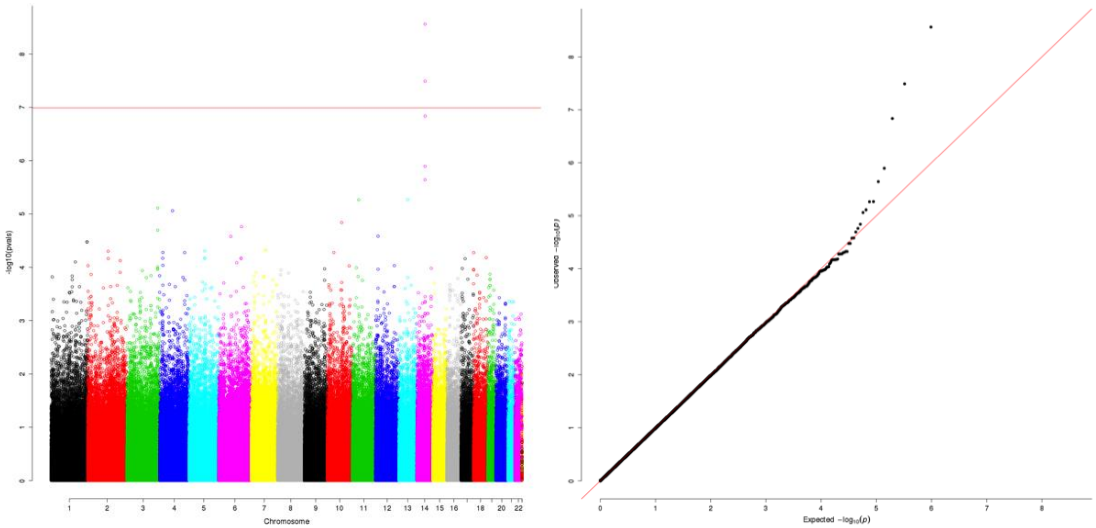
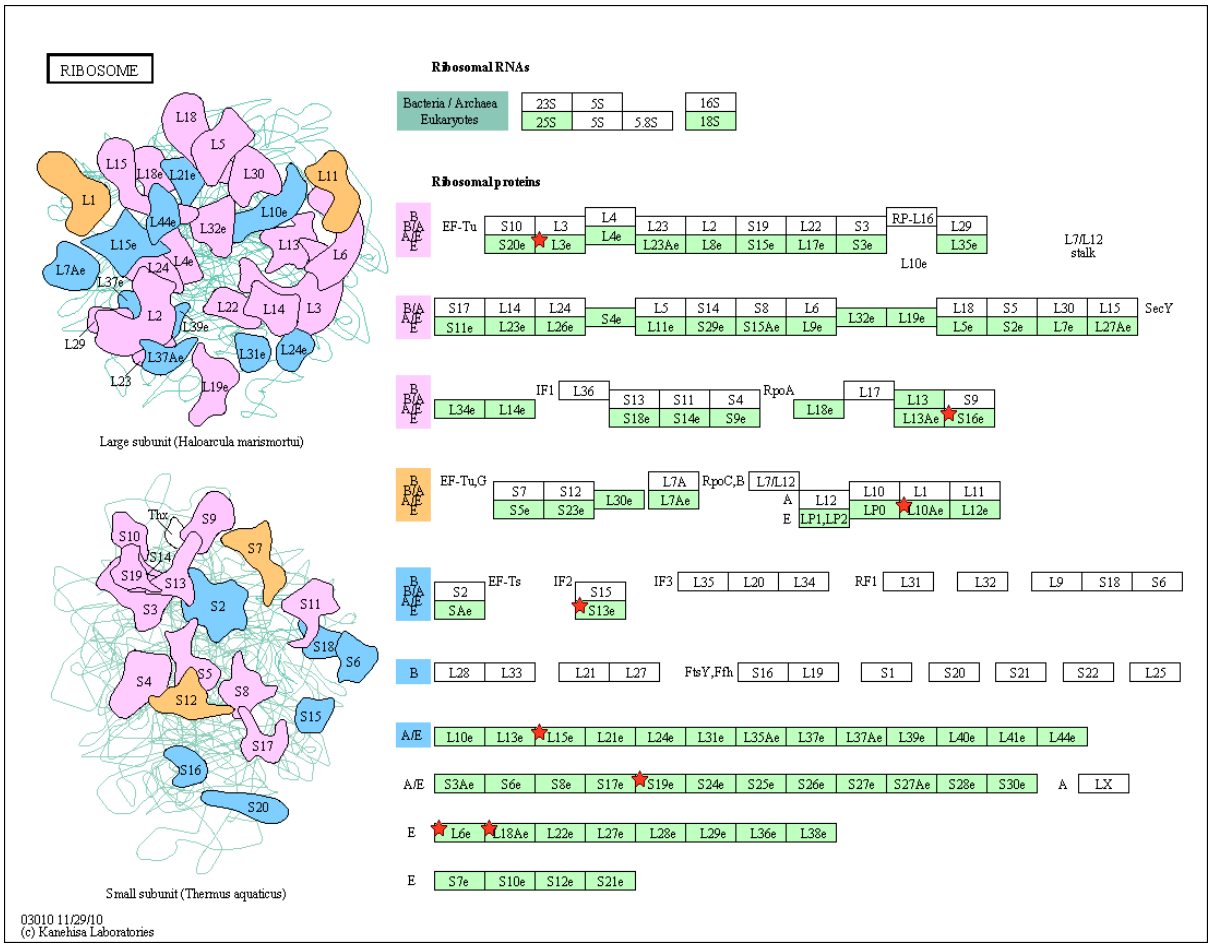
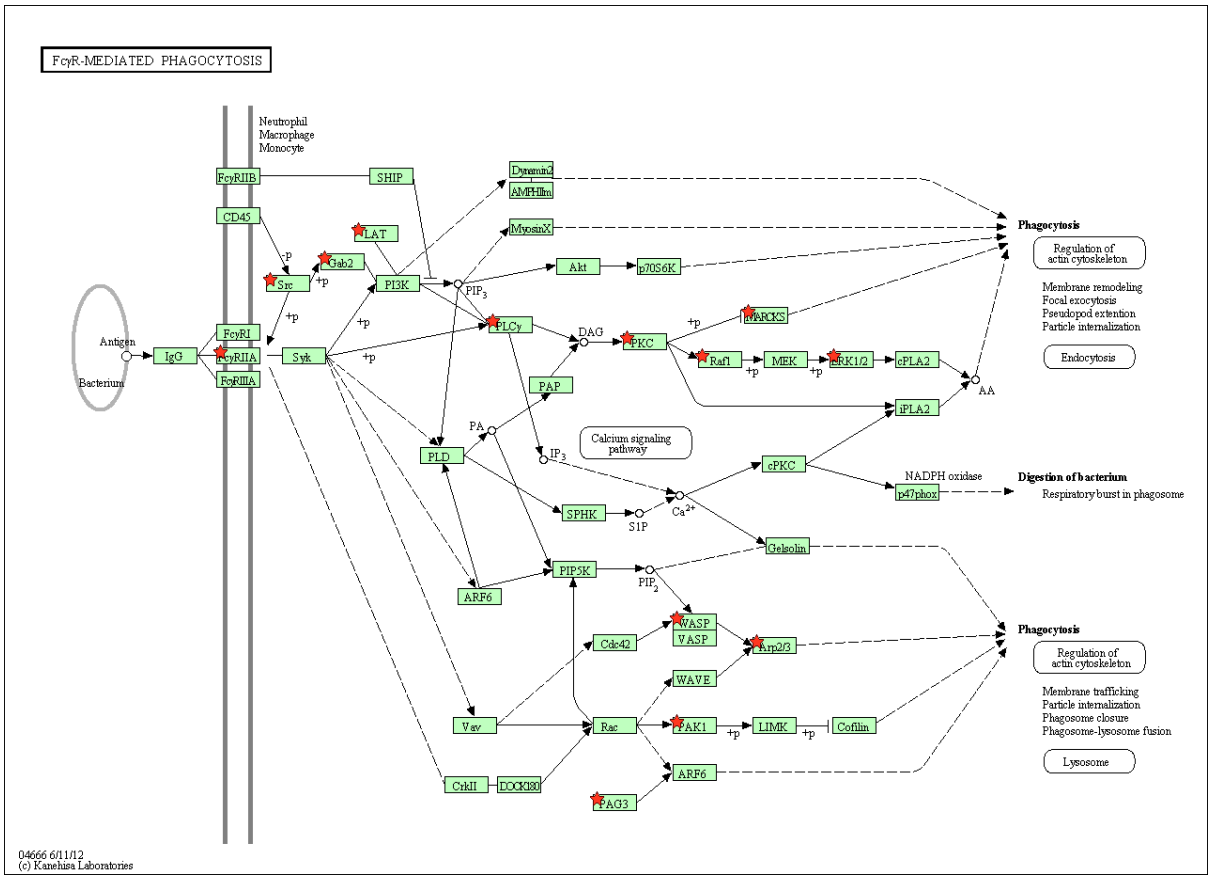


Figure S5 Manhattan and QQ plots. Manhattan and QQ plots for each PC with a significant SNP associated in Table S1. The significance value cut-off is drawn as a red line drawn on the Manhattan plots and is based on the Bonferroni correction for each PC. The QQ plot shows the expected p-values vs. the observed p-values in the study and the lambda value gives a numerical estimation of any inflation in the statistics.

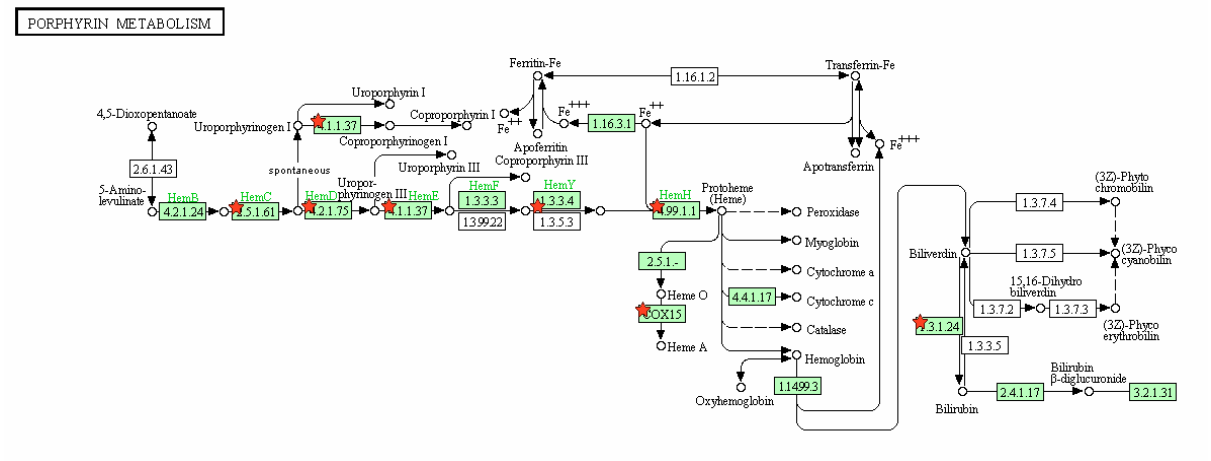
PC1


RIBOSOME
Large subunit (Haloarcula marismortui)
Small subunit (Thermus aquaticus)
Ribosomal RNAs
Bacteria / Archaea
Eukaryotes
23S
5S
5.8S
16S
18S
Ribosomal proteins
EF-Tu
RP-L16
L10e
L7/L12 stalk
SecY
IF1
RpoA
EF-Tu,G
L7A
L7Ae
RpoC,B
L7/L12
L12
LP1,LP2
EF-Ts
IF2
IF3
RF1
FtsY,Ffh
Thx
03010 11/29/10
(c) Kanehisa Laboratories

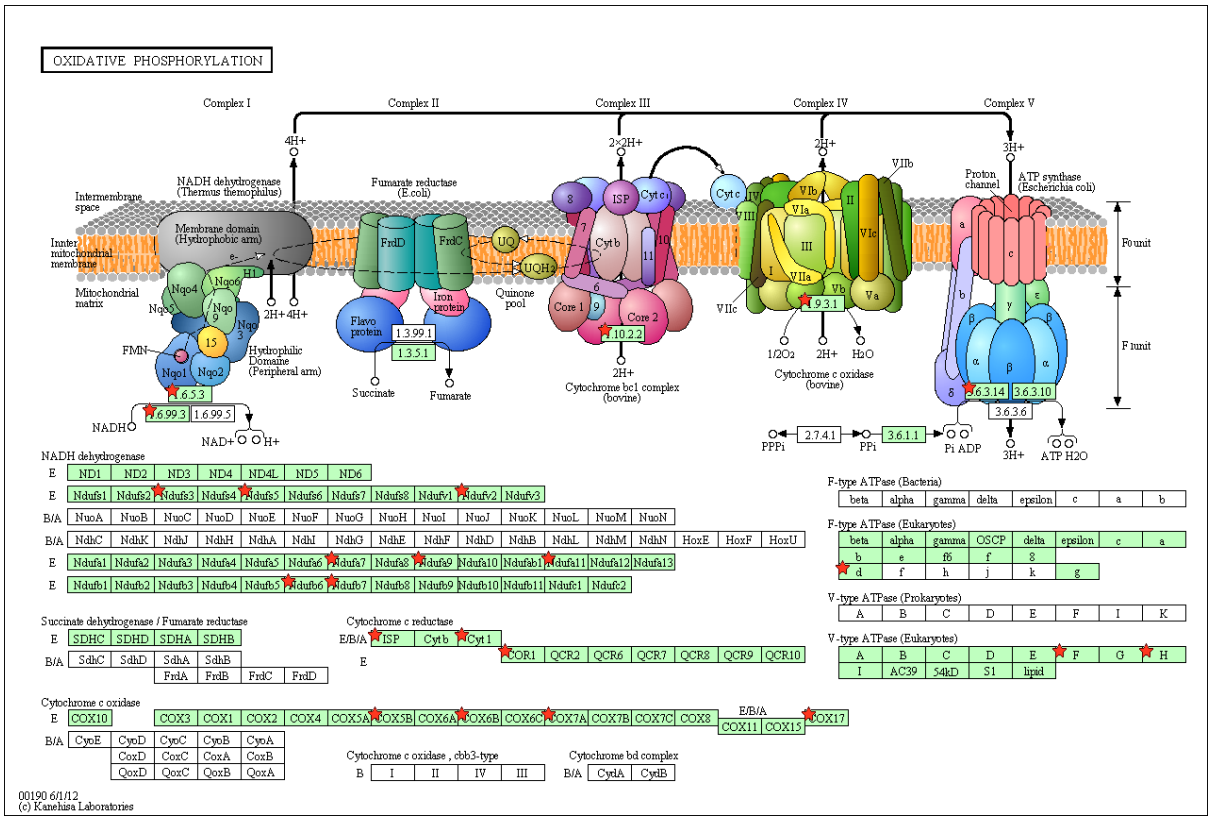
PC3


FcγR-MEDIATED PHAGOCYTOSIS
Neutrophil
Macrophage
Monocyte
FcγRIIB
SHIP
CD45
LAT
Gab2
Src
PI3K
Dynamin2
AMPHIIm
MyosinX
Akt
p70S6K
Phagocytosis
Regulation of actin cytoskeleton
Membrane remodeling
Focal exocytosis
Pseudopod extention
Particle internalization
Endocytosis
Antigen
IgG
Bacterium
FcγRI
FcγRIIA
FcγRIIIA
Syk
PLCγ
DAG
PKC
WASPCKS
Raf1
MEK
ERK1/2
cPLA2
AA
PAP
PA
iPLA2
PLD
Calcium signaling pathway
IP3
cPKC
NADPH oxidase
p47phox
Digestion of bacterium
Respiratory burst in phagosome
SPHK
S1P
Ca2+
Gelsolin
PIP5K
PIP2
ARF6
WASP
VASP
Cdc42
Arp2/3
Phagocytosis
Regulation of actin cytoskeleton
WAVE
Membrane trafficking
Particle internalization
Phagosome closure
Phagosome-lysosome fusion
Vav
Rac
PAK1
LIMK
Cofilin
Lysosome
CrkII
DOCK180
PAG3
04666 6/11/12
(c) Kanehisa Laboratories

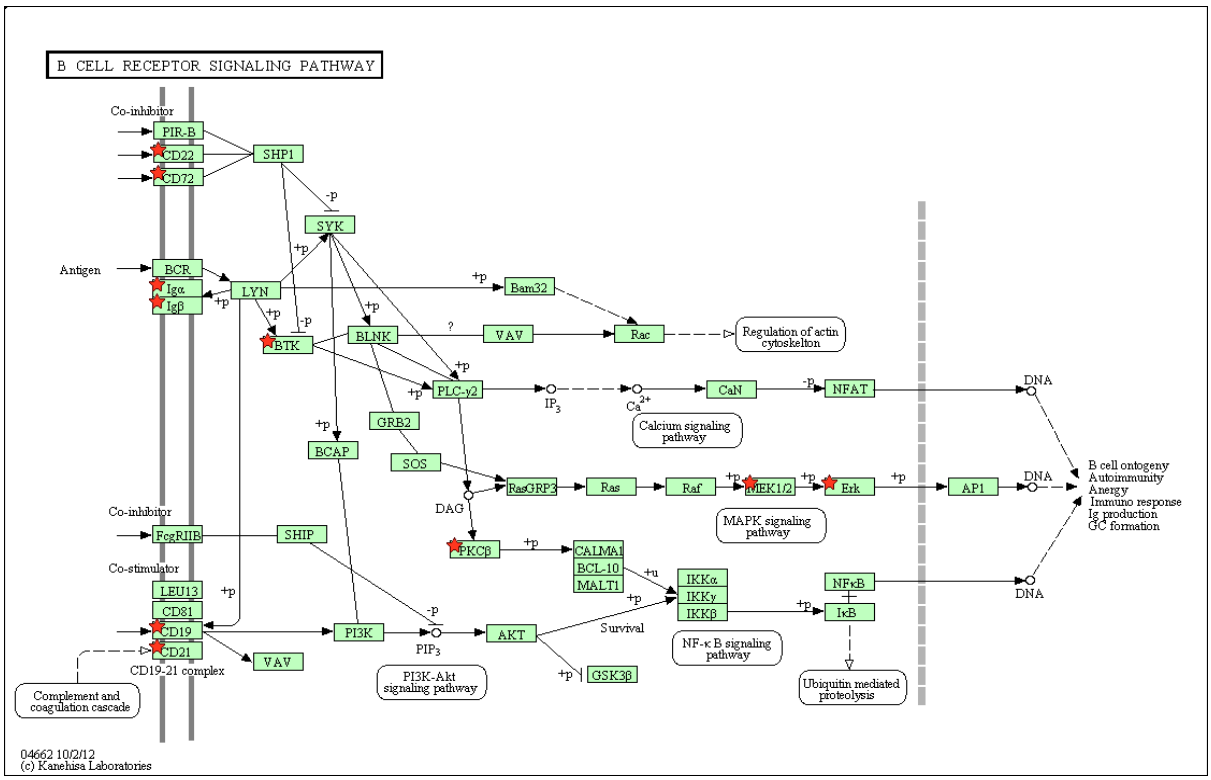
PC7

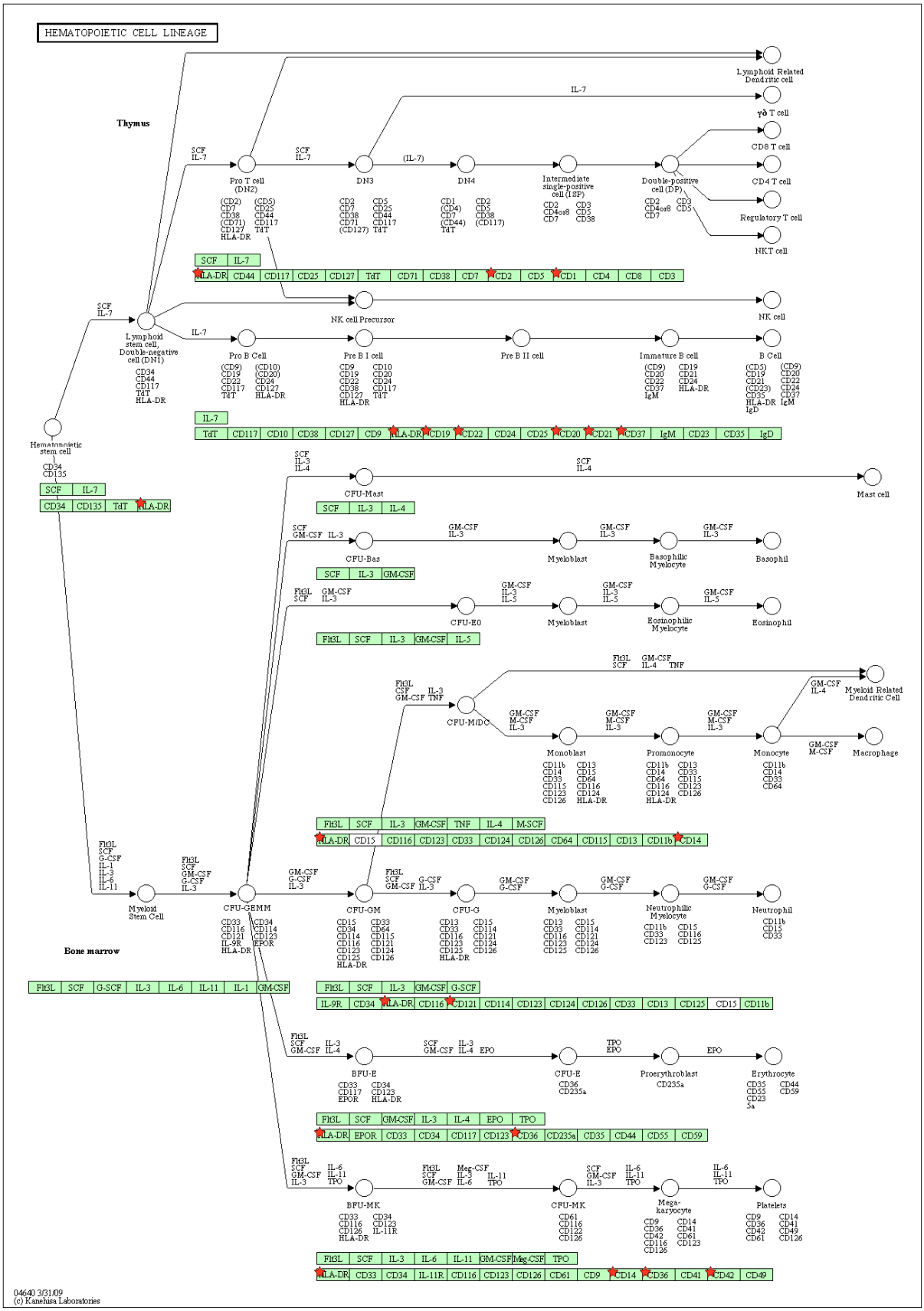

PORPHYRIN METABOLISM
Ferritin-Fe
1.16.1.2
Transferrin-Fe
Uroporphyrin I
4,5-Dioxopentanoate
Uroporphyrinogen I
4.1.1.37
Coproporphyrin I
Fe++
Fe+++
1.16.3.1
Coproporphyrinogen I
Apoferritin
Coproporphyrin III
Apotransferrin
2.6.1.43
spontaneous
Uroporphyrin III
5-Amino-levulinate
HemB
4.2.1.24
HemC
2.5.1.61
HemD
4.2.1.75
Uropor-phyrinogen III
HemE
4.1.1.37
HemF
1.3.3.3
1.3.99.22
HemY
1.3.3.4
1.3.5.3
HemH
4.99.1.1
Protoheme (Heme)
Peroxidase
Myoglobin
Cytochrome a
Cytochrome c
Catalase
Hemoglobin
2.5.1.-
Heme O
4.4.1.17
COX15
Heme A
1.14.99.3
Oxyhemoglobin
Biliverdin
1.3.7.4
(3Z)-Phyto chromobilin
1.3.7.5
(3Z)-Phyco cyanobilin
15,16-Dihydro biliverdin
1.3.7.2
1.3.7.3
(3Z)-Phyco erythrobilin
1.3.1.24
1.3.3.5
Bilirubin β-diglucuronide
2.4.1.17
3.2.1.31
Bilirubin

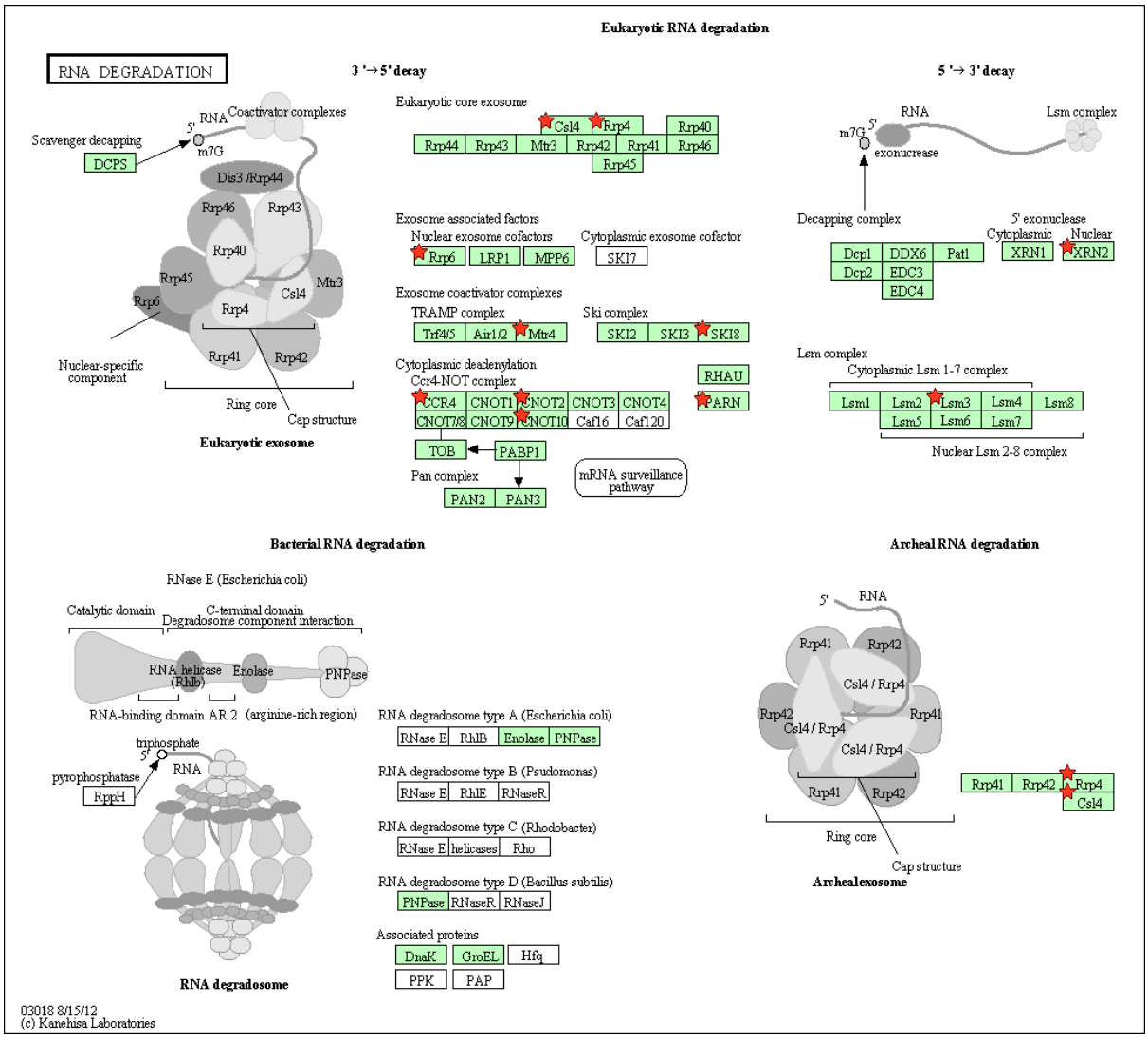
PC8

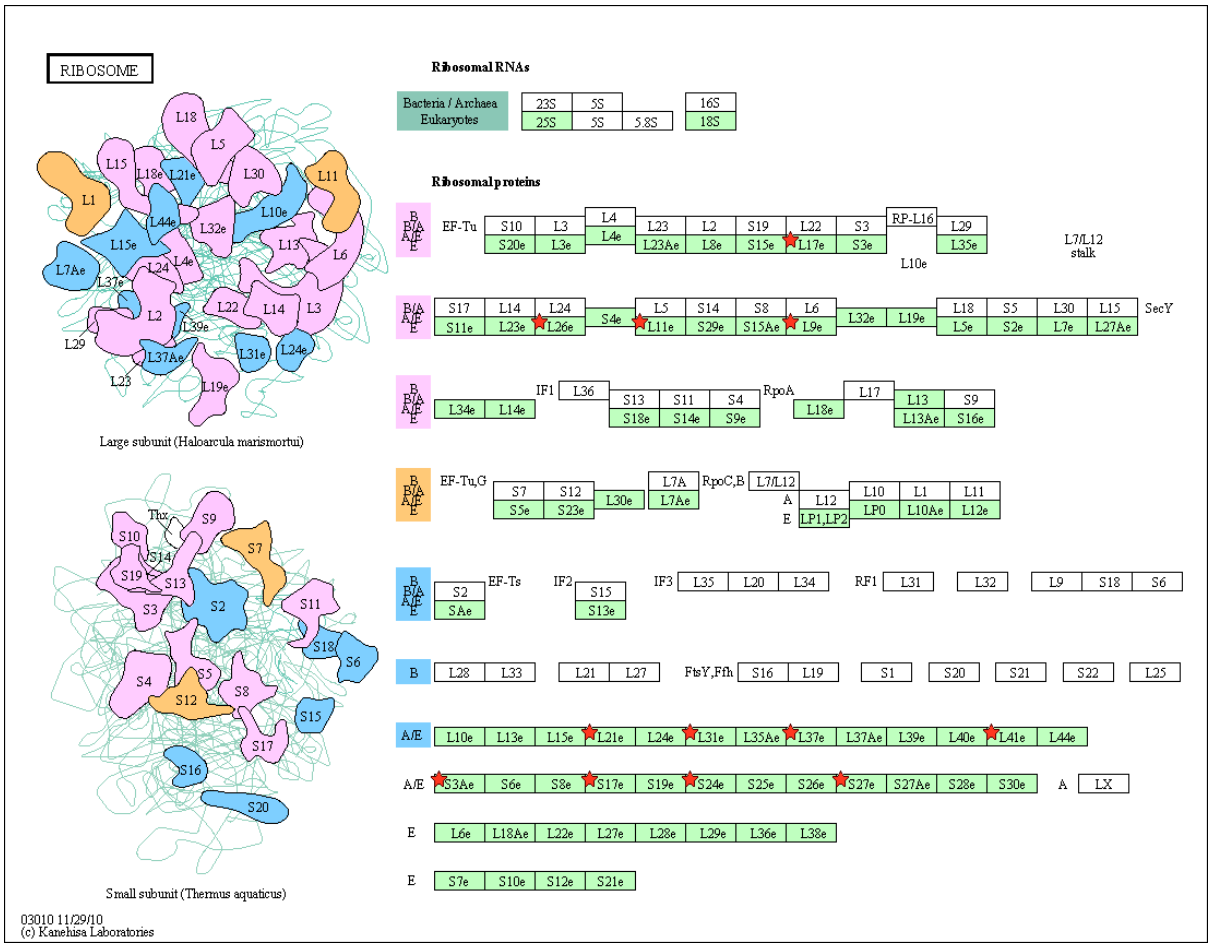

OXIDATIVE PHOSPHORYLATION
Complex I
Complex II
Complex III
Complex IV
Complex V
NADH dehydrogenase (Thermus thermophilus)
Fumarate reductase (E.coli)
Cytochrome bc1 complex (bovine)
Cytochrome c oxidase (bovine)
Proton channel
ATP synthase (Escherichia coli)
Intermembrane space
Inner mitochondrial membrane
Mitochondrial matrix
Membrane domain (Hydrophobic arm)
Hydrophilic Domain (Peripheral arm)
Quinone pool
Iron protein
Flavo protein
Succinate
Fumarate
NADH
NAD+
FMN
F0 unit
F1 unit
NADH dehydrogenase
Succinate dehydrogenase / Fumarate reductase
Cytochrome c reductase
Cytochrome c oxidase
Cytochrome c oxidase , cbb3-type
Cytochrome bd complex
F-type ATPase (Bacteria)
F-type ATPase (Eukaryotes)
V-type ATPase (Prokaryotes)
V-type ATPase (Eukaryotes)
00190 6/1/12
(c) Kanehisa Laboratories

PC12


B CELL RECEPTOR SIGNALING PATHWAY
Co-inhibitor
PIR-B
CD22
CD72
SHP1
SYK
Antigen
BCR
Igα
Igβ
LYN
Bam32
BTK
BLNK
VAV
Rac
Regulation of actin cytoskelton
PLC-γ2
CaN
NFAT
DNA
Calcium signaling pathway
GRB2
BCAP
SOS
RasGRP3
Ras
Raf
MEK1/2
Erk
AP1
DAG
MAPK signaling pathway
B cell ontogeny
Autoimmunity
Anergy
Immuno response
Ig production
GC formation
Co-inhibitor
FcgRIIB
SHIP
PKCβ
CALMA1
BCL-10
MALT1
IKKα
IKKγ
IKKβ
NFκB
IκB
Co-stimulator
LEU13
CD81
CD19
CD21
CD19-21 complex
PI3K
AKT
PIP3
Survival
GSK3β
NF-κ B signaling pathway
PI3K-Akt signaling pathway
Ubiquitin mediated proteolysis
Complement and coagulation cascade
04662 10/2/12
(c) Kanehisa Laboratories


HEMATOPOIETIC CELL LINEAGE
Thymus
Bone marrow
Lymphoid Related Dendritic cell
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
NK cell Precursor
NK cell
Pro T cell (DN2)
DN3
DN4
Intermediate single-positive cell (ISP)
Double-positive cell (DP)
Lymphoid stem cell, Double-negative cell (DN1)
Pro B Cell
Pre B I cell
Pre B II cell
Immature B cell
B Cell
Hematopoietic stem cell
CFU-Mast
Mast cell
CFU-Bas
Myeloblast
Basophilic Myelocyte
Basophil
CFU-E0
Eosinophilic Myelocyte
Eosinophil
CFU-M/DC
Monoblast
Promonocyte
Monocyte
Myeloid Related Dendritic Cell
Macrophage
Myeloid Stem Cell
CFU-GEMM
CFU-GM
CFU-G
Neutrophilic Myelocyte
Neutrophil
BFU-E
CFU-E
Proerythroblast
Erythrocyte
BFU-MK
CFU-MK
Mega-karyocyte
Platelets
0464O 3/31/09
(c) Kanehisa Laboratories


Eukaryotic RNA degradation
RNA DEGRADATION
3 '→ 5' decay
5 '→ 3' decay
Coactivator complexes
RNA
5'
Scavenger decapping
DCPS
m7G
Dis3 /Rrp44
Rrp46
Rrp43
Rrp40
Rrp45
Csl4
Mtr3
Rrp6
Rrp4
Rrp41
Rrp42
Nuclear-specific component
Ring core
Cap structure
Eukaryotic exosome
Eukaryotic core exosome
Csl4
Rrp4
Rrp40
Rrp44
Rrp43
Mtr3
Rrp42
Rrp41
Rrp46
Rrp45
Exosome associated factors
Nuclear exosome cofactors
Cytoplasmic exosome cofactor
Rrp6
LRP1
MPP6
SKI7
Exosome coactivator complexes
TRAMP complex
Ski complex
Trf4/5
Air1/2
Mtr4
SKI2
SKI3
SKI8
Cytoplasmic deadenylation
Ccr4-NOT complex
RHAU
CCR4
CNOT1
CNOT2
CNOT3
CNOT4
PARN
CNOT7/8
CNOT9
CNOT10
Caf16
Caf120
TOB
PABP1
Pan complex
PAN2
PAN3
mRNA surveillance pathway
RNA
Lsm complex
m7G
5'
exonucrease
Decapping complex
5' exonuclease
Cytoplasmic
Nuclear
Dcp1
DDX6
Pat1
XRN1
XRN2
Dcp2
EDC3
EDC4
Lsm complex
Cytoplasmic Lsm 1-7 complex
Lsm1
Lsm2
Lsm3
Lsm4
Lsm8
Lsm5
Lsm6
Lsm7
Nuclear Lsm 2-8 complex
Bacterial RNA degradation
Archeal RNA degradation
RNase E (Escherichia coli)
Catalytic domain
C-terminal domain
Degradosome component interaction
RNA helicase (RhlB)
Enolase
PNPase
RNA-binding domain
AR 2 (arginine-rich region)
triphosphate
5'
RNA
pyrophosphatase
RppH
RNA degradosome
RNA degradosome type A (Escherichia coli)
RNase E
RhlB
Enolase
PNPase
RNA degradosome type B (Psudomonas)
RNase E
RhlE
RNaseR
RNA degradosome type C (Rhodobacter)
RNase E
helicases
Rho
RNA degradosome type D (Bacillus subtilis)
PNPase
RNaseR
RNaseJ
Associated proteins
DnaK
GroEL
Hfq
PPK
PAP
5'
RNA
Rrp41
Rrp42
Csl4 / Rrp4
Rrp42
Rrp41
Csl4 / Rrp4
Csl4 / Rrp4
Rrp41
Rrp42
Ring core
Cap structure
Archealexosome
Rrp41
Rrp42
Rrp4
Csl4
03018 8/15/12
(c) Kanehisa Laboratories


RIBOSOME
Large subunit (Haloarcula marismortui)
Small subunit (Thermus aquaticus)
Ribosomal RNAs
Bacteria / Archaea
Eukaryotes
23S
5S
5.8S
16S
18S
Ribosomal proteins
EF-Tu
RP-L16
L10e
L7/L12 stalk
SecY
IF1
RpoA
EF-Tu,G
L7/L12
EF-Ts
IF2
IF3
RF1
FtsY,Ffh
03010 11/29/10
(c) Kanehisa Laboratories

PC25

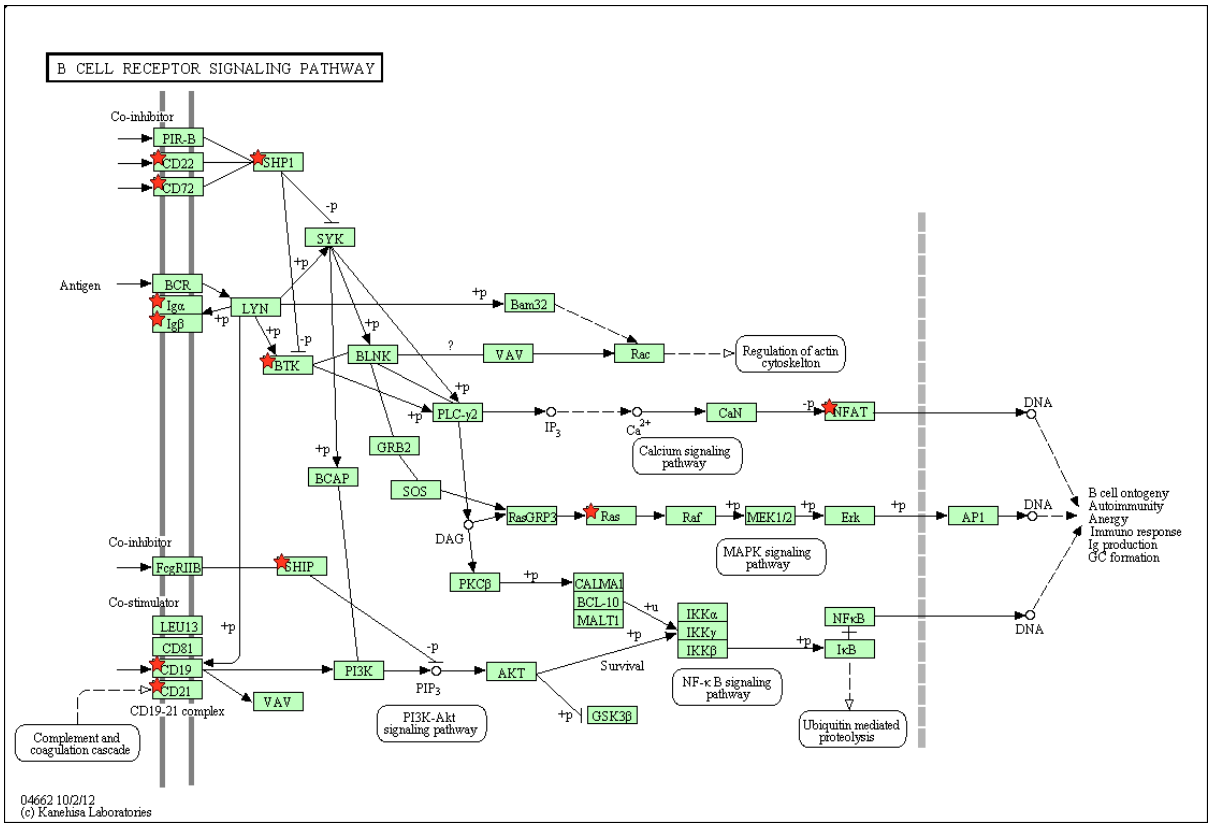


Figure S6 Pathway diagrams of enriched biological networks. Figures generated from the KEGG pathway database. Pathway analysis for PC1-50 was performed using DAVID Bioinformatics Resources 6.7, Functional Annotation Tool. These pathways were significant after multiple correction (FDR) (Table S2). Components highlighted with red stars represent probes present within the corresponding PC. PC1 shows enrichment for ribosomal components. PC3 is enriched for Fc gamma R-mediated phagocytosis. PC7 is enriched for porphyrin metabolism. PC8 shows enrichment for enzymatic subunits involved oxidative phosphorylation, PC12 is enriched in components involved in the B-cell receptor signaling pathway and hematopoietic cell lineage. PC13 is enriched for RNA degradation. PC18 shows enrichment for ribosomal components and PC25 shows enrichment for B-cell reception signaling.

Table S1 Results from the GWAS for each PC. Probes that were found to be significant after a Bonferroni correction (0.05/488,462 SNPs) on each PC are listed in this table. Though none of these are significant after correcting for all PCs, they are significant at an empirical p-value of 0.05 for each PC after 1000 permutations. PC – principal component, CHR – chromosome, SNP – SNP ID, BP – base pair, BETA – regression coefficient, STAT – Coefficient T-statistic, P – Asymptotic p-value for t-statistic, EMP – empirical p-value after 1000 permutations.

PC	CHR	SNP	BP	BETA	STAT	P	EMP
22	10	rs11004899	56960734	3.754	5.487	8.14E-08	0.036
35	2	rs1516174	51724845	2.826	5.43	1.089E-07	0.049
81	16	rs9673242	14078070	-3.404	-5.524	6.69E-08	0.021
81	16	rs1004637	14113245	-3.404	-5.524	6.69E-08	0.021
100	16	rs7190803	77375823	-1.444	-5.535	6.33E-08	0.021
106	2	rs10497190	158347486	-1.741	-5.585	4.87E-08	0.021
106	13	rs17072974	21351926	2.275	5.501	7.53E-08	0.027
106	13	rs12428031	21355249	2.275	5.501	7.53E-08	0.027
110	11	rs10501384	59950456	2.555	5.482	8.31E-08	0.033
110	11	rs17542525	59958103	2.555	5.482	8.31E-08	0.033
119	8	rs4596672	88124581	-1.641	-5.488	8.23E-08	0.032
119	8	rs2974279	88144159	-1.385	-5.517	6.95E-08	0.028
156	1	rs825113	221564768	1.762	5.503	7.45E-08	0.026
157	2	rs11674634	132055980	-1.305	-6.005	5.02E-09	0.004
175	4	rs6848983	298010	1.736	5.71	2.50E-08	0.005
214	5	rs1279627	55966337	-1.095	-5.562	5.50E-08	0.019
225	10	rs7919814	109720733	-1.055	-5.502	7.52E-08	0.03
245	8	rs17128272	19257994	-1.516	-5.449	9.85E-08	0.042
323	9	rs10813262	30474037	1.542	5.538	6.22E-08	0.044
323	9	rs4878432	30490252	1.542	5.538	6.22E-08	0.044
323	9	rs7866981	30548222	1.568	5.639	3.65E-08	0.034
324	14	rs10498517	64832534	1.619	6.112	2.74E-09	0.002
324	14	rs4902382	64834310	1.44	5.662	3.24E-08	0.017

Table S2 Pathway analysis for the first 50 PCs Pathway analysis for PC1-50 was performed using DAVID Bioinformatics Resources 6.7, Functional Annotation Tool. PC – principal component, Term – name of KEGG pathway, Count – count of probes in each hit, % – percentage of all probes submitted for that PC that are present within the pathway, P – the p-value that is calculated using a modified Fischer's exact test for enrichment, FDR – correction of p-values and using the Benjamini-Hochberg FDR method.

PC	Term	Count	%	P	FDR
1	Ribosome	8	7.9	1.00E-06	4.50E-05
3	Fc gamma R-mediated phagocytosis	14	2.7	4.30E-05	5.40E-03
7	Porphyrin metabolism	7	1.5	1.60E-04	2.00E-02
8	Proteasome	12	2.6	2.70E-07	4.00E-05
	Oxidative phosphorylation	18	3.9	1.20E-06	8.70E-05
	Huntington's disease	21	4.6	1.90E-06	9.00E-05
	Parkinson's disease	15	3.3	8.50E-05	3.10E-03
	Alzheimer's disease	15	3.3	1.10E-03	3.00E-02
12	Hematopoietic cell lineage	13	2.5	1.40E-05	2.00E-03
	B cell receptor signaling pathway	10	1.9	5.60E-04	3.80E-02
	Antigen processing and presentation	10	1.9	1.20E-03	5.30E-02
	Graft-versus-host disease	7	1.3	1.30E-03	4.40E-02
	Non-small cell lung cancer	8	1.5	1.50E-03	4.00E-02
	Asthma	6	1.2	2.00E-03	4.40E-02
13	RNA degradation	11	2.3	1.30E-05	1.80E-03
	Oxidative phosphorylation	15	3.1	7.80E-05	5.60E-03
	Ribosome	11	2.3	5.00E-04	2.40E-02
18	Ribosome	14	2.9	9.40E-07	1.00E-04

24	B cell receptor signaling pathway	14	2.9	2.00E-07	2.70E-05
25	B cell receptor signaling pathway	11	2.2	1.10E-04	1.60E-02
26	Primary immunodeficiency	8	1.7	7.90E-05	1.20E-02
32	Oxidative phosphorylation	13	2.7	2.90E-04	3.70E-02