

Chapter 3

Using latent factors to enhance power in mapping expression QTLs for ageing

Collaboration note. *This chapter contains work in collaboration with Andrew Brown. The manuscript of this work is accepted by G3 subject to minor revisions. I am the joint lead author with Andrew Brown. My contribution includes processing microarray probe intensities, normalizations, learning both global and pathway factors, association analysis, and manuscript writing.*

3.1 Overview

Ageing is a multifactorial process, reflecting how the physical state of an organism accumulates changes and gene expression has been a field of increasing interest in studying this process. Microarrays and more recent RNA-seq technologies allow the simultaneous quantification of cell population average mRNA abundance for thousands of genes. These technologies have proved useful in providing diagnostic profiles for certain diseases ([Reis-Filho and Pusztai, 2011](#)). In the case of ageing, consistent patterns of age-related changes in gene expression have been observed across several

tissues and species (Lu et al., 2004), such as over-expression of inflammation and immune-response genes and under-expression of genes involved in energy metabolism in older samples (de Magalhães et al., 2009). Given this commonality of function amongst genes which show age related changes in expression, we decided to investigate ageing in the context of biological knowledge on the function of genes, as provided by pathway annotations.

Array expression experiments generate high dimensional structured data sets, in which there are correlated patterns across large numbers of genes. Some of these are due to known technical or biological effects such as batch effects and cell growth stage, which when not the focus of the analysis can be removed by including them as covariates. However, even after this, there is typically substantial structural correlation. In previous studies, these can be represented by linear components of expression measurements, or factors, that can be inferred using methods such as principal components analysis (PCA) or factor analysis to create global phenotypes (Leek and Storey, 2007; Parts et al., 2011). When the aim is to discover local effects, such as *cis* genetic regulation, these global phenotypes can be treated as nuisance variables and removed from further analysis. This has been seen to increase power in analysis (Montgomery et al., 2010; Pickrell et al., 2010; Stegle et al., 2010). Conversely, if the aim is to differentiate between a case and control condition using expression, then global phenotypes could be more effective classifiers than local phenotypes (Hastie et al., 2000).

A recent study applied factor analysis methods in a two stage procedure to generate phenotypes representing expressions of groups of genes (Stegle et al., 2010). After regressing out global factors, as in Parts et al. (2011), expression levels for groups of functionally related genes, as defined by annotations from pathway databases, were treated as new expression datasets and the same factor analysis methods were used to construct pathway factors. The factors constructed on pathway sets of genes were taken as concise summaries of common expression variation across each pathway. We test these factors values below as phenotypes, and so refer to them as phenotype factors as in some cases just phenotypes.

Here, we apply this method to gene expression data from abdominal skin tissues from 647 samples. Unlike previous studies which have concentrated on genetic variants

which regulate multiple genes within a pathway (Parts et al., 2011), we focused on discovering associations between gene expression and an environmental variable age. We obtain our pathway gene sets from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa et al., 2004). Subsequently, by looking for associations between these new pathway phenotypes and age, we discover groups of functionally related genes with a common response to ageing which could be used as biomarkers describing molecular changes with age.

With data from a twin cohort containing both monozygotic and dizygotic twins, we can estimate proportions of variance explained by age, genetic variation, common environmental variation, and unique environmental variation (noise). Stochasticity in gene expression, which will form part of the unique environment component, is believed to play a role in the ageing process (Bahar et al., 2006). By investigating sources of variation within the pathway phenotypes, we find that they are more robust than the expression of individual genes with less unique environment variation. This explains some of our success at discovering associations with age.

3.2 Expression profiling

The data analyzed here are part of the MuTHER project (Multiple Tissue Human Expression Resource - <http://www.muther.ac.uk/>, Nica et al., 2011) and were downloaded from the ArrayExpress archive, accession no. E-TABM-1140. In summary, the study included 856 Caucasian female individuals (336 monozygotic (MZ) and 520 dizygotic (DZ) twins) recruited from the TwinsUK Adult twin registry (Moayyeri et al., 2013a). The age at sampling ranged from 39 to 85 years with a mean age of 59 years. Punch biopsies (8mm) were taken from relatively photo-protected infra-umbilical skin. Subcutaneous adipose tissue was dissected from each biopsy and the remaining skin tissue was weighed and stored in liquid nitrogen. Expression profiling of this skin tissue was performed using Illumina Human HT-12 V3 BeadChips where 200ng of total RNA was processed according to the protocol supplied by Illumina. All samples were randomized prior to array hybridization and the technical replicates were always hybridised on different beadchips. Raw data were imported to the Illumina

Beadstudio software and probes with fewer than three beads present were excluded. Log₂-transformed expression signals were then normalized separately per tissue with quantile normalization of the replicates of each individual followed by quantile normalization across all individuals as previously described (Grundberg et al., 2012). Post-QC expression profiles were subsequently obtained for 647 individuals. The Illumina probe annotations were cross-checked by mapping the probe sequence to the NCBI Build 36 genome with MAQ (Li et al., 2008). Only uniquely mapping probes with no mismatches and either an Ensembl or RefSeq ID were kept for analysis. Probes mapping to genes of uncertain function (LOC symbols) and those encompassing a common SNP (1000G release June 2010) were further excluded, leaving 23,555 probes used in the analysis.

Box 1: Modeling

We model phenotype y_i of individual i (age A_i) as follows:

$$\text{(Full)} \quad y_i = \mu + \alpha A_i + \beta_i + \gamma_i + \epsilon_i$$

$$\beta_i \sim N(0, \sigma_{FAM}^2)$$

$$\gamma_i \sim N(0, \sigma_{MZ}^2)$$

$$\epsilon_i \sim N(0, \sigma^2)$$

$$\text{(Null)} \quad y_i = \mu + \beta_i + \gamma_i + \epsilon_i$$

$$\beta_i \sim N(0, \sigma_{FAM}^2)$$

$$\gamma_i \sim N(0, \sigma_{MZ}^2)$$

$$\epsilon_i \sim N(0, \sigma^2)$$

To correctly model the twin structure we enforce that $\beta_i = \beta_j$ when i and j are twins, and $\gamma_i = \gamma_j$ when i and j are monozygotic twins (capturing the increased genetic correlation of monozygotic twins).

From the null model we can define heritability (h^2), proportion of environmental variance explained by age (p_a) and the proportion of variance explained by the unique environment (p_e) as:

$$h^2 = \frac{2\sigma_{MZ}^2}{\sigma_{FAM}^2 + \sigma_{MZ}^2 + \sigma^2 + \alpha_i^2 \text{var}(A_i)}$$

$$p_a = \frac{\alpha_i^2 \text{var}(A_i)}{\sigma_{FAM}^2 - \sigma_{MZ}^2 + \sigma^2 + \alpha_i^2 \text{var}(A_i)}$$

$$p_e = \frac{\sigma^2}{\sigma_{FAM}^2 + \sigma_{MZ}^2 + \sigma^2 + \alpha_i^2 \text{var}(A_i)}$$

Note that for p_a the genetic variance is removed from the denominator.

3.3 Gene expression pathway factors

In a two step approach, factor analysis methods were first used to discover patterns of common variation across the entire dataset. The software package PEER was applied using the default settings and using technical measurements (experimental batch, RNA quality and concentration) as covariates to create 5 components which in total explained 35.7% of the variation in the dataset. Secondly, the effects of the five global factors together with the technical covariates including batch, RNA concentration and RNA quality were removed from the whole gene expression data sets. The residuals after this process were then grouped to pathway subsets according to the KEGG annotation. For each pathway, we created five pathway phenotypes using PEER with the default settings.

In a two step approach, factor analysis methods were first used to discover patterns of common variation across the entire dataset. The software package PEER (Parts et al., 2011) was applied using the default settings and using technical measurements (experimental batch, RNA quality and concentration) as covariates to create 5 factors, which in total explained 35.7% of the variation in the dataset. For each individual, a factor is a weighted sum of all the gene expression measurements of that individual. The weights are chosen so that the factors iteratively explain the maximum amount of variation in the dataset subject to certain prior assumptions; these factors produce concise summaries of consistent patterns of expression for large numbers of genes.

We then used KEGG pathway annotation (186 pathways) as prior information to group genes into pathways. This allows inference of PEER factors for each pathway that we refer to as phenotype factors, in contrast to the global factors previously described. As before, these factors are weighted sums of gene expression measurements, but in this case only of genes within the pathway. Since global factors have been removed from the dataset prior to calculation of phenotype factors, these factors are unlikely to capture global effects on gene expression, but instead pathway specific patterns of expression. If a large enough module of genes within the pathway is co-expressed then likely one factor will also show the same pattern of co-expression across individuals. Equally, groups of genes could show opposite patterns of expression; this

antagonistic gene expression could be reflected as factor values which correlate with one set of genes and are anti-correlated with the other across individuals. Individual genes can contribute positively or negatively to the weighted sum (indicated by the sign of the corresponding weight), meaning that a positive correlation between age and phenotype factor can be induced by negative correlations with individual genes.

We grouped the expression data set into 186 pathway subsets. For each pathway we created five pathway phenotypes using PEER with the default settings. We consider the learnt pathway factor values across individuals as five new phenotypes which can be investigated for associations with age (analysis performed as described in Box 1. An alternative strategy would be to choose different numbers of factors based on the cumulative amount of variance explained. For the sake of simplicity and as a proof of principle, in this analysis we chose to use five factors as they explained a substantial amount of the variance in expression without too large a multiple testing burden.

3.4 Pathway factor and phenotype association

Association tests were performed: i) between each pathway factor and chronological age, and ii) between single genes and chronological age using the linear mixed models defined in Box 1. These models have been implemented by the *lme4* package (Bates et al., 2014) in R (Computing, R Foundation for Statistical Vienna, 2008). For each phenotype a likelihood ratio test of the full model, which includes the age term, and the null model (without modeling age) produced evidence of an age effect. P values produced by this analysis were assessed for significance allowing for multiple testing using a Bonferroni adjusted threshold. A permuted dataset was created, which maintained the twin structure by permuting singletons, dizygotic and monozygotic twins separately and ensuring that twin pairs were kept together.

Significant associations between pathway phenotypes and age were further investigated to trace the particular genes within the pathway driving the signal. We report genes with a Bonferroni significant P value which accounts for the number of genes within the pathway that was tested.

3.5 Heritability analysis

To compute heritability, proportion of environmental variance explained by age, and the proportion of variance explained by unique environment, we fitted the full model from Box 1. Then the genetic component to variation was estimated as twice the additional correlation of MZ twins relative to DZ twins. The environmental component to the phenotype was the sum of the contribution from the fixed age effect, the random noise term, and the shared environmental component, again estimated from the difference between MZ and DZ. Estimates of these proportions are constrained to lie between 0 and 1 inclusive.

3.6 Single-gene based pathway enrichment analysis

We compared the significant pathways found by our factor analysis methods to those found by looking for enrichment of single gene associations with age. Firstly we tested each gene for association with age using the methods described in Box 1 and produced a list of Bonferroni significant genes $P < 0.05$ (this list contained 682 differentially expressed genes). For each pathway, we applied a Fisher's exact test to infer whether the proportion of significantly associated genes within the pathway was greater than would be expected by chance, which is estimated as being proportional to the pathway size. We also investigated whether using an FDR cut-off for significant age associations would produce more significant pathways or power would be diluted by including too many false positives. When re-running the analysis using a less stringent threshold (3,487 genes were associated with age with $FDR < 0.05$) we found fewer significant pathways, and results correlated less well with the results of the factor based analysis (Spearman correlation of 0.36 ($P = 5.1 \times 10^{-7}$) compared to 0.49 for Bonferroni, $P = 2.1 \times 10^{-12}$). A complete list of all significant single gene age associations ($FDR < 0.05$, 3,487 genes), with estimate of effect size and direction, can be found in Appendix Table [A.3](#).

3.7 Results

The first stage of the analysis was to remove the effect of both known and unknown nuisance variables from the gene expression data. Using PEER software, we estimated five global factors which explained 35.7% of the variation in the complete gene expression data. As the aim of this analysis was to find pathway specific responses to ageing, we treated these global factors as nuisance covariates and regressed these out of the data, together with batch and RNA quality which are known experimental confounders. Data were then divided into subsets of genes within 186 KEGG pathways. For each pathway, five factors were estimated using PEER as described above, which explained on average 17.5% of the residual variation of all genes within this pathway after removing the global factors. For the 186 KEGG pathways, this produced 930 phenotypes which were tested for association with age. In total, 69 significant associations ($P < 5.38E-5$, the Bonferroni adjusted threshold) from 57 distinct pathways were identified. The most significant 20 pathways are listed in Table 3.1, and a list of all 57 significant pathways can be found in Appendix Table A.2.

We also explored an alternative method for finding pathway related to ageing, looking for enrichment in the number of significantly associated genes falling into a particular pathway, analogous to the method used in the DAVID methodology (Huang et al., 2009). This discovered a total of 7 significant pathways (Appendix Table A.3). Thus, applying factor analysis methods to discover significantly associated pathways uncovered eight times as many hits. All pathways discovered by single gene enrichment methods were also discovered using factor analysis. There is strong concordance between P values discovered by the two methods (Spearman correlation = 0.49, $P = 2.1 \times 10E12$). Figure 3.1 shows a Q-Q plot of P values for both methods against the theoretical P values under the complete null hypothesis. We see enrichment of significant P values for both methods, but this is not present when analysing the permuted data with factor analysis methods (green dots). This suggests that age plays a widespread role in the expression of these pathways, as enrichment of P values is not due to invalid model assumptions and can be observed using two different methods.

To investigate which genes drove the significant pathway associations, we exam-

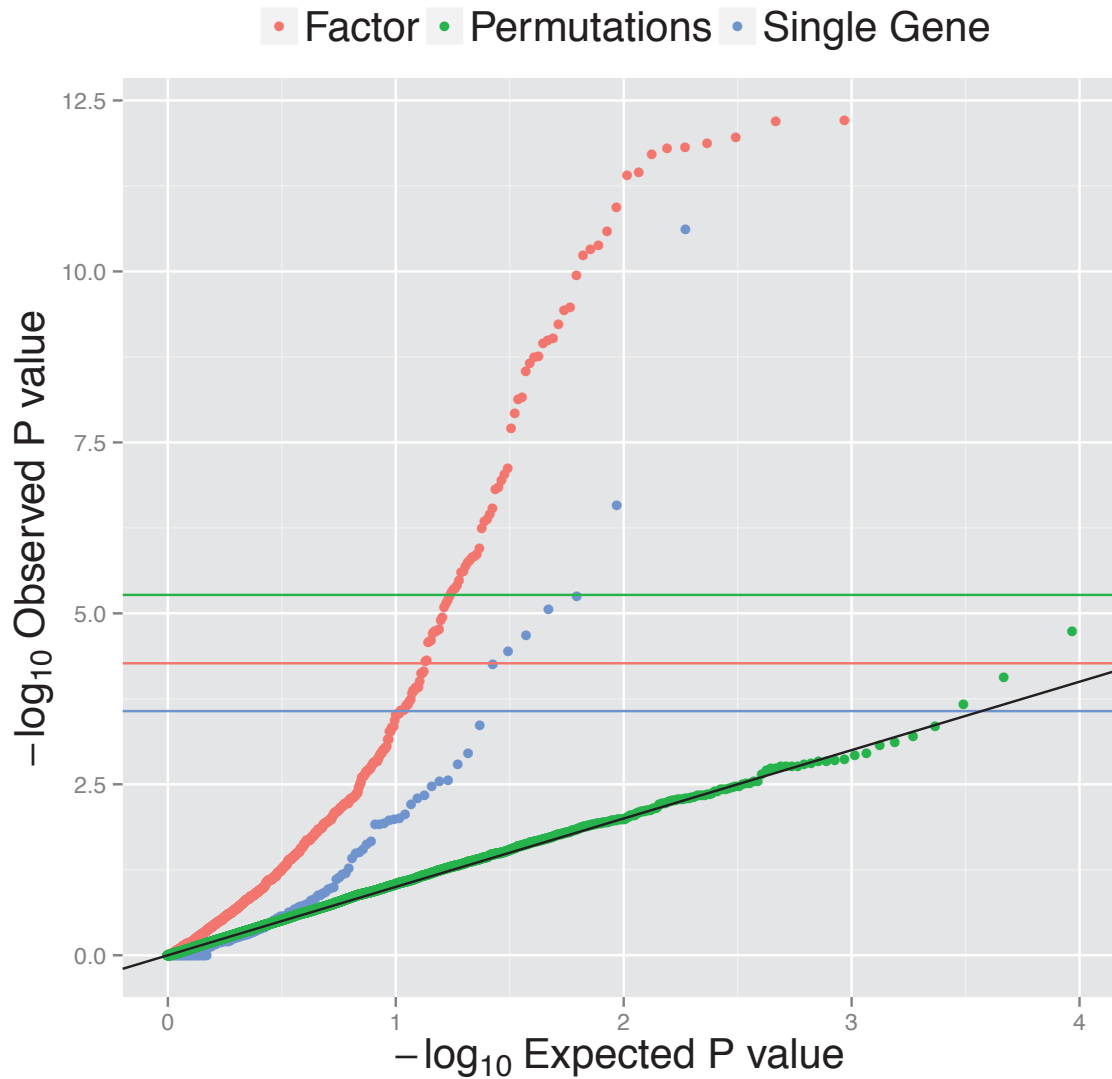


Figure 3.1: Q-Q plot of observed P values against theoretical P values for factor analysis (red dots) and single-gene based methods (in blue). Permutations (in green) shows the results of a combined analysis of 10 permuted datasets. Horizontal lines shown Bonferroni significance thresholds accounting for different numbers of tests (186 tests for single gene measures in blue, 930 for factor analysis in red, and 9300 for the combined 10 permutation analysis in green).

ined how many genes within a significant pathway showed significant age associations (Table 3.1 and Appendix Table A.2). On average 16% of genes within the pathways have $P < 0.05$ after adjusting for the number of genes in the pathway, with a minimum of 1 gene and maximum of 24. The proportion is similar between pathways of different sizes, in contrary to the traditional pathway enrichment analysis, where there is bias towards large pathways.

Table 3.1: List of 20 pathways most significantly associated with age, together with the the number of significantly associated genes ($P < 0.05$, corrected using Bonferroni for the total number of genes in the pathway), the total number of genes, and the heritability of the pathway factor.

KEGG_ID	Pathway	P value of pathway factor	Number of significant genes	Number of genes in pathway	Heritability
00900	Terpenoid Backbone Biosynthesis	6.23E-13	6	13	0.00
00980	Metabolism of Xenobiotics by Cytochrome P450	6.47E-13	6	54	0.09
01040	Biosynthesis of Unsaturated Fatty Acids	1.11E-12	6	17	0.25
00100	Steroid Biosynthesis	1.33E-12	12	14	0.41
00650	Butanoate Metabolism	1.51E-12	8	27	0.39
04146	Peroxisome	1.56E-12	17	64	0.45
00830	Retinol Metabolism	1.93E-12	6	48	0.45
00010	Glycolysis Gluconeogenesis	3.59E-12	12	49	0.42
00051	Fructose and Mannose Metabolism	3.99E-12	8	32	0.32
00290	Valine Leucine and Isoleucine Biosynthesis	1.15E-11	3	11	0.00
00561	Glycerolipid Metabolism	2.63E-11	6	38	0.34
00620	Pyruvate Metabolism	4.20E-11	11	35	0.37
00770	Pantothenate and COA Biosynthesis	4.76E-11	4	16	0.48

00280	Valine Leucine and Isoleucine Degradation	5.79E-11	10	35	0.51
00020	Citrate Cycle TCA Cycle	1.12E-10	8	23	0.43
04916	Melanogenesis	3.34E-10	10	93	0.00
04910	Insulin Signalling Pathway	3.70E-10	13	122	0.45
00565	Ether Lipid Metabolism	5.89E-10	3	27	0.00
00350	Tyrosine Metabolism	9.44E-10	4	32	0.34
00640	Propanoate Metabolism	1.03E-09	6	26	0.59

Different KEGG pathways can contain overlapping sets of genes, as they can describe related biological function. Because of this, our significant associations with age for different pathways could be related due to a common underlying effect on a given set of genes. To explore whether the observed age-associations are unique to their pathway, or common to multiple pathways, we calculated the Spearman correlation between those phenotypes. There are 24 pathway phenotypes with a correlation greater than 0.8 with at least one other phenotype (Appendix Table A.4). These phenotypes frequently relate to metabolism, and form a highly connected set (Figure 3.2). We infer from this that there could be a common effect of age acting on all these phenotype factors.

We next explored how different sources of variation in the different phenotypes analysed here affect our ability to discover age associations. We calculated the heritabilities, the proportion of environmental variance explained by age, and the proportion of variance explained by the unique environment (Box 1) for i) KEGG pathways, ii) global factors (which we have treated as nuisance covariates) and iii) for individual genes (Figure 3.3, global factor histograms are not shown as there are too few phenotypes). The relative differences in sources of variation between global and pathway factors, and individual genes are shown in Figure 3.4. We see that as we move away from local phenotypes (individual genes) to pathway phenotypes and then to global

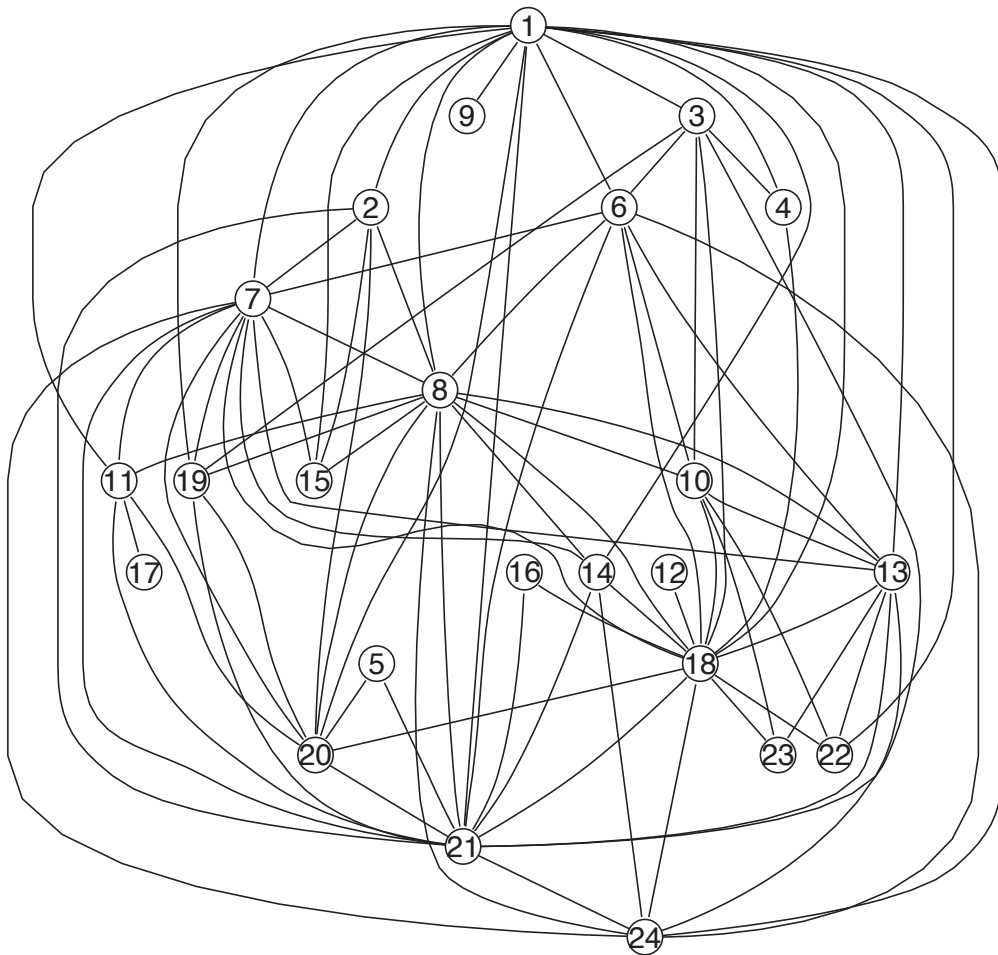


Figure 3.2: Network of connected factor phenotypes. Twenty four of the 69 age-associated factor phenotypes have a Spearman correlation of at least 0.8 with at least one other phenotype. These phenotypes show a highly connected structure, likely meaning there are common age effects driving these associations. A key for identifying which pathways correspond to the nodes can be found in Appendix Table A.4.

phenotypes, the proportion of variation explained by unique environment decreases. This is because that there is a stochastic component to each single gene's expression: by taking a weighted average of a number of genes, we average away this component. If all else were to remain constant, this reduction in stochastic noise would simultaneously increase heritability (as the total variance decreases), and boost the ability to discover associations with biological meaning, such as age. We see in the first panel of Figure 3.4 that the relative contribution of unique environment to pathway phenotypes is smaller than the contribution to genes. This also partly explains the results shown in the second and third panels: a greater proportion of variance is explained by age and genetic factors (heritability) for pathway factors than individual gene measurements.

When considering global factors, as expected the unique environment is greatly reduced. However, there is not a strong influence of ageing and heritability in this case is still moderate. This is likely because age and genetics do not act in a consistent way across large sets of genes. Leek and Storey, 2007 argued that global factors can capture experimental noise and batch effects. This is consistent with our findings. Heritabilities and proportion of variance explained by age for all pathways are reported in Appendix Table A.5.

We further looked for novel genetic associations with these pathway phenotypes, not seen as single gene expression associations. However, this was unsuccessful despite the increased heritability in pathway factors. This is likely due to the genetic architecture of gene regulation. Genes are regulated both in *cis*, where a nearby variant effects the expression of a single gene, and in *trans*, where a long range regulatory effect can hit multiple genes (Grundberg et al., 2012). The genetics of pathway phenotypes is a combination of *cis* effects on individual genes and *trans* effects, potentially affecting multiple genes in the pathway. However, *trans* variants typically have much smaller effect size: the increase in the reliability of pathway phenotypes is insufficient to compensate for the lower power to discover *trans* effects. Thus, the only associations discovered were when single genes loaded heavily enough on a pathway to indirectly reflect the *cis* association.

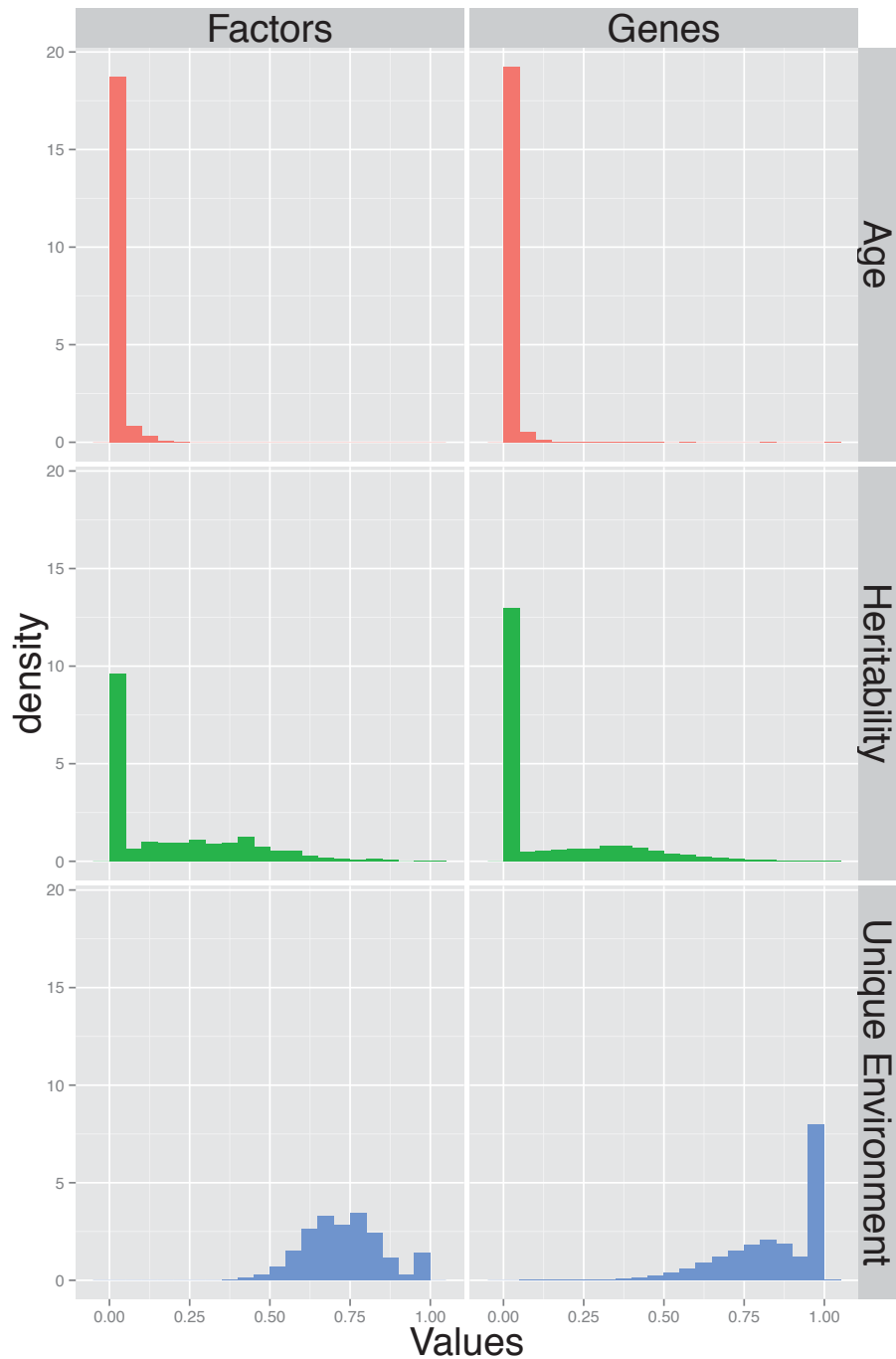


Figure 3.3: Histograms showing the proportion of environmental variation explained by age, heritability, and the proportion of variance explained by the unique environment for pathway factors and the individual gene measurements. The calculations correspond to equations in Box 1. Note that the proportions are not sum to one as they are not normalized by a same denominator: for age the variance explained by the genetic factors is removed.

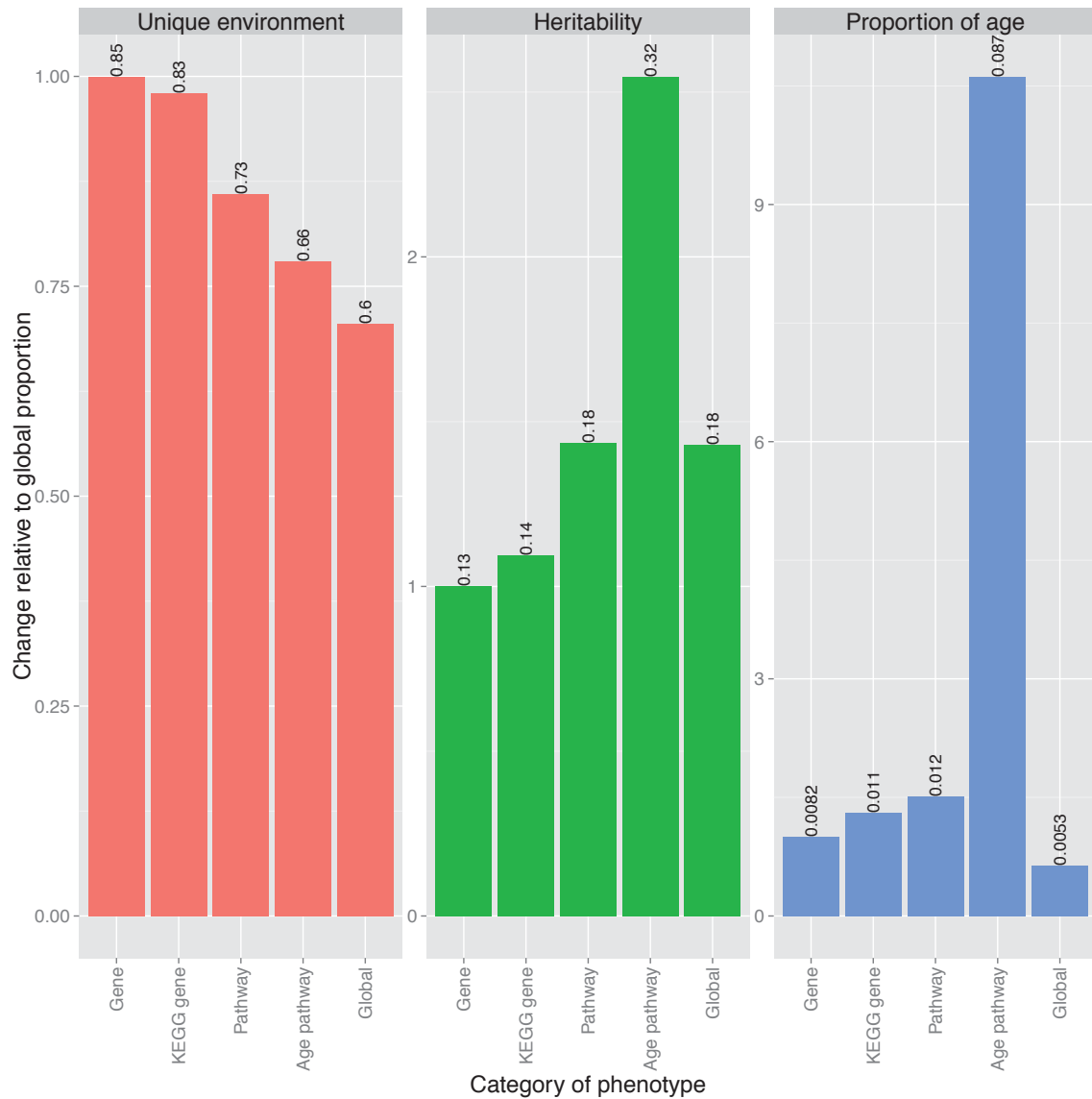


Figure 3.4: The relative importance of sources of variation to global, pathway and gene phenotypes. Measures of variation shown are the proportion of variance explained by unique environment, proportion of variance explained by genetics (heritability) and the proportion of environmental variation explained by age. The five categories are individual genes; genes that are in pathways annotated by KEGG; pathway factors; age associated factors and global factors. To show more clearly the differences in relative importance of these measures to different classes of phenotypes, all proportions are scaled such that contribution to gene phenotypes equals one. Numbers above the bars give the absolute, unscaled proportions.

3.8 Discussion

We have seen that both the heritability and the proportion of environmental variance explained by age is greater for pathway phenotypes than for individual genes. Consistent with this, a previous study found a greater proportion of associations for the pathway phenotypes than using single gene tests using this same dataset (Glass et al., 2013, 23% compared to 7% of phenotypes are significantly associated with age when using the same 0.05 FDR threshold adopted in that paper). This can be explained by our findings on the influence of unique environment on pathway phenotypes relative to single genes.

Stochasticity in gene expression, which contributes to the unique environment component that we measure, has been seen to increase with age. For example, animal model studies (Herndon et al., 2002; Bahar et al., 2006) have reported increased cell-to-cell variation in gene expression with age and tissue specific decline of functions associated to stochastic events. Others have found genes associated with longevity to be strongly regulated in older animals with low levels of stochasticity and higher levels of heritability (McCarroll et al., 2004; Vinuela et al., 2012). The aim of our analysis was to find mean effects, rather than variance effects (though both effects are often seen together). By reducing the unique environment variable component using pathway factor analysis methods, we arguably focus much more on a systematic longevity changes with age rather than the environmental stochasticity. However, it is difficult to make inference about causality with gene expression: we cannot know whether we are observing changes in expression which are driving the ageing process, or markers for it.

Of the 57 significant pathways, we frequently see four types of pathway, all of which have been previously linked with ageing: i) insulin signaling ; ii) sugar and fatty acid metabolism; iii) xenobiotic metabolism; and iv) cancer related pathways.

We find the insulin signaling pathway (hsa04910) to be highly associated with age in our data ($P=3.7E-10$). Much evidence has accumulated for the influence of the insulin signaling pathway on longevity, originating in *C. elegans*, where lowered insulin/IGF-1 signalling (IIS) can lead to a significant increase in life span (Friedman and Johnson,

1988). This effect has also been seen in the fruit fly *D. melanogaster* (Clancy et al., 2001) and in mice (Holzenberger et al., 2003). Outside of model organisms, it has been observed that variants in FOXO transcription factors related to this pathway can affect longevity in humans (Willcox et al., 2008), although its gene expression does not show significant association with age.

In addition to those related to insulin, our list of age-associated pathways includes many that are involved in metabolism or glycolysis. Examples of these include biosynthesis of unsaturated fatty acids (hsa00980), butanoate metabolism (hsa00650), glycolysis gluconeogenesis (hsa00010), fructose and mannose metabolism (hsa00051) and valine leucine and isoleucine biosynthesis (hsa00290) ($P \leq 1.15E-11$). It has previously been suggested that metabolism related pathways play roles in ageing and ageing related diseases (Barzilai et al., 2012). In particular, Houtkooper et al. (2011) showed that glucose and compounds involved in the metabolism of glucose were biomarkers of ageing in liver and muscle tissue in mice.

Other ageing related pathways include those involved in the metabolism of xenobiotics allow cells to deactivate and excrete unexpected compounds. One example is glutathione metabolism (hsa00480, $P=1.45E-7$), a well known anti-oxidant which protects against cell damage by reactive oxygen species (Pompella et al., 2003).

Finally, similarities between cancer and ageing have been noticed (Finkel et al., 2007). For example, cellular senescence, when a cell loses the ability to divide, can form a break on cancer development; clearing such cells can delay the development of age-associated disorders (Baker et al., 2011). There are a number of pathways in our list that have been linked to cancer, in particular skin cancer, possibility because this was done using skin tissue. These include melanogenesis (hsa04916, $P=3.34E-10$), the PPAR signaling pathway (hsa03320, $P=1.83E-9$), the hedgehog signaling pathway (hsa04340, $P=1.12E-7$) and glioma (hsa05214, $P=4.26E-7$)

In addition to age, other phenotypes have been linked to expression patterns of multiple genes. For example, BMI has been linked to expression patterns in adipose tissue of multiple genes within a group which share a common *trans* master regulator, and such phenotypes could mediate between expression and diseases such as type 2 diabetes (Small et al., 2011). Principal components and factor analysis has also

been suggested as a way to build classifiers for binary traits (Hastie et al., 2000), perhaps to predict prognosis of disease from gene expression data. The ability of pathway phenotypes to provide reliable measures of expression with direct biological interpretation means they could also be applied in both situations, to understand the relationship between expression and such phenotypes.

Our analysis shows that factor analysis applied to gene expression data effectively reduces stochastic noise in summaries of gene expression patterns, giving more power to discover associations. These phenotypes are substantially more heritable than individual genes. Using them we can improve our ability to identify biological processes underpinning ageing. This is consistent with the idea that removing latent factors that exert broad effects on gene expressions increases power in associations. We show that the same idea can be used to create pathway factors that are robust and interpretable. Finally, our analysis reveals pathways that have been seen to be important in longevity from a number of previous studies, as well as novel pathways that can be further investigated.