Divergent selection experiments: a powerful tool for unravelling the genetic regulation of complex traits.

Ibanez-Escriche N.1,2, Jackson I.1, Rawlik K.1, Bünger L.3 Stewart M.4, Canela-Xandri O.1, Prendergast J.1, and Tenesa A.1

1The Roslin Institute. The University of Edinburgh, UK.
2Institute for Animal Science and Technology. Universitat Politècnica de València, Spain.
noeibes@dca.upv.es (Corresponding Author)
3Independent Consultant at AnGeCo; 48 The Spinney, Edinburgh, EH177LD
4MRC Harwell, Mary Lyon Centre, Harwell Campus, Oxfordshire, OX11 0RD, UK.

Summary

In mammals, mice growth divergent selection experiments have historically been a key tool to understand how selection shapes phenotypic and genetic variation. However, how selection for a particular trait operates at the transcriptomic level in mammals is still not well known. A significant genome-wide conservation of expression levels has been observed between humans and mice. Nevertheless, knowledge of whether the expression of these genes is regulating equivalent complex traits in both species and human is still limited. In this study, we use embryos of four inbred lines of mice derived from long-term divergent selection for growth to identify genes with transcript expression correlated to selection. We also explore whether those genes are associated with human traits linked to the selected one. Results showed 3871 transcripts differential expressed (DE) between lines and replicating in both experiments at False discovery rate <0.05. The genes of these transcripts showed a clear enrichment for embryo weight, embryo growth retardation and body weight in mouse and for height and body mass index (BMI) in GWAS studies in human. From these genes, we identified 417 transcripts with expression correlated to embryo weight after adjusting for line. These genes showed an enrichment with abnormal long bone morphology in mouse and BMI in human. We computed the heritability of each one of these genes for BMI and height in human using the UK Biobank data (~130,000 individuals genotyped). The results reveal novel genes explaining a high percentage of the total heritability for BMI or height, and some of them with pleiotropic effects. In conclusion, results indicate that directional selection drives correlated changes at the RNA level. It implies that transcript’s expression can be a good predictor of the phenotypes. Moreover, the results of this study support the idea of genome functional conservation between mammals. It makes divergent selection experiments a powerful tool for unravelling the genetic regulation of complex traits.

Keywords: Divergent selection experiments, RNA-Seq, differential expression, growth

Introduction

It is well known that gene expression is a key factor shaping organism’s phenotype (Jacob & Monod 1961). Conversely, a poorly understood question that is a central focus of many fields, from animal and plant breeding, to evolution and human genetics is to realize how selection for a particular trait operates at the transcriptomic level in mammals. Divergent selection experiments can be a key tool since difference between divergent lines in the same environment are mainly driven by selection (Bunger et al., 2001). There is some evidence that regulation of some genes has evolved under directional selection (Gallego et al., 2012).
However, whether directional selection drives correlated changes at the RNA level genes has never been shown. This is fundamental, because if true, it will allow us to better understand the connections between the molecular level and the resulting phenotype which could have important implications. For instance, phenotypes could be predicted based on transcriptomic expression.

In humans establishing the link between the transcript expression and a target trait is very challenging. The environmental changes are difficult to monitor and the target tissues are not easily accessible. However, it is known that regardless of the cell or tissue type of the sample there is a significant genome-wide conservation of expression levels between human and mouse (Pervouchine et al., 2015). Furthermore, similarity of exon structure and splicing in terms of the number and order of exons per gene, exon length, precise boundaries and sequence has been found between humans and mice (Breschi et al., 2017). Nevertheless, the knowledge of whether the expression of transcripts is regulating equivalent complex traits in mouse and human is still limited. This insight is important when using animal models to elucidate genetic regulation of complex traits in mammals.

The aim of this work was to study whether selection shapes transcripts’ expression and whether these transcripts have related effect in human.

**Material and methods**

**Biological material**

Four inbred mouse lines, Dummerstorf High (DUHi), Berlin Low (BELi) line, Roslin High (ROHi) and Low (ROL) resulting from at least 20 generations of divergent selection for BW (Bünger et al., 2001) followed by 14 generations of inbreeding were used in two different experiments. The first compromises 2 DUHi x DUHi, 2 DUHi x BELi, 3 BELi x BELi timed crosses. From each cross, 4 embryos at E13.5 were collected and weighted. The embryo weight differences between DUHi x DUHi and BELi x BELi crosses were highly significant (p-value<1.e-16). The second experiment compromises 4 ROHi x ROHi, 4 ROLi x ROLi, 4 ROHi xROLi and 3 ROLi x ROHi timed crosses. From each cross 4 embryos at E12.5 were collected and measures of size and weight (dehydrated) were taken. The embryo weight differences between ROHi x ROHi and ROLi x ROLi crosses were also significant (p-value= 0.018) and the correlation between embryo weight and size was 0.7.

**Transcript quantification and expression analysis**

RNA sequencing yielded on average 47.1 and 104.9 million paired-end reads per library in experiment 1 and 2, respectively. Adapter and barcode trimming and quality filtering were carried out using the trimmomatic-0.33 (Bolger et al., 2014). Maximum likelihood transcript read count estimates for each sample were obtained with Kallisto v0.42.4 (Nicolas et al., 2016), using the Mus Musculus transcriptome assembly GRCm38 as a reference transcriptome. In each experiment, transcripts with less than 10 reads within experiment were filtered out. The differential transcripts expression (DE) analysis between High and Low lines were performed using limma v3.28.17 (Ritchie et al., 2015). The normalization was applied using the voom function from limma package (Law et al., 2013). It performs a LOWESS regression to estimate the mean-variance relation and transforms the read counts to the appropriate log form for linear modelling. In order to identify transcripts correlated to embryo weight, meta-analysis has been employed for the two experiments using
all crosses and a linear model with the embryo weight as response variable and the cross and transcript expression as fixed and covariate effect, respectively.

**Enrichment analysis**

The enrichment of differentially expressed genes was analysed on the mouse genomic informatics (MGI) Mammalian Phenotype 2017 data base using Enrichr software (http://amp.pharm.mssm.edu/Enrichr/) and on the human genome wide association studies (GWAS) catalog data base (https://www.ebi.ac.uk/gwas/gwas_catalog_v1.0.1-associations_e88_r2017-04-03.tsv) using a hypergeometric distribution.

**Human gene replication**

Data for the individuals genotyped in phase 1 of the UK Biobank genotyping program to replicate the genes with transcripts correlated to embryo weight were used. We filter out SNPs with imputed quality threshold < 30 and MAF<0.001. Measures for height and body mass index (BMI) (http://biobank.ctsu.ox.ac.uk/crystal/index.cgi) were obtained. and we removed outliers from height and BMI, defining outliers as males and females who were outside ±3 standard deviations. A linear mixed model approach was used to estimate the genetic parameters for each gene for which expression was correlated with the embryo weight. Specifically, we used a Bayesian ridge regression on markers model (Gianola, 2013). The posterior distribution of the genetic variance was computed based on the sample-variance of the posterior genomic values (De los Campos et al., 2015; Sorensen et al., 2001). A total of 80,000 iterations with a burn-in of 10,000 were run for the analyses. A permutation within Markov chain Monte Carlo (McMC) approach (Che & Xu 2010) was used to determine the relevance of the gene’s genetic variance for the Bayesian model.

**Results and discussion**

The study workflow is shown in figure 1. A total of 3871 transcripts (3145 genes) were differentially expressed at a false discovery rate (FDR) <0.05 and with the same direction within the two independent RNA-Seq experiments. These genes showed a clear enrichment for embryo weight, embryo growth retardation and body weight in mouse and height and BMI in GWAS in human (Table 1 and 2). These results are in agreement with the response for growth in mouse and correlated response on embryo weight and body fat percentage previously reported in these lines (Bünger et al., 2001). This suggests that difference between the expressions of these transcripts between divergent lines were driven by selection. From this set of transcripts, we identify those transcript’s whose expression was correlated to embryo weight adjusted by line (i.e. Figure 2). These genes were enriched with those linked to abnormally long bone in mice (P<1.e-4) and BMI in human (P<5.e-4). Next, for those genes with a clear homolog in human (351 genes), we computed the heritability of height and BMI traits explained for each gene. Results show that 30 and 15 % of the genes explain a significant percentage of the total heritability for height and BMI based on the permutation analysis (P<0.05). The genes previously associated in GWAS studies were confirmed. Moreover, we identify novel variants also explaining a large proportion of the heritability for height (21) and BMI (19). Subsequently, we validate these novel genes with the results of the giant meta-analysis carried out by Wood et al. (2014) for height and by Locke AE et al. (2015) for BMI. 9 and 5 genes have SNPs with p-values<1.e-3 for height and BMI, respectively. Notably, the SEMA3F, PREX1 and SMARCC1 genes that explaining a
large percentage of the heritability in both height and BMI. Additionally, mutations in mice of SEMA3F have been associated with decreased body lean mass and body size and the PREX1 gene with decreased body weight. SMARCC1 is related to the transcription ligand-dependent activation of the ESR1/SP pathway, which play a role in cell cycle regulation and proliferation. The results indicate that these genes have pleiotropic effects for height and BMI and are in agreement with the correlated response on fat percentage to selection for growth in these mouse lines (Bünger et al., 2001).

In conclusion, these results indicate that directional selection drives correlated changes at the RNA level. It implies that a transcript’s expression can be a good predictor of the phenotypes. Moreover, the validation of these genes in human traits related to the selected one support the idea of functional conservation between mammals. It makes divergent selection experiments a powerful tool for unravelling gene regulation of complex traits in mammals.

Table 1. The top five traits of the Mouse Genomic informatics (MGI) database enriched for the genes with transcripts DE at FDR <0.05 in both experiments.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Genes reported</th>
<th>Genes Matching</th>
<th>P-value</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embryonic lethality</td>
<td>674</td>
<td>164</td>
<td>1.6E-09</td>
<td>3.3E-06</td>
</tr>
<tr>
<td>Decreased embryo size</td>
<td>472</td>
<td>121</td>
<td>1.6E-09</td>
<td>1.0E-05</td>
</tr>
<tr>
<td>Exencephaly</td>
<td>182</td>
<td>60</td>
<td>4.5E-9</td>
<td>6.2E-6</td>
</tr>
<tr>
<td>Embryonic growth retardation</td>
<td>435</td>
<td>109</td>
<td>1.9E-07</td>
<td>1.3E-04</td>
</tr>
<tr>
<td>Decreased body weight</td>
<td>1189</td>
<td>236</td>
<td>3.5E-05</td>
<td>9.2E-03</td>
</tr>
</tbody>
</table>

Table 2. The top five traits of the GWAS catalog data base enriched for the genes with transcripts DE at FDR <0.05 in both experiments.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Genes reported</th>
<th>Genes Matching</th>
<th>P-value</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>408</td>
<td>96</td>
<td>1.2E-10</td>
<td>3.9E-08</td>
</tr>
<tr>
<td>Body mass index</td>
<td>293</td>
<td>68</td>
<td>8.4E-08</td>
<td>1.3E-05</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>219</td>
<td>47</td>
<td>5.3E-05</td>
<td>5.6E-03</td>
</tr>
<tr>
<td>Major depressive disorder</td>
<td>81</td>
<td>22</td>
<td>9.0E-05</td>
<td>5.6E-03</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>114</td>
<td>28</td>
<td>1.1E-04</td>
<td>5.6E-03</td>
</tr>
</tbody>
</table>

1Traits with at least 15 genes reported on GWAS studies
Figure 1. Workflow analysis of this study.
Figure 2. Embryo weight (mg) vs normalized expression of ENSG00000133110 transcript (POSTN gene) weight. Points with blue colour indicate experiment 1 (DUH and BEL cross lines) and red experiment 2 (ROH and ROL cross lines). The point’s shapes indicate the four different crosses between the lines selected for high growth (H) and the lines selected for low growth (L).
List of References


