The Detection of Covariation of mRNA Levels of Large Sets of Genes across Multiple Human Populations

Covariation of Human mRNA Levels

Yu Quan[1], Chao Xie[1,2], Rohan B. H. Williams[2], Peter Little[1,2,3*]

[1] Life Sciences Institute, National University of Singapore, Centre for Life Sciences, 28 Medical Drive, Singapore 117456
[2] Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore, 28 Medical Drive, Singapore 117456
[3] Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117596

* Corresponding author

E-mail: peter.little@nus.edu.sg (PL)

Abstract

In this study we report that, measured in human lymphoblastoid cell lines (LCLs), different individuals exhibit correlated changes in the mRNA levels of multiple groups of genes; the largest group of genes such variation contains 2168 members and over 52% of all genes expressed in these cell lines are members of one or more groups. We show that these gene groups are replicable across different human populations and across different methods of analysis, supporting the view that the covariation is caused by biological and not experimental batch effects. This view is also supported by enrichment in the covarying genes of single and combinations of transcription factors (TFs) binding to cognate promoter regions, enrichment of genes shown to be sensitive to the knockdown of individual TFs and by enrichment of functional pathways and finally of protein-protein interactions. The properties of the groups of genes are therefore most readily explained by the influence of cumulative variations in the effectors of gene expression that act in *trans* on cognate genes. We suggest that covariation has functional outcomes by showing that covariation of 83 genes involved in the spliceosome pathway accounts for 21–43% of the variation in the alternative splicing patterns of genes expressed in human LCLs.

Author Summary

Enormous effort has gone into understanding the changes in gene expressions that underpin human disease and welfare and the research has been largely focused on the local genetic variations near and within genes that influence when and how much mRNA is made by the adjacent gene. In our study we focus on how variations in the many different proteins (and other molecules) that are involved in controlling genes might result in changes in a gene's mRNA output. Because controlling proteins control multiple genes, we predict we should be able to observe the influence by looking for groups of, rather than single, genes that co-ordinately exhibit changes in amount of mRNA. Using a novel approach to detect co-ordinated changes of mRNA levels we identify large groups of genes that are apparently affected by the shared effect of changes in control. Some groups can contain thousands of genes and importantly many of these genes encode proteins that operate together to perform specific functions in cells. The simultaneous change of control of many genes involved in a given cell function is likely to have a more dramatic effect than changes of just a single gene. We report evidence that supports this view and defines for the first time sets of genes whose simultaneous change in control might explain some major differences between humans, including their health.

Introduction

Genetic variation that influences gene expression in humans is thought to be a major mechanism contributing to human phenotypes [1]. The variation in mRNA levels in normal human beings is very substantial. The mRNA levels of the around 14,000 genes expressed and measured by mRNA-Seq in lymphoblastoid cell lines (LCLs) derived from 88 Yoruban individuals (data from Lappalainen et al. [2]) has a median of 4.3 fold change comparing the minimum to maximum expression levels. Presently, the overall inter-individual genetically derived variation in the expression level of a gene is best explained by a mix of *cis*- and *trans*-acting eQTLs (expression quantitative trait loci) defined as associated genetic variations local (*cis*) or distant (*trans*) to the affected gene and Huan et al. [3] reported that *cis*-acting eQTLs account for 33–53% of mRNA variability of a single gene whereas individual *trans*-eQTLs explain only 2–7%. This information derives from detecting the influence of a single eQTL, be it in *cis* or in *trans*, on a single gene, *trans*-eQTLs being defined simply by virtue of remote location from the cognate gene. However, it is also clear that the proteins and RNA elements involved in *trans*-acting controls on gene expression invariably have multiple gene or mRNA targets and so a common feature of variation in *trans* influences should be that they will cause simultaneous changes to the mRNA levels of multiple genes. The impact of variation in these pleiotropic regulators can be substantial; Lovén et al. [4] showed that engineered overexpression of the *MYC* gene results

in increases in expression of 90% of all genes expressed in the cells and Cusanovich et al. [5] used individual RNAi knockdowns of 59 transcription factors (TFs) to show that between 39 and 3,892 genes were differentially expressed as a consequence of knockdown of individual TFs. The knockdown of each TF mRNA level ranged 52–92% of normal in the target lymphoblastoid cell line and analysis of mRNA encoding TFs measured in LCLs of multiple normal human beings reveals very similar levels of variation (see S1 Fig). This observation suggests that variation in TF mRNA levels in normal humans might result in *trans* influences upon multiple genes. Moreover, genetic variations on coding sequences of TFs can affect their DNA-binding activity [6]. Such simultaneous variation could be of substantial importance to phenotypic variation in humans, simply because many *trans* controls, particularly TFs, have been implicated in control of genes organised within pathways of function [5,7–9] suggesting that *trans* variation might have pathway specific influences.

Detecting *trans* genetic influences presents two substantial challenges to conventional eQTL strategies: firstly, very large sample sizes are required to overcome the large multiple testing implicit in analysing each mRNA trait against all genomic SNPs, and secondly the influence of a single *trans*-eQTL upon a target gene is low [3].

An alternative to the gene/eQTL centric approach is to directly detect the outcome of *trans* influence by detecting the correlated variation of the mRNA levels of groups of target genes in multiple individuals [8,10]. Covariation of mRNA levels in such an analysis will necessarily be due to cumulative effects of *trans*-acting variations, be they caused by TFs or any other *trans* factors, and the detection does not lead to the identity of the underlying *trans* eQTLs but instead identifies the cumulative influence of all *trans* influences on mRNA levels within the individuals under study. Importantly, detecting covariation of mRNA level in a set of genes as an analytic approach is statistically tractable in extant sample sets because it is not confounded by the *trans*-eQTL multiple testing problem, nor is it necessarily limited to identifying the small effect associated with single *trans*-eQTLs.

However the major confounding difficulty in detecting covariation due to *trans* influences is to distinguish this from the covariation in mRNA levels that is induced by a wealth of technical artefacts, collectively referred to as batch effects, which are inherent to the analytic methods used in mRNA level estimation. Such effects include systematic differences in mRNA isolation and yield, purity, cDNA library or target mRNA construction and ascertainment of yield by both mRNA-Seq and by microarrays [11]. Many methods have been developed [12] to control for such batch effects but in most cases these are explicitly designed to detect and eliminate large-scale covariation in mRNA levels since the methods are intended to provide accurate mRNA measure on a gene by gene basis. If *trans* influences are indeed present on human gene expression then the likely outcome would be covariation of mRNA levels of multiple genes that, based upon the data of Cusanovich et al. [5], could number in the thousands. Goldinger et al. [13] used eQTL methodology to identify *trans* effects in human mRNA samples analyzed by microarrays and were able to show that principal component analysis (PCA), a statistical procedure commonly used to control batch effects, could also remove *trans* genetic influences. Formally, batch effects or biological causes may have similar effect upon levels of mRNA and cause correlated changes in expressions of many genes; methods that might identify batch effects may therefore also identify shared, *trans*, influences.

In this paper we report on an analysis based upon detection of large scale covariation of mRNA level using PCA to detect simultaneous variation of large sets of genes – *trans* target genes (TTGs) – by analysis of covariation of mRNA levels across multiple normal individuals.

Our observations lead us to suggest that covariation of mRNA levels in sets of human genes is common and could contribute directly to human phenotypic variability at both the individual and population level. Further, we show the covariation of a set of 83 genes that are involved in mRNA splicing has a significant influence upon splicing patterns of human genes in LCLs.

Experimental design

Our experimental strategy is based upon using PCA to detect groups of genes whose mRNA levels covarying. We use three published mRNA level data sets of human LCLs derived by mRNA-Seq, the first from Lappalainen et al. [2] containing 462 individuals from 5 populations: 91 Caucasian (CEU), 89 Yoruba (YRI), 95 Finns (FIN), 94 British (GBR) and 93 Toscani (TSI); the second from Pickrell et al. [14] containing 69 Yoruban individuals; and from Montgomery et al. [15] containing 60 Caucasian. We also analyse a microarray data set from Stranger et al. [16] containing 109 Caucasian. Full details are in Materials and Methods.

Our approach is based upon the view that *trans* genetic influences should affect multiple genes and that we do not know the numbers of genes, nor the scale of variation, that are likely to be affected. The mRNA level of any gene is conceptually controlled by multiple *cis*- and *trans*-acting elements and so the mRNA level in any given individual will be set by the particular combinations of *cis* and *trans* variables that that individual contains. Each *trans* variable, such as the mRNA level of a TF, will act in concert with the mRNA level of many other TFs in the individual, contributing to the final level of mRNA in the individual. Given each TF affects multiple genes what we expect across multiple individuals is different patterns of shared variation affecting overlapping (or even unique) sets of genes and these will be defined by the interaction of multiple TFs whose activity is varying independently in the different individuals.

Identifying *trans* target genes

Based upon this view, we would expect that such expression data assayed from multiple individuals in a population should demonstrate covariation associated with these shared, *trans*, genetic regulatory influences, along with contributions from both other biological regulatory influences and batch and related technical artefacts. Such shared variation is readily captured by PCA, and building on the findings of Goldinger et al. [13] and related works [17,18], we would predict that different principal components will contain differing proportion of contribution from each of these multiple sources of variation. Under this assumption, we can build our analytical approach as follows. We can apply PCA on the mRNA level data and the first principal component (PC) will capture some shared variance, which in principle could be caused by a shared *trans* effect. Similarly, the second PC will capture variance that might be due to a second shared *trans* influence and this process can be repeated sequentially. Given we have argued that *trans* influences are associated with correlated mRNA level variation in multiple genes, we use correlation statistics to interpret the effect of removal of variance by PCA. The process we have developed is detailed in Fig 1. We first set an absolute correlation level above which we deem the correlation to be significant using a permutation approach (see below and S2 Fig for justification of correlation thresholds). To illustrate the process, let us assume we have mRNA level data on multiple genes derived from LCLs of multiple individuals. In the data prior to PCA analysis (mRNA level matrix X_1), we calculate all pairwise mRNA level correlations (correlation matrix C_1) and identify the genes whose mRNA levels are correlated better than the threshold (Fig 1A, Fig 1B). Following the removal of PC1 (yielding matrix X_2) we recalculate the correlation (matrix C_2) and again find genes correlated above the threshold. In Fig 1C, the Venn diagram is composed of two sets of genes: those genes that are correlated above the threshold in the source data (C_1) and genes that are correlated above the threshold in the PC1 removed data (C_2). For the genes displayed in the sector marked PC1* in Fig 1C, we suggest that PC1 has removed variance from these genes that corresponds to shared, and possibly *trans*, influences and so reducing the correlation below threshold. We then remove PC2 (X_3) and again recalculate the correlation of all pairs of genes (C_3) and identify the genes that fall into the Venn diagram sector marked PC2* in Fig 1C. Again, we suggest PC2 has removed shared, possibly *trans*, variance influencing this set of genes. Thus in each PC there are a set of genes that are better correlated than the threshold but that in the next PC no longer meet the threshold. Each set of genes we define as "*trans* target genes (TTGs)" whose covariation is removed by PC1 or PC2 (Fig 1C). We use a

171 nomenclature that is based upon the covariation detectable in one set of data (PC_n) and removed by
172 the next PC (PC_{n+1}), and we call these genes the PC_n *trans* target genes or *trans* target genes of PC_n.

173 Any given gene may have multiple pairwise correlations above or below the threshold so we
174 define the gene pairs that meet these criteria as *trans* target gene pairs (TTGPs). Alternatively, we
175 can classify as *trans* target genes only those genes that have one or more above threshold
176 correlation(s) in the PC_n but no above threshold correlations in PC_{n+1} and this is a more extreme
177 selection of correlated behaviour. Using this approach, we can define multiple sets of *trans* target
178 gene pairs and *trans* target genes from a full set of sequential PCA analyses.

179

180 Fig 1. Schematic of experimental design to identify *trans* target genes. (A) PCA is applied to a set of
181 appropriate, pre-processed mRNA level measurements (matrix X_1) derived from multiple human
182 LCLs. PCA is used sequentially and after each principal component has been removed we generate a
183 new mRNA level matrices X_n to create 20 matrices X_1–X_{20} (X_1–X_3 are shown here). (B) Pair-wise mRNA
184 level correlations are calculated for each mRNA level matrix to yield a matrix of correlations of all
185 possible gene pairs (matrices C_1–C_{20}, C_1–C_3 are shown here). Gene pairs with correlations above the
186 threshold are highlighted in orange. (C) *Trans* target genes (TTGs) of PC1 are defined as genes
187 included in gene pairs with correlations above the threshold in correlation matrix C_1, but not
188 contained in any gene pairs with correlations above the threshold in correlation matrix C_2. Similarly
189 TTGs of PC2 are defined as genes included in correlated gene pairs from C_2 but not in correlated
190 gene pairs from C_3.

191

192 At this stage we cannot determine if the origin of covariation that defines our TTGs is either
193 due to a batch effect, shared variation of biological origin, including *trans* influences, or some
194 combination of the two. To determine if batch effects can, or cannot, best explain the behaviour of
195 the TTGs we carry out multiple replication studies to show that it is implausible that the effects we
196 observe are due to experimental artefacts. We also employ multiple analyses of different biological
197 properties of the TTGs to show that the distribution of these properties is also incompatible with
198 experimental batch effects.

199

200 **Replication analyses**

201 Replication analyses are carried out using the same procedure on two sets of data and
202 comparing the overlap of *trans* target genes identified by each. We use 3 classes of analysis

203 1. Within and between population replication.
204 2. Technical replication by analysis of the same mRNA samples with library generation and
205 sequencing in 2 independent labs.
206 3. Cross platform replication using mRNA samples of the same individuals analysed using different
207 analysis techniques.

208 Our reasoning is that technical batch effects affecting *trans* target genes are less likely to be
209 replicable between laboratories and across different platforms but that *trans* target genes should
210 replicate within and between different populations.

211

212 **Biological properties**

213 We then ask if the *trans* target genes have biological properties that are not expected to be
214 found in sets of genes that are simply influenced by batch effects. We use four approaches: (1) we
215 test for overabundance of TFs bound to the census promoter region of *trans* target genes, as binding

might be expected to be enriched in the case of *trans* influence compared to groups of covarying genes generated by batch effects; (2) we test for overabundance of genes that are sensitive to the knockdown of individual TFs, which are expected to be enriched in TTGs due to *trans* influence of TFs; (3) we test for overabundance of functional descriptors (KEGG pathways [19]) in the *trans* target genes: as gene groupings artefactually formed by batch effects are expected to be independent of any annotated biological function; and finally (4) we look at the protein-protein interactions (PPIs) of the proteins encoded by *trans* target genes and ask if they interact more than random expectation: again, batch effects are not expected similarly to influence genes that encode proteins with more interactions than random expectation. Related approaches have been used previously [5,8,9] to support a biological rather than batch origin of changes in gene expression.

Finally we show that inter-individual variation in expression of a set of genes encoding proteins involved in mRNA splicing and contained in a *trans* target gene set is significantly correlated with variation in splice patterns within the same individuals, suggesting that the shared variation of mRNA level in *trans* target genes may have phenotypic consequences.

These multiple lines of evidence lead to the conclusion that mRNA levels of substantial numbers of genes are influenced collectively in *trans* and that the covariation has phenotypic effects upon at least one property of human cells.

Results

Defining correlation thresholds

We set a correlation threshold for our analyses by systematically permuting the gene names of the expression values in each expression matrix created by sequential removal of PCs. We reason that permutation reveals background noise, and therefore any correlations observed in unpermuted data are likely to derive from either biological or background noise. We applied 1000 permutations and recorded the maximum absolute correlation in each matrix. In S2 Fig we graph for each PC removal step the correlation values corresponding to P=0.001, P=0.01 and P=0.05 based on 1000 permutations (see Materials and Methods). We do this for all sets of mRNA level matrices used in our analyses (S1 Table); for the Lappalainen et al. [2] data sets this is an absolute correlation more than 0.64. We examined the number of gene pairs with correlation above the threshold in mRNA level data with sequential removal of PCs (S3A Fig), and observe that the numbers decrease until about PC20, plateau until about PC40 then steadily increase. We also observe (S3B Fig) that in randomly permuted mRNA level data there are no gene pairs correlating above the threshold in PC1–20, followed by a steady increase of numbers of correlating gene pairs with subsequent removal of PCs. We therefore limit our analyses to the first 20 PCs.

Defining *trans* target genes

To identify *trans* target genes we used the RNA-Seq data for 88 individuals each drawn from the Caucasian, Yoruban and combined Finns, British and Toscani population from Lappalainen et al. [2] and carried out the process shown in Fig 1. In Fig 2 we show the number of genes within *trans* target gene sets as consecutive PCs are removed (54–2137, 75–2168 and 75–2155 genes for Caucasian, Yoruban and combined Finns, British and Toscani population, respectively). The pairwise mRNA level correlations in *trans* target genes are complex, and genes may have very different numbers of pairwise correlations above the threshold and not all genes in the TTGs are equally correlated to each other. To illustrate this, the gene-pair correlations are conveniently treated as a network with each gene as a node connected by edges representing correlations above the threshold. Summary statistics of these networks from all three populations are in S2 Table and as an example of the complexity of the correlations the gene *PPIAP29* identified in PC1 TTGs in the Lappalainen et al. [2] combined Finns, British and Toscani populations has 997 pairwise absolute correlations better than 0.64 — the highest number of connections in our data. An example (PC7

TTGs from Caucasian individuals Lappalainen et al. [2]) of the correlation network is shown in S4 Fig which is a network defined by 1578 genes and 6268 gene pairs with absolute correlation ≥ 0.64. This network contains around 0.5% of the 1,244,253 total possible gene pairs definable between member genes and 56% of all correlations are poor observed to lie between -0.2 and 0.2 (S4A Fig). The network is non-homogeneous and has obvious multiple clusters (S4B Fig). Generally the correlations that define TTGs correspond to 0.3–1.4% of all of the gene pairs in the sample (S2 Table).

Fig 2. Numbers of genes in *trans* target genes defined by principal components. The x-axis is sequential PCs removed from mRNA level data to identify *trans* target genes and the y-axis is numbers of genes identified as *trans* target genes defined by the relevant PC. Left panel, CEU: Caucasian individuals; centre panel, YRI: Yoruba individuals; right panel, FIN+GBR+TSI: Finns, British and Toscani mixed individuals. mRNA level data is from Lappalainen et al. [2].

This complexity notwithstanding, we would expect *trans* target gene identity to have some consistency across each of these analyses and so we identified for each population the union of genes that are identified as being a member of a group of TTGs defined by any PC up to 20 PCs; this list is the "combined *trans* target genes" of the population, which cover 52.7%, 55.8% and 61.5% of all expressed genes in LCLs from Yoruba, Caucasian and combined Finns, British and Toscani individuals, respectively (S3 Table). We then tested for the identity of overlapped genes between the combined TTGs of each population and in Table 1 we show the overlap in identity of genes within the TTGs detected within a population, in this case the 88 Caucasian and 88 Finns, British and Toscani (Lappalainen et al. [2]) is 84% and 75% overlap with an odds ratio of 13.6 (Table 1, Row 1), between 2 populations, in this case Caucasian and Yoruban is 74% and 79% overlap and has an odds ratio of 11.5 (Table 1, Row 2). In Fig 3A we show replication analyses for the TTGs defined by PC1–10 and note that there is significant replication of TTGs but the best overlaps are not always of TTGs defined by the same PC in the two analyses. This is to be expected for two reasons: firstly because batch effects will be different in the two independent analyses and thus there is no necessary equivalence between PCs, and secondly because *trans* effects might only be partially shared between and within populations of different individuals. In parallel we performed identical analyses for *trans* target gene pairs and observed significant overlaps of gene pairs within and between populations (Fig 3B). Nevertheless, there is a substantial replication of TTGs across populations but at this stage we do not believe these analyses eliminate the possibility that we are observing batch effects since these samples derive from a single, albeit multicentre, analysis.

Table 1. Overlap of *trans* target genes from different populations and different analyses.

Datasets[a]		Number of Genes			Percentage			
Dataset1	**Dataset2**	**TTG1[b]**	**TTG2[c]**	**Overlap[d]**	**TTG1[e]**	**TTG2[f]**	**P value[g]**	**Odds**
CEU	FIN+GBR+TSI	8386	9363	7029	83.8%	75.1%	<1e-300	13.6
CEU	YRI	8386	7863	6207	74.0%	78.9%	<1e-300	11.5
UNIGE	CNAG_CRG	3844	4893	2844	74.0%	58.1%	<1e-300	13.5
YRI Yale	YRI Argonne	8710	8590	7763	89.1%	90.4%	<1e-300	38.7
Illumina HiSeq 2000	Illumina GAII	2747	4538	1905	69.3%	42.0%	6e-134	3.2
RNA-Seq	Microarray	8062	8341	4286	53.2%	51.4%	6e-46	2.2

[a] Dataset1 and dataset2 are two set of combined *trans* target genes that are compared to establish the extent to which the sets overlap. CEU, Caucasian individuals; YRI, Yoruba individuals; FIN+GBR+TSI, Finns, British and Toscani individuals all from Lappalainen et al. [2]; UNIGE, RNA-Seq data from 52 individuals carried out by University of Geneva, and CNAG_CRG, RNA-Seq data from the same 52 individuals performed by the Centro Nacional de Análisis Genómico and Centre for Genomic Regulation (Lappalainen et al. [2]). YRI Yale, RNA-Seq data from 68 Yoruba individuals from

305 Yale sequencing center and YRI Argonne, RNA-Seq data from the same 68 Yoruba individuals from
306 Argonne sequencing center (Pickrell et al. [14]). Illumina HiSeq 2000, RNA-Seq data from 47
307 Caucasian individuals using the Illumina HiSeq 2000 platform (Lappalainen et al. [2]); Illumina GAII,
308 RNA-Seq data from the same 47 Caucasian individuals using the Illumina GAII platform (Pickrell et al.
309 [14]); RNA-Seq, RNA-Seq data from 80 Caucasian individuals (Lappalainen et al. [2]); Microarray,
310 microarray estimates on mRNA levels from the same 80 Caucasian individuals (Stranger et al. [16]).
311 [b] Numbers of unique genes in combined *trans* target genes identified from dataset1.
312 [c] Numbers of unique genes in combined *trans* target genes identified from dataset2.
313 [d] Numbers of genes observed in common between TTG1 and TTG2.
314 [e] The percentage of common genes in TTG1.
315 [f] The percentage of common genes in TTG2.
316 [g] *P* value for Fisher's exact test for the observed numbers of genes in common.
317

318 **Fig 3. Overlap of genes identified in *trans* target genes and *trans* target gene pairs defined in**
319 **different populations.** (A) Upper left panel, comparison of numbers of genes identified by first 10
320 PCs as *trans* target genes from both Caucasian (CEU) and Yoruba (YRI) individuals. Upper right panel,
321 similar analysis of Finns, British and Toscani (FIN+GBR+TSI) and CEU individuals. Axes are
322 sequentially numbered PCs; first row or column (label N) is the total number of genes in the TTG and
323 the intersection (labels 1-10) is the number of genes in common. Lower 2 panels, corresponding *P*
324 values from Fisher's exact test for the overlap of *trans* target genes in the 2 populations in the upper
325 2 panels. The most significant overlaps in a column are highlighted in orange in all panels. (B)
326 Identical analysis as (A) but measuring the overlap of specific gene pairs rather than genes in the
327 relevant populations.
328

329 **Technical replication of *trans* target genes**

330 To start to distinguish the confounding influence of batch effects upon *trans* target genes we
331 carried out a series of replication studies. We reason that the same analyses carried out by separate
332 laboratories will likely contain batch effects but these are not expected to be identical between the
333 two laboratories. These results based upon the analysis of combined TTGs summarised in Table 1
334 and full details for the relevant information derived from PC1–10 are in S5A Fig**.** We tested for
335 technical replication by using samples from Lappalainen et al. [2] who report the mRNA-Seq analysis
336 of the same 52 mRNA samples with library generation and sequencing in 2 independent
337 laboratories, and we observe replication with 74% and 58% overlap and an odds ratio of 13.5 (Table
338 1, Row 3). Similarly 68 Yoruban mRNA samples were sequenced by 2 separate laboratories reported
339 by Pickrell et al. [14], and we observe replication with 89% and 90% overlap and an odds ratio of
340 38.7 (Table 1, Row 4). In parallel we carried out identical analyses based upon *trans* target gene pairs
341 (S5B Fig). In all cases we observe substantial and significant replication of genes in the different TTGs
342 and TTGPs. These data suggest that the origin of the shared behaviour of *trans* target genes is not
343 readily explained by simple batch effects unless these are associated with particular types of
344 sequencing machines and so shared between laboratories.

345 To eliminate batch effects due to machine type, we compared *trans* target genes of the
346 same individuals analysed in different labs using different analysis machines or techniques. We
347 compare TTGs from Illumina GAII and HiSeq2000 analysis of the same 47 individuals analysed by
348 Pickrell et al. [14] and Lappalainen et al. [2], respectively (Table 1 Row 5 and S6 Fig) and we
349 compared the TTGs from the same 80 individual's mRNA analysed on microarrays and Illumina
350 HiSeq2000 by Stranger et al. [16] and Lappalainen et al. [2] respectively (Table 1, Row 6, S6 Fig).
351 Based upon the Fisher's exact test the genes that are contained in *trans* target gene sets defined by
352 any PC in these data sets replicate across the analyses with 69% and 42% overlap and an odds ratio
353 for enrichment (3.2). We even observe replication between mRNA-Seq and microarray analyses with
354 51% and 53% overlap and an odds ratio of 2.2, which is for data derived from two very different

355 methodologies. In parallel we conducted identical analyses of *trans* target gene pairs (S5 Fig) and
356 were again able to observe significant replication of the pairwise correlations of mRNA level levels.
357 The list of TTGs for each dataset mentioned above are provided in S3 Table and the IDs of the
358 individuals analysed in S4 Table.

359

360 **Batch effects**

361 The fact we see replication of *trans* target genes across multiple analyses, populations and
362 platforms makes it difficult to suggest that experimental batch effects alone are the source of the
363 shared behaviour. Later we will show that the genes contained within the *trans* target gene sets
364 indeed have a number of biological properties that cannot be accounted for by experimental
365 variations but we accept that batch effects must have an influence upon overall correlation of mRNA
366 levels. We therefore repeated our analyses but first used two methods to remove technical
367 artefacts. PEER (probabilistic estimation of expression residuals) from Stegle et al. [20] is a software
368 that is widely used for removing batch effects from mRNA-Seq data. PEER removes 61.7%, 63.0%
369 and 64% of the variance of the mRNA levels of LCLs from Yoruban, Caucasian, and the Finns, British
370 and Toscani population samples from Lappalainen et al. [2], and we identify 651, 693 and 644
371 combined *trans* target genes, compared with 7863, 8386 and 9363 combined *trans* target genes
372 identified without PEER, with an overlap of 6.1%, 6.2% and 5.4% of genes, respectively. PEER is very
373 effective at removing variance and so it is unsurprising that this leads to a very substantial drop in
374 *trans* target gene numbers but this analysis does not identify the sources of the shared variance and
375 so is likely to be removing both technical and biological sources of correlation.

376 In the second case we used an approach that was based upon variation that was due to GC
377 base composition biases in mRNA-Seq. This method was developed by Pickrell et al. [10] to identify
378 sample to sample deviation from expectation of mRNA-Seq read counts mapped to exons separated
379 by GC base composition and to correct for such deviation. For the Yoruban mRNA level data [2] this
380 approach removes 6% of the variance and after GC bias correction we replicate 95% of combined
381 *trans* target genes suggesting that variable GC bias is not an explanation for correlated behaviors of
382 genes.

383 In both cases these methods are designed to leave an expression level of an individual gene
384 as unaffected as possible by shared influences; in reality we believe it is most probable that removal
385 of shared variations or correction of deviation from random or other expectation is likely to be
386 removing both technical artefacts and biological signal from some genes and, inevitably, as greater
387 variance is removed, correlation in mRNA levels will necessarily decrease. As a consequence, we rely
388 on our PCA approach to remove variance systematically, recognizing that the sources of this variance
389 cannot be established from identification of correlations alone.

390 Collectively our analyses support the view that the behaviour of *trans* target genes is
391 unlikely to be best explained by batch effects as we see replication across both platforms and
392 laboratories, individuals and populations. We therefore set out to examine if *trans* target genes have
393 biological properties that are not associated with the technology of mRNA-Seq at a frequency
394 greater than random expectation: we argue that association of such biological properties with *trans*
395 target genes is supportive evidence for a biological explanation for correlations in mRNA levels.

396

397 **Some TFs show over abundant binding to census promoter regions of *trans* target genes**

398 Allocco et al. [21] have shown that the expression level of genes that share TFs bound to
399 their cognate promoter are better correlated than those that share fewer. We therefore analysed
400 the TF binding in the 1000 bp up- and down-stream of the transcription start site of sets of the *trans*
401 targets genes reasoning that if the behaviour of *trans* target genes is indeed of biological origin we

402 might expect particular TFs to be more associated with *trans* target genes than random expectation.
403 We use the binding site data from the 50 TFs (listed in S5 Table) whose binding has been established
404 in a lymphoblastoid cell line by the ENCODE project [22] and use hypergeometric test, with
405 appropriate correction for multiple tests, to identify enrichment for binding of any of the TFs to the
406 cognate promoter regions of the combined *trans* target genes using RNA-Seq data of 3 populations
407 containing 88 Yoruban, 88 Caucasian and 88 individuals from the British, Finns, Toscani populations
408 (data of Lappalainen et al. [2]). These results are displayed in Table 2. 38 of the 50 TFs have
409 significantly enriched binding to the promoter regions of genes in the combined TTGs defined from
410 one or more populations and in S6 Table we show the enrichment within the TTGs on a PC by PC
411 basis. In S7 Fig we show that there is substantial similarity in overabundance of TF binding to the
412 combined TTGs of 3 populations.

413

414 **Table 2. Enrichment of individual transcription factors bound to promoter regions of *trans* target**
415 **genes.**
416

	CEU		YRI		FIN+GBR+TSI	
TF symbol	**N**	**P**	**N**	**P**	**N**	**P**
ATF3	826	1.8e-05*	782	1.0e-07*	893	2.0e-05*
BCLAF1	1044	3.7e-05*	974	1.7e-05*	1131	2.0e-05*
BRCA1	222	1.7e-03	216	1.3e-05*	236	1.4e-02
CHD2	1691	1.5e-10*	1615	1.0e-17*	1827	3.6e-12*
EBF1	2164	3.3e-04*	2067	5.5e-07*	2428	2.5e-09*
EGR1	3425	1.4e-08*	3187	1.7e-10*	3776	9.2e-16*
ELF1	5114	5.3e-16*	4833	5.0e-29*	5595	8.9e-23*
ETS1	1344	1.4e-08*	1271	9.3e-12*	1472	1.5e-11*
FOS	866	4.3e-03	833	3.2e-05*	916	1.5e-02
GABPA	2468	2.7e-09*	2343	8.7e-16*	2683	3.7e-12*
IRF4	938	3.3e-05*	888	4.0e-07*	1025	6.3e-06*
MAX	501	1.3e-01	493	8.0e-04*	548	1.5e-01
MEF2A	1060	1.0e-04*	1017	5.2e-08*	1163	2.9e-05*
MEF2C	456	2.1e-03	432	4.7e-04*	497	3.5e-03
MYC	1176	4.8e-13*	1111	1.8e-15*	1260	6.5e-12*
NR2C2	345	4.2e-04*	328	7.8e-05*	355	2.1e-02
NRF1	2394	3.0e-07*	2233	1.2e-09*	2604	1.4e-09*
PAX5	2680	8.8e-08*	2508	2.2e-09*	2923	2.0e-09*
PBX3	1324	1.2e-06*	1249	8.0e-08*	1401	5.9e-04*
POU2F2	2684	7.0e-20*	2539	3.6e-27*	2884	1.3e-17*
RELA	750	1.9e-06*	702	1.2e-07*	799	8.9e-05*
RFX5	1038	3.7e-03	997	3.3e-07*	1146	5.4e-06*
RXRA	632	2.0e-06*	599	1.9e-07*	677	1.4e-05*
SIN3A	4057	1.4e-18*	3811	7.0e-26*	4370	4.6e-19*
SIX5	2137	5.3e-14*	2056	6.2e-23*	2301	2.1e-13*
SP1	3426	7.4e-14*	3277	1.9e-25*	3716	6.8e-16*
SPI1	1976	1.8e-05*	1873	2.8e-09*	2201	5.9e-09*
SRF	1775	7.7e-10*	1658	6.4e-12*	1896	6.0e-08*
TAF1	2635	7.5e-31*	2446	8.5e-31*	2797	2.0e-25*
TBP	2886	2.7e-26*	2692	1.2e-31*	3101	7.8e-26*
TCF12	2083	1.2e-10*	1946	2.5e-12*	2261	1.7e-11*
USF1	1604	4.9e-04*	1533	2.3e-07*	1765	1.9e-06*
USF2	1454	1.2e-02	1393	9.9e-06*	1586	1.6e-03
WRNIP1	344	3.0e-07*	327	1.7e-08*	351	2.1e-04*
YY1	4273	1.9e-19*	4008	8.8e-25*	4617	6.6e-20*

	CEU		YRI		FIN+GBR+TSI	
TF symbol	**N**	**P**	**N**	**P**	**N**	**P**
ZBTB33	795	4.2e-04*	756	1.6e-05*	852	1.6e-03
ZEB1	1922	7.7e-03	1821	2.9e-05*	2136	2.5e-05*
ZNF143	2506	2.4e-06*	2421	3.4e-14*	2768	7.4e-11*

The table lists the TFs that are found to be significantly associated to combined *trans* target genes in at least one population. CEU, Caucasian individuals YRI, Yoruba individuals; FIN+GBR+TSI, Finns, British and Toscani individuals all from Lappalainen et al. [2]. N: The number of *trans* target genes with individual transcription factors bound to promoter region defined as ±1000 base pairs of transcription start site; P: *P* value for hypergeometric test for enrichment of binding.
* $P < 1.0 \times 10^{-3}$ (Bonferroni corrected significance threshold).

In summary, multiple TFs compared to random expectation appear to be over abundantly associated with census promoter regions of *trans* target genes, which have been identified from multiple populations. We recognise of course that classification of genes as having a bound TF is necessarily an inference based upon the ENCODE [22] analysis of a single LCL; Although there is inter-individual variation in TF binding (*e.g.* 7.5% of binding regions of *NFKB* were found different among individuals by Kasowski et al. [23]), the scale of the overlaps suggests that such effects would have rather limited influence upon our analysis. At the very least, we would not expect enrichment of inferred TF binding to cognate promoters to be associated with batch effects, adding further support to the view that there is a biological basis for *trans* target genes behaviours.

Genes known to be sensitive to single TF knockdown are enriched in *trans* target genes

Our results suggest the possibility that the variation in the amount or activity of the appropriate TFs may be involved in generating the correlated behaviour of *trans* target genes, but numerous lines of experimental evidence show that for most genes binding of a single TF to a cognate promoter region is neither necessary nor sufficient for controlling mRNA level [24–26]. Consequently, our detection of overabundance of TF binding to some *trans* target gene sets does not necessarily imply a causal relationship between variation in those overabundant TF bindings and variation in their target genes' mRNA. Cusanovich et al. [5] tested the relationship directly by studying the effects of knockdown of expression of 59 TFs and chromatin modifiers (TFs used in this study are listed in S7 Table). They were able to identify sets of genes whose expression levels were directly influenced by knockdown of single TFs. If the relationship of *trans* target genes with TF is causal of shared behaviour, we might expect the genes shown by Cusanovich et al. [5] to be enriched in relevant *trans* target genes. In Table 3 we show that there is a statistically significant overlap of genes identified as combined *trans* target genes with genes defined as being sensitive to knockdown of 29 different TFs (see S8 Table for enrichment in PC by PC *trans* target genes). Of the 29 TFs, 9 are also analysed for binding in the ENCODE data: of these 9 TFs, 8 (*IRF4*, *RELA*, *POU2F2*, *PAX5*, *SP1*, *TCF12*, *USF1*, *YY1*) demonstrate both overabundance of binding sites, and enrichment for targets of TF knockdowns.

Table 3. Enrichment of differentially expressed genes from transcription factor knockdown in *trans* target genes.

	CEU		YRI		FIN+GBR+TSI	
TF Knockdown	**N**	**P**	**N**	**P**	**N**	**P**
ARNTL2	443	2.0e-04*	447	2.1e-09*	501	3.1e-08*
CEBPG	269	2.0e-02	275	2.0e-06*	286	2.2e-02
CEBPZ	263	1.4e-03	238	6.4e-03	287	7.3e-04*
CREBBP	1516	2.7e-08*	1416	1.6e-09*	1624	6.7e-07*

DIP2B	508	7.5e-04*	470	3.8e-04*	548	3.7e-04*
E2F1	270	8.2e-04*	244	4.1e-03	285	3.4e-03
ESRRA	375	1.5e-06*	333	2.5e-04*	373	3.8e-02
HOXB7	506	6.9e-04*	466	1.2e-03	531	8.6e-03
IRF4	2203	4.7e-12*	2062	1.3e-14*	2371	1.4e-10*
IRF5	546	7.2e-05*	517	6.9e-06*	566	1.3e-02
NFE2L1	525	2.5e-07*	473	1.9e-05*	528	1.1e-02
NFKB2	663	2.1e-04*	606	2.4e-03	703	4.7e-03
NR1D2	362	7.5e-04*	328	6.0e-03	376	9.9e-03
NR2F6	694	4.8e-06*	624	1.8e-03	762	4.5e-08*
NR3C1	281	9.1e-04*	272	3.6e-06*	303	8.0e-04*
PAX5	1380	2.2e-06*	1292	1.1e-07*	1469	1.2e-04*
POU2F1	411	8.4e-06*	385	9.2e-07*	443	2.4e-06*
POU2F2	478	1.1e-02	453	3.6e-04*	536	6.3e-05*
RAD21	1539	1.4e-05*	1430	1.8e-05*	1667	7.7e-05*
RELA	240	5.0e-02	236	5.4e-04*	250	1.5e-01
SP1	2023	2.0e-05*	1952	1.7e-13*	2212	4.2e-07*
STAT2	52	3.3e-04*	40	1.2e-01	50	6.1e-03
TCF12	537	5.0e-04*	494	2.3e-03	577	1.3e-03
TFDP1	694	5.4e-07*	647	1.1e-06*	729	8.3e-05*
TFDP2	468	6.0e-06*	442	7.3e-08*	488	5.8e-04*
TFE3	427	1.7e-03	401	2.4e-04*	462	9.1e-04*
USF1	310	5.1e-03	297	9.2e-04*	327	2.8e-02
YY1	1025	2.2e-02	957	6.0e-03	1135	8.9e-05*
ZHX2	298	4.4e-04*	269	5.7e-03	311	1.8e-03

The table presents genes sensitive to knockdown of individual TFs from Cusanovich et al. [5] that are significantly enriched in combined *trans* target genes from at least one population. CEU, Caucasian individuals; YRI, Yoruba individuals; FIN+GBR+TSI, Finns, British and Toscani individuals all from Lappalainen et al. [2]. N: The number of *trans* target genes that overlap with genes exhibiting differential expression following TF knockdown by Cusanovich et al. [5]. P: *P* value of hypergeometric test.

* $P < 9.6 \times 10^{-4}$ (Bonferroni corrected significance threshold).

Combinations of TFs are overabundant in promoter regions of *trans* target genes

Despite these analyses, it is unlikely that single TF binding is the only likely biological contribution to the variable expression of the *trans* target genes. As discussed earlier, multiple TFs generally bind to the census promoter region of genes [22,27] and it would therefore be useful to test for combinations of TFs binding to the census promoter. A systematic analysis of all possible combinations of the 50 TFs in the ENCODE data set is statistically impossible due to the huge numbers of combinatorial possibilities. To overcome this we used two approaches. Wang et al. [27] show that 20 pairs of TFs co-bind in the genome of a human LCL: we can ask if these pairs are more commonly found in promoter regions of *trans* target genes than would be expected at random. In the second approach we remove PCs sequentially until we have reached a minimum number of pairwise correlated genes and then search for shared binding of TFs in promoter regions of this minimal gene set; we reason that a correlated expression in just a pair of genes is the most minimal evidence of *trans* activity and therefore the shared TFs might be important in mediating *trans* influences. In both cases we use the hypergeometric test to detect overabundance of binding of the combinations of TFs, ignoring their relative location, to the cognate promoter region in both the combined and the PC defined *trans* target genes of the Yoruban, the Caucasian and the British, Finns, Toscani populations from Lappalainen et al. [2].

In the first approach, of the 20 TF pairs identified in Wang et al. [27], 14 (see Table 4) are enriched in the combined *trans* target gene set in one or more populations of which 8 pairs are enriched in all three populations S9 Table shows the enrichment of the PC defined *trans* target gene sets.

Table 4. Enrichment of pairs of transcription factors bound to promoter regions of *trans* target genes.

	CEU		YRI		FIN+GBR+TSI	
TF Pair	**N**	**P**	**N**	**P**	**N**	**P**
ATF3, NRF1	402	2.0e-03*	383	4.0e-05*	454	5.5e-06*
ELF1, YY1	3860	1.1e-19*	3636	3.5e-27*	4176	3.4e-21*
GABPA, YY1	2010	2.2e-10*	1903	9.6e-15*	2178	7.3e-12*
IRF4, JUND	20	2.8e-01	16	4.7e-01	22	4.2e-01
IRF4, MAX	130	1.1e-01	118	2.1e-01	136	4.5e-01
IRF4, MEF2A	441	3.5e-03	421	1.7e-04*	478	1.6e-02
IRF4, MEF2C	231	5.3e-02	215	3.4e-02	246	1.7e-01
IRF4, RELA	271	3.9e-03	254	4.6e-04*	293	6.9e-03
IRF4, TCF12	510	6.0e-04*	463	3.1e-03	546	3.5e-03
JUND, MEF2C	9	5.9e-01	10	2.1e-01	9	8.9e-01
JUND, SPI1	11	4.2e-01	11	8.8e-02	13	3.6e-01
MAX, SPI1	193	2.1e-01	187	5.6e-02	211	2.1e-01
MAX, USF1	354	1.3e-01	352	7.2e-04*	387	9.7e-02
MAX, USF2	361	2.0e-01	358	1.6e-03*	396	1.3e-01
MEF2A, SPI1	466	2.8e-04*	439	2.1e-05*	513	2.5e-05*
MEF2C, SPI1	186	9.9e-03	181	2.9e-04*	204	5.4e-03
PBX3, SP1	1072	1.4e-05*	1018	1.2e-07*	1129	1.5e-03*
RELA, SPI1	394	2.8e-05*	357	1.1e-04*	416	2.5e-03*
SPI1, TCF12	793	1.5e-05*	730	2.8e-05*	885	7.7e-09*
USF1, YY1	1186	1.1e-04*	1140	3.9e-09*	1301	6.1e-07*

The table lists the pairs of TFs taken from Wang et al. [27] that are found to be significantly associated to combined *trans* target genes in at least one population. CEU, Caucasian individuals YRI, Yoruba individuals; FIN+GBR+TSI: Finns, British and Toscani individuals all from Lappalainen et al. [2]. N: The number of *trans* target genes with individual pairs of transcription factors bound to promoter region defined as ±1000 base pairs of transcription start site; P: *P* value for hypergeometric test for enrichment of bindings.

* $P < 2.5 \times 10^{-3}$ (Bonferroni corrected significance threshold).

In the second approach we sequentially removed PCs from the Lappalainen et al. [2] Yoruban, Caucasian and British, Finns, Toscani data sets. Respectively removal of 30, 31 and 31 PCs yielded a minimum of 2142, 1892 and 1719 gene pairs with absolute correlation better than 0.64 of which 303, 178 and 258 had between 2 to 25 TF binding sites in common. 51 combinations were common to all 3 populations and so we then tested these for significance of the overabundance of these combinations of bound TFs in the promoter regions of combined and PC defined *trans* target genes of the 3 populations. In S10 Table we identify significant enrichments of combinations of TFs within both the combined and PC defined *trans* target genes. In the combined *trans* target genes 23 of the 51 combinations of TFs are enriched, 17 in all 3 populations, 1 in Yoruban and British, Finns and Toscani and 5 in just the Yoruban. These data suggest the hypothesis that combinations of TFs might be contributing to the correlated variation in expression of *trans* target genes and also again reinforce the view that these biological properties are not those expected of technical batch effects.

508 ***Trans* target genes encode proteins with related functions**

509 Our data collectively suggest groups of genes are covarying in humans under complex *trans*
510 control. We wish to understand if these genes might be associated with biological functions firstly to
511 further prove the view that covariation has a biological origin and secondly to start to understand
512 what might be, if any, the phenotypic consequences of *trans* variation. We wish to ask two questions
513 as to the biological properties of proteins encoded in both combined and individual groups of *trans*
514 target genes. Firstly, do *trans* target genes encode proteins that physically interact more commonly
515 than random expectation, and secondly does this indicate a possible functional interaction that can
516 be detected by enrichment of functional annotation within the KEGG pathway database?

517 We tested for physical interaction of encoded proteins within the combined *trans* target
518 genes identified from Yoruban, the Caucasian and the British, Finns and Toscani populations from
519 Lappalainen et al. [2] using protein-protein interactions from Lage et al. [28] and using a
520 permutation test for significance. We compared the number of PPIs seen in the various *trans* target
521 gene sets with expectation created by 1000 random samples of genes (see Materials and Methods).
522 These analyses show that the combined *trans* target gene sets from all 3 populations encode
523 proteins that interact more commonly than expected; the Youruban set contains 4241 combined
524 *trans* target genes with 36,927 protein-protein interactions ($P < 0.001$); the Caucasian set of 4554
525 proteins with 41,714 PPIs ($P < 0.001$) and the Finns, British, Toscani with 4969 proteins with 45,376
526 PPIs ($P < 0.001$)

527 In S11 Table we show similar PC by PC analysis of *trans* target genes indicating that for all
528 three populations there are multiple significant enrichments for PPIs within multiple PC defined
529 *trans* target gene sets. The finding that the proteins encoded by *trans* target genes are more likely to
530 interact with each other than expectation again suggests that the variation in mRNA levels is unlikely
531 to be due to batch effects within sample analyses. More interestingly in many cases proteins that
532 physically interact have been shown to be contained within pathways of biological activity [29] and
533 suggests therefore that *trans* target genes might be associated with specific pathways of biological
534 activity.

535 We therefore tested for this by asking if pathways defined within the KEGG database are
536 enriched in *trans* target genes. In Table 5 we show the results of enrichment analysis of KEGG
537 function terms in the combined *trans* target genes using hypergeometric test (see Materials and
538 Methods) and in S12 Table we present the results of a similar analysis on a PC by PC analysis of *trans*
539 target genes. We see enrichment for 10 functions in the combined *trans* target gene sets: oxidative
540 phosphorylation, pyrimidine metabolism, ribosome, spliceosome, proteasome, cell cycle, protein
541 processing in endoplasmic reticulum, Alzheimer's disease (AD), Parkinson's disease (PD) and
542 Huntington's disease (HD). In these pathways 31–109 genes are shared in one population (Table 5)
543 and 30–86 genes are shared within all 3 populations (S8 Fig). These results suggest that in humans
544 there are significant numbers of genes that are functionally related and covarying and opens the
545 possibility that *trans* variation in genes involved in the same pathway might have phenotypic
546 outcomes.

547

548 **Table 5. Enrichment of KEGG pathways in *trans* target genes.**

	CEU		YRI		FIN+GBR+TSI	
KEGG pathway	**N**	**P**	**N**	**P**	**N**	**P**
Oxidative phosphorylation	79	2.2e-03	79	5.2e-05*	87	1.9e-04*
Pyrimidine metabolism	55	3.4e-02	50	4.9e-02	64	7.0e-05*
Ribosome	71	4.3e-11*	67	4.2e-11*	69	1.1e-07*
Spliceosome	83	1.5e-05*	83	8.5e-08*	81	7.8e-04
Proteasome	31	3.1e-04	33	2.4e-06*	33	1.9e-04*
Cell cycle	82	1.8e-04*	69	1.2e-02	81	2.0e-03

	CEU		YRI		FIN+GBR+TSI	
KEGG pathway	**N**	**P**	**N**	**P**	**N**	**P**
Protein processing in endoplasmic reticulum	90	6.2e-02	90	3.6e-03	108	9.4e-05*
Alzheimer's disease	91	6.5e-05*	89	6.0e-06*	98	5.6e-06*
Parkinson's disease	82	2.0e-05*	80	5.6e-07*	87	1.5e-05*
Huntington's disease	104	2.1e-04*	97	6.9e-05*	109	2.0e-04*

549 KEGG pathways that are enriched in the combined *trans* target genes in at least one population.
550 CEU, Caucasian individuals; YRI, Yoruba individuals; FIN+GBR+TSI: Finns, British and Toscani
551 individuals all from Lappalainen et al. [2]. N: The number of *trans* target genes annotated with a
552 KEGG pathway; P: *P* value of hypergeometric test.
553 * $P < 2.8 \times 10^{-4}$ (Bonferroni corrected significance threshold).

554

555 **Expressions of spliceosome genes correlate with changes in splice patterns**

556 Testing of phenotypic outcomes of variation in the *trans* target genes enriched in particular
557 pathways in most cases would require further case control or biochemical analyses but the
558 hypothesis that expression variation within the spliceosome set of genes results in changes to splice
559 patterns can readily be tested by analysis of the alternative splicing in the same mRNA samples that
560 we have used in our analyses. Variation in mRNA levels will result in some variation in protein levels
561 and Battle et al. [30] showed the median Spearman's correlation between LCL mRNA levels and
562 protein levels across 62 Yoruban individuals for 4322 genes is 0.14. Even though this correlation is
563 relatively weak the mean variation for the 83 spliceosome genes in the combined *trans* target genes
564 is 3.2 fold for mRNA levels and 3.0 fold for protein levels (see S13 Table). Given this variability we
565 sought to ask if covariation in the mRNA levels of the spliceosome genes might contribute to
566 changes in splicing. We focused upon the RNA-Seq data of 88 individuals from the Yoruban and
567 Caucasian sample sets from Lappalainen et al. [2] and we used the software MISO [31] which uses
568 annotations of alternative splicing events of human genes (see Materials and Methods). We
569 calculated the "percent spliced in" (PSI, or Ψ) for each alternative splicing event in the LCL. PSI values
570 were calculated for two classes of alternative splicing: retained introns (RI), where an intron is
571 retained in the mRNA transcript relative to other transcripts from the same gene where the intron is
572 removed, and skipped exons (SE) where an exon is removed by splicing in some but not all
573 transcripts of the same gene. In the Yoruban samples MISO detects 1792 RI events in 1196 genes
574 and 11641 SE events in 4922 genes and for the 88 Caucasian samples there were 1609 RI events in
575 1109 genes and 10661 SE events in 4663 genes. For each of the Yoruban and Caucasian individuals
576 (in brackets) we calculated a per individual average of all 1792 (1609) RI PSI values and for SE we
577 used the alternative splicing events corresponding to the largest 1792, top 15.4%, (1609, top 15.1%)
578 standard deviations of PSI values and calculated a per individual average PSI of this set of SE events.
579 These average measurements are a course measure of variability in splicing and we then ask how
580 variability of the expressions of spliceosome genes might be correlated with the variability of RI and
581 SE PSI values. In total in the 88 Yoruban data set there are 83 (out of 107 with detectable levels)
582 spliceosome genes in the combined *trans* target genes and for the Caucasian individuals there are 83
583 out of 110 detectable genes: 73 genes are found in the combined *trans* target genes of both
584 populations. There are both positive and negative correlations of mRNA levels of the 83 genes and
585 so to capture the variability we used the PC1 eigenvector as a convenient summary and in Fig 4 we
586 plot for each individual the average PSI value for the individual against the value for that individual in
587 the PC1 eigenvector. We observe significant negative correlation in both Yoruban and Caucasian
588 samples (data in brackets) of the PC1 eigenvector with average PSI of both RI and SE events: for RI
589 the Pearson's correlation coefficient is -0.49 (-0.66) and for SE -0.45 (-0.48), and for RI the R^2 is 0.24
590 (0.43) and for SE 0.21 (0.23). In contrast an identical comparison based upon the 24 (Yoruban) and
591 27 (Caucasian) spliceosome genes that are not contained in *trans* target genes sets yields (Caucasian
592 value again in brackets) a correlation of -0.29 (-0.37) and R^2 of 0.085 (0.14) for RI, and a correlation

593 of -0.31 (-0.17) and R^2 of 0.094 (0.03) for SE events. These data suggest that for both populations
594 variation of *trans* target genes from the spliceosome pathway has a significant impact upon
595 differential splicing of genes even when the average PSI for multiple genes is used as a summary
596 measure of differential splicing; the *trans* target genes variation accounts for 43% of the variation in
597 the Caucasian RI events and given the variation is an average PSI for 1609 events this is a substantial
598 influence.

599

600 **Fig 4. Association between variation in expression of spliceosome genes and alternative splicing.**
601 Left panels: analysis of Yoruba (YRI) and right side Caucasian (CEU) individuals. Upper panel is
602 analysis of retained introns and lower panel of skipped exons in the relevant population. The y-axis is
603 the per individual average percent spliced in (Ψ) plotted against the x-axis, PC1 eigenvector of the
604 expressions of 83 spliceosome genes contained in TTGs identified in that population. For each panel,
605 a regression line is shown, indicating negative association. Pearson's correlation coefficient r, R^2 and
606 *P* value for t-test with a null of zero correlation, are listed at the top of each panel.
607

608 ## Discussion

609 In this paper we have used multiple lines of evidence to show that correlation in mRNA
610 levels has a biological rather than purely technical origin. In summary we show this by using PCA of
611 multiple sets of mRNA-Seq data to identify genes with correlated mRNA levels. We then use four
612 lines of evidence to show covariation is unlikely to be due to batch effects: 1) they can be replicated
613 within and between populations and across different analytic platforms including mRNA-Seq and
614 microarrays. 2) they contain an overabundance of cognate TFs bound to their census promoter
615 regions and they are enriched for genes known to be sensitive to knockdown of single TFs. 3) the
616 proteins they encode are more likely to physically interact with each other than expected. 4) they
617 comprise groups of genes encoding shared functions more frequently than expected. These multiple
618 lines of evidence lead us to believe that the *trans* target genes are indeed covarying under the
619 influence of variation in their *trans*-acting controllers and that the cumulative *trans* effects can
620 account for a significant proportion of the individual genes' expression variation.

621

622 ### The properties of *trans* target genes align with what is known about control of mRNA
623 ### levels

624 The control of mRNA levels can occur at any point from initiation of transcription, to splicing,
625 mRNA stability and degradation and these process all involve multiple protein and/or RNA
626 interactions in *trans* [32]. If we focus upon transcriptional control by TFs then any given gene can be
627 controlled by multiple TFs and the action of any given TF can be activating or repressive conditional
628 on proximity to promoter and/or interacting with or co-binding to other TFs. Variation of a TF can
629 therefore be expected to influence sets of genes that based upon Cusanovich et al. [5] can number
630 from a few dozen to over 3000. However the simultaneous variation of multiple TFs will necessarily
631 have complicated outcomes conditional upon the overlap of influence; presently we have no
632 understanding of the extent of these outcomes but it is likely that the mRNA of subsets of genes will
633 be increased, decreased or remain unaltered in any one individual under the simultaneous variation
634 of two TFs. Ultimately the combinatorial nature of the control of gene expression by TFs means that
635 *trans* influences could influence anywhere from several thousand down to a single gene and that
636 these effects could be significantly different both between individuals and between populations; this
637 behaviour is indeed what we observe in our replication studies where we observe significant
638 overlap, but not identity, of *trans* target genes.

639

640 **How much variation is associated with combined *trans* influences?**

641 That combined *trans* target genes are less variable than non-*trans* target genes (defined as
642 all expressed genes other than combined *trans* target genes) with median coefficient of variation
643 0.33 compared to 0.41 (P = 3.7x10^{-85}, Wilcoxon test) but comparison of the variance sequentially
644 removed by PCA for the two sets of genes reveals that *trans* target genes lose variance faster than
645 non-*trans* target genes until PC9 when the reverse becomes true for the RNA-Seq data of Yoruban,
646 the Caucasian and the British, Finns and Toscani populations from Lappalainen et al. [2]. The total
647 variance of expressions of *trans* target genes accounted by the first 9 PCs is 63.8%, 62.2% and 64.3%,
648 respectively and the total variance of expression of non-*trans* target genes accounted for by the first
649 9 PCs are 37.8%, 37.8% and 37.5%, respectively. Therefore, the difference in average variance lost
650 between TTGs and non-TTGs is 26.0%, 24.4% and 26.8% for the 3 populations, respectively. It is
651 reasonable that the difference reflects the effects of the shared variation which, by virtue of its
652 scale, is recognized earlier in the PCA sequential analysis and suggests that at least 24% of mRNA
653 level variation appears to be due to combined *trans* effects.

654 **Comparison to other studies**

655 Goldinger et al. [13] systematically studied the influence of PCA on mRNA levels and eQTLs
656 using microarray data and tests of SNP association and heritability of expression. They concluded
657 "*Most importantly, we show that the first 50 PCs, which have been removed in previously published*
658 *articles to correct for batch effects (references omitted), contain …… a considerable proportion of*
659 *genetic variation influencing gene expression……*". We note that microarray and mRNA-Seq have
660 quite different distributions of variance explained per PC: for example PC1 explains a much greater
661 proportion of variance from microarray data than it does from mRNA-Seq data. This means our
662 findings cannot be directly related to those of Goldinger et al. [13] both because the variance
663 structures are very different and also because we can have no insight into the causation of shared
664 influences—a necessary feature of our approach. However, our findings similarly suggest batch
665 effects and biological signal can be confounded and techniques to de-noise mRNA-Seq data need to
666 be applied very carefully. As we discuss above, the PEER software of Stegle et al. [20] can effectively
667 removed most of the covariation that underpins *trans* target genes' behavior.
668

669 The approach we have developed has advantages and disadvantages over present more
670 gene centric eQTL approaches. The advantages are, firstly, the overall level of an mRNA is
671 determined by many influences and it is the actual level that ultimately may be associated with
672 biological outcome. In contrast, eQTL based approaches dissect the causative variations influencing
673 mRNA levels and given these are individually small effects and subject to epistatic interactions this
674 can be a challenging approach to detailing the final mRNA level variation. Secondly our approach
675 does not generate the multiple testing problem of *trans* eQTL analyses and so can be applied to
676 relatively small data sets. Thirdly, the correlated mRNA changes within functionally related genes
677 may well prove to have a greater predictive effect upon phenotypic outcomes than the single *cis*- or
678 *trans*-eQTL associations. The disadvantages to our approach are, firstly that we require mRNA level
679 analysis of phenotypically relevant tissues (a problem which is increasingly shared by tissue specific
680 eQTL analyses) and secondly that we do not gain mechanistic understanding at the level of
681 association between sequence variation and effector regulators or targets. We also recognize that
682 the LCL resources we have used here are tissue culture cell lines and therefore subject to culture
683 induced changes (see for example Yuan et al. [33] and Choy et al. [34]) which may confound analysis.

684

685 **Spliceosome mRNA variation and alternative splicing**

686 The finding that collective variation in the mRNA level of the spliceosome genes may
687 account for 21–43% of the variation in alternative splicing of multiple genes enables us to ask if
688 collective behavior of TTGs might have a stronger effect than variation in expression of individual

689 genes. This is not the case; we compare the 21–43% collective influence to the influence of single
690 spliceosome genes using a similar analysis framework. We calculated the correlation of variation in
691 the mRNA levels of individual genes to the average PSI values of RI events across the same
692 individuals and we observe correlations in both the Caucasian and the Yoruban data (see S14 Table)
693 that range from -0.698 to 0.652 and R^2 of up to 0.487 with about 1/3 of the single gene correlations
694 being positive; these figures replicate with a correlation between the two populations of 0.72. The
695 very complex structure of *trans* target gene covariation combined with the complex biochemistry of
696 splicing events where changing levels of proteins could stimulate or decrease splicing activities,
697 suggests that for any set of genes in a pathway the overall effect of expression variation is as likely to
698 reduce the impact of variation in individual genes as it is to enhance. This highlights the interesting
699 possibility that the control of gene expression in *trans* should be additionally viewed as a pathway
700 homeostatic mechanism.

701

702 **Alzheimer's, Parkinson's and Huntington's disease and oxidative phosphorylation**

703 The enrichment in the combined *trans* target genes for the KEGG functions Alzheimer's
704 disease (62%, 62% and 68% of 158 KEGG annotated genes, Caucasian, Yoruban and British, Finns,
705 Toscani populations respectively), Parkinson's diseases (65%, 63% and 69% of 127 genes) and
706 Huntington's disease (59%, 55%, and 62% of 177 genes) is particularly striking (see S15 Table for
707 details). In the Yoruban samples of Battle et al. [30] the mRNA levels vary a median of 3.0, 2.9 and
708 2.9 fold change and protein levels a median of 4.2, 4.4 and 3.9 fold change for AD, PD and HD genes,
709 respectively. The enrichment is driven in good measure by 58 genes functionally classified as
710 oxidative phosphorylation which is itself an enriched function (62%, 62% and 68% of 128 genes).
711 These 58 genes are detected as enriched in the combined *trans* target genes of all 3 populations and
712 a further 12 oxidative phosphorylation genes are found in TTGs from one or two populations only. In
713 total an additional 17 genes involved in oxidative phosphorylation are found in one or more of the 3
714 disease functions in TTGs from one or more populations yielding a total of 88 genes involved in
715 oxidative phosphorylation out of 265 genes in the combined list of genes involved in one or more
716 disease categories. The involvement of the mitochondrion in the etiology of all three diseases has
717 been extensively documented (see Wang et al. [35], Franco-Iborra et al. [36], Turner and Schapira
718 [37] for reviews of mitochondrial dysfunction and AD, PD and HD respectively and Biffi et al. [38] for
719 genetic evidence for involvement in AD) and our detection of correlated changes raises the question
720 of whether this variation results in significant change to mitochondrial function and therefore an
721 increased susceptibility to, or severity of, disease outcome. It has also been widely discussed (see
722 Sun et al. [39] for a recent review) that aging processes are associated with changes to mitochondrial
723 function and properties. This suggests that identifying an effect of the natural covariation of genes
724 involved in the oxidative phosphorylation pathway could be a significant contribution to
725 understanding the natural variation of the human aging process.

726 For the remaining functional categories, pyrimidine metabolism, ribosome, spliceosome,
727 proteasome, cell cycle, protein processing in endoplasmic reticulum there is relatively little overlap
728 with the exception of four proteins involved in ubiquitin conjugating and a ubiquitin ligase shared
729 between protein processing in endoplasmic reticulum and Parkinson's disease (see Ross et al. [40]
730 for a review of the role of ubiquitin and mitochondrial damage in PD and AD) and 10 RNA
731 polymerase II subunits shared between pyrimidine metabolism and Huntington's disease.

732 In each case testing of the relationship of variation in mRNA levels to ultimate biochemical
733 phenotype will be challenging but we believe that we have defined some potential targets that could
734 contribute to natural human variation that in turn might have significant impact upon our health.

735

Materials and Methods

Datasets of mRNA level in human LCLs

1. Lappalainen et al. [2] mRNA-Seq data from LCLs using the Illumina HiSeq2000 platform. Reads per kilobase per million mapped reads (RPKM) values of genes expressed in LCLs from Caucasian (CEU), Yoruba (YRI) and Finns, British and Toscani (FIN+GBR+TSI) individuals were obtained from ArrayExpress [41] under accession number E-GEUV-1. We selected expressed genes defined as non-zero RPKM values across all LCL samples yielding 15,016 genes from CEU, 14,918 genes from YRI and 15,231 genes from FIN+GBR+TSI. Where available, RPKM values of duplicate samples from the same LCL were averaged for each gene. We inspected log10 transformed RPKM distributions for all samples and excluded 1 YRI, 2 FIN and 1 GBR samples with outlier distributions. We included 88 individuals from YRI, 91 individuals from CEU and 279 individuals from FIN+GBR+TSI. 88 individuals were then randomly selected from CEU and FIN+GBR+TSI dataset to make the same sample size for 3 datasets. To correct for confounding effect of gender on mRNA level measurements, linear regression of each mRNA level profile was then applied using R function *lm*.

2. Pickrell et al. [14] mRNA-Seq data from LCLs from 69 Yoruba individuals sequenced at Yale (YRI Yale) and Argonne (YRI Argonne) sequencing center using the Illumina GAII platform, and Montgomery et al. [15] mRNA-Seq data of LCLs from 60 Caucasians (CEU) using the Illumina GAII platform. To quantify mRNA level of genes, for both RNA-Seq data from Pickrell et al. [14] and Montgomery et al. [15] reads were mapped to the human reference genome (GRCh37) using BWA [42] v0.5.9 and Samtools [43] v0.1.12a, and counted using *Rsamtools* [44] *and GenomicRanges* [45] based on human gene annotations from Ensembl Genes 69 [46]. RPKM values were then calculated for each gene. We detected expressed genes including 12,171 genes from YRI Yale, 12,385 genes from YRI Argonne and 9,418 genes from CEU dataset. RPKM values of duplicate samples from the same LCL were averaged for each gene. We inspected log10 transformed RPKM distributions for all samples and excluded 1 YRI Yale, 1 YRI Argonne and 6 CEU samples with outlier distributions. RPKM values with the remaining 68 individuals from YRI Yale, the same 68 individuals from YRI Argonne and 54 individuals from CEU dataset were included in the following analysis. Linear regression of each mRNA level profile in Pickrell et al. [14] was carried out against sample RNA concentration and gender and for Montgomery et al. [15] against gender, in both cases using R function *lm.*

3. Stranger et al. [16] microarray data from LCLs from 109 Caucasian individuals using Illumina's Human-6 Expression BeadChip version 2. Normalized mRNA level values were obtained from ArrayExpress [41] under accession number E-MTAB-198. We detected 15,211 expressed genes from the microarray mRNA level dataset. Linear regression of each mRNA level profile on gender was performed using R function *lm*.

Principal component analysis

To calculate mRNA level matrices X_n with sequentially removed principal components, we performed principal component analysis on mRNA level data for each dataset using R function *svd*.

mRNA correlation analysis

We calculated correlations of all gene pairs from all mRNA level matrices as appropriate using absolute Spearman's correlation coefficient $|\rho|$ and R function *cor*. Gene correlations were considered as significant as described in the paper and listed in S2 Fig and S1 Table. Gene correlation networks were built by connecting gene pairs with correlation above the threshold by edges using *igraph* [47].

782 **Identification of *trans* target genes**

783 We defined *trans* target genes (TTGs) as any gene that can be observed in one or more gene
784 pairs correlating above the threshold from mRNA level data X_n, but not in any gene pairs correlating
785 above the threshold from mRNA level data X_{n+1}. *Trans* target gene pairs (TTGPs) were defined as any
786 gene pair with correlation above the threshold from mRNA level data X_n, but the same gene pair
787 correlating below the threshold from mRNA level data X_{n+1}.

788 **Replication analyses**

789 We used Fisher's exact test using R function *fisher.test* to test whether there is significant
790 overlap of genes or gene pairs between *trans* target genes identified from different data sets,
791 Bonferroni corrected for the number of comparisons.

792 **PEER and GC-content bias correction**

793 PEER was used in processing mRNA level data with 22 unobserved factors (25% of sample
794 size) as recommended in Stegle et al. [20]. To investigate the effect of GC-content bias on the
795 covariation of mRNA level we applied the GC-content bias correction procedure as described in
796 Pickrell et al. [14]

797 **Identification of binding sites of transcription factors**

798 TF binding data for human LCL GM12878 was from the ENCODE (Encyclopedia of DNA
799 Elements) project [22]; hg19 coordinates of TF binding regions were obtained from the "Txn Factor
800 ChIP" track (the wgEncodeRegTfbsClusteredV2.bed.gz file). We used binding regions for 50 of the TFs
801 (see S5 Table) that are consistently expressed in all 3 RNA-Seq datasets. We defined a census
802 promoter region as the 1000 bp upstream to 1000 bp downstream of gene TSS. Overlaps between TF
803 binding regions and gene promoter regions were detected using *GenomicRanges* [45] and gene
804 annotations from Ensembl Genes 69 [46].

805 **Enrichment analysis for TF binding, TF knock down and KEGG pathways**

806 Enrichment tests were performed by upper-tailed hypergeometric test using R function
807 *phyper,* Bonferroni corrected for the number of tests, to determine whether binding of individual,
808 pairs and combinations of TFs are enriched in promoter regions of TTGs, or whether KEGG pathways
809 or genes that are differentially expressed from the TF knockdowns of Cusanovich et al. [5], are
810 significantly enriched in *trans* target genes. We retrieved Ensembl gene IDs for 229 KEGG pathways
811 [48] using *org.Hs.eg.db* [49]. To identify TTGs with mitochondrial localization, the list of 1158
812 mitochondrial human genes were obtained from MitoCarta2.0 [50].

813 **Analysis of protein-protein interactions**

814 We identified protein-protein interactions from Lage et al. [28]. To test for significance we
815 sampled without replacement the same number of genes as the group of TTGs under test and
816 counted the number of PPIs among them and performed 1000 iterations to estimate a null
817 distribution of numbers of PPIs among the relevant number of proteins. A P value was assigned to
818 the observed number of PPIs in TTGs according to its ranking in the null distribution, Bonferroni
819 corrected for the numbers of groups of genes tested.

820 **Analysis of alternative splicing**

821 RNA-Seq reads for LCLs from 88 Yoruba and 88 Caucasian individuals [2] were mapped to
822 genomic regions of retained intron (RI) or skipped exon (SE) events using TopHat2 [51]. Annotations
823 of RI and SE events were obtained from MISO [31] (specifically the *miso_annotations_hg19_v2* file).
824 Only uniquely mapped reads were considered in the following analysis. To quantify the level of
825 alternative splicing, MISO was used to detect alternative splicing events and calculate the percent
826 spliced in (ψ) values. To test the association between expressions of spliceosome genes in the
827 combined TTGs of the YRI and CEU RNA-Seq data [2] we calculated using R function *svd* the PC1
828 eigenvector of expression profiles of the spliceosome genes to represent their shared expression

829 pattern across multiple individuals. Separately we calculated for each sample the median ψ value of
830 all detected RI or SE events. Associations between the sample median ψ of RI or SE events and the
831 PC1 eigenvector of expression profiles of spliceosome genes were calculated using Pearson's
832 correlation coefficient r, which was tested using R function *cor.test* at the significant threshold $P <$
833 1.25×10^{-2} (Bonferroni corrected for 4 tests).

834

835

836 ## Acknowledgments

837 We would like to thank Teo Yik Ying for advice and insight.

838

839 ## References

840 1. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more
841 likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6:
842 e1000888. doi:10.1371/journal.pgen.1000888

843 2. Lappalainen T, Sammeth M, Friedländer MR, Hoen PA, Monlong J, Rivas MA, et al.
844 Transcriptome and genome sequencing uncovers functional variation in humans. Nature.
845 2013;501: 506–511. doi:10.1038/nature12531

846 3. Huan T, Liu C, Joehanes R, Zhang X, Chen BH, Johnson AD, et al. A systematic heritability
847 analysis of the human whole blood transcriptome. Hum Genet. 2015;134: 343–358.
848 doi:10.1007/s00439-014-1524-3

849 4. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting global gene
850 expression analysis. Cell. 2012;151: 476–482.

851 5. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in
852 transcription factor binding. PLoS Genet. 2014;10: e1004226.
853 doi:10.1371/journal.pgen.1004226

854 6. Barrera LA, Vedenko A, Kurland J V., Rogers JM, Gisselbrecht SS, Rossin EJ, et al. Survey of
855 variation in human transcription factors reveals prevalent DNA binding changes. Science.
856 2016;351: 1450–1454.

857 7. Wu C, Delano DL, Mitro N, Su S V, Janes J, McClurg P, et al. Gene set enrichment in eQTL data
858 identifies novel annotations and pathway regulators. PLoS Genet. 2008;4: e1000070.
859 doi:10.1371/journal.pgen.1000070

860 8. Rotival M, Zeller T, Wild PS, Maouche S, Szymczak S, Schillert A, et al. Integrating genome-
861 wide genetic variations and monocyte expression data reveals trans-regulated gene modules
862 in humans. PLoS Genet. 2011;7: e1002367. doi:10.1371/journal.pgen.1002367

863 9. Brynedal B, Choi J, Raj T, Bjornson R, Stranger B, Neale B, et al. Large-scale trans-eQTLs affect
864 hundreds of transcripts and mediate patterns of transcriptional co-regulation. bioRxiv. 2016;
865 56283. doi:10.1101/056283

866 10. Cowley MJ, Cotsapas CJ, Williams RBH, Chan EKF, Pulvers JN, Liu MY, et al. Intra- and inter-
867 individual genetic differences in gene expression. Mamm Genome. 2009;20: 281–295.
868 doi:10.1007/s00335-009-9181-x

869 11. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the

870 widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet.
871 2010;11: 733–739. doi:10.1038/nrg2825

872 12. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in
873 analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS
874 One. 2011;6: e17238. doi:10.1371/journal.pone.0017238

875 13. Goldinger A, Henders AK, McRae AF, Martin NG, Gibson G, Montgomery GW, et al. Genetic
876 and nongenetic variation revealed for the principal components of human gene expression.
877 Genetics. 2013;195: 1117–1128. doi:10.1534/genetics.113.153221

878 14. Pickrell JK, Marioni JC, Pai A a, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding
879 mechanisms underlying human gene expression variation with RNA sequencing. Nature.
880 2010;464: 768–772. doi:10.1038/nature08872

881 15. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al.
882 Transcriptome genetics using second generation sequencing in a Caucasian population.
883 Nature. 2010;464: 773–777. doi:10.1038/nature08903

884 16. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis
885 regulatory variation in diverse human populations. PLoS Genet. 2012;8: e1002639.
886 doi:10.1371/journal.pgen.1002639

887 17. Qin S, Kim J, Arafat D, Gibson G. Effect of normalization on statistical and biological
888 interpretation of gene expression profiles. Front Genet. 2013;3: 160.
889 doi:10.3389/fgene.2012.00160

890 18. Fehrmann RSN, Jansen RC, Veldink JH, Westra H-J, Arends D, Bonder MJ, et al. Trans-eQTLs
891 reveal that independent genetic variants associated with a complex phenotype converge on
892 intermediate genes, with a major role for the HLA. PLoS Genet. 2011;7: e1002197.
893 doi:10.1371/journal.pgen.1002197

894 19. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information,
895 knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014;42: D199–
896 D205. doi:10.1093/nar/gkt1076

897 20. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression
898 residuals (PEER) to obtain increased power and interpretability of gene expression analyses.
899 Nat Protoc. 2012;7: 500–507. doi:10.1038/nprot.2011.457

900 21. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-
901 regulation and gene function. BMC Bioinformatics. 2004;5: 18. doi:10.1186/1471-2105-5-18

902 22. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, et al. Architecture of the
903 human regulatory network derived from ENCODE data. Nature. 2012;489: 91–100.
904 doi:10.1038/nature11245

905 23. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, et al. Variation in
906 transcription factor binding among humans. Science. 2010;328: 232–235.
907 doi:10.1126/science.1183621

908 24. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control.
909 Nat Rev Genet. 2012;13: 613–626. doi:10.1038/nrg3207

910 25. Spivakov M. Spurious transcription factor binding: Non-functional or genetically redundant?
911 BioEssays. 2014;36: 798–806. doi:10.1002/bies.201400036

912 26. Farnham PJ. Insights from genomic profiling of transcription factors. Nat Rev Genet. 2009;10:

913 605–616. doi:10.1038/nrg2636

914 27. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and
915 chromatin structure around the genomic regions bound by 119 human transcription factors.
916 Genome Res. 2012;22: 1798–1812. doi:10.1101/gr.139105.112

917 28. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, et al. A human phenome-
918 interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol.
919 2007;25: 309–316. doi:10.1038/nbt1295

920 29. Ideker T, Sharan R. Protein networks in disease. Genome Research. 2008. pp. 644–652.
921 doi:10.1101/gr.071852.107

922 30. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, et al. Impact of regulatory
923 variation from RNA to protein. Science. 2015;347: 664–667. doi:10.1126/science.1260793

924 31. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments
925 for identifying isoform regulation. Nat Methods. 2010;7: 1009–1015.
926 doi:10.1038/nmeth.1528

927 32. Williams RBH, Chan EKF, Cowley MJ, Little PFR. The influence of genetic variation on gene
928 expression. Genome Res. 2007;17: 1707–1716. doi:10.1101/gr.6981507

929 33. Yuan Y, Tian L, Lu D, Xu S. Analysis of genome-wide RNA-sequencing data suggests age of the
930 CEPH/Utah (CEU) lymphoblastoid cell lines systematically biases gene expression profiles. Sci
931 Rep. 2015;5: 7960. doi:10.1038/srep07960

932 34. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, De Jager PL, et al. Genetic analysis of
933 human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. PLoS
934 Genet. 2008;4: e1000287. doi:10.1371/journal.pgen.1000287

935 35. Wang X, Wang W, Li L, Perry G, Lee H, Zhu X. Oxidative stress and mitochondrial dysfunction
936 in Alzheimer's disease. Biochim Biophys Acta - Mol Basis Dis. 2014;1842: 1240–1247.
937 doi:10.1016/j.bbadis.2013.10.015

938 36. Franco-Iborra S, Vila M, Perier C. The Parkinson Disease Mitochondrial Hypothesis: Where
939 Are We at? Neurosci. 2016;22: 266–77. doi:10.1177/1073858415574600

940 37. Turner C, Schapira AH V. Mitochondrial matters of the brain: The role in Huntington's disease.
941 J Bioenerg Biomembr. 2010;42: 193–198. doi:10.1007/s10863-010-9290-y

942 38. Biffi A, Sabuncu MR, Desikan RS, Schmansky N, Salat DH, Rosand J, et al. Genetic variation of
943 oxidative phosphorylation genes in stroke and Alzheimer's disease. Neurobiol Aging. 2014;35:
944 1956-e1. doi:10.1016/j.neurobiolaging.2014.01.141

945 39. Sun N, Youle RJ, Finkel T. The Mitochondrial Basis of Aging. Mol Cell. 2016;61: 654–666.
946 doi:10.1016/j.molcel.2016.01.028

947 40. Ross JM, Olson L, Coppotelli G. Mitochondrial and ubiquitin proteasome system dysfunction
948 in ageing and disease: Two sides of the same coin? Int J Mol Sci. 2015;16: 19458–19476.
949 doi:10.3390/ijms160819458

950 41. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress
951 update-simplifying data submissions. Nucleic Acids Res. 2015;43: D1113–D1116.
952 doi:10.1093/nar/gku1057

953 42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
954 Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

44. Morgan M, Pagès H, Obenchain V, Hayden N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. R package version 1.24.0. 2016.

45. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9: e1003118. doi:10.1371/journal.pcbi.1003118

46. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic Acids Res. 2012; gks1236. doi:10.1093/nar/gks1236

47. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Syst. 2006;1695: 1–9.

48. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38: D355–D360. doi:10.1093/nar/gkp896

49. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. R package version 2.10.1; 2013.

50. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. 2015;44: D1251-7. doi:10.1093/nar/gkv1003

51. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14: R36. doi:10.1186/gb-2013-14-4-r36

52. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10: 252–263. doi:10.1038/nrg2538

Supporting Information

S1 Fig. Variability of mRNA level of transcription factors in human LCLs. Variability of mRNA levels of 777 TFs in LCLs from 88 Caucasian individuals [2]. TF genes were from Vaquerizas et al. [52] and selected with RPKM > 0.1 in all LCLs. Maximum percent change of individual TFs were calculated by comparing its minimum and maximum mRNA levels in all LCLs.

S2 Fig. Determination of correlation threshold for gene pairs. (A) To assign *P* values to gene-pair correlations we constructed a null distribution of gene-pair correlations using 1000 permutations of appropriate mRNA level data. For each permutation we recorded the maximum absolute correlation value $|\rho|$. $P < 0.01$ can be assigned to $|\rho| > 0.64$ (dashed red line) indicating that 1% of correlation values from the null distribution are larger than 0.64. (B–D) We plot on a PC by PC basis $|\rho|$ corresponding to $P < 0.001$ (red), $P < 0.01$ (green) and $P < 0.05$ (blue) using mRNA level data of LCLs [2] from Yoruba (B) from Caucasian (C) and from Finns, British and Toscani (D). The assignment of $P <$

996 0.01 to |ρ|>0.64 is consistent across mRNA level data with sequential removal of top 20 PCs for all 3
997 datasets.
998

999 **S3 Fig. Numbers of correlating gene pairs in mRNA level data following sequential removal of**
1000 **principal components.** (A) The numbers of gene pairs with |ρ| > 0.64 are plotted against sequential
1001 removal of 79 PCs (x-axis 1–80). (B) We permuted mRNA level data and counted gene pairs with |ρ|
1002 > 0.64 with sequential removal of 79 PCs (x-axis 1–80). We interpret these data to mean there is
1003 substantial correlation of mRNA level in unpermuted data which is based upon covariation
1004 recognised and removed in PC1–20. In contrast, in permuted data there are no gene pairs
1005 correlating above the threshold in PC1–20 and then a steady increase of numbers of gene pairs in
1006 PC20-80.
1007

1008 **S4 Fig. Correlation of mRNA levels of PC7 TTGs.** (A) Histogram of Spearman's correlation coefficient
1009 for 1,244,253 gene pairs from 1578 genes contained in 6268 *trans* target gene pairs (TTGPs) from
1010 PC7 mRNA level data of lymphoblastoid cell lines from Caucasian individuals [2]. Correlations that
1011 define *trans* target gene pairs are highlighted in red. Correlations of most gene pairs are poor (56%
1012 of all correlations are between -0.2 and 0.2). (B) mRNA correlation network of PC7 TTGPs. Only the
1013 largest connected component containing 1195 genes and 6003 TTGPs (95.8% of all TTGPs) is shown
1014 and we do not display 265 TTGPs contained in 144 connected components containing 2 to 38 genes
1015 (S2 Table). We note a complex clustering of pairwise correlations.
1016

1017 **S5 Fig. Overlap of *trans* target genes and *trans* target gene pairs from technical replicates.** (A) Top
1018 left panel, *trans* target genes numbers identified in first 10 PCs and the overlap gene numbers
1019 between TTGs from the same 52 individuals analysed by University of Geneva (UNIGE) and by the
1020 Centro Nacional de Análisis Genómico and Centre for Genomic Regulation (CNAG_CRG) taken from
1021 Lappalainen et al. [2]. Lower panel, their corresponding *P* values from Fisher's exact test for the
1022 overlap. Top right panel, *trans* target genes numbers identified and overlap gene numbers between
1023 TTGs from the same 68 Yoruba individuals analysed by the Yale (YRI Yale) and the Argonne (YRI
1024 Argonne) sequencing center [14]. Lower panel, the corresponding *P* values from Fisher's exact test
1025 for the overlap. Overlaps with the best *P* value in a column are highlighted. (B) Overlap of *trans*
1026 target gene pairs identified from the same datasets as shown in (A). The overlaps are extensive and
1027 significant and represent replication between laboratories of *trans* target gene pairs.
1028

1029 **S6 Fig. Overlap of *trans* target genes and *trans* target gene pairs from analysis on different**
1030 **platforms.** (A) Top left panel, *trans* target genes numbers identified in first 10 PCs and the overlap
1031 gene numbers between TTGs from the same 47 Caucasian individuals analysed on the Illumina
1032 HiSeq2000 platform [2] and the Illumina GAII platform (data from Pickrell et al. [14]). Lower panel
1033 displays the corresponding *P* values from Fisher's exact test for the overlap. Overlaps with the lowest
1034 *P* value in a column are highlighted in orange. Top right panel, *trans* target genes numbers identified
1035 in first 10 PCs and the overlap gene numbers between TTGs from the same 80 Caucasian individuals
1036 analysed by RNA-Seq (data from Lappalainen et al. [2]) and microarray (data from Stranger et al.
1037 [16]). Lower panel displays the corresponding *P* values from Fisher's exact test for the overlap;
1038 overlaps with the lowest *P* value in a column are highlighted in orange. (B) Identical 4 panels
1039 displaying the analysis of *trans* target gene pairs (TTGPs). In all cases we see replication of the
1040 identity of both TTGs and TTGPs across the technical replicates with significant changes to the PCs
1041 that define sets of *trans* target genes and *trans* target gene pairs. This is expected since the technical
1042 artefacts associated with each platform are different.
1043

1044 **S7 Fig. Overlap of genes, in *trans* target gene sets identified in different populations, whose census**
1045 **promoter regions contain bound TFs.** Numbers of combined *trans* target genes are shown that
1046 contain a bound TF in their promoter regions, given in capital letters (gene symbols) below each
1047 Venn diagram. Combined *trans* target genes are identified using RNA-Seq data of LCLs from Yoruba
1048 individuals (blue circle), Caucasian individuals (red circle) and Finns, British and Toscani mixed
1049 individuals (green circle) from Lappalainen et al. [2].
1050

1051 **S8 Fig. Overlap of genes identified in *trans* target genes from different populations that are**
1052 **enriched in different KEGG pathways.** 10 KEGG pathways that are enriched in combined *trans* target
1053 genes are shown in Venn diagrams; numbers are the number of genes annotated in the relevant
1054 KEGG term. Combined *trans* target genes are identified using RNA-Seq data of LCLs from Yoruba
1055 individuals (blue circle), Caucasian individuals (red circle) and Finns, British and Toscani mixed
1056 individuals (green circle) from Lappalainen et al. [2].
1057

1058 **S1 Table. Summary of correlation thresholds for different datasets.** Correlation thresholds for
1059 different datasets were calculated using 1000 permutations as described in the main text and
1060 Methods. We note that the smaller the sample size the larger the correlation threshold value. CEU:
1061 RNA-Seq data of LCLs from Caucasian individuals [2]; YRI: RNA-Seq data of LCLs from Yoruba
1062 individuals [2]; FIN+GBR+TSI: RNA-Seq data of LCLs from Finns, British and Toscani mixed individuals
1063 [2]. UNIGE, RNA-Seq data of LCLs from 52 individuals from University of Geneva [2]. CNAG_CRG,
1064 RNA-Seq data of LCLs from the same 52 individuals from Centro Nacional de Análisis Genómico and
1065 Centre for Genomic Regulation [2]. YRI Yale, RNA-Seq data of LCLs from 68 Yoruba individuals from
1066 Yale sequencing center [14]. YRI Argonne, RNA-Seq data of LCLs from the same 68 Yoruba individuals
1067 from Argonne sequencing center [14]. Illumina HiSeq 2000, RNA-Seq data of LCLs from 47 Caucasian
1068 individuals using Illumina HiSeq 2000 platform [2]. Illumina GAII, RNA-Seq data of LCLs from the
1069 same 47 Caucasian individuals using Illumina GAII platform [15]. RNA-Seq, RNA-Seq data for LCLs
1070 from 80 Caucasian individuals [2]. Microarray, microarray data of LCLs from the same 80 Caucasian
1071 individuals [16].
1072

1073 **S2 Table. Properties of gene correlation network formed by *trans* target gene pairs.** Properties of
1074 correlation networks of mRNA levels of TTGs of PC1–20 identified using RNA-Seq data of LCLs from
1075 Caucasian (CEU), Yoruba (YRI) and combined Finns, British and Toscani (FIN+GBR+TSI) from
1076 Lappalainen et al. [2].
1077

1078 **S3 Table. Ensembl gene IDs of TTGs identified from all mRNA level datasets.**

1079

1080 **S4 Table. Human individual IDs from all mRNA level datasets.**

1081

1082 **S5 Table. 50 transcription factors and their corresponding antibodies used in ChIP-Seq.** ChIP-Seq
1083 data from Gerstein et al. [22] is used to classify genes in the analyses presented here.
1084

1085 **S6 Table. Enrichment of individual transcription factors bound to cognate promoter regions of**
1086 ***trans* target genes.** The table displays the enrichment of individual TFs bound to cognate promoter
1087 regions of *trans* target genes defined by a given PC. Census promoter region of a gene is defined as
1088 1000 bp upstream to 1000 bp downstream region of that gene TSS. TF binding locations were taken
1089 from ChIP-Seq data of Gerstein et al. [22]. TTGs of PC1–20 were identified using RNA-Seq data of

LCLs from Yoruba (YRI), Caucasian (CEU), and combined Finns, British and Toscani (FIN+GBR+TSI) from Lappalainen et al. [2].

S7 Table. 52 transcription factors taken from Cusanovich et al. [5] **and used to identify differentially expressed genes following knockdown of individual TF genes.**

S8 Table. Enrichment of genes identified as differentially expressed following TF knockdown in *trans* target genes. Differentially expressed (DE) genes following TF knockdown were taken from Cusanovich et al. [5]. TTGs of PC1–20 were identified using RNA-Seq data of LCLs from Caucasian (CEU), Yoruba (YRI), and Finns, British and Toscani (FIN+GBR+TSI) from Lappalainen et al [2].

S9 Table. Enrichment of co-binding transcription factors bound to promoter regions of *trans* target genes. 20 co-binding TF pairs identified in human LCL were taken from Wang et al. [27]. TTGs of PC1–20 were identified using RNA-Seq data of LCLs from Caucasian (CEU), Yoruba (YRI), and Finns, British and Toscani (FIN+GBR+TSI) from Lappalainen et al. [2]. Census promoter region of a gene is defined as 1000 bp upstream to 1000 bp downstream region of that gene TSS.

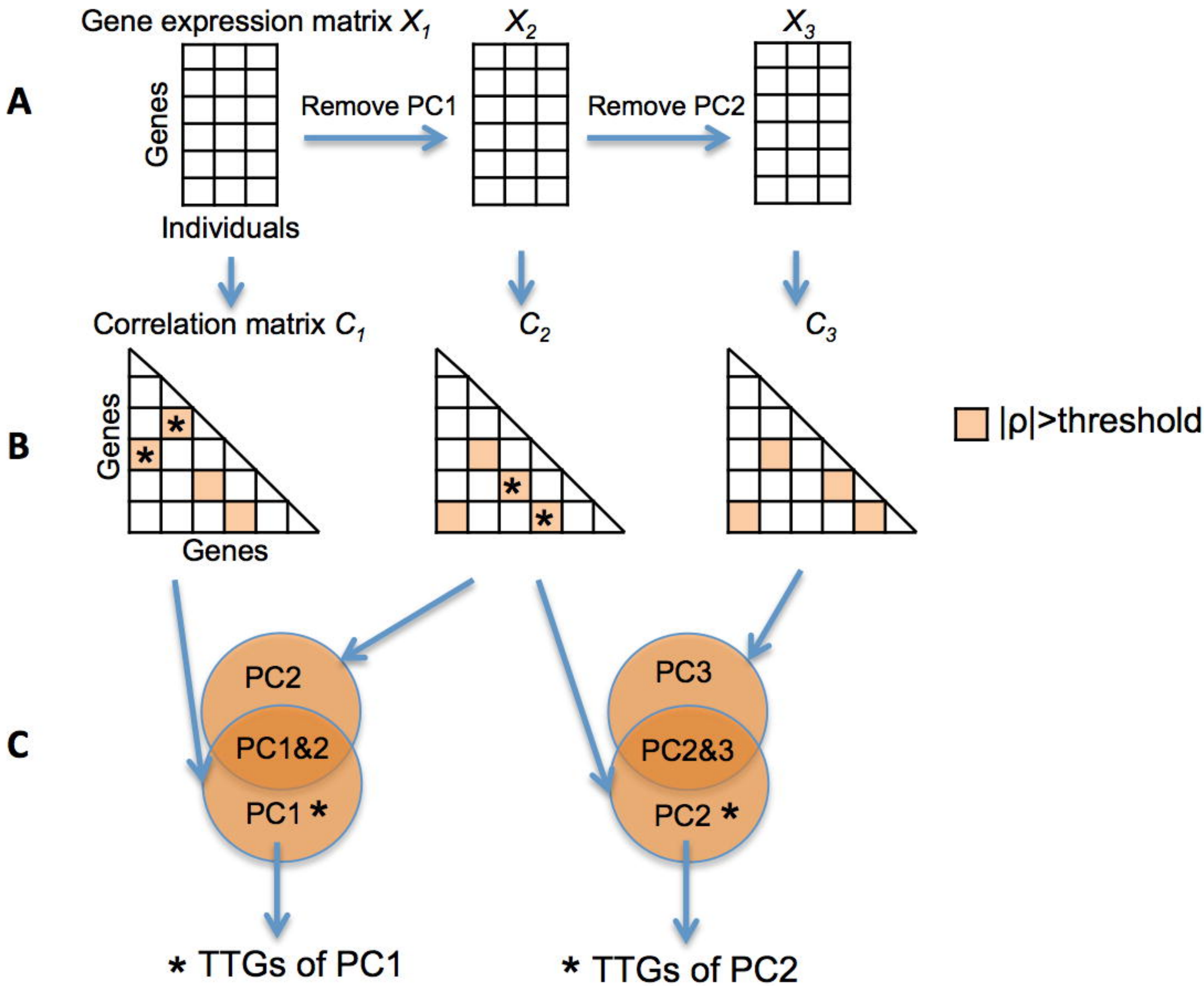
S10 Table. Enrichment of combinations of transcription factors bound to promoter regions of *trans* target genes. (A) Enrichment of combinations of TFs bound to promoter regions of combined TTGs. Combinations of TFs were identified as in the main text. Combined TTGs were identified using RNA-Seq data of LCLs from Caucasian (CEU), Yoruba (YRI), and Finns, British and Toscani (FIN+GBR+TSI) from Lappalainen et al. [2]. The hypergeometric test was used to test the enrichment and only those with a significant enrichment at $P < 1.3 \times 10^{-3}$ (Bonferroni corrected threshold) in at least one dataset are shown. N: The number of overlapped genes; P: *P* value of hypergeometric test. (B) Enrichment of combinations of TFs bound to promoter regions of TTGs of PC1–20, which were identified using RNA-Seq data of LCLs from Caucasian (CEU), Yoruba (YRI), and Finns, British and Toscani (FIN+GBR+TSI) from Lappalainen et al. [2]. Only significant enrichments at $P < 9.8 \times 10^{-5}$ (Bonferroni corrected threshold) are shown.

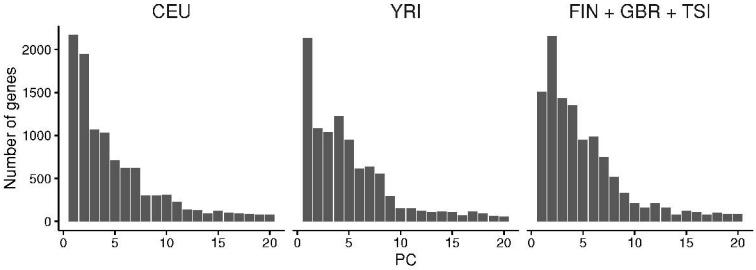
S11 Table. Proteins encoded by *trans* target genes have more protein-protein interactions than expected. We test whether proteins encoded by TTGs of PC1–20 have more PPIs than chance expectation using a permutation test. PPIs were taken from Lage et al. [28]. TTGs of PC1–20 were identified using RNA-Seq data of LCLs from Caucasian (CEU), Yoruba (YRI), and Finns, British and Toscani (FIN+GBR+TSI) from Lappalainen et al. [2].

S12 Table. Enrichment of KEGG pathways in *trans* target genes. Gene annotations for KEGG pathways were obtained from Kanehisa et al. [48]**.** TTGs of PC1–20 were identified using RNA-Seq data of LCLs from Caucasian (CEU), Yoruba (YRI), and Finns, British and Toscani (FIN+GBR+TSI) from Lappalainen et al [2].

S13 Table. Expressions of spliceosome genes in LCLs from Yoruba individuals. 83 spliceosome genes from combined TTGs, their minimum RPKM, maximum RPKM, fold change between the minimum (min) and the maximum (max) RPKM is taken from mRNA-Seq of 88 Yoruba individuals (Lappalainen et al. [2]), and their minimum and maximum protein yield and fold change is taken from analysis of 62 Yoruba individuals in Battle et al. [30].

1136 **S14 Table. Association between individual spliceosome genes and alternative splicing of retained**
1137 **intron events.** Both average Ψ (percent spliced in and estimate of isoform expression, see main text)
1138 of retained intron events and mRNA levels of individual spliceosome genes from TTGs were
1139 calculated using RNA-Seq data of LCLs from two populations (CEU: Caucasian; YRI: Yoruba) from
1140 Lappalainen et al. [2]. The genes are those annotated in the spliceosome pathway and present in the
1141 combined TTGs derived from 20 PCs. r, Pearson's correlation coefficient between mRNA levels of a
1142 spliceosome gene and average Ψ of retained intron events calculated for each individual. R^2,
1143 proportion of the variance of average Ψ that can be explained by mRNA levels of a spliceosome
1144 gene.
1145
1146 **S15 Table. Overlap of genes among 10 enriched KEGG pathways, mitochondria localized genes and**
1147 **TTGs.**


A
Gene expression matrix X_1
X_2
X_3
Genes
Individuals
Remove PC1
Remove PC2
B
Correlation matrix C_1
C_2
C_3
Genes
Genes
|ρ|>threshold
C
PC2
PC1&2
PC1 *
PC3
PC2&3
PC2 *
* TTGs of PC1
* TTGs of PC2


CEU
YRI
FIN + GBR + TSI
Number of genes
2000
1500
1000
500
0
0
5
10
15
20
PC

A

Numbers of genes

YRI TTGs \ CEU TTGs	N	1	2	3	4	5	6	7	8	9	10
N		2168	1944	1067	1032	709	623	624	302	299	306
1	2137	705	380	130	148	131	70	53	28	22	6
2	1084	189	253	205	74	55	55	47	18	21	24
3	1042	185	256	116	183	63	71	51	23	17	12
4	1225	232	283	169	163	97	97	71	29	26	19
5	947	120	153	141	107	82	63	89	59	43	41
6	611	59	88	59	76	63	110	74	34	25	18
7	637	67	86	72	91	99	57	118	28	29	23
8	557	34	50	52	46	54	62	76	41	58	59
9	295	28	42	29	31	14	33	26	15	26	32
10	149	8	10	15	17	6	16	16	6	9	14

Numbers of genes

FIN+GBR+TSI TTGs \ CEU TTGs	N	1	2	3	4	5	6	7	8	9	10
N		2168	1944	1067	1032	709	623	624	302	299	306
1	1506	170	483	96	107	39	45	35	15	24	10
2	2155	927	369	149	121	90	43	44	29	17	17
3	1434	212	361	230	171	71	61	91	19	11	23
4	1353	200	243	209	146	110	84	81	44	25	28
5	952	105	136	108	142	124	82	52	36	20	21
6	988	98	137	91	106	116	144	146	42	57	59
7	747	62	94	85	149	88	95	67	35	48	24
8	519	51	35	65	44	36	49	76	30	37	55
9	330	18	35	28	43	21	29	31	20	23	32
10	215	9	29	17	20	16	22	19	10	22	13

P values

YRI TTGs \ CEU TTGs	1	2	3	4	5	6	7	8	9	10
1	5e−133	8e−13	1	0.4	5e−04	1	1	1	1	1
2	7e−04	1e−23	7e−43	0.5	0.3	0.06	0.4	0.8	0.6	0.3
3	2e−04	9e−28	2e−07	1e−35	0.02	1e−05	0.1	0.3	0.8	1
4	2e−07	1e−26	2e−19	2e−18	5e−08	7e−11	0.002	0.2	0.4	0.9
5	0.9	7e−04	4e−19	4e−08	1e−08	7e−05	8e−14	2e−15	2e−07	2e−06
6	1	0.1	0.007	2e−07	2e−09	1e−41	5e−17	8e−08	6e−04	0.07
7	1	0.3	4e−05	9e−12	5e−27	3e−08	3e−45	1e−04	4e−05	0.005
8	1	1	0.02	0.1	3e−07	9e−13	3e−20	8e−13	1e−25	2e−26
9	1	0.3	0.04	0.01	0.5	2e−07	3e−04	0.001	2e−10	6e−15
10	1	1	0.1	0.02	0.7	4e−04	5e−04	0.08	0.003	2e−06

P values

FIN+GBR+TSI TTGs \ CEU TTGs	1	2	3	4	5	6	7	8	9	10
1	1	3e−99	0.8	0.3	1	1	1	1	0.9	1
2	4e−293	2e−10	0.6	1	0.9	1	1	1	1	1
3	0.3	1e−41	2e−35	3e−14	0.3	0.4	1e−05	1	1	0.9
4	0.2	9e−10	5e−31	1e−09	7e−10	3e−05	1e−04	4e−04	0.6	0.4
5	1	0.04	9e−08	3e−21	4e−28	2e−11	0.01	7e−05	0.4	0.3
6	1	0.1	0.004	8e−07	2e−21	5e−44	2e−45	2e−06	1e−13	3e−14
7	1	0.6	4e−06	2e−35	1e−16	4e−24	6e−10	1e−06	2e−13	0.02
8	1	1	1e−06	0.05	0.007	2e−08	1e−23	7e−08	5e−12	4e−25
9	1	0.9	0.2	2e−05	0.08	7e−05	1e−05	7e−06	1e−07	1e−13
10	1	0.5	0.4	0.1	0.05	9e−05	0.002	0.01	4e−10	5e−04

B

Numbers of gene pairs

YRI TTGPs \ CEU TTGPs	N	1	2	3	4	5	6	7	8	9	10
N		248461	116515	71164	12847	8673	9064	6268	1685	2473	2483
1	180288	34931	6945	6252	2096	1583	480	811	80	227	197
2	80458	8313	5639	29948	1746	1419	3173	932	240	519	679
3	20235	2439	2940	781	635	250	280	193	57	58	54
4	18737	2292	1734	1426	785	698	745	551	70	107	128
5	9169	628	808	1189	392	481	329	348	126	170	306
6	7609	494	467	978	504	405	1790	613	177	78	114
7	8769	1563	464	580	787	1022	300	1355	97	144	107
8	3422	228	220	412	154	225	332	361	274	219	177
9	1158	45	135	193	29	81	54	50	49	94	101
10	446	37	22	55	20	35	18	20	10	22	43

Numbers of gene pairs

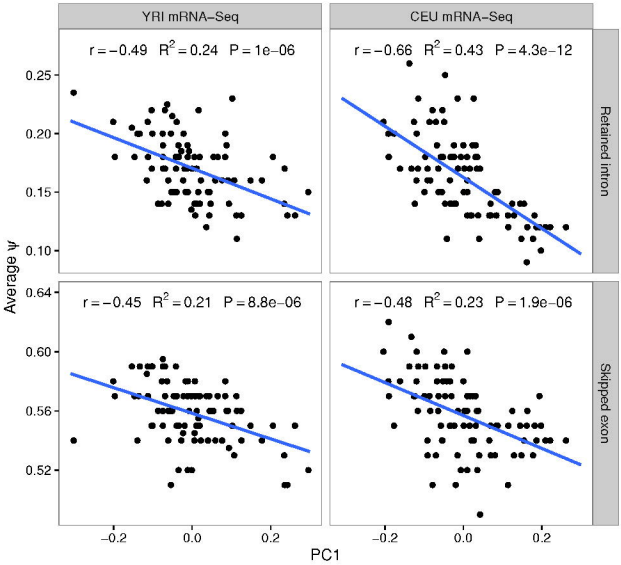
FIN+GBR+TSI TTGPs \ CEU TTGPs	N	1	2	3	4	5	6	7	8	9	10
N		248461	116515	71164	12847	8673	9064	6268	1685	2473	2483
1	338259	25013	37539	20803	1115	674	695	351	119	336	229
2	265877	70685	5181	4016	922	660	238	256	56	87	151
3	107669	7920	7257	24208	2477	1430	2299	974	135	397	559
4	33074	2345	2479	2347	762	705	540	514	103	150	235
5	20750	1458	839	2853	1181	1003	1806	548	244	168	123
6	17010	2349	752	1345	1015	1463	713	1648	155	298	224
7	5798	210	429	706	344	255	1058	363	248	103	103
8	3906	169	215	1340	100	316	228	210	96	236	363
9	1100	31	164	96	57	40	83	75	48	69	68
10	890	22	123	181	18	35	52	53	31	77	50

P values

YRI TTGPs \ CEU TTGPs	1	2	3	4	5	6	7	8	9	10
1	<1e−300	<1e−300	0.01	<1e−300	<1e−300	1	1e−84	0.3	2e−31	7e−22
2	1	1	<1e−300	8e−59	3e−30	<1e−300	2e−10	0.7	1e−51	5e−88
3	1e−213	<1e−300	7e−77	4e−160	6e−19	1e−18	7e−08	1e−04	2e−08	5e−04
4	5e−213	5e−252	1e−130	1e−89	9e−61	8e−67	3e−54	0.02	5e−14	1e−13
5	2e−09	1e−116	9e−215	4e−31	1e−46	0.6	0.6	0.2	2e−15	8e−75
6	1	0.001	3e−55	0.02	1	<1e−300	2e−50	1e−37	1e−08	5e−22
7	7e−292	2e−27	4e−07	2e−120	2e−189	1e−11	<1e−300	3e−13	3e−42	8e−37
8	2e−29	5e−13	3e−24	2e−18	1e−30	2e−70	4e−112	1e−94	3e−72	2e−61
9	4e−04	7e−34	8e−45	6e−04	6e−28	5e−10	1e−11	8e−11	3e−29	1e−27
10	4e−11	0.009	1e−09	8e−05	2e−17	0.003	3e−06	0.1	2e−05	3e−14

P values

FIN+GBR+TSI TTGPs \ CEU TTGPs	1	2	3	4	5	6	7	8	9	10
1	1	<1e−300	<1e−300	6e−103	7e−39	1e−20	0.1	0.01	5e−44	4e−19
2	<1e−300	<1e−300	5e−125	2e−143	2e−124	0.7	5e−07	0.04	8e−04	1e−20
3	1e−10	4e−70	<1e−300	9e−278	9e−48	1e−215	6e−40	1	9e−26	1e−102
4	2e−147	<1e−300	5e−258	2e−69	2e−38	0.004	5e−14	0.4	2e−04	5e−42
5	0.007	1e−09	9e−175	3e−78	3e−18	<1e−300	8e−21	1e−44	2e−18	8e−16
6	<1e−300	3e−49	1e−08	5e−120	9e−223	5e−04	<1e−300	6e−11	4e−70	5e−66
7	2e−19	2e−28	0.9	2e−42	0.001	1e−253	9e−90	7e−96	7e−17	2e−20
8	1e−18	0.7	2e−176	2e−08	3e−76	2e−19	7e−24	5e−04	2e−22	7e−99
9	2e−05	3e−38	0.5	1e−17	3e−06	9e−10	2e−14	2e−08	5e−13	7e−24
10	0.002	6e−35	2e−25	0.002	3e−08	3e−05	2e−10	5e−04	1e−20	2e−15


YRI mRNA−Seq
CEU mRNA−Seq
Retained intron
Skipped exon
r = −0.49 R2 = 0.24 P = 1e−06
r = −0.66 R2 = 0.43 P = 4.3e−12
r = −0.45 R2 = 0.21 P = 8.8e−06
r = −0.48 R2 = 0.23 P = 1.9e−06
Average ψ
0.25
0.20
0.15
0.10
0.64
0.60
0.56
0.52
−0.2
0.0
0.2
PC1