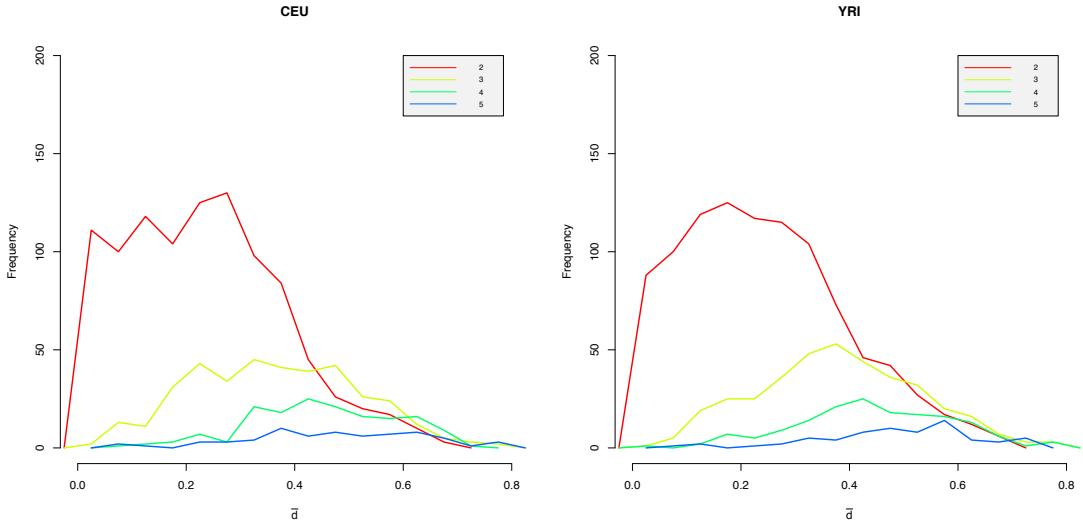
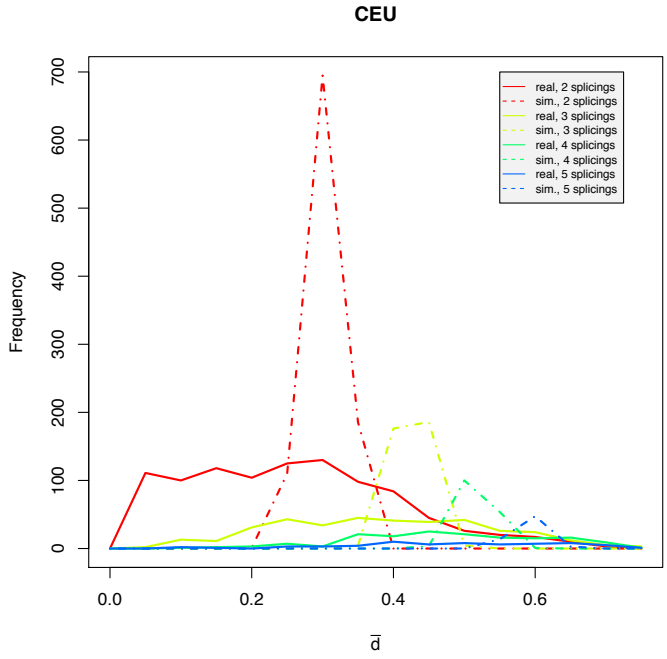
Supplementary materials

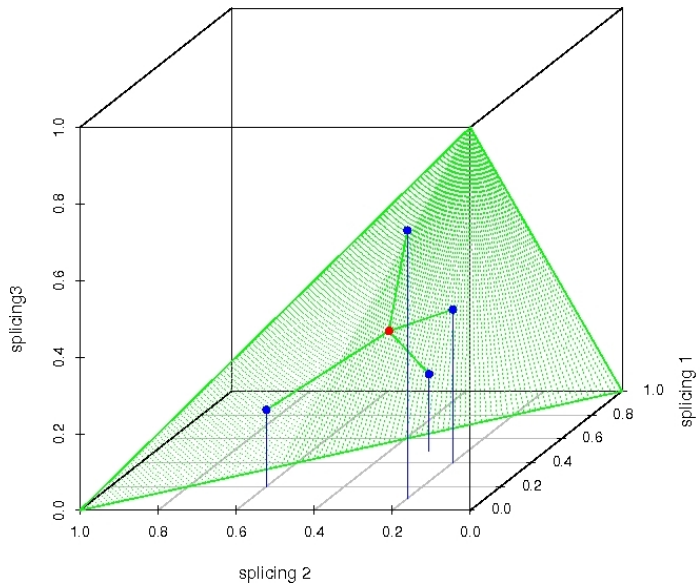
Supplementary Figures and Tables



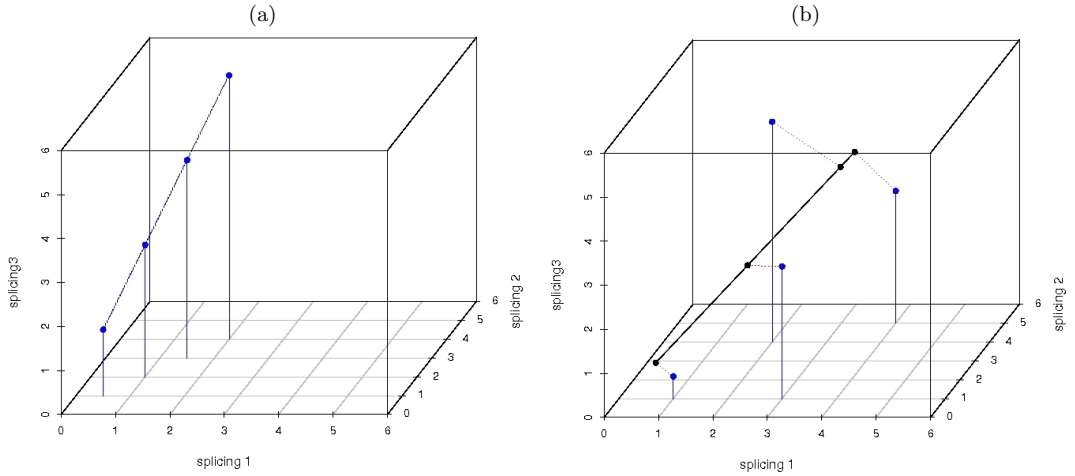
Supp. Figure 1| Variability in alternative splicing depending on the number of splice forms per gene. Only distributions for genes with 5 or less splicing forms have been represented.



Supp. Figure 2| Simulated distribution of $\overline{d}$ and comparison with the observed distribution. Only genes with 5 or less splicing forms have been represented. Simulated distributions are plotted with dashed lines.



Supp. Figure 3| Geometric interpretation of the mean of Hellinger distances to the centroid as a measure of splicing variability. In this example we consider a gene with 3 transcripts and a population of 4 individuals. Each dimension corresponds to the relative abundance (ratio) of one transcript. The 4 points (blue) represent the individuals (in order to simplify, only the projection to the x-z plane is shown). Because relative transcript abundances add to one, the four points are restricted to the equilateral triangle (green area). The centroid $\boldsymbol{c}_m$ (red) is the point that minimizes the sum of the Hellinger distances to the individual points. The dispersion of these points around the centroid, that we compute as the average distance of the points to the centroid, $\overline{d}$, can be used as an indicator of the variability in alternative splicing.



Supp. Figure 4| Geometric interpretation of the multiplicative model. In the examples, we consider a gene with 3 transcripts sampled on 4 individuals. As in **Supplementary Figure 3**, the dimensions correspond to the abundances of alternative transcript forms, but in this case the abundances are absolute, not relative, and there is no restriction to the points (blue – corresponding to individuals) to lie in a subspace of the n-dimensional space. **(a) Example of a gene with constant (1:2:3) splicing ratios.** In this case, all variation of transcript abundances across individuals is caused by variations in gene expression, and there is absolute no difference in splicing ratios. In such a case, the points lie on an straight line. Obviously the variability computed on the line is identical to the variability computed on the n-dimensional space. **(b) Example of a gene with variable splicing ratios.** In the method that we have developed, we infer the least squares line that best fits the points, and then project the points into this line. Now, in contrast with the case in **Supplementary Figure 4(a)**, the variability of the points within the line, V_{ls} – which can be assumed to correspond to the variability in gene expression – captures only a fraction of the total variability of the points in the n-dimensional space, V_t. The fraction V_{ls}/V_t is assumed to measure thus the contribution of variation in gene expression to the total variation in transcript abundance.

	genes expressed in...			
	50% of the population (n=5128)		**at least 1 individual** (n=7525)	
	CEU	**YRI**	**CEU**	**YRI**
variability in gene expression (cv)	0.79	0.70	1.17	1.05
standard deviation (σ)	0.42	0.39	0.85	0.80
variability in alternative splicing ratios ($\overline{d}$)	0.39	0.39	0.40	0.40
standard deviation (σ)	0.18	0.18	0.18	0.18
variability in gene expression *vs* variability in alternative splicing (V_{ls}/V_t)	0.68	0.64	0.71	0.68
standard deviation (σ)	0.23	0.24	0.22	0.23

Supp. Table 1| Summary of calculated variability for two extended gene sets.

Supplementary methods

Laboratory *vs* biological effects

The comparison of the two populations investigated here is confounded by the difficulty of separating biological from laboratory effects. In order to estimate the impact of such effects in our conclusions, we have compared the gene expression values obtained by Pickrell et al. (2010) and Montgomery et al. (2010) with gene expression values obtained using expression arrays in the studies of Zhang et al. (2009) and Li et al. (2010). There are 44 CEU and 53 YRI individuals in Zhang et al. (2009) that we have also used in our study. The average correlation of gene expression between these studies is 0.44 for the CEU and 0.45 for the YRI. Similarly, there are 43 CEU and 54 YRI individuals in Li et al. (2010) that we have also used in our study. In this case, the average correlation of gene expression between these studies is 0.46 both for the CEU and the YRI. While the correlations are not outstanding, the fact that they are very similar in both populations in both studies suggests, although does not demonstrate, that lab effects are not outstanding, either.

Some considerations about $\overline{d}$

As mentioned in the methodology section, the meaning of the values of $\overline{d}$ can be more easily interpreted when compared with the values in a reference distribution. We have therefore conducted a Montecarlo simulation of $\overline{d}$, where the proportions of the splicing isoforms have been drawn from a uniform distribution over the standard n-simplex, assuming additionally independent values between individuals. As mentioned in the main text, these simulated distributions are symmetrical showing a Gaussian-like shape. In probabilistic terms, under the uniform simulations on the simplex where the distribution is a particular case of the Dirichlet distribution, the expectation of a given splicing ratio is equal to $1/n$ (for n splice forms) and the variance is equal to $(n-1)/(n^2(n+1))$. As $\overline{d}$ is a mean of k distances to the spatial median, the normality guaranteed asymptotically by the central limit theorem can explain the symmetric shape of $\overline{d}$ centered over a value that depends on n. Our Montecarlo study shows also that for the majority of the genes, this variance is smaller than that of the uniform model. This simply means that for these genes the space of possible conformations of splicing ratios is restricted compared to the relatively free conformations under the uniform model. As an illustrative example, let us assume a gene with two splice forms with $(p_1, 1-p_1)$ mean population frequencies and random normal noise for p_1 with zero mean and 0.0001 variance. The vector of frequencies will be a very stable, but not constant, vector of frequencies across individuals, and the resulting $\overline{d}$ will be smaller than that from a random uniform distribution with variance (2-1)/4/3=0.08333.

Analysis of V_{ls}

We have defined the total variation (V_t) as the sum of the variances in the abundances of the alternative splice forms across the k individuals. It is the trace of the covariance matrix of the abundances of the alternative transcript isoforms. If $\Sigma = [\sigma_{ij}]_{n\times n}$ is the sample covariance matrix of the n-vector splicing counts, then:

$$V_t = tr(\Sigma) = \sum_{i=1}^{n} \sigma_i^2 \tag{3}$$

The relation between V_t and the variance of the global expression λ of the gene is easily derived taking into

account that $\lambda = \sum_{i=1}^{n} x_{i\cdot}$ where $x_{i\cdot}$ is the expression of the i transcript. If $u = (1, 1, ..., 1)$ is the vector on n ones, the variance of the global expression $var(\lambda)$ can be expressed in a short form as $var(\lambda) = u\Sigma u^t$ (where u^t stands for the transpose of u). Or in a more extended form:

$$var(\lambda) = \sum_{i=1}^{n} var(x_{i\cdot}) + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} cov(x_{i\cdot}, x_{j\cdot})$$

The sign of the covariances can only be conjectured. For a given gene, some of the pairs of expression transcripts $(x_{i\cdot}, x_{j\cdot})$ will show positive covariance while others will show it negative. Only for the obvious case of independence between the abundances of different splice forms, $var(\lambda)$ and V_t are guaranteed to coincide. This is however a quite unrealistic assumption, given that the abundance of the splice forms depends on overall gene expression, and some degree of dependence is usually to be expected, resulting in a Σ with non nulls covariances.

In this section we focus on the variation V_{ls} over the line of $\mathbb{R}^n$ with non-negative coefficients that minimizes the distance between the original and the k projected points. When a multiplicative model is assumed this is the line which better fits the n values of the alternative splice abundances of the k individuals. To our effects it is not necessary to obtain the explicit values of the projected points. We have demonstrated that the linear algebra required in order to obtain their variability can be reduced to compute only the left singular vector of the raw data matrix X, then to compute its norm respect to the covariance matrix.

More technically, the singular value decomposition (SVD) is a classical factorization of a general rectangular matrix. Following the previously introduced notation, if we define $X = [x_{ij}]_{n\times k}$ as the $n \times k$ matrix where the columns are the $\mathbf{x}_j$ counts of the k individuals, the SVD states that $X = U\Delta V^t$, where U is an $n \times n$ orthogonal matrix (multiplied by its transpose results the identity matrix), Δ is an $n \times k$ diagonal matrix with non-negative real numbers on the diagonal arranged in descending order, V an $k \times k$ orthogonal matrix and V^t denotes the transpose of V. The columns of U are the left singular vectors.

If u_1 is the first singular vector, Σ the sample covariance matrix and V_{ls} the variability of the projected data, we can express the later as:

$$V_{ls} = \mathbf{u}_1^t \Sigma \, \mathbf{u}_1 \tag{4}$$

This computation can be easily obtained by the basic package of R, (Team, 2008). Note that our result avoids the full singular value decomposition reducing considerably the computational effort, a clear advantage to other non-negative matrix factorizations based approaches.

Let us briefly proof this result. Our model specifies that for the j individual, the expression level for each transcript i is the product of a global expression parameter λ_j of this individual multiplied by the relative expression level p_i of each transcript (assumed to be constant over the k individuals).

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \cdots \\ x_{nj} \end{pmatrix} = \lambda_j \begin{pmatrix} p_1 \\ p_2 \\ \cdots \\ p_n \end{pmatrix} = \lambda_j \mathbf{p} \tag{5}$$

We use the least square criteria in order to estimate $\mathbf{p}$ and λ_j, restricted to $\lambda_j \geq 0$ and $p_i \geq 0$. If we note $\boldsymbol{\lambda} = [\lambda_j]_{k\times 1}$ as the vector of global expressions, in matrix notation, the least square solution must satisfy:

$$\min_{\boldsymbol{\lambda}\geq 0, \mathbf{p}\geq 0} \|X - \mathbf{p}\boldsymbol{\lambda}^t\|_F \tag{6}$$

where $\|A\|_F = \sqrt{\sum_{i,j=1}^{m,q} a_{ij}^2}$ is the Frobenius norm of the matrix $A = [a_{ij}]_{m\times q}$. If d_1, $\mathbf{u}_1$ and $\mathbf{v}_1$ denote respectively the greater *singular value* of X and its corresponding left and right *singular vectors*, taking into account that X is a non-negative rectangular matrix, the Perron-Frobenius theorem (Gantmacher, 1959) ensures that d_1 and all the components u_{1i} of $\mathbf{u}_1$ and v_{1j} $\mathbf{v}_1$ are non-negative. The Eckart-Young theorem (Eckart and Young, 1936) ensures that:

$$\min_{\boldsymbol{\lambda}\geq 0, \mathbf{p}\geq 0} \|X - \mathbf{p}\boldsymbol{\lambda}^t\|_F = \|X - d_1\mathbf{u}_1\mathbf{v}_1^t\|_F \tag{7}$$

The explicit expressions of $\mathbf{p}$ and $\boldsymbol{\lambda}$ are respectively $\mathbf{p} = (\sum_{i=1}^n u_{1i})^{-1}\mathbf{u}_1$ and $\boldsymbol{\lambda} = (d_1 \sum_{i=1}^n u_{1i})\mathbf{v}_1$, but the computation of V_{ls} requires strictly only the values of $\mathbf{u}_1$. Notice that the direct approach using non-negative factorization of X ($X = WH$ with $W \geq 0$, $H \geq 0$) is avoided.

Now we can compute how much variability is explained by the multiplicative model, taking into account that a singular vector has norm 1, $\|\mathbf{u}_1\|_F = 1$:

$$V_{ls} = \mathbf{u}_1^t \Sigma\, \mathbf{u}_1 \tag{8}$$

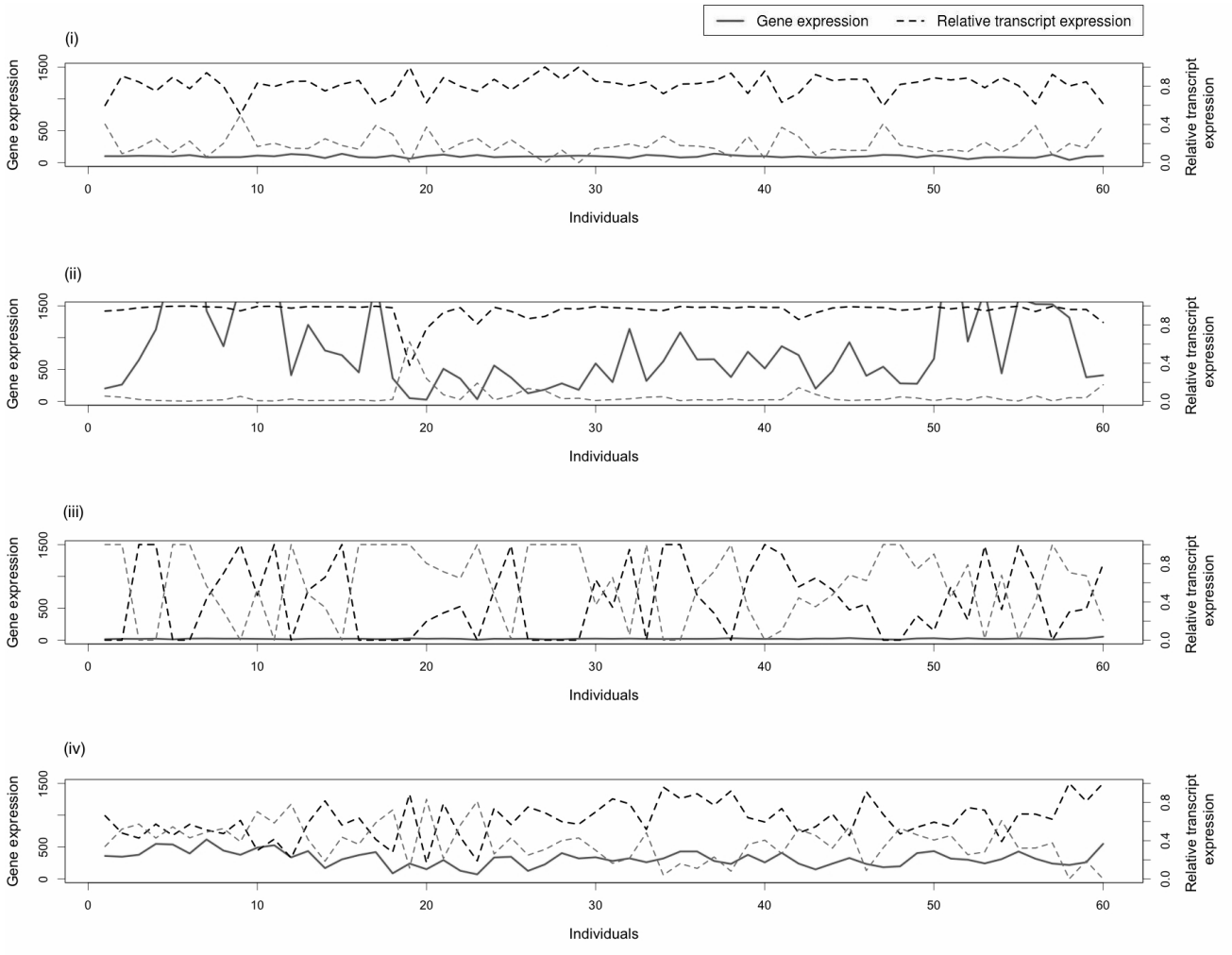
The ratio $V_{ls}/V_t * 100$ represents the percentage of total variability explained when a fixed vector of proportional expressions of the n splicing variants is assumed.

Data availability

Flux Capacitor quantifications for genes and transcripts in the Caucasian and Yoruban populations, as well as the scripts used in this paper, can be found online in `http://big.crg.cat/bioinformatics_and_genomics/SplicingVariability`.

8 Figures

(a)

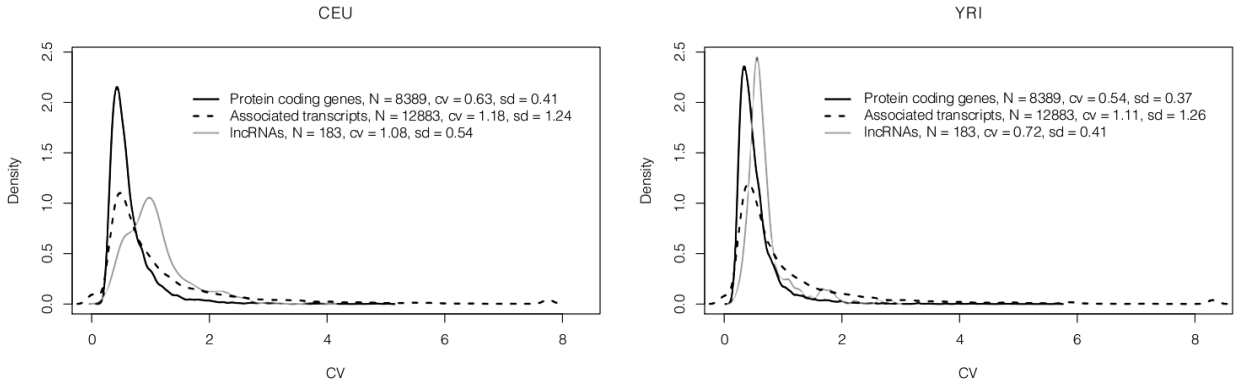


(b)

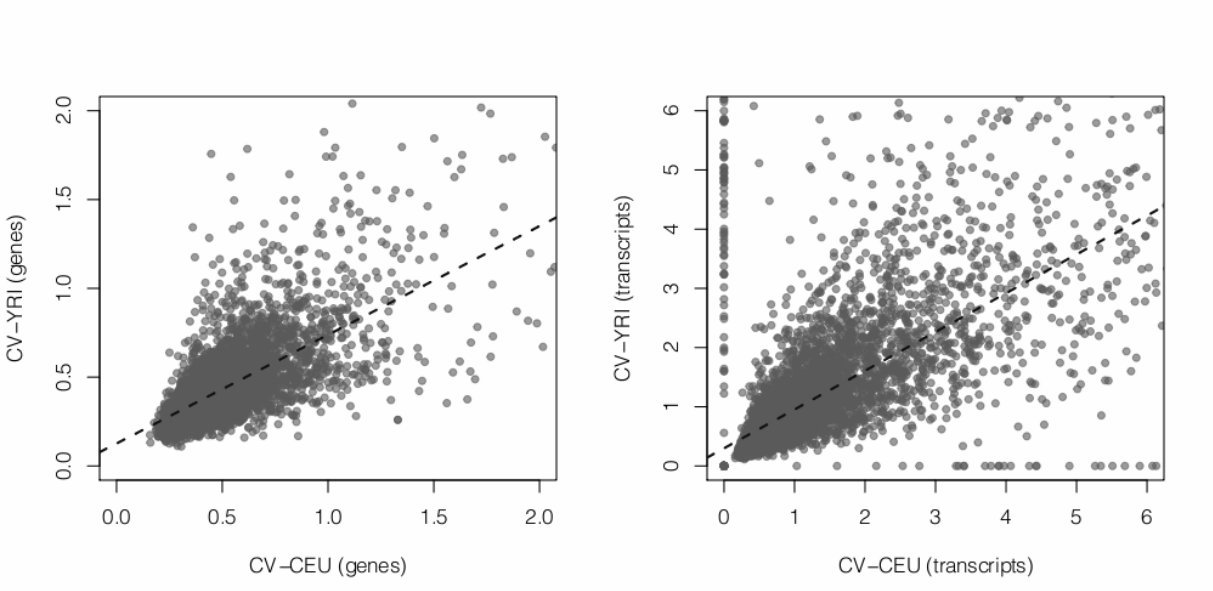
Gene	Mean expression (RPKM)	Standard deviation	cv	$\overline{d}$	V_{ls}/V_t	V_t
VPS28	97.30	19.43	0.20	0.11	0.69	484.68
PTMA	830.46	730.54	0.88	0.08	1.00	515319.80
CCDC43	20.73	6.79	0.33	0.46	0.16	140.76
HNRNPM	322.23	119.00	0.37	0.17	0.50	14611.10

Supp. Figure 5

(a)

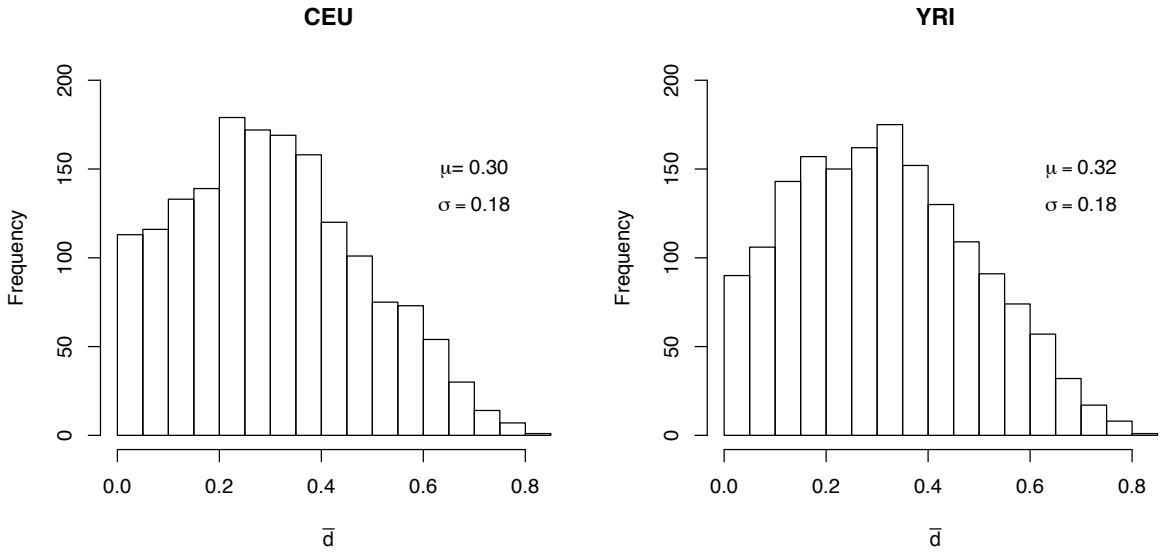


(b)

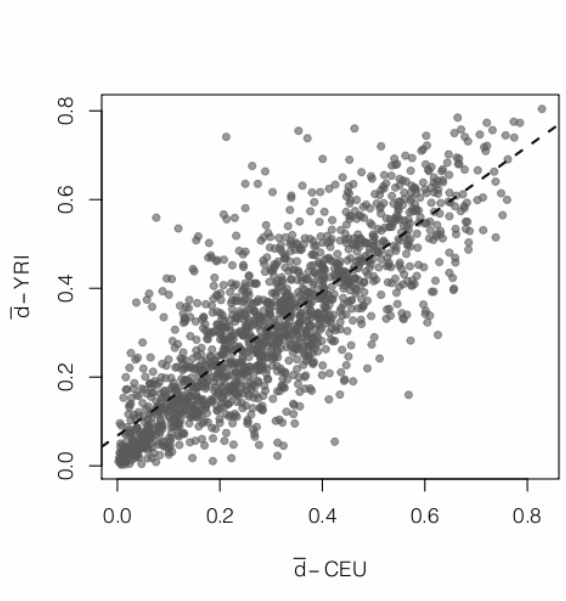


Supp. Figure 6

(a)

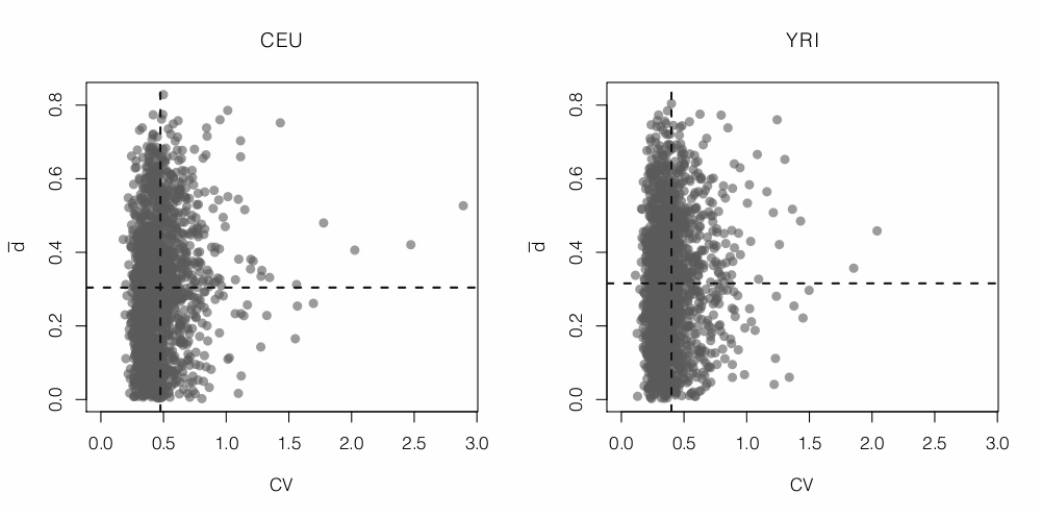


(b)



Supp. Figure 7

(a)



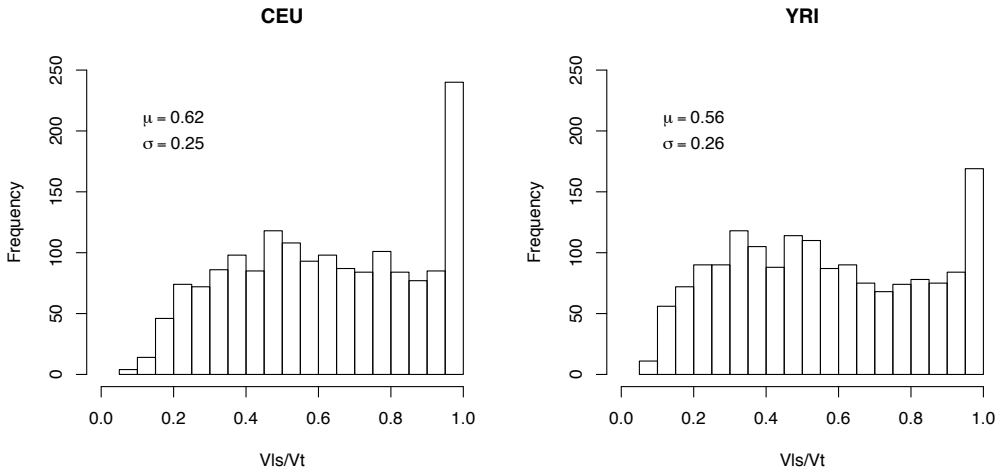
(b)

3	**4**
1	**2**

(c)

area	CEU	YRI	common	
1	576	587	397	68.92 %
2	276	288	144	52.17 %
3	461	455	282	61.98 %
4	341	324	181	55.86 %
total	1654	1654	181	60.70 %

Supp. Figure 8



Supp. Figure 9

References

Eckart C, Young G. 1936. The approximation of one matrix by another of lower rank. Psychometrika 1: 211–218.

Gantmacher F. 1959. The theory of matrices, Vol. 2. AMS Chelsea Publishing, Providence, RI.

R Development Core Team. 2008. R: A language and environment for statistical computing. http://cran.r-project.org/doc/manuals/refman.pdf.