**When in Rome—population specific adaptations in personal transcriptomes of human individuals**

**Supplementary Figures and Tables**

Figure S1: Distance metrics based on gene expression 2

Figure S2: Number of transcripts expressed by genes 3

Figure S3: Population-specific transcript variability 4

Figure S4: Variants in splice sites 5

Figure S5: The U2-2 /WDR74 locus (chr11:62,609,041) 6

Figure S6: Characteristics of novel introns and sites 7

Figure S7: Exon boundary extensions 8

Figure S8: Modality of exon boundary extensions 9

Figure S9: Exon boundaries of non-/coding transcripts 10

Figure S10: Putative cleavage sites 11

Table S1: Gene discovery *vs*. experimental attributes 12

Table S2: Pairwise comparison of DE genes 12

Table S3: Different thresholds on DE genes 12

Table S4: Mutual intersection of DE genes 13

Table S5: Genes with population-specific major transcripts 13

Table S6: Functional annotation of genes with variable or population-specific expression/splicing 13

Table S7: Overview of AS events implied by split-mappings 14

**Supplementary Figures**

|  |
| --- |
| **a** |
| > 20 y.o.  3 y.o.  7 y.o.  5 y.o.  CEU  FIN  GBR  TSI  YRI |
|  |
| |  |  | | --- | --- | | **b** | **c** | | TSI  GBR  FIN  YRI  CEU | TSI  GBR  FIN  YRI  CEU | |
| Figure S1: Distance metrics based on gene expression (a) Tree-representation of the distances between populations when considering their mutual differences as by pairwise DE comparison. The branch lengths obtained by standard Neighbor-Joining Tree reconstruction scale with the estimated “age” of the underlying cell lines. (b) The onion plot demonstrates that such direct pairwise comparison of DE genes identifies CEU as outgroup, whereas a combined comparison between all populations is able to recover more population phylogenetic signals (c). |

|  |
| --- |
|  |
| Figure S2: Number of transcripts expressed by genes The blue squares show the comparison between the number of transcripts detected for each gene in each specific population (y-axis) and the median number of transcripts found for the corresponding gene in all populations (x-axis); the darker the blue, the more instances of genes coincide at the same (x,y) coordinate. Similarly, the superimposed red dots depict the same comparison exclusively for the subset of genes that are outliers in their transcript dispersion (Fig.2b). As can be seen by the high number of genes falling along the diagonal, for most genes the number of transcripts found in a specific population agrees well with the median number of transcripts amongst all populations. Furthermore, the majority of genes that are outliers in transcript dispersion (red spots) cluster tightly around the diagonal, dissipating variations in the number of predicted genes as the major reason for deviations in splicing dispersion coefficients. |

|  |
| --- |
|  |
| Figure S3: Population-specific transcript variability (a) The Bhattacharyya Distance reflecting the variability of differential transcript usage shows that the LCL samples differ slightly in their variability in each of the populations, with the distributions of TSI and especially of CEU shifted to higher degrees of variability. (b) Splice junction mappings as well reveal that intron-usage in YRI individuals and in the Northern European populations (FIN + GBT) is more conservative compared to Central-/South- Europeans. (c) Multidimensional scaling based on pairwise comparisons of genes by their splicing ratios identifies YRI as an out-group of the dataset. (d) Subjecting all genes with a minimum between-population variation of 0.05 to a statistical test (i.e., Mann-Whitney) on gene expression contribution to the in between-population variation yields highly significant p-values <0.001 exclusively for EUR-YRI comparisons. |

|  |
| --- |
| Fig_SX_ss-scores.png |
| Figure S4: Variants in splice sites (a) The boxplots summarize the log-odd score distributions computed for splice donor sites (green), splice acceptor sites (red) and exons (blue). Splice sites have been considered to be constitutive (light green/red boxplot) or alternative (dark green/red boxplot) based on the transcript annotations in Gencode. In contrast, we empirically determined exons to be constitutive (light blue) if they exhibit a PSI value of >0.9 in >90% of the individuals, otherwise they are considered alternative exons (dark blue). Thermodynamics scores for sites and also for entire exons are usually lower when they are used alternatively. (b-d) The scatter plots depict examples of switch-like splice site choice according to a SNP. For each example, the distribution of read counts at the extended (x-axis) respectively shortened (y-axis) exon boundary is reported across individuals carrying exclusively the reference allele (green), across individuals with homozygote SNP variant alleles (blue), respectively across heterozygote individuals (red). |

|  |
| --- |
|  |
| |  |  | | --- | --- | | Figure S5: The U2-2 /WDR74 locus (chr11:62,609,041) The splice donor of the first intron in the WDR74 transcript harbors the highest number of annotated variants, i.e., 7 SNPs (red) within 9 considered nucleotides. (a) The sequence of the human reference genome, with the underlined sequences of the annotated donor site as well as alternative donor sites found up- and down-stream at 62,609,027 and 62,609,047. (b) The Gencode reference annotates a 5’-start of the U2-2 snRNA that is identical to the first WDR74 exon, and the annotated 3’-end of U2-2 is 50 nt close to the population-hypervariable donor site. (c) Split-mappings support the usage of alternative donor sites around the considered donor. (d) The genomic sequence between the 3’-end of U2-2 and the variable donor site contains a motif that is required for correct 3’-end processing of human snRNAs transcribed by RNA polymerase II. Consistently, we find no conservation of the relevant sequence stretch in other mammals. (e) The predicted U2-2 expression varies substantially in individuals carrying variants of the investigated splice donor, however, results do not suggest direct anti-/correlation of expression levels with the computed splice site strength. |  | |

|  |
| --- |
|  |
| Figure S6: Characteristics of novel introns and sites (a) Splice site scores of novel donor (grey boxplot) and acceptor sites (white boxplot) are generally lower than those of their annotated counterparts. (b) Most of the novel introns and sites correspond to minority events, observed exclusively in a small fraction of individuals. (c) The distribution of read support levels across different individuals found for novel introns (grey dots) is similar to the one observed for known introns (black dots), although former values are shifted towards lower read support levels. (d) Thresholds on the number of samples that reproduce a novel intron (x-axis) also increase the sequence coverage observed on average (y-axis); the number of introns above a correspondingly chosen threshold drops rapidly (top bars). (e) The median length of introns decreases with about quadratic character in the first 90 samples (~1 population size, left red dashed line), then linearly in the next ~180 samples, and remains about constant across the last 90 samples (i.e., introns that are supported by at least one sample from all populations, right red dashed line). (f) As observed for alternative choices of annotated sites, genetic variants can also switch the expression levels at novel exon boundaries. |

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | |  | |  | | Figure S7: Exon boundary extensions The histogram of novel exon extensions/truncations confirms intrinsic splice site motif biases at (a) position +4 for splice donors, and correspondingly at position (b) –3 at acceptor sites. Furthermore, exon boundary variations generally describe a periodicity of 3, implying influences by the coding sequence. About 10nt upstream of the acceptor site, coinciding with the common location of the branch point, we observe a region that is repressed of acceptor sites, probably due to the general absence of adenine residues, which could be confounded with the designated branchpoint. (c) Exon extensions at acceptor sites are larger than at donor sites (median 21 *vs*. 7nt), probably because of the higher frequency of in-frame stops implied by the canonical splice donor motif. (d) Exons with one variable flank (left boxplot) are significantly longer (median 130nt, p-value <e-16 Wilcoxon) than exons with variations observed at both flanks (median 116nt, right boxplot). | |

|  |
| --- |
|  |
| Figure S8: Modality of exon boundary extensions The histograms to the left show the number of exon extensions as a function of the extended length (x-axis), comparing the exon alternative with lower coverage (“minor”) with the one of higher coverage (“major”). The bar-/boxplots to the right summarize the distributions formed according to the modulo-3 values, and p-values for the differences between modulo-0/1 respectively modulo-0/2 are provided. (a) The absolute number of events exhibits a 3-periodicity, which is (b) reflected by the coverage of the minor transcript, however, (c) also by the coverage of the major transcript, revealing that transcripts with potential length changes that are a multiple of 3 are generally higher expressed transcripts that that harbor alternatives with other length changes. (d) Minor transcripts that exhibit frame-preserving length changes are observed in relatively more samples than frame-shifting length changes. When compared to the number of samples where the corresponding major transcript form is found, a preferred fixation of alternative splice sites at distances that are a multiple of 6 becomes evident. |

|  |
| --- |
|  |
| Figure S9: Exon boundaries of non-/coding transcripts (a) Novel exon boundaries in transcripts with an annotated CDS are enriched for length changes that are a multiple of 3 (left boxplot), as compared to frame-shifting events (middle respectively right boxplot). (b) Coding transcripts are significantly (KS test p-value <0.0004) higher expressed than transcripts without CDS annotation. |

|  |
| --- |
|  |
| Figure S10: Putative cleavage sites (a) The boxplots summarize of expression level distributions (as by RPKM values) measured for transcripts that overlap the genomic region of predicted cleavage sites (PCSs) and grouped by the number of supporting reads (top scale). There is no evident connection between the usage of PCSs and the transcript expression level, confirming that cleavage site usage is not dominated by expression. (b) Sequence logos for poly-A signals and their variants: overall the information content is lower in unknown poly-A motifs (i.e., none of the 13 earlier described CPSF binding motifs, bottom panels) as compared to known ones (upper panels), and lower in variant motifs (right panels) than in the poly-A motifs represented by the reference (left panels). (c) Cleavage sites are observed more frequently upstream of poly-A signals in alleles that produce unknown CPSF binding motifs (dark bars) compared to alleles that describe a known poly-A signal (white bars). |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***ID*** | ***Nr. of Samples*** | ***Total Reads***  ***[Million]*** | ***Nr. Genes >1 RPKM*** | ***LCL age***  ***[years]*** |
| FIN | 95 | 387 | 18,232 | 3 |
| GBR | 94 | 394 | 18,145 | 3 |
| TSI | 93 | 373 | 18,521 | 5 |
| YRI | 89 | 379 | 18,766 | 7 |
| CEU | 91 | 374 | 19,378 | >20 |

# Table S1: Gene discovery *vs*. experimental attributes

For each population, the number of individual samples after QC and the corresponding total read volume from the experiment, the number of genes >1 RPKM, and the approximate time since creation of the respective cell line is reported. Apparently, the number of genes detected across all individuals of a population varies and correlates best with the estimated cell line “age”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Population*** | ***CEU*** | ***FIN*** | ***GBR*** | ***TSI*** | ***YRB*** |
| ***CEU*** | - | 1,176 | 1,456 | 1,189 | 751 |
| ***FIN*** | 570 | - | 744 | 518 | 866 |
| ***GBR*** | 723 | 228 | - | 273 | 801 |
| ***TSI*** | 551 | 116 | 65 | - | 478 |
| ***YRI*** | 236 | 350 | 344 | 167 | - |

# Table S2: Pairwise comparison of DE genes

The table summarizes the number of differentially expressed (DE) genes that are predicted by pairwise comparison. The upper-right triangle of the matrix reports the number of DE genes with a >log-2fold change, the lower-left triangle correspondingly the number of genes that exhibit a >log-3fold change.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***pop ID*** | ***log-2FC*** | ***up-regul.*** | ***down-regul.*** | ***log-3FC*** | ***up-regul.*** | ***down-regul.*** | ***specific*** | ***up-regul.*** | ***down-regul.*** |
| ***CEU*** | 861 | 557 | 304 | 334 | 236 | 98 | 291 | 144 | 147 |
| ***FIN*** | 773 | 71 | 702 | 332 | 21 | 311 | 231 | 25 | 206 |
| ***GBR*** | 890 | 76 | 814 | 417 | 13 | 404 | 259 | 23 | 236 |
| ***TSI*** | 531 | 520 | 11 | 248 | 248 | 0 | 86 | 84 | 2 |
| ***YRI*** | 405 | 132 | 273 | 116 | 23 | 93 | 163 | 78 | 85 |

# Table S3: Different thresholds on DE genes

We compare the number of statistically significant DE genes at a threshold of log-2fold change (log-2FC), and to the number at a threshold of log-3FC, respectively, to the subset of log-2FC DE genes that are specific for the corresponding population (i.e., that do not intersect with the DE predictions in other populations). In addition to the total number of DE genes, the number of up- and down-regulations is provided in each category.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Pop*** | ***CEU*** | ***FIN*** | ***GBR*** | ***TSI*** | ***YRB*** |
| ***CEU*** | - | 0.61 | 0.58 | 0.59 | 0.85 |
| ***FIN*** | 1004/1634 | - | 0.57 | 0.63 | 0.77 |
| ***GBR*** | 1009/1751 | 953/1663 | - | 0.56 | 0.79 |
| ***TSI*** | 826/1392 | 818/1304 | 793/1421 | - | 0.85 |
| ***YRI*** | 1076/1266 | 908/1178 | 1027/1295 | 1076/1266 | - |

# Table S4: Mutual intersection of DE genes

The lower-left triangle of the matrix shows the ratio of the number of non-intersecting DE genes (as predicted by comparing each population against all other ones) by the number of genes in the union of both compared DE gene sets. In the upper right triangle of the matrix, the distance measure obtained by this division is reported.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Population*** | ***1 FPKM*** | ***5 FPKM*** | ***­10 FPKM*** |
| ***CEU*** | 56 | 40 | 28 |
| ***FIN*** | 35 | 27 | 15 |
| ***GBR*** | 76 | 59 | 45 |
| ***TSI*** | 30 | 26 | 31 |
| ***YRI*** | 114 | 74 | 64 |

# Table S5: Genes with population-specific major transcripts

The table summarizes the number of genes that exhibit consistent usage of a major transcript within one specific (rows) but not within the other populations. Observations for each of the five populations are reported at different expression thresholds, i.e., 1 FPKM, 5 FPKM and 10 FPKM (columns).

# Table S6: Functional annotation of genes with variable or population-specific expression/splicing

(a) Genes that exhibit a particularly high/low degree of variability in their transcript usage dispersion coefficients, as well as (b) differentially expressed genes are enriched for functional terms that refer to the cell’s outer membrane. In contrast, (c) the 5% genes with least variability in their expression levels, and (d) genes showing population-specific transcript usage are rather depleted of effects at the cell surface, but show significant enrichments for proteins localized in inner compartments (i.e., cytosol, organelles and the nucleus).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| nr. novel sites | 0 | 1 | 1 | 1 | 1 | 2 |
| description | exon skipping | 5’ extension | 3’extension | novel donor | novel acceptor | novel introns |
| events total | 1,133 | 945 | 1,439 | 4,662 | 4,926 | 4,458 |
| pattern #1 |  |  |  |  |  |  |
| events #1 | 823 | 604 | 1,167 | 4,223 | 4,628 | 1,012 |
| pattern #2 |  |  |  |  |  |  |
| events #2 | 202 | 229 | 150 | 339 | 276 | 1,043 |
| more patterns |  |  |  |  |  |  |
| more events | 108 | 112 | 122 | 82 | 22 | 2,403 |

# Table S7: Overview of AS events implied by novel introns

The table provides a summary about novel alternative splicing events inferred by superimposing novel introns on the transcript structures of the Gencode reference. Events have been grouped according to the number splice sites that are novel according to the reference, and by the qualitative location of such novel splice sites within the corresponding gene (i.e., internal or 5'-/3'-extension). A brief description for every category is given in the 2nd row, along with the total number of instances observed in that category (3rd row). Then, the two most dominant patterns are shown along with the corresponding number of different events in each pattern (rows 4-7). The last rows summarize by a pictogram the pattern of remaining minority patterns of events in the respective category.