Publication Type

Article

Title

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

Authors

Pedro G. Ferreira[1,2]*, Jean Monlong[1,3]*, Mar Gonzàlez-Porta[4]*, Matthias Barann[5]*, Tuuli Lappalainen[2,6,7], Emilio Palumbo[1], Marc R Friedländer[1], Anaïs Gouin[1,8,9], Manuel A Rivas[10], The Geuvadis Consortium, Emmanouil T Dermitzakis[2,6,7], Roderic Guigó[1,11], Michael Sammeth[1#]

* These authors contributed equally to this work
# Corresponding author
1 Center for Genomic Regulation (CRG), 08003 Barcelona, Catalonia, Spain
2 Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland
3 Department of Human Genetics, McGill University, Montréal H3A 1B, Canada
4 European Bioinformatics Institute, EMBL-EBI, Hinxton, United Kingdom
5 Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, D-24105 Kiel, Germany
6 Institute for Genetics and Genomics in Geneva (G3), University of Geneva, 1211 Geneva, Switzerland
7 Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland
8 Institut National de la Recherche Agronomique (INRA), UMR Institute de Génétique Environnement et Protection des Plantes (IGEPP), 35653 Le Rheu cedex, France
9 INRIA, GENSCALE team, IRISA-INRIA, 35042 Rennes cedex, France
10 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom
11 Pompeu Fabra University (UPF), 08003 Barcelona, Catalonia, Spain

Abstract

Transcriptomes are defined by the interplay between gene expression and RNA processing—most importantly by splicing—and constitute the atomary cellular phenotype at the molecular level. We employed the Geuvadis RNA-Seq experiments in Lymphoblastoid Cell Lines of known genotypes to rigorously identify subtle, but systematic, population variations in expression levels and exon-intron structures—especially between individuals of European and of African origin. Our results show that, mechanistically, many of these differences are caused by effects of genetic variants on the functionality of splice site motifs; beyond the paradigm of trans-acting splicing regulators established by multitudinous reports, we thus provide the first comprehensive collection of alternative exons and splice sites controlled by DNA variants rather than by factors that enact on the RNA level. Furthermore, we rigorously investigate novel transcriptional elements discovered by RNA-Seq—i.e., splice sites, introns, and cleavage sites—which we include in our investigation of population-genetic transcriptome adaptations.

Introduction

The Geuvadis project [1] provides high resolution RNA-Seq data from lymphoblastoid cell lines (LCLs) obtained from individuals whose genomes have been sequenced by the 1000 Genomes Project [2], comprising five distinct populations—namely, the African Yoruba in Ibadan/Nigeria (YRI) and the European Finns (FIN), British (GBR), Toscani (TSI) and Utah residents with Northern and Western European ancestry (CEU). These data are particularly

well suited to study individual differences in the regulation of transcription and RNA processing in human cells.

The RNA composition of a cell line defining its phenotype at the molecular level is primarily the result from thermodynamics of gene abundance, a steady-state balance between specific transcription and degradation rates, as well as from processes determining the composition of the transcripts: transcription initiation, splicing and 3'-end formation, among others.

Previous studies of transcriptome variation in human populations focused on fewer populations of smaller sample size [3–5], and/or interrogated gene expression by micro-arrays [3,6]. The Geuvadis dataset provides expression estimates in ≥89 individuals from each of the aforementioned populations by RNA-Seq, which provide more accurate quantifications of genes and even of single transcripts [5,7,8], thus allowing to monitor not only quantitative variations in gene expression, but also qualitative variations in the usage of alternatively spliced transcripts allowing to determine the predominantly expressed transcript from each gene (i.e., "major transcript" [8]).

Moreover, the availability of personal genomes for all Geuvadis individuals allows us to investigate the impact of genetic variants (SNPs and indels) on molecular mechanisms of RNA processing, and especially on the substrates of splicing processes. (Alternative) splicing has been identified as the key mechanism to generate the complexity observed in mammalian transcriptomes [9], and malfunctioning can lead to severe illnesses [10,11], but little is known about the mechanisms and extent to which genetic factors shape the transcriptome landscape.

It is known that as a prerequisite for the splicing of an intron, specific factors recognize the intronic sequence adjacent to the upstream exon, the *splice donor*, by proper RNA-RNA pairing, whereas sequence attributes of the *splice acceptor* intron end at the downstream exon are recognized by less specific protein-RNA interactions [12]. Traditional research on *in silico* gene finding elaborated computational models for the thermodynamic affinity of these RNA motifs recognized by corresponding factors of the splicing machinery [13]. Such models can be employed to predict immediate effects of variants within the splice donor/acceptor sequences on the splicing process. Furthermore "exon definition", the concerted binding of splicing factors at both sides of especially short mammalian exons, could in theory play a role in propagating the effect of genetic polymorphisms from one intron to the neighboring one [14].

RNA-Seq also has demonstrated to improve our understanding of the transcriptome compared to annotations based on traditional EST data [15,16] by allowing *de novo* detection of upstream transcription initiation, unannotated splicing structures and alternative cleavage sites [8,17,18]. The combined sequencing depth of the Geuvadis project with >1 billion paired-ends reads, each of 2*75nt in length, provides an excellent resource to characterize individual transcript expression variation throughout human populations, and thus to complete our picture of LCL transcriptomes.

Results

Variability in Gene Expression

We first investigated gene expression patterns in lymphoblastoid cell lines (LCL) derived from different individuals and populations. Based on the

Geuvadis RNA-Seq dataset of 462 individuals that passed the carefully chosen quality control criteria [19], we identify population *marker genes*, i.e., those that are shared by >90% of the individuals of the population they are marking, which then are distinguished qualitatively whether they are *population-specific* genes, or genes shared by a certain subset of populations, respectively *ubiquitous* genes that are shared across all five studied populations. As a reference annotation, we used Gencode [20].

The number of expressed genes ≥ 1 RPKM varies from 18,145 to 19,378 per population and does not scale with the number of samples sequenced, nor the total/average sequencing depth, but with the estimated, but with the estimated age of the cell lines in each of the populations (Tab.S1) : i.e., >20 years for CEU, ~7 years for YRI, ~5 years for TSI and ~3 years for FIN and GBR samples (Coriell Cell Repositories, pers. communication). Our observations agree well with earlier reports about increasing levels of mRNA expression with progressing cell culture "age" and passaging [21–23]. Overall, the rate of gene detection declines but does not saturate (Fig.1a). Also, the number of genes cumulatively observed in a comparable number of replica samples is much lower, which underlines the information gained by sequencing an additional sample. Furthermore, the number of detected genes increases by leaps and bounds across technical replicates from individuals of different populations, suggesting population-specific expression patterns (Fig.1a).

Although the number of genes cumulatively detected above 1 RPKM in the five investigated populations altogether corresponds about to the highest number of genes found in a single population (Tab.S1 and Fig.1a), the

proportion of ubiquitous genes is highly dependent on the expression cut-off: at the levels of 1, 5, and 10 FPKM, we observe a steadily decreasing rate of ubiquitous genes, respectively from 92%, 84% and 77% of ubiquitous genes, respectively (Fig.1b). Although the majority of marker genes are shared by multiple populations—as can be seen from the steadily increasing number of markers observed from 2 to 5 populations (Fig. 1b)—the number of population-specific genes cannot be explained by statistical dilution as there are less genes shared between 2 populations than between 3 or only 1 population(s). When increasing the expression threshold from 1 to 10 FPKM, the number of ubiquitous genes is decreasing rapidly whereas the number of non-ubiquitous genes (i.e., genes shared between 1—4 populations) stays rather constant. Altogether, these findings suggest that population specificity is a true biological feature of transcriptomes.

To further dissect population-specific transcription rates of ubiquitous genes, we next focused on genes that are predicted to be differentially expressed (DE) between population pairs. The number of DE genes varies more than 5-fold for two-fold log expression level differences (and 10-fold for three-fold differences), between 273 (65) and 1,456 (723) observed DE genes (Tab.S2). As expected, population-specific gene expression accumulates in lowly abundant genes, with marginal differences in the median abundance of DE genes predicted for each population (Fig.1c).

In order to address the biological relevance, we performed a PCA analysis on the DE genes identified by such all-against-all pairwise comparisons (Fig.1d).

As expected by our observations on the number of expressed genes, CEU samples exhibit the largest differences (i.e., the highest number of DE predictions) to the other populations, and also the number of DE genes found for the other populations correlates with the cell line age (Tab.S2); therefore, clustering approaches at first identify CEU rather than YRI as outgroup of the dataset (Fig.S1a and b).

In order to deconvolute the underlying biology of populations from cell-culture driven biases, we compared each population against the other ones combined (Fig.S1c). Although the number of thus predicted DE genes still varies amongst populations, with a persistent trend of genes more frequently observed to be up-regulated in older cell lines, we now observe <2-fold differences in the number of predictions across all populations without any obvious correlation between cell line age and number of predicted DE genes, and less polarity between up- and down-regulated genes Tab.S3. A distance measure based on pairwise intersections of DE gene sets separates YRI from the European populations, thus better reflecting population history (Tab.S4 and Fig.1e) [2,24].

Variability in Splicing

We then assessed population-specific transcript usage. Under the assumption that genes usually express one main transcript [8], we first investigated how often this major transcript is different within and between populations. Similar to our previous analysis we first focused at an expression threshold of 1 FPKM on ubiquitously expressed genes that preserve the major

transcript consistently within one of the populations (i.e., in ≥90% of the individuals). The majority of those genes (70%) also exhibit an ubiquitous usage of the major transcript throughout all populations, which constitutes a lower fraction than we recorded for ubiquitously expressed genes. However, observations of ubiquitous major transcripts are more consistent for lowly and highly expressed genes (68% and 64% ubiquitous at 5 and 10 FPKM, respectively, Fig.2a). Interestingly, we observe at all tested expression thresholds a relatively higher number of genes that show YRI-specific major transcripts (Tab.S5).

To estimate variations of the transcript usage in a more continuous manner, we considered the dispersion of relative transcript ratios within each gene: Fig.2b shows that most of the expressed genes—i.e., together 61,600 pairwise gene comparisons—agree very well in their population-specific transcript dispersion with the median variability observed across all populations (R= 0.99). At a dispersion threshold of 0.1, merely a minority of <1% of the expressed genes exhibit a comparatively high or low variability. Reassuringly, these outliers in splicing variability are not accompanied by predictions of comparatively more/less expressed transcripts Fig.S2. As a proxy for transcript variability, the distribution of Bhattacharyya distances is comparatively lower in Nothern European and African (GBR, FIN and YRI) individuals than in the Central- and South-European populations (CEU and TSI, Fig.S3a). Consequently, we also observe relatively higher support levels for Gencode introns in Northern European populations and in YRI (Fig.S3b), which could be linked to earlier

observations on differences in the degree of alleles sharing of the corresponding populations [2].

We next studied genes with significantly different splicing patterns and, in agreement with our previous analysis regarding the usage of the major transcript, we find at a false discovery rate (FDR) of 1% most genes with population-specific transcript patterns in YRI individuals: 193 genes in YRI *vs.* 13 genes in CEU, and none in the other populations. As splicing patterns have demonstrated to be determinant of the corresponding cell type [9,25], large differences in the qualitative transcript usage would have been surprising given that our study contains exclusively LCL samples. In contrast to previously identified splicing dispersion outliers, these genes with *bona fide* population-specific splicing exhibit usual dispersion coefficients (colored dots in Fig.2b).

In spite of the time of LCL cultivation, an analysis by multidimensional scaling of the pairwise differences in splicing ratios identifies YRI to be less similar to the other investigated populations (Fig.S3c *vs.* Fig.1c and d). Putting our former observations about population-specific variation of gene expression (Tab.S2) on a common scale with these latter results about population-specific transcript usage, we find that differential transcript usage has a relatively higher contribution exclusively to the delineation of the YRI population (Fig.S3d and Fig.2c) [1].

Finally we assessed whether genes that we classified to have a particularly high or low variability in their expression and/or splicing exhibit specific functions. Genes that we identified to have extreme splicing dispersion

coefficients are enriched for proteins at the outer cell membrane (Tab.S6a and Fig.2d). Interestingly, we observe very similar patterns of functional enrichment in DE genes (Tab.S6b and Fig.2d), although there is little overlap between these two sets of genes (93 out of 1030 respectively 595 genes). Given that a cell is to preserve the attributes and functionality of its respective cell type, our findings agree well with observations that individual adaptations are mainly effected through modifications at the surface [21–23]. As a complementary confirmation, we observe a core of expression-invariant genes with functions in inner compartments of the cell, i.e. in vesicles, in organelles and in the nucleus (Tab.S6c and Fig.2d). Strikingly, functional terms annotated for the products of genes with population-specific splicing patterns coincide well with those observed in genes that vary little in their expression (Tab.S6d and Fig.2d), reconfirming that our observations of population-specific splicing are not related to differences in transcription rates.

Summarizing all, our results imply that population-specific transcript use is a different indicator than population-genetic gene expression, and, importantly, in contrast to our previous studies of DE genes, our results obtained from transcript usage phenotypes exhibit less evident cell line biases, suggesting that RNA processing—predominantly splicing—reflects biological patterns of population differences better than gene expression.

Splicing in Personal Transcriptomes

In order to investigate the molecular mechanisms that cause transcriptome variations between populations and especially between continents, we employed genotype data from the 1000 Genomes Project [26]. We focused on variants that directly impact on the affinity of annotated splice sites and, following traditional approaches in gene finding [27], we employ computational splice site models that consider an informative sequence of 9nt for splice donors, including the GT dinucleotide, and 27nt for splice acceptors—including the AG dinucleotide and additionally the typical area of the preceding poly-pyrimidine tract [28]. To estimate the splicing efficiency of different variants, the thermodynamics of splicing is modeled as log-odds scores based on the frequency of observed donor and acceptor sequence motifs [13]. Under this model, sequences with a higher degree of similarity to the consensus bind more tightly to the corresponding splicing factors [29,30], and therefore are more frequently observed as authentic splice sites [31,32].

Confirming earlier reports that modification of splicing can be driven by less efficient thermodynamic binding of splicing factors to the sequence [33], our thermodynamic model predicts lower scores for alternative splice sites in Gencode as compared to the scores of sites annotated as constitutive (Fig.S4a) [33]. Complementarily, when testing our bioinformatics models empirically by calculating from the Geuvadis RNA-Seq data so-called PSI-scores ("percent spliced in") that reflect the degree of inclusion of an exon in mature transcripts [18,34], we also observe that the distribution of splicing scores at the flanks of skipped exons (PSI score <0.9 in >90% of the individuals) is significantly lower

than *bona fide* constitutive exons (PSI score >=0.9 in >90% of the individuals, Fig.S4a).

We found >10% (51,342 out of 477,880) of annotated splice sites are affected by at least one variant in the splicing motif, equally in constitutive and alternative splice sites. The majority (i.e., 47,453) of these sites are affected by exclusively one variant, but we found that splice sites can comprise up to 7 known polymorphisms (Fig.3a). Interestingly, the splice site with the highest degree of genetic variation falls close to U2-2, the RNA part of the central splicing factor U2, 14nt downstream of a characteristic sequence at +23nt that has been reported to be required for 3'-end formation of human snRNA genes transcribed by RNA polymerase II [35]. However, although variants seem to influence the predicted U2-2 expression significantly, we did not observe a direct correlation with the computed splicing scores Fig.S5.

Most variants in splice site regions are single nucleotide polymorphisms (SNPs), with a repression of indels (2% vs. 3.6% annotated indels in general) likely due to purifying selection against large genomic perturbations in functional elements, although coding sequences exhibit an even higher depletion (<0.5% indels) [2]. The distribution of variant frequencies in splice site sequences is negatively correlated to the information content in the corresponding splice site consensus motif, and the dinucleotides involved in the splicing reaction are mostly depleted of sequence variants (Fig.3b). However, we also observe few examples with RNA-Seq evidence for alternative splice site usage triggered by SNPs that impact the functionality of splice site dinucleotides: in these cases homozygote individuals exhibit exclusively the use of the one or the other exon

boundary whereas heterozygote individuals provide evidence of both splice sites being used (Fig.S4b-d).

Considering the differences in our splice site score predictions between the sequence of the reference and the alternative allele, we classify the variants in five classes: most alternative alleles (~52%) decrease the predicted score and correspondingly are classified as *deteriorating variants*, however, ~32% *enhancing variants* are observed to increase the score; less frequently, genetic variants lead to destruction (~10% *"inhibiting variants"*) or activation (< 0.5%*"activating variants"*) of splice site functionality. In the remaining ~5% of the alternated splice site motifs, our computational models do not predict any thermodynamic effect by the computed score ("neutral variants").

This classification is based on the effect of the *nonreference* allele, which is usually the novel derived allele, with the reference genome corresponding to the ancestral state. However, this is not always the case. For variants in each of these classes, we measured the global derived allele frequency (DAF, i.e., the frequency of the non-ancestral allele, Fig.3c) [2]. The distribution of activating variants differs substantially from all the other classes, with 72% of DAFs >0.1 in contrast to less than 13% DAFs >0.1 in other variant classes. This implies that sites in the activating nonreference class mostly are rare cases where the reference genome contains a low-frequency derived allele that disrupts the splice site and where the nonreference allele represents “normal” active state of the corresponding site. As expected, the deteriorating/inhibiting variants are enriched in very low allele frequencies likely due to purifying selection (Fig.3c).

In order to analyze how the predicted splice score of variants correlates with our RNA-seq data, we studied the PSI scores of alternatively included exons (0.2< PSI< 0.8 in >80% of the individuals). We found that exons with potentially splicing-deteriorating/inhibiting variants at their flanks exhibit already low inclusion levels even in individuals carrying the reference allele (median PSI score 0.37), whereas splicing-enhancing/activating variants target preferentially the flanks of exons that are already relatively highly included when employing the reference allele (median PSI score 0.76). As expected, exon inclusion gradually decreases/increases when including more alleles with a predicted negative/positive effect on splicing (Fig3d) [1].

Discovery of Novel Transcript Elements

To estimate up to which degree the Geuvadis experiment can complement current knowledge about transcript annotation in LCL, we superimpose split-mappings to Gencode exon-intron structures. We found >64 million reads supporting ~2/3 of the annotated introns (222,862 out of 337,247 introns) and additionally ~14.7 million split-mappings that provide evidence for >1 million intron variations within the proximity (30nt) of annotated exon boundaries. Although the overall distribution of predicted introns follows largely the one of introns annotated in the reference, a mixture of two log-normal distributions caused by distinct groups of short (~100nt) and long (~1,600nt) introns [36], there are outliers of extremely short and long split-mappings Fig.4a. Furthermore, the unannotated splice sites inferred by split-mappings exhibit realistic, but significantly weaker, splice site scores (Fig.S6a). Obviously, novel elements are

of rather etiolated character because they mostly constitute minority events of the population (Fig.S6b), observed at comparatively low expression levels (Fig.S6c). However, a conservative set of 3,545 novel introns is found in each population and exhibits a read coverage distribution comparable to annotated introns (Fig.S6d); also their median lengths are comparatively close to the one of annotated introns (2,217nt *vs.* 1,589nt, Fig.S6e).

In spite of the generally low abundance and coverage, we find biological phenomena of annotated introns also reflected by the attributes of novel introns: at the boundary of exon extensions created by novel sites we observe a shift in the amount of nucleotide diversity (Fig.4b)—similar to the ones of annotated counterparts (Fig.3b), albeit more diluted. As for annotated exon-intron structures (Fig.S4b-d), we find evidence for genetic control of splice site functionality at novel exon boundaries, albeit with lower read support (Fig.S6f). Fig.S7a and b show positive biases in the location of novel exon boundaries as expected by characteristics of the splice site motifs, at +4 for donor [37] and at –3 at acceptor sites [38]. Additionally, we observe alternative splice site creation repressed in a region of about -10nt before the acceptor dinucleotide, which corresponds to the distance of a typical branch point and is expected to be depleted of adenine bases as they could be confounded with the true branch site and change the splicing mechanism [39]. Nevertheless, we observe alternative acceptor sites at larger distances to the annotated exon boundary than alternative donor sites (Fig.S7c), probably due to higher chances of in-frame

stop codons created by splice donor sequences [40]. Novel introns mostly modify only one end of annotated exons, but we identify 36,426 exons that exhibit bilateral splice site variation. These exons are significantly shorter (p-value <e-16 Wilcoxon) than exons with exclusively one variable flank, in agreement with exon definition mechanisms along short exons (Fig.S7d) [41].

Novel introns found by split-mappings also exhibit selective constraints with respect to protein-coding functionality of RNA: introns confirmed by at least half of the individuals exhibit to change the length of the sequence included in a transcript by a multiple of 3 more frequently than expected (Fig.S8a). Consistently, we also find relatively more transcripts with an annotated CDS to be affected by frame-preserving exon boundary changes (Fig.S9a). This could be either due to the generally higher expression level of protein-coding genes (Fig.S9b), or due to nonsense-mediated decay, and/or by alternative splice sites in the genome being shaped to preferentially produce alternative transcripts that preserve the frame. Our data (Fig.S7) supports the latter hypothesis, suggesting that genomic rather than cellular control causes the phenomenon: possibly caused by high levels of nuclear RNA in the whole cell RNA extractions[42], the exon coverage of non-3 shifts is not decreased (Fig.S8b and c), and, in fact, persistent novel alternatives exhibit often a periodicity of 6 (Fig.S8d and e), highlighting the mutual importance of neighboring codons as reported by dicodon frequency biases [43].

Next we analyzed the overall landscape of alternative splicing events depicted by the set of 3,545 novel introns that are supported by at least one

individual from each population. Tab.S7 shows the exhaustive classification of all AS patterns inferred by these novel introns in the context of Gencode exon-intron structures [40], with most introns (69%) linking novel splice sites to an existing one, less frequently introns employ 2 novel sites (25%), and novel combinations of exclusively 2 existing sites are rather exceptional (5%). In spite of substantial differences in motif complexity, split-mappings predict about equally many novel donor and acceptor sites (4,662 *vs* 4,926). We find comparatively many events that indicate alternative exons beyond the transcript extremities, especially at the 3'-end (1,616 5'- *vs* 3,210 3'-events, Tab.S7).

To further investigate 3'-end modifications, we extracted read mappings that align partly with the genomic sequence and partly with poly-A tail retrieving in total 52,436 putative cleavage sites (PCSs). The number of PCSs found with higher read support is decreasing rapidly (Fig.4c), but independent of the expression rate from the underlying transcript (Fig.S10a). Focusing on a conservative subset of 21,102 PCSs supported by ≥2 reads, which are still more than twice as many cleavage sites as identified in a complementary study [4], we observe that 71.4% of them fall within annotated 3'UTRs, and 66% even within 50nt of the 41,542 3'-transcript ends annotated in Gencode reference (Fig.4d). Reassuringly, for 96.3% of PCS predictions we find within 50nt a hexamer sequence that agrees with one of the 13 known consensus motifs of the CPSF (cleavage / poly-adenylation specificity factor) binding site, the so-called poly-A signal. Motif frequencies of known poly-A signals found at PCSs are very similar to those previously reported (upper left panel in Fig.S10b) [44], but information

content is reduced when considering genetic variants, even if they produce one of known poly-A signals (Fig.S10b, upper right panel). Also unknown poly-A motifs in the reference sequence exhibit relatively lower information content (Fig.S10b, lower left panel), consequently, variants that create unknown poly-A signal motifs are rather degraded and unlikely to serve as substrate for CPSF (Fig.S10b, lower right panel). Such deteriorating variant effects are recorded more frequently in poly-A signals immediately downstream of a PCS (Fig.S10c and Fig.4d), supporting an hypothesis that suboptimal CPSF binding is possible in cases where upstream alternatives for 3'-end termination are available.

Discussion

Our comprehensive study demonstrates that, overall, the cellular phenotype of lymphoblastoids is well conserved across the investigated 462 individuals from five populations, both in terms of quantitative levels of transcription rates as well as by the qualitative patterns of transcript usage (Fig.2a,b). However, we identified hundreds of genes that show population-specific deviations in their expression levels (Fig.1b) or transcript variability (Fig.2b). Functionally, genes with population-specific transcription rates are enriched for proteins impacting the cell surface, whereas patterns of distinct transcript usage are controlling many proteins inside the nucleus, organelles and other inner compartments (Fig.2c)[45].

From a technical point of view, we describe methods to avoid biases of gene expression known to be caused by the age of the cell lines, i.e., by normalization of DE analyzes. Whereas, our results demonstrate a general robustness of transcript comparisons to cell line age. This enhances our understanding of changes that happen during LCL cultures: despite modifications of transcription rates, the established splicing program that mainly defines the cell type [18] does not appear to be altered during cell line aging.

The comparison of the relative contribution of expression levels and splicing on population differences showed an interesting pattern: although relatively few genes show significant changes in their splicing patterns between populations, the contribution of transcript usage to inter-continental differences is significantly increased (Fig.2c). Our observations shed new light on recent findings about the role of (alternative) splicing in evolutionary adaptation of mammalian species [9,25], suggesting that adaptations in human populations can trigger modifications in the splicing patterns as well. However, in our study such differences are limited to populations from different continents (i.e., Europeans as compared to Africans), whereas intron-exon structures within European sub-populations do not differ significantly from each other (Fig.S3d).

For the splicing variation in human populations that is driven by genetic variants in splice sites (Fig.S4b-d), we demonstrate that our computational models can predict changes in the thermodynamics splicing potential due to these variants. Most of the variants segregating in the population appear to have relatively modest effects on splicing, but we pinpoint a set of propagating variants that actually describe a functional splice site in contrast to the splicing

defective reference (Fig.3c). Importantly, we are able to show an enrichment of variants with negative effects on splice sites where already individuals with the reference allele show low inclusion levels of the corresponding exons, and similarly we demonstrate that splicing enhancing variants are mostly targeting exons that are already predominantly included in reference alleles (Fig.3d). In analogy to the exon-gain and exon-loss mechanisms described for species evolution [46], this suggests gradual effects of variants fine-tuning the inclusion level of alternative exons.

Furthermore, we demonstrate that *de novo* split-mappings—when reproduced by sufficient individuals/populations—exhibit insert size distributions, sequencing coverage and population support similar to their annotated counterparts (Fig.S6d,e and Fig.4a). In contrast to expression thresholds, employing reproducibility across samples as a fidelity criterion still allows for assessing novel splicing events also in rather lowly expressed genes, highlighting the unprecedented possibilities provided by the plentitude of individuals in the Geuvadis dataset.

Based on those *bona fide* introns, we report the compendium of alternative exon-intron structures not described in the Gencode v12 reference, which exhibits mainly patterns of modifications at existing exon boundaries and alternative 5'-/3'-exons (Tab.S7). The extensions/truncations of known exons show an enrichment of changes by a multiple of 3nt, and our results suggest that this is not exclusively caused by degrading mechanisms—NMD—of frame-

shifting transcripts, but by the splicing motifs in the genome being biased to prefer shifts that preserve the reading frame (Fig.S8).

Moreover, our analysis reveals that also our current picture of cleavage sites is rather incomplete and we describe the to our knowledge currently most complete atlas of putative cleavage sites in human cell lines, derived from RNA-Seq evidence from hundreds of heterogeneous individuals. Most of these empirically found cleavage sites fall close to annotated cleavage sites and exhibit an upstream poly-A motif (Fig.4c). Genetic variants also impact poly-A motifs, and we observe loss of function especially in poly-A signals that are downstream of predicted cleavage sites, i.e., where alternative mechanisms of 3'-end formation exist (Fig.4d).

Altogether, this study demonstrates the power of large-scale RNA-sequencing analysis to understand population variation in the transcriptome, shaped by genomic motifs for RNA processing and fine-tuned by genetic variants in them. Our results demonstrate already a vast diversity of annotated and unannotated transcript features, and it is likely that in the future application of our approaches to different cell types and tissues will add another important dimension to transcriptome complexity. Our improved understanding of this key cellular phenotype opens doors for better characterization of cellular function and its role in human variation.

Online Methods

Gene Expression

Ubiquitous Genes: Our study is based on the read mappings and transcript quantifications available from the Geuvadis project [1] for all 462 samples that passed QC [19]. Gene expression is estimated by the sum of length- and volume-normalized quantifications of its annotated transcripts (RPKM values based on paired-end reads, i.e. FPKM values). Genes are classified according to their expression behavior within a population; all genes that are shared by >90% of the individuals in a population are considered representative *marker genes*. We further distinguish markers that are observed exclusively in one population ("population-specific" genes), from those that are shared between 2, 3, 4, or 5 populations ("ubiquitous genes").

Differentially Expressed Genes: Due to variations in the number of genes detected in each population pool Tab.S1, we limited our DE analysis to 16,568 (~73% of all detected genes) autosomal genes that are ubiquitously expressed and show ≥5 counts per million mapped reads in ≥1 sample. For predicting differentially expressed (DE) genes, we employ a Bioconductor package based on the Poisson-Tweedie family of distributions, which has been demonstrated to be suited for datasets with >15 samples [47]. This software normalizes read counts according to the Trimmed Mean of M-values (TMM) method [48]. Genes were deemed significantly differentially expressed between populations at an FDR < 0.05 and a log-2fold respectively log-3fold change.

Transcript Expression

Major Transcripts: To identify preferentially expressed forms of a gene, we ranked the transcripts annotated for each gene according to their expression

level, and identified the highest expressed one. Subsequently, in analogy to our previous study on population-specific genes, all major transcripts expressed above a certain threshold (i.e., 1, 5, or 10 FPKM) were assessed whether they are expressed in ≥90% of the individuals of a population (i.e., "major transcripts") [8]. We estimated the degree up to which such major transcripts are shared between populations and labeled transcripts that are identified as the major form of their respective gene in all 5 populations as "ubiquitous".

Splicing Dispersion: To assess alternative splicing variability between populations by splicing dispersion, the expression levels of transcripts are represented using their relative abundance, i.e. *splicing ratio* in the space $[0,1]^T$ (R^T space) with T being the number of transcripts expressed from a gene. The centroid of the samples can then be used to test for the homogeneity of the different population variability using an analysis of the variance-like framework [49]. As an estimation of the splicing variability, we computed the dispersion as the average Hellinger distance of the samples from the centroid. After correction for multiple testing employing the Benjamini-Hochberg algorithm, the dispersion of a gene is considered to be different between two populations if the computed p-value is <0.01 (FDR of 1%) and the change in dispersion is >0.1 [7].

Components of the Population Classificator: By ANOVA decomposition, the total inter-individual variability of transcript expression levels (Vt) is separated into within-population variability and between-population variability. Computing the coefficient of determination based on these two variability measures provides an estimation of the population classificator effects. On average, the between-population variability of a gene represents 3% of the transcript expression variability, from which changes originating from

fluctuations in gene expression and splicing are estimated by the population classificator. Samples are projected to a model of constant splicing [7] (straight line in R^T) which allows to re-compute the between-population variability considering gene expression effects exclusively: lower values than the former between-population variability indicate the degree of contribution of alternative splicing to the delineation of populations.

Functional Analysis: Computation of overrepresented GO terms in the set of DE genes respectively genes with differential transcript dispersion has been computed by the David functional classification program [50]. Terms that are significantly overrepresented compared to background gene data have been classified according to their cellular location in one of the following classes: surface, organelle/vesicle, nucleus, plasm, ubiquitous (i.e., present in multiple of the other categories).

Population Genetics of (Alternative) Splicing

Splice Site Scores: We computed scores for donor and acceptor sites by employing a first order Markov Model to score dinucleotide transitions, as implemented in the gene predictor geneID [13]. In brief, transition probabilities are estimated by known human splice site sequences, and motivated by traditions of scoring metrics based on probabilities, a score is assigned to each splice site by the sum of log-likelihoods from all transitions in a splice site sequence: at the 5'-end of introns a region [-2;7] around the exon boundary is considered to score the donor potential, whereas at the 3'-end of introns a corresponding sequence stretch of [-24;3] is analyzed to score the splice acceptor dinucleotide with preceding poly-pyrimidine tract. Naturally, by the differences in the number of

considered positions and also in their respective information content, scores derived for donors and acceptors differ from each other. Annotated splice sites usually obtain scores in the range [-10;10], a score of $<-10^3$ indicates that there currently exists no known functional site with the corresponding sequence.

PSI Scores: To estimate the exon inclusion level from RNA-Seq reads, we computed the so-called Percentage Splice Index (PSI) similar to an earlier proposed approach [18,34] where 3 types of quantifications are used per exon: (A) the number of reads that map within a certain exon, (B) the number of split-mappings to exon-exon junctions between the considered exon and both adjacent exons, and (C) the number of split-mappings to the exon-exon junction from the adjacent exon upstream to the adjacent exon downstream. Then,

$$PSI = A + B / (A + B + C)$$

is computed, where a value of 0 means that the tested exon is not included, whereas a value of 1 indicates that the exon is constitutively spliced in. In our study, we focused on 64,120 exons that are alternatively spliced in all 462 samples, i.e., with observations for A, B, and C > 0 as a precondition to compute the PSI score.

Discovery of Novel Transcript Elements

Novel Intron Forms: We rescued *bona fide* introns and splice sites that are not annotated in the Gencode v12 reference transcriptome by analyzing split-mapped RNA-Seq reads. Novel introns are identified by split-mappings with one end in 30nt proximity of an annotated exon boundary, considering only properly paired mappings with a mapping quality of at least 150, an edit distance ≤6, and an insert-size of ≤1,000,000nt.

Prediction of Putative Cleavage Sites: To identify putative cleavage sites (PCSs), we employed reads containing a poly-A tail or a poly-T head that are indicative of the cleavage site in poly-adenylated mRNAs. After trimming the reads for these subsequences, filtering by a minimum informative length (i.e., >25nt after trimming) and removing low complexity reads (i.e., read sequences with an [A] and [T] content ≥80%), we obtained ~24M reads of which 685,351 map uniquely to the genome and indicate a PCS.

Poly-A Signals: In order to investigate which motifs are enriched in the sequences around these PCS, we used a recursive approach similar to an earlier proposed method [51], where we scanned the sequences for the 13 motifs previously identified as the binding site for the CPSF poly-adenylation factor by the order of their know frequency.

Acknowledgments

This project was supported by the European Commission 7th Framework Program, Project N. 261123 (GEUVADIS). The research leading to these results has received funding from the European Community's FP7 HEALTH grants CAGEKID (grant agreement 241669).

Author Contributions

Designed the analysis: PGF, JM, MGP, MB, RG, MS

Analyzed the data: PGF, JM, MGP, MB, TL, MS

Contributed to data processing and analysis: EP, MF, AG, MAR

Coordinated the analysis: MS

Drafted the paper: MS

Revised the manuscript: PGF, JM, MGP, MB, TL, MR, RG

References

1. Lappalainen, T., Sammeth, M., Friedländer, M. R. & Ac, P. Transcriptome and genome sequencing uncovers functional variation in human populations.

2. García, M. V. i An integrated map of genetic variation from 1,092 human genomes. *Nature* **135**, 0–9 (2012).

3. Storey, J. D. *et al*. Gene-expression variation within and among human populations. *American journal of human genetics* **80**, 502–9 (2007).

4. Pickrell, J., Marioni, J. & Pai, A. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).

5. Montgomery, S. B. *et al*. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–7 (2010).

6. Stranger, B. E. *et al*. Population genomics of human gene expression. *Nature genetics* **39**, 1217–24 (2007).

7. Gonzàlez-Porta, M., Calvo, M., Sammeth, M. & Guigó, R. Estimation of alternative splicing variability in human populations. *Genome research* **22**, 528–38 (2012).

8. Djebali, S. *et al*. Landscape of transcription in human cells. *Nature* **489**, 101–8 (2012).

9. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* **338**, 1593–1599 (2012).

10. Garcia-Blanco, M. a, Baraniak, A. P. & Lasda, E. L. Alternative splicing in disease and therapy. *Nature biotechnology* **22**, 535–46 (2004).

11. Rivas, M., Beaudoin, M. & Gardet, A. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature* … (2011).at <http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.952.html>

12. Zhang, X. H.-F., Leslie, C. S. & Chasin, L. a Dichotomous splicing signals in exon flanks. *Genome research* **15**, 768–79 (2005).

13. Blanco, E., Parra, G. & Guigó, R. Using geneid to Identify Genes. *Current protocols in* … (2007).at <http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0403s18/full?globalMessage=0>

14. Berget, S. Exon Recognition in Vertebrate Splicing. *Journal of Biological Chemistry* (1995).at <http://www.jbc.org/content/270/6/2411.short>

15. Mironov, A. a., Fickett, J. W. & Gelfand, M. S. Frequent Alternative Splicing of Human Genes. *Genome Research* **9**, 1288–1293 (1999).

16. Kan, Z., Rouchka, E. & Gish, W. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research* 889–900 (2001).doi:10.1101/gr.155001.1

17. Mortazavi, A., Williams, B. A. B., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature …* **5**, 1–8 (2008).

18. Wang, E. T. *et al*. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–6 (2008).

19. 't Hoen, P. A. C. *et al*. Reproducible transcriptome sequencing across different laboratories.

20. Harrow, J., Frankish, A. & Gonzalez, J. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome …* (2012).at <http://genome.cshlp.org/content/22/9/1760.short>

21. Neumann, E. *et al*. Cell culture and passaging alters gene expression pattern and proliferation rate in rheumatoid arthritis synovial fibroblasts. *Arthritis research & therapy* **12**, R83 (2010).

22. Zschenker, O., Streichert, T., Hehlgans, S. & Cordes, N. Genome-wide gene expression analysis in cancer cells reveals 3D growth to affect ECM and processes associated with cell adhesion but not DNA repair. *PloS one* **7**, e34279 (2012).

23. Baker, B. M. & Chen, C. S. Deconstructing the third dimension: how 3D culture microenvironments alter cellular cues. *Journal of cell science* **125**, 3015–24 (2012).

24. Altshuler, D. M. *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–8 (2010).

25. Barbosa-Morais, N. L. *et al*. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science* **338**, 1587–1593 (2012).

26. Garcia-Blanco, M. Alternative splicing: therapeutic target and tool. *Alternative Splicing and Disease* 6–7 (2006).at <http://www.springerlink.com/index/p814r6kljh428wg5.pdf>

27. Guigó, R., Knudsen, S., Drake, N. & Smith, T. Prediction of gene structure. *Journal of molecular biology* **226**, 141–57 (1992).

28. Coolidge, C. J., Seely, R. J. & Patton, J. G. Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic acids research* **25**, 888–96 (1997).

29. Nelson, K. K. & Green, M. R. Mechanism for cryptic splice site activation during pre-mRNA splicing. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 6253–7 (1990).

30. Zamore, P., Patton, J. & Green, M. Cloning and domain structure of the mammalian splicing factor U2AF. *Nature* (1992).at <http://ukpmc.ac.uk/abstract/MED/1538748>

31. Ohshima, Y. & Gotoh, Y. Signals for the selection of a splice site in pre-mRNA: Computer analysis of splice junction sequences and like sequences. *Journal of molecular biology* (1987).at <http://www.sciencedirect.com/science/article/pii/0022283687906474>

32. Brunak, S. ren, Engelbrecht, J. & Knudseny, S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of molecular biology* (1991).at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.4527&rep=rep1&type=pdf>

33. Ast, G. How did alternative splicing evolve? *Nature reviews. Genetics* **5**, 773–82 (2004).

34. Shapiro, I. M. *et al*. An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS genetics* **7**, e1002218 (2011).

35. Yuo, C., Ares, M. & Weiner, A. Sequences Required for 3' End Formation of Human U2 Small Nuclear RNA. *Cell* (1985).at <http://www.sciencedirect.com/science/article/pii/S009286748580115X>

36. Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 11193–8 (2001).

37. Hiller, M., Huse, K. & Szafranski, K. Phylogenetically widespread alternative splicing at unusual GYNGYN donors. *Genome …* 1–16 (2006).doi:10.1186/gb-2006-7-7-r65

38. Hiller, M. *et al*. Single-nucleotide polymorphisms in NAGNAG acceptors are highly predictive for variations of alternative splicing. *American journal of human genetics* **78**, 291–302 (2006).

39. Gooding, C. *et al*. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome biology* **7**, R1 (2006).

40. Sammeth, M., Foissac, S. & Guigó, R. A general definition and nomenclature for alternative splicing events. *PLoS computational biology* **4**, e1000147 (2008).

41. Buratti, E., Baralle, M. & Baralle, F. E. Defective splicing, disease and therapy: searching for master checkpoints in exon definition. *Nucleic acids research* **34**, 3494–510 (2006).

42. Brandhorst, B. & McConkey, E. Stability of nuclear RNA in mammalian cells. *Journal of molecular biology* (1974).at <http://www.ncbi.nlm.nih.gov/pubmed/22003578>

43. Fickett, J. The gene identification problem: an overview for developers. *Computers & chemistry* **20**, (1996).

44. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome research* **10**, 1001–10 (2000).

45. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology* **5**, e1000598 (2009).

46. Modrek, B. & Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature genetics* **34**, 177–80 (2003).

47. J. R. Gonzalez & Esnaola, M. tweeDEseq: RNA-seq data analysis using the Poisson-Tweedie family of distributions. at <http://www.creal.cat/jrgonzalez/software.htm>

48. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, R25 (2010).

49. Anderson, M. J. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**, 245–53 (2006).

50. Huang, D. W., Sherman, B. T. & Lempicki, R. a Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44–57 (2009).

51. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic acids research* **33**, 201–12 (2005).

Figure Legends

Figure 1: Gene expression in LCL from 5 different populations. (a) Gene discovery observed by the cumulatively discovered number of quantified genes (>1 RPKM) as a function of the number of sequenced samples, for 462 non-redundant individuals (green) and for 5 replicated samples—one per population—sequenced 8 times each (blue; roman numbers). The lower panel shows a zoomed area of the larger panel with all individuals. The order of the samples has been permuted 30 times and thick bars represent the median, whereas the colored areas are the 25^{th} and 75^{th} percentiles of the corresponding distribution. The increasing curve demonstrates how almost every sample expresses some genes not observed in others, and the green curve of different samples being above the replicate samples shows that part of this increase is due to population diversity rather than increased total sequencing depth. Furthermore, gene discovery exhibits volatile increases when crossing population boundaries, indicating population-specific genes. (b) Ubiquitous genes (i.e., expressed above a certain threshold in >90% of the individuals of all 5 populations; white bars) constitute the largest fraction across all the investigated expression thresholds (1, 5, and 10 FPKM). There is a comparatively small accumulation of population-specific genes (dark green bars), however, their relative fraction is substantially higher than expected considering the decreasing trend in the number of genes that are shared between less populations, and importantly these fractions remain about the same also at higher expression thresholds. (c) The curves show the log-log expression profiles of transcript expression levels recorded in each population. Differentially expressed (DE) genes accumulate in the lower abundant expression ranks. There are minor, but significant, differences in the distributions of DE expression levels in between the populations. (d) Multivariate analysis of comparisons based on pairwise comparisons of populations by their gene expression levels indicates CEU to be an outlier of the dataset, most likely due to the unequally older cell lines of CEU samples. (e) When considering all 5 populations together for the prediction of

DE genes, population-specific signals that are less biased by cell line age and more information from the biology of the underlying populations can be recovered.

Figure 2: Population-specific splicing. (a) Although we observe relatively less ubiquitous major transcripts (grey bars) than there are ubiquitously expressed genes (white bars), the proportion of former is more similar in lowly and highly expressed genes (with thresholds at 1, 5, respectively 10 FPKM) when compared to the latter. (b) The scatter plot summarizes the splicing dispersion coefficients of all genes, i.e., the variability amongst the relative transcript expression levels of each gene. Comparing the splicing dispersion observed in a specific population (y-axis) with the median dispersion of the corresponding gene in all 5 populations (x-axis), it can be seen that for most genes the transcript variability observed in a particular population correlates well with the overall variability (black dots). Colored dots highlight genes with significant differences in their population-specific splicing patterns, which obviously do mostly not coincide with dispersion outliers (i.e., dots beyond dashed lines). (c) Comparing the degree to which DE genes and genes with population-specific dispersion coefficients determine the corresponding population, we find that population specificity reflected by differential transcript usage is generally out-ruled by population-specific gene expression (median <20%). However, the contribution of population-specific splicing is significantly higher in African individuals (YRI) compared to the European stereotype (EUR). (d) Functional annotations in the GO Cellular Component category coincide well between genes with particularly high/low splicing dispersions and DE genes: despite of minor overlap (10%), genes of both sets affect primarily the cell's surface. In contrast, the 5% genes with the most constant expression levels as well as the disjoint set of genes with significant population-specific splicing ratios encode protein products that localize predominantly in the nucleus, organelles and vesicles.

Figure 3: Splicing in different populations and individuals. (a) The histogram summarizes the number of genetic variants that fall into the area of a splice site considered by our model; frequencies are decreasing exponentially with higher number of variants in the same splice site, we observe ~1 order of magnitude

less instances for every additional variant. (b) Variants are repressed in exonic stretches as compared to introns, primarily due to restrictions by the coding sequences they harbor. Splice site dinucleotides are largely exempt of annotated variants, and population-genetic effects in the rest of the splicing motif scale about inversely to the information content of the site consensus sequence. (c) Derived allele frequencies of activating splice site variants are distributed inversely compared to alleles of other types of modifications, with a substantially higher proportion of derived alleles that are highly abundant throughout the investigated populations. Furthermore, deteriorating/inhibiting classified variants accumulate at low allele frequencies as compared to the frequencies of enhancing/neutral variants. (d) Population-genetic effects on the inclusion level observed for alternative exons (0.2< PSI < 0.8 in >75% of the population) are not random: variants with negative splicing effects target exons that are already mostly skipped in the reference allele (median PSI ~0.4, green curve), whereas variants with predicted positive effects evolve at the splice sites of exons that are mostly included in the reference allele (median PSI >0.75, yellow line). The variant effect predicted by the model then gradually increases the observed PSI ex-/inclusion level in genotypes with negative/positive alleles at one respectively both sides of the corresponding exon.

Figure 4: Discovery of novel transcriptional elements. (a) The length distribution of novel introns observed by split-mappings (grey line) follows largely the shape of the distribution observed for introns annotated in Gencode v12 (black line), with two main peaks for short introns (~100nt) and long introns (~1,600nt). However, novel introns are shifted towards longer lengths and exhibit several outliers created by very short and very long split-mappings. (b) At exon boundaries extended by novel split-mappings (grey line), an increase in the population-genetic variant rate similar to the one at the exon flanks they extend (black line) can be observed. (c) Most predictions of putative cleavage sites (PCSs, black bars) fall within annotated 3'UTRs (dark grey bars), and a subset of them within a 50nt proximity to annotated cleavage sites (light grey bar). The relative overlap between PCSs and annotated 3'UTR regions, particularly within the region of the annotated cleavage sites, improves at higher

read coverage thresholds. (d) The boxplots summarize distances of a PCS to the closest annotated poly-A signal, i.e., positive values indicate that the PCS is downstream of the poly-A motif. Variants that produce known CPSF binding site motifs are found in poly-A signals that are predominantly upstream of PCSs (light grey boxplot), whereas variants that create unknown CPSF binding motifs are found in poly-A signals that lie downstream of the closest PCS (dark grey boxplot).