Alternative Splicing QTLs in European and African Populations

Halit Ongen[1,2,3,*] and Emmanouil T. Dermitzakis[1,2,3,4,5,*]

With the advent of RNA-sequencing technology, we can detect different types of alternative splicing and determine how DNA variation regulates splicing. However, given the short read lengths used in most population-based RNA-sequencing experiments, quantifying transcripts accurately remains a challenge. Here we present a method, Altrans, for discovery of alternative splicing quantitative trait loci (asQTLs). To assess the performance of Altrans, we compared it to Cufflinks and MISO in simulations and Cufflinks for asQTL discovery. Simulations show that in the presence of unannotated transcripts, Altrans performs better in quantifications than Cufflinks and MISO. We have applied Altrans and Cufflinks to the Geuvadis dataset, which comprises samples from European and African populations, and discovered (FDR = 1%) 1,427 and 166 asQTLs with Altrans and 1,737 and 304 asQTLs with Cufflinks for Europeans and Africans, respectively. We show that, by discovering a set of asQTLs in a smaller subset of European samples and replicating these in the remaining larger subset of Europeans, both methods achieve similar replication levels (95% for both methods). We find many Altrans-specific asQTLs, which replicate to a high degree (93%). This is mainly due to junctions absent from the annotations and hence not tested with Cufflinks. The asQTLs are significantly enriched for biochemically active regions of the genome, functional marks, and variants in splicing regions, highlighting their biological relevance. We present an approach for discovering asQTLs that is a more direct assessment of splicing compared to other methods and is complementary to other transcript quantification methods.

Introduction

In eukaryotes, alternative splicing is involved in development, differentiation,[1] and disease[2] in a tissue-specific manner. Splicing events can be categorized under skipped exon, retained intron, alternative 3′ or 5′ splice sites, mutually exclusive exons, alternative first or last exons, or tandem UTR categories. Before the invention of microarray technology, the proportion of multi-exonic genes undergoing alternative splicing was estimated at approximately 50%.[3] However, as the technology improved, these estimates increased to 74% with microarrays[4] and to almost 100% with RNA sequencing.[5] Although RNA sequencing has been a very powerful tool in discovering unique transcription in tissues and diseases[6] and also in elucidating the regulation of transcription,[7–10] accurately quantifying transcripts remains a challenge due to the short read length used in most population-based studies. Currently there are multiple transcript quantification methods available including de novo quantification methods like Cufflinks[11] and Scripture[12] and annotation-based methods like MISO[13] and Flux Capacitor.[8] However, both approaches have inherent flaws because de novo methods make the assumption that the most parsimonious solution best describes the underlying transcriptome and annotation-based methods assume complete knowledge of the transcriptome, both of which are unlikely to be true.

In this study we present a method for relative quantification of splicing events from RNA-sequencing data called Altrans. Our approach is an annotation-based method, which makes the least number of assumptions from the annotation. To this end we chose to simplify the problem and quantify relative frequencies of observed exon pairings in RNA-sequencing data for all categories of splicing events. This approach assumes only correct knowledge of the exons in the transcriptome and is agnostic to the isoform structures defined in an annotation, which would, in theory, make it more accurate and sensitive in the presence of unknown isoforms. We tested the performance of Altrans versus two well-established transcript quantification methods, Cufflinks[11] and MISO,[13] and benchmarked our method in two ways. First, we conducted a simulation study and assessed the concordance of the measured quantifications by each method with the simulated quantifications. Second, we assessed the relative power of discovering alternative splicing quantitative trait loci (asQTLs) for each method. For the asQTL analyses, we chose the Geuvadis dataset, since it was, at the time of analyses, the largest publically available population-based RNA-sequencing study. The Geuvadis dataset comprises 462 individuals in the 1000 Genomes project[14] from five populations—the CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI), and Yoruba (YRI)—and contains data for whole-genome DNA sequencing and deep mRNA sequencing in the lymphoblastoid cell line (LCL)[7] and is thus an ideal dataset for our purposes.

Material and Methods

Altrans Method for Relative Quantification of Splicing Events

Altrans is a method for the relative quantification of splicing events. It is written in C++ and requires a BAM alignment file[15]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland; [2]Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, 1211 Geneva, Switzerland; [3]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland; [4]Center of Excellence for Genomic Medicine Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia; [5]Biomedical Research Foundation Academy of Athens, Athens 11527, Greece
*Correspondence: halit.ongen@unige.ch (H.O.), emmanouil.dermitzakis@unige.ch (E.T.D.)

http://dx.doi.org/10.1016/j.ajhg.2015.09.004. ©2015 The Authors

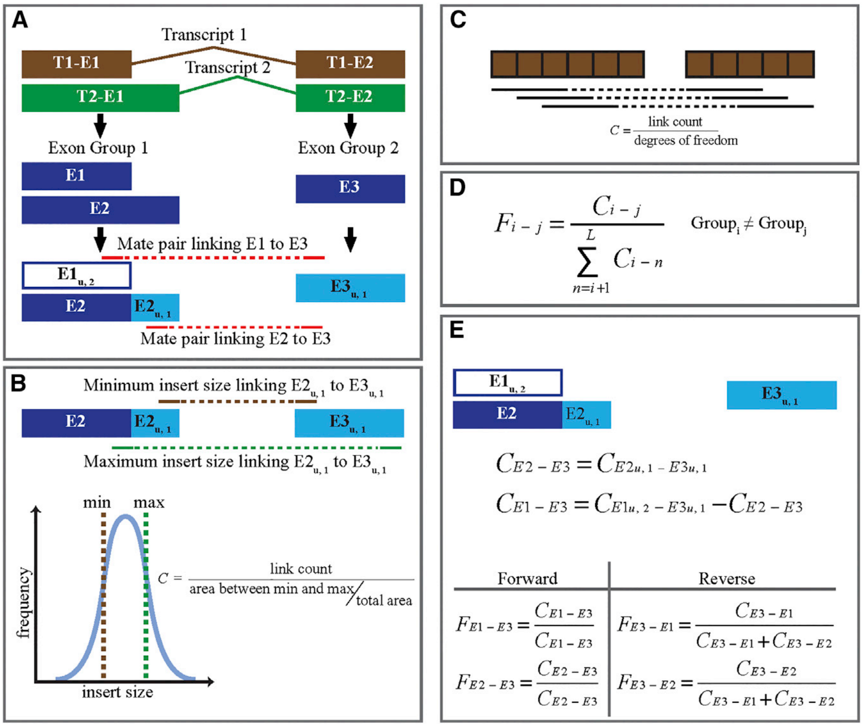
A
Transcript 1
Transcript 2
T1-E1
T1-E2
T2-E1
T2-E2
Exon Group 1
Exon Group 2
E1
E2
E3
Mate pair linking E1 to E3
E1u, 2
E2
E2u, 1
E3u, 1
Mate pair linking E2 to E3
B
Minimum insert size linking E2u, 1 to E3u, 1
E2
E2u, 1
E3u, 1
Maximum insert size linking E2u, 1 to E3u, 1
min
max
frequency
insert size
C = link count / (area between min and max / total area)
C
C = link count / degrees of freedom
D
F i−j = C i−j / Σ n=i+1 L C i−n
Group i ≠ Group j
E
E1u, 2
E2
E2u, 1
E3u, 1
CE2−E3 = CE2u,1−E3u,1
CE1−E3 = CE1u,2−E3u,1 − CE2−E3
Forward
Reverse
FE1−E3 = CE1−E3 / CE1−E3
FE2−E3 = CE2−E3 / CE2−E3
FE3−E1 = CE3−E1 / (CE3−E1 + CE3−E2)
FE3−E2 = CE3−E2 / (CE3−E1 + CE3−E2)

Figure 1. Schematic of the Altrans Algorithm

(A) Overlapping exons are grouped into exon groups where identical exons belonging to multiple transcripts are treated as one unique entity. Two transcripts, shown as connected brown and green boxes, result in two exon groups and three exons shown as blue boxes. Next, the unique regions of each exon, depicted as light blue boxes and a subscript u followed by the level of the exon, are identified. Because E2 has a region that is not shared by any other exon, it is assigned a "level" of 1, and the reads aligning to $E2_{u,1}$ can be unambiguously assigned to E2. E1 does not have a unique portion, and therefore the level 1 exon, E2, is removed from the exon group and the whole of E1 becomes a unique portion, shown as an empty blue box, with a level of 2. These unique regions are used when assigning mate pairs to links as shown with the red lines where the solid portions of the line are the sequenced mates and the dashed part represents the inferred insert.

(B) The default method for calculating link coverage. Link coverage is necessary to normalize the observed counts for the length of the unique portions being linked and the insert size. The theoretical minimum and maximum insert sizes linking the two unique portions, represented as brown and green lines, respectively, are calculated and given the empirically determined insert size distribution, and the area under the curve between the minimum and maximum insert sizes is estimated. The link coverage equals the number of mate pairs linking the two unique portions over the ratio of this area to the area of the whole insert size distribution.

(C) The degrees of freedom method for determining link coverage. Here given a read length and insert size of 3 and two exons that are 6 and 5 bases long, there are three mate pair alignments that can link these two exons. Therefore, the degrees of freedom refer to the theoretical number of positions where a mate pair (given, in this case, 3+3+3 = 9 bp long fragment size) exists that links these exons on the mRNA, shown as black lines. The link coverage is the number of mate pairs linking the exons over the degrees of freedom.

(D) The equation to calculate F value for a link.

(E) A worked example of calculation of the F values. First the coverage of E2 to E3 link (CE2 − E3) is determined from level 1 unique regions ($CE2_{u,1} - E3_{u,1}$), which is then subtracted from the coverage attained from the pseudo-unique E1 to E3 link ($CE1_{u,2} - E3_{u,1}$) in order to calculate the true E1 − E3 coverage (CE1 − E3). In the forward direction, E1 and E2 become primary exons and in the reverse direction E3 is the primary exon and the corresponding F values are calculated as shown.

from an RNA-seq experiment and an annotation file in GTF format containing exon locations. The BAM file is read using the BamTools API.[16] Altrans utilizes paired end reads, where one mate maps to one exon and the other mate to a different exon, and/or split reads spanning exon-exon junctions to count "links" between two exons. For reads aligning to multiple locations in the genome with the same mapping quality, only the primary alignment, i.e., the one reported in the BAM file, is considered and alternative alignments that are reported as tags in the BAM file are ignored. The first exon in a link is referred to as the "primary exon." The algorithm is as follows:

1. Group overlapping exons from annotation into exon groups. Because we are quantifying splicing events and not individual transcripts, transcript level information is ignored and exons with identical coordinates belonging to multiple transcripts are treated as one unique exon (Figure 1A).
2. In order to assign reads to overlapping exons, identify unique portion(s) of each exon in an exon group. Exons with immediate unique portions, where there is no other overlapping exon and where a read can be unambiguously assigned to the exon ($E2_{u,1}$ in Figure 1A), are called "level 1 exons." For exons with no unique positions, remove the level 1 exons from the exon group to determine regions that identify the remaining exons uniquely, where again there is no other overlapping exon after the removal of the level 1 exons ($E2_{u,2}$ in Figure 1A), and increment the level of these exons. In the rare cases where an exon shares its start position with one exon and its end position with another, causing it to have no unique portion, then this exon is removed from the analysis in order to be able to assign unique portions to the remaining exons in the same exon group. In cases where a larger exon overlaps and fully contains two smaller exons, the insert size distribution is used to probabilistically assign links between the two smaller exons (please refer to the Altrans manual for a more detailed annotation of these rare cases). Iterate through this process of removing exons that have unique regions until all exons in a group have unique portions (Figure 1A).
3. Use these unique portions to assign mate pairs or split reads to links (Figure 1A). Links assigned to unique portions that exist only after the removal of overlapping exons are putative assignments and "deconvolution" of these is handled in the next step.
4. Because all the exons have ambiguous unique portions and share regions with other exon(s), reads aligning here might

belong to multiple exons. In order to unambiguously quantify links between these exons, we calculate "link coverage" for all pairs of exons in a given window size. The default method divides the link counts with the probability of observing such a link given the insert size distribution, which is empirically determined from pairs aligning to long exons (Figure 1B). The second method involves calculating the number of degrees of freedom linking two exons given the empirically determined most frequent insert size and read length (Figure 1C). This is an alternative model to calculating link coverage, but we recommend using the default model unless you have a very tight distribution of insert sizes. The link coverage metric ensures that the link counts are normalized for the specific insert size distribution of the experiment, which has direct effect on the observed link counts. Hence, link coverage allows us to quantify an exon link from the unique portions only, i.e., this value should be equivalent to the one we would calculate if we were able to measure the whole exon. The coverage between level 1 exons can be calculated directly using the unique portions, whereas links between higher-level exons are calculated by iteratively subtracting coverage of all the other lower level links from the coverage of these links (Figure 1E).

5. Calculate the quantitative metric, F value, for one exon link as the coverage of the link over the sum of the coverages of all the links that the primary exon makes (Figures 1D and 1E). Using this fraction rather than link counts or coverage ensures that the metric is independent of global effects on gene expression.
6. Repeat step 5 in both 5′-to-3′ (forward) and 3′-to-5′ (reverse) directions to capture splice acceptor and donor effects, respectively (Figure 1E).

Please refer to the Altrans manual where the method is annotated in more detail and examples of how to run Altrans are provided. The program also allows the user to calculate an F value from all the links that a primary exon makes regardless of the direction. Along with the F values, the raw link counts are also outputted, which allow filtering of results eliminating low count links. These raw counts can also be normalized and subsequently reread by the program to calculate the F values. Memory usage and speed heavily depend on the complexity of the annotation and the number of reads in the alignment file. For a sample alignment with 50 million reads and an annotation with 539,748 unique exons, Altrans ran for 20 min and consumed 784 MB of RAM on a single 2.2 GHz core under Linux.

Conversion of Transcript Quantifications to Link Quantifications

We convert the transcript quantifications generated with Cufflinks to relative link quantifications. This is achieved by assigning the same quantification to all linked exons of a transcript based on the measured quantification of the said transcript. We then apply the same method of relative link quantification used in the Altrans algorithm, specifically steps 5 and 6 in the previous section, to calculate the F value for all the links a primary exon makes.

Simulation Analysis

In order to benchmark the link quantifications generated by Altrans, we conducted a simulation analysis using the Flux Simulator software.[17] We simulated an RNA-sequencing experiment with 50 million reads with the GENCODE v.12 annotation[18] reflecting cases where we have a perfect annotation describing all the observed transcripts in the data. Additionally, we introduced novel transcripts, made up of existing exons of a gene, into the annotation. This was achieved by creating novel combinations of exons of a gene while checking for compatibility of the randomly selected exons (non-overlapping, order matches the genomic order, and where a UTR or first or last exon is not an internal exon) and keeping the distribution of number of exons of the random transcripts similar to that of the known transcripts. We then simulated 5 cases with 50 million reads where the novel transcripts accounted for 5%, 10%, 25%, 50%, and 75% of all transcripts, reflecting cases where the annotation is not perfect. Altrans, Cufflinks,[11] and MISO[13] were run on these 6 simulated datasets using the standard GENCODE v.12 annotation. In each simulation, the "correct" quantification of a transcript is taken as the RNA molecule count that the Flux Capacitor used to simulate reads for a given transcript. We have converted these "correct" transcript quantifications and the measured transcript quantifications of Cufflinks and MISO to exon link quantifications, as described in the previous section, and correlated the simulated expected link quantifications with the measured link quantifications for the three programs in the six simulation scenarios, using links where there were overlapping reads or links that were quantified in both the simulation and the given program. We measured the concordance between the simulated and measured quantifications via Spearman's correlation. The estimates of novel splicing in a dataset are done through counting the number of uniquely mapping split reads. We then take junctions that are represented by at least eight split reads and check whether this junction is present in the annotation.

cis-Alternative Splicing QTL Discovery by Each Method in the Geuvadis Dataset

The RNA-seq reads were aligned to the human reference genome (GRCh37) using the GEM aligner[19] and alignments were filtered for properly paired and uniquely mapping reads (mapping quality greater than or equal to 150). Genotypes originated from 1000 Genomes phase 1 data, which is based on 1,092 individuals with 5× whole-genome sequencing data, 80× exome sequencing data, and high-quality genotyping. The genotype data were filtered for variants with MAF $< 5\%$ and HWE $p < 1 \times 10^{-6}$ for each population separately and were corrected for population stratification using the first three and two eigenvectors for Europeans and Africans, respectively.[7] The Altrans link counts were normalized using the first 15 principal components calculated from these link counts. We first looked at all pairwise links between exon groups considering the union of all exons in the exon group as one entity and filter so that we keep only pairs of exon groups that have 15 links in 80% of the samples. Then we count the links between exons of the initial exon group and exons of the terminal exon group and keep only links where the exon in the initial exon group made at least ten links with any of the exons in the terminal exon group in at least 30% of the samples. Cufflinks quantifications were run using the annotation with the –GTF option. In the case of Cufflinks, the transcript quantifications were converted to link quantifications and we assessed links originating from the same genes where there were Altrans quantifications. The *cis*-window for asQTL discovery was 1 Mb flanking the transcription start site of each gene. The associations were run with the FastQTL package.[20] The observed nominal p values were calculated by correlating the

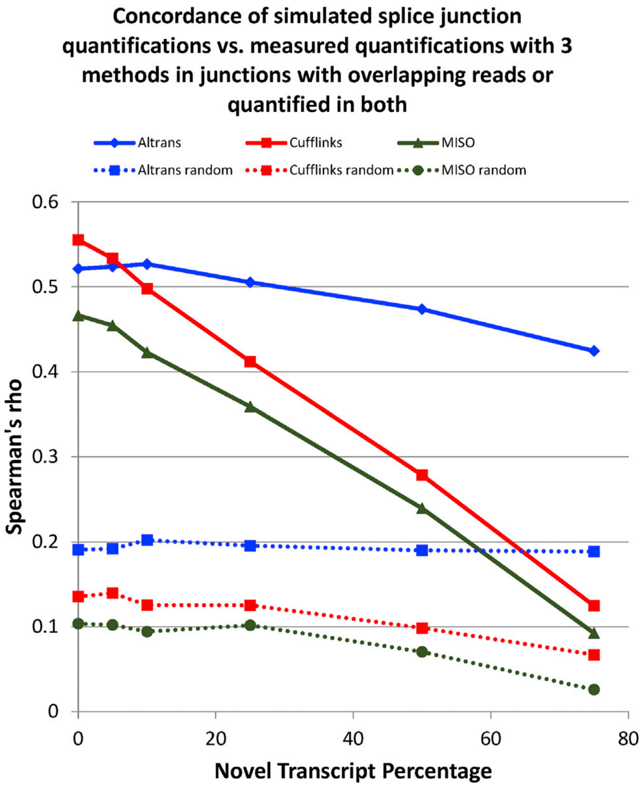

Figure 2. Simulation Results

Using Flux Simulator, we ran six simulations with varying levels of unannotated transcripts. Subsequently, we ran quantifications with three methods with the known GENCODE v.12 annotation. We compared the simulated versus measured link quantifications via Spearman's rank correlation. These comparisons are shown as colored solid lines. In order to produce a null random distribution for each method, we took the link quantifications for each gene, permutated these for 100 times within the links of this gene, and measured the correlation of these random assignments with the simulated ones. By using this sampling method stratified by genes, we account for the variability of number of isoforms per gene. These correlations for random assignments are shown as dashed lines. We observe that as the percentage of novel transcripts increase, the performance of Cufflinks and MISO suffer, whereas this is not the case for Altrans, which results in best quantifications with increased levels of unannotated transcripts.

genotype and link quantifications, which were Gaussian transformed. Subsequently, we ran permutations for each link separately to assign empirical p values to each link. The permutation scheme involved permuting all links of a given gene together 1,000 times and in each permutation iteration, we record the most significant p value from an association between any variant in the cis- window and any link of a given gene, thereby accounting for the dependencies among the link quantifications of a gene, allowing us to find significant asQTL genes. From this distribution of null p values we use an approximation using the beta distribution to estimate the extremes of the null p value distribution, and using this we calculate an adjusted p value. These adjusted p values are then corrected for multiple testing using the qvalue R package.[21]

Classification of Splicing Events

The alternative splicing events were classified into ten categories: alternative 3′ splice site, alternative 3′ UTR, alternative 5′ splice site, alternative 5′ UTR, alternative first exon, alternative last exon, mutually exclusive exon, skipped exon, tandem 3′ UTR, and tandem 5′ UTR. For more information on these events, refer to Wang et al.[5] We then classify each primary exon into these classes based on all of the observed links of the primary exon. This means that a primary exon can be involved in multiple splicing events. From these classifications, we then calculate the proportion of each splicing class in the pool of significant primary exons. This method of classification was chosen because each link quantification is dependent on the quantification of all the other links that a primary exon makes.

Functional Enrichment of asQTLs

To compare the asQTL variants to a null distribution of similar variants without splicing association, we sampled genetic variants in the same *cis*-window of 1 Mb surrounding the transcription start site (TSS) and matched them to alternative splicing variants with respect to relative distance to TSS (within 5 kb) and minor allele frequency (within 2%). The variant effect predictor (VEP)[22] tool from Ensembl was modified to produce custom tags that were STOP_GAINED, SPLICE_DONOR, SPLICE_ACCEPTOR, and FRAME_SHIFT. This modified version of VEP was applied to the imputed genotypes using the GENCODE v.12[18] annotation. To this we added information of overlap with chromatin states[23] and the Ensembl regulatory build,[24] which constituted our functional annotation. The enrichment for a given category was calculated as the proportion between number regulatory associations in a given category and all regulatory variants over the same proportion in the null distribution of variants. The p value for this enrichment is calculated with the Fisher exact test.

Results

Simulation Results

The general overview of the Altrans algorithm is provided in Figure 1. We first aimed to compare the results between Altrans, Cufflinks, and MISO using simulations. We compared six scenarios, one where the given annotation perfectly described the transcripts in the simulations and five others with 5%, 10%, 25%, 50%, and 75% novel transcripts absent from the annotation (see Material and Methods). Subsequently we quantified the six simulation results with both algorithms using the known annotation in all cases. For MISO we have quantified transcript abundances. This was done to assess how methods performed in cases of complete versus incomplete transcriptome knowledge. The transcript quantifications generated by Cufflinks and MISO were transformed into link quantifications to make them comparable to those generated by Altrans.

The results of the simulation analysis are shown in Figure 2. We observe that Cufflinks performs better than Altrans when the annotation is perfect, but as the percentage of novel transcripts in the simulations increases, Altrans performs better because it suffers less from the imperfect annotation used in the quantification. In comparison, MISO performs less well than both methods. In order to produce a null random distribution for each method, we took the link quantifications for each gene

Table 1. Number of Genes Tested and asQTLs Discovered at FDR = 1% in Each Population and by Each Method

Population	Number of Genes: Altrans	asQTLs Altrans	Number of Genes: Cufflinks	asQTLs Cufflinks	Overlap	Overlap p Value
EUR	7,443	1,427	7,148	1,737	780	1.3×10^{-4}
YRI	7,720	166	7,391	304	76	1.2×10^{-4}

The overlap column lists the common genes between the methods and the p value refers to this overlap arising by chance.

and permutated these for 100 times within the links of this gene. We then measured the correlation of these random assignments with the simulated ones and find that Cufflinks and MISO fall to the levels of random assignment of link quantifications as the novel transcripts increase in the simulations. We estimated the proportion of novel transcripts by using split read mappings from a well-studied LCL transcriptome RNA-sequencing experiment[7] and a less well-studied pancreatic beta cell transcriptome RNA-sequencing experiment.[25] We observe that in the LCLs on average 25.8% (SD = 3.5%) and in the beta cells 34.7% (SD = 9.3%) of the junctions are not found in the GENCODE v.12 annotation. Therefore we conclude that in RNA-sequencing experiments where the annotation does not fully reflect the underlying isoform variety, Altrans is a sensitive method for quantifying exon junctions.

cis-Alternative Splicing QTL Discovery and Replication between Populations

The Geuvadis dataset comprises 373 European (EUR) and 89 African (YRI) samples and the *cis*-asQTL discovery was conducted separately in each population as described in the Material and Methods section. At an FDR threshold of 1%, we find 1,472 and 1,737 asQTL genes in the European population with Altrans and Cufflinks, respectively. For the Africans these numbers are 166 and 304, respectively (Table 1). There is a significant overlap between the methods in the asQTL genes, with Altrans finding approximately 45% of the genes identified by Cufflinks in the Europeans and about 25% in the Africans (Table 1). The relative decrease of overlap between the methods in the African population is due to the decreased samples size, hence power, in this cohort compared to the Europeans. When we plot the significant asQTLs distances from the TSS, we observe that for both methods the asQTLs that are shared between the two populations and asQTLs with stronger effects tend to be closer to the TSS than population-specific and weaker asQTLs (Figure 3A). As expected, given the sample sizes of each population, majority of the asQTLs genes in Europeans at this FDR threshold are unique to this populations (91% for Altrans and 86% for Cufflinks) whereas most of the African asQTLs genes are also found in the Europeans (81% for Altrans and 82% for Cufflinks) (Figure 3B). Using a more sensitive π_1 approach,[21] we estimate that 72% of the Altrans asQTLs in Europeans are replicated in Africans and 94% of the African asQTLs are replicated in Europeans. In the case of Cufflinks, these estimates are 78% and 93%, respectively (Figure 3C).

We have taken the correlation coefficient as a proxy to the effect size of an asQTL and compared the absolute value distribution of the correlation coefficients of significant asQTLs identified by each method in both populations (Figure S1). Cufflinks asQTLs have significantly higher effect sizes than Altrans asQTLs (Mann-Whitney $p < 1.69 \times 10^{-5}$, indicating that Altrans is identifying associations with smaller effect sizes compared to Cufflinks and, together with changes in sample size, this contributes to slight decrease of replication of European Altrans asQTLs in Africans, compared to Cufflinks. Of note, when we discover asQTLs in Africans (smaller sample size) and replicate in Europeans (larger sample size), both methods achieve very high levels of replication (94% and 93% for Altrans and Cufflinks, respectively). In order to test the replication of asQTLs by each method independent of sample size and different populations, we have selected 91 European individuals belonging to the CEU population and replicated the findings of this cohort in the larger 282 remaining European samples. When we calculate the π_1 statistic in this analysis, we observe that both methods attain very similar levels of replication (π_1 = 95% for both methods) (Figure S2).

Differences between Methods

Given that both methods replicate at similar levels and Cufflinks finds more asQTLs, one can make the argument that this could be the method of choice. However, almost half of the asQTLs that are discovered with Altrans are unique to Altrans. Although the methodology in identifying splicing QTLs in the original Geuvadis analysis differs significantly from the process described here, we also checked the asQTL gene level overlap between the published lists of splicing QTLs[7] and the ones identified here (Figure S3). We find that Altrans detects 258 out of the 620 asQTLs identified in the Europeans in the original study, and Cufflinks finds 348 overlapping asQTLs. The union of both methods used here identifies 395 genes as significant asQTLs out of the 620 in the original discovery. In the African population, the overlap proportions are similar, with Altrans finding 16 out of 83 asQTLs as also significant, whereas Cufflinks finds 35 common genes, and the union of Altrans and Cufflinks overlaps with 38 asQTLs in the original study. This is a confirmation of the complementary nature of asQTL discovery methods.

We investigated the Altrans-specific asQTLs further. First we find that the majority of the Altrans-specific asQTLs originate from links between exons that are not annotated in the GENCODE v.12 annotation and therefore were

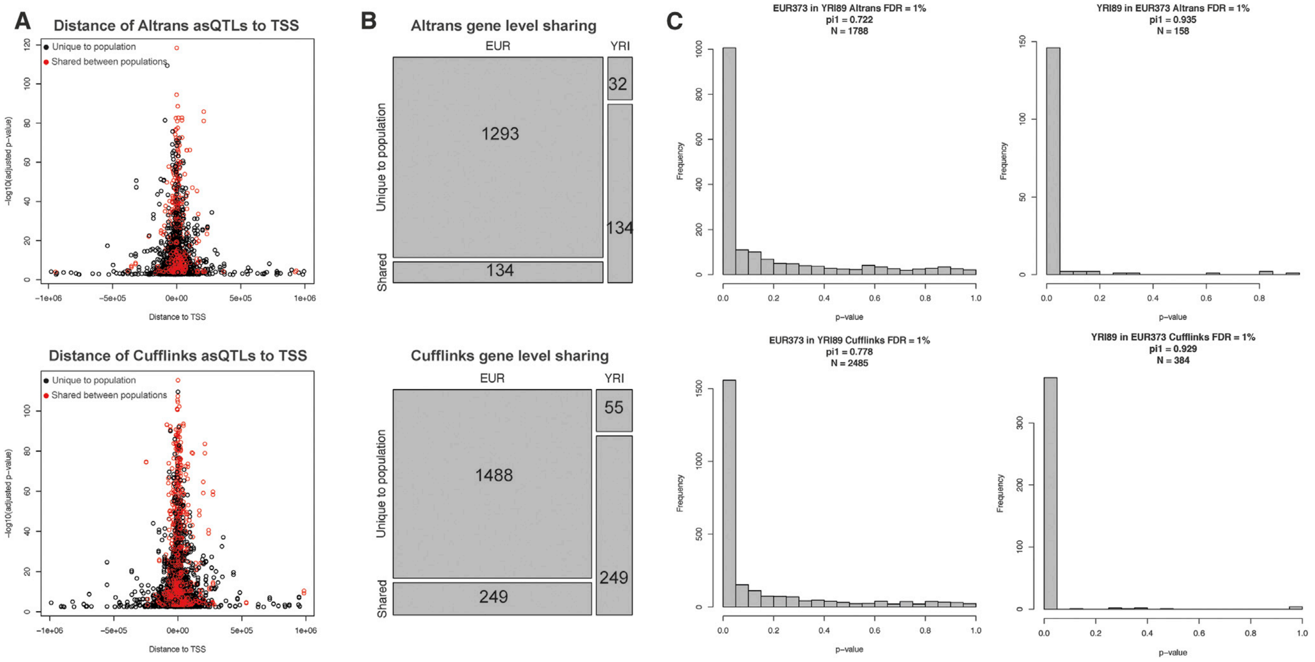

Figure 3. asQTL Discovery
(A) The relative distance of asQTLs to the TSS versus the p value.
(B) Mosaic plots of gene level sharing of asQTLs for each method at FDR = 1%.
(C) The p value distributions of a variant-link pair tested in the other population for each method. From these p value distributions, the π_1 statistic is calculated that estimates the proportion of true positives.

never tested by Cufflinks (89% and 83% not annotated for Europeans and Africans, respectively; Figure S3). Next, we assessed whether Altrans-specific discoveries replicate, and to do so we tested the Altrans-specific discoveries originating from the 91 CEU individuals in the remaining Europeans, and these associations achieve a π_1 statistic of 93%, indicating a high true positive rate in Altrans-specific asQTLs (Figure S4A). We also estimate that 63% of the Altrans-specific asQTLs in Europeans are replicated in Africans and 95% of the African Altrans-specific asQTLs are replicated in Europeans.

Moreover, we compared the types of splicing events that are found to be significant by both methods (Figure S5) and observed that there are differences between the two methods. The majority (66%) of the signal that Altrans captures is due to exon skipping events followed by alternative 5′ and 3′ UTRs (15% and 11%, respectively). In comparison, Cufflinks has a more uniform distribution of significant event types, with the most common being alternative 5′ UTR (23%), followed by exon skipping (15%) and alternative first exons (14%). This difference in types of significant splicing events each method finds highlights their relative merits in identifying different types of splicing events and is one of the reasons for method-specific significant results. We have tested whether the exon skipping events identified by Altrans replicate between CEU discovery and remaining Europeans, and across populations, and we achieve high π_1 values of 98% for CEU discovery replicated in remaining Europeans (Figure S4B), 70% for Europeans replicated in Africans, and 96% in Africans replicated in Europeans, which confirms that these events are enriched for true positives.

Replication of Discoveries by One Method in the Other Method

We wanted to assess how discoveries of one method compared to the other. For each significant variant-link pair in one population by one method, we calculated the p value of the same variant-link pair in the same population based on quantifications by the other method. For this we had to select common links identified by each method, and therefore many genes are not being tested for replication across methods. From these p value distributions, we calculated the π_1 statistic, which indicates the proportion of true positives (Figure S6). We estimate that 94% of Altrans asQTLs in Europeans and 90% Altrans asQTLs in Africans are replicated by Cufflinks quantifications in the corresponding population, for the common links between the two methods. In contrast, replication in the other direction, Cufflinks asQTLs in Altrans, is lower: 57% and 51% for Europeans and Africans, respectively. When we are testing Altrans results in Cufflinks, we are testing 507 and 77 genes for Europeans and Africans, respectively, and when testing Cufflinks in Altrans, these values are 1,260 and 230, respectively. We then multiply the corresponding π_1 values with these number of genes tested to get an estimate of the number of genes that replicate across methods and divide these with the corresponding number of asQTL genes found in the original discovery

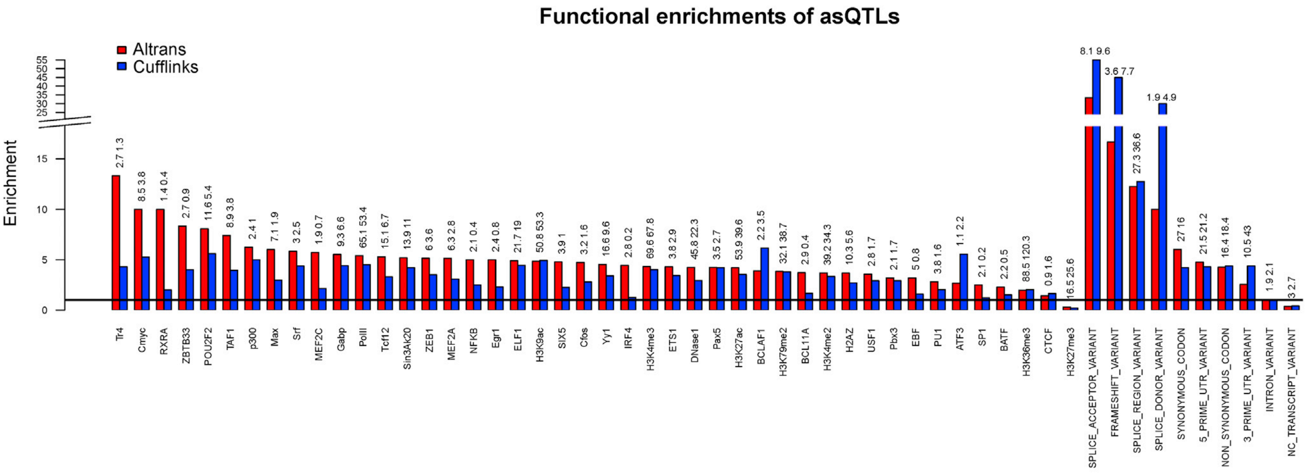

Figure 4. Functional Enrichments of asQTLs Discovered by Altrans and Cufflinks
All variants identified in separate populations are merged. The null (frequency and distance matched) is represented as the black horizontal line. The numbers above each bar are the −log10 p values of the enrichment, Altrans enrichment p value followed by Cufflinks p value.

(e.g., for European Cufflinks in Altrans: 1,260 × 0.57 / 1,737 = 41%). In doing so we estimate the percentage of genes that are "discoverable" by the other method. This percentage is similar across methods and is in the Europeans 33% and 41% for Altrans and Cufflinks, respectively. In the Africans these values are 42% and 39%. This is due to the different space of alternative splicing that each method is best at quantifying and is another confirmation of the complementary nature of these methods.

Functional Relevance of asQTLs

In the absence of a known and true set of asQTLs, we can use the functional annotation of the human genome generated by the ENCODE project to assess whether the asQTLs discovered are likely to be biologically active. If the identified asQTLs are "real," then we would expect them to lie in biochemically functional regions of the genome more often than expected by chance. We have tested this by overlapping asQTLs with functional annotations provided by the Ensemble regulatory build[24] and comparing this overlap to that of random set non-asQTL variants, which were matched to the asQTLs based on relative distance from TSS and allele frequency (Material and Methods). We find significant enrichments for many transcription factor peaks (median 5.2× median $p = 4.41 \times 10^{-8}$ for Altrans and median 4.4× median $p = 2.26 \times 10^{-7}$ for Cufflinks), DNase1 hypersensitive sites (4.2× $p = 1.53 \times 10^{-46}$ for Altrans and 2.9× $p = 5.05 \times 10^{-23}$ for Cufflinks), chromatin marks for active promoters (median 4.3× median $p = 1.51 \times 10^{-51}$ for Altrans and median 4.0× median $p = 5.15 \times 10^{-54}$ for Cufflinks), as well as strong enhancer marks (median 3.9× median $p = 3.46 \times 10^{-40}$ for Altrans and median 3.4× median $p = 2.76 \times 10^{-35}$ for Cufflinks) in asQTLs identified by both methods (Figure 4). We also observe a significant depletion in repressor marks (3.3× $p = 3.51 \times 10^{-17}$ for Altrans and 5.0× $p = 2.28 \times 10^{-26}$ for Cufflinks). All together these results confirm the functional relevance of asQTLs and indicate that we are capturing true biological signal. Furthermore, we also observe strong significant enrichments for variants that are in splice acceptor (33.3× $p = 8.36 \times 10^{-9}$ for Altrans and 55× $p = 2.25 \times 10^{-10}$ for Cufflinks) and donor (10× $p = 0.01$ for Altrans and 30× $p = 1.34 \times 10^{-5}$ for Cufflinks) sites as well as variants in splice regions (12.3× $p = 4.83 \times 10^{-28}$ for Altrans and 12.8× $p = 2.31 \times 10^{-37}$ for Cufflinks), which also indicates that we are capturing variants involved in splicing machinery.

Discussion

Here we present a method, Altrans, for relative quantification of splicing events (Figure 1) to be used in population genetics studies in discovery of asQTLs. Because the phenotype is splicing ratios of exon links calculated from mapping of RNA-sequencing reads without modeling of transcript structure, it is a more direct estimation of splicing. We have assessed the performance of the Altrans algorithm versus the Cufflinks method both on simulated and biological data. The simulation analysis indicates that when the annotation perfectly describes the underlying isoform variety, Cufflinks performs better than Altrans. Because there is no easy way to generate junction annotations that is used by MISO, and because we needed to have a common annotation in all analysis (we could not use the junctions provided in the MISO website), we chose to quantify transcripts rather than junctions. Although MISO might perform better if we had quantified junctions, the analysis performed is equivalent to the one with Cufflinks, and still MISO underperforms compared to Cufflinks. The reason Altrans is worse when compared to Cufflinks in the presence of a perfect annotation is that

Cufflinks quantifies transcript rather than exon links, i.e., it uses the total length of the transcript in quantifications, whereas Altrans uses only the observed reads that are linking a pair of exons. When we convert the transcript quantifications of Cufflinks into link quantifications, this means that all the links in a transcript will "borrow" information from other links of the transcript, whereas in Altrans all the links will be independently measured from the observed reads overlapping the link. Moreover, when the perfect annotation is available using transcript quantifications, as in the case of Cufflinks, then Cufflinks is a more accurate approach. However, the simulations also show that when there are novel transcripts, i.e., isoforms that are not represented in the annotation, the accuracy of transcript quantifications decreases for Cufflinks and MISO whereas Altrans quantifications do not suffer as much as the transcript quantifications. We estimate that in less well-studied transcriptomes like the human pancreatic beta cell transcriptome,[25] the proportion of the links between exons that are novel would be high enough that using the known annotation can result in unreliable estimates.

It is important to assess the performance of a method using biological data, and we applied Altrans and Cufflinks to the Geuvadis dataset[7] with the specific aim of identifying asQTLs. We find 1,427 and 1,737 asQTL genes in the European population and 166 and 304 asQTLs in the Africans with Altrans and Cufflinks, respectively. Using two subsets from the European samples, we show that Altrans and Cufflinks achieve similar levels of replication. Altrans-specific asQTLs accounts for 45% of this method's discovery, which we show is mainly due to it quantifying junctions that are not annotated in the reference. Moreover, these Altrans-specific asQTLs replicate as well as the common genes, indicating that they are probably true positives. The other reason for the method-specific asQTLs is the different types of alternative splicing events each method captures (Figure S4). Altrans is more powerful in capturing exon skipping events, whereas Cufflinks appears to be as powerful in capturing events in the ends of transcripts. This is an expected result given how each method works. Because Altrans is examining reads that link multiple exons, it will perform relatively poorly when a read pair has to extend over constitutive parts of exon groups if constitutive parts are larger than the insert size of the experiment, because there will be very few reads joining these types of exons. On the other hand, because Cufflinks uses all reads over a transcript, it will not fail to quantify these types of events accurately and this is reflected in the types of events each algorithm identifies. Furthermore, when we compare replication of results of one method by the other method and account for the overlap of observed links between the two methods, we find similar levels of overlap between the detectable discoveries for each method.

The relevance of the asQTLs identified by both methods is confirmed by their significant overlap with functional annotations. This result, in the absence of a comprehensive list of asQTLs, shows that asQTLs that we are capturing reside in biochemically active regions of the genome, which reaffirms that we are capturing real biological signal.

RNA sequencing allows us to comprehensively measure transcript diversity in different cells types at the population scale. However, quantifying alternative splicing from short read length RNA sequencing remains a challenge. This problem will be alleviated when technologies that would permit sequencing of full-length transcripts, like nanopore sequencing,[26] become available, reliable, and are cost effective in population studies. Currently all methods have to infer quantifications of transcripts or splice junctions, and each method in doing so has its relative merits. Here we present a different approach to this problem, called Altrans, and show that it is sensitive and performs comparably to other methods. We show that it is capable of identifying thousands of asQTLs, many of which are missed by other methods. We believe it will prove useful in the search for alternative splicing QTLs in population genetics studies.

Supplemental Data

Supplemental Data include six figures and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.09.004.

Acknowledgments

This research is supported by grants from European Commission SYSCOL FP7 (UE7-SYSCOL-258236), European Research Council, Louis Jeantet Foundation, Swiss National Science Foundation, and the NIH-NIMH (GTEx). We thank Alfonso Buil for statistical support. The computations were performed at the Vital-IT Swiss Institute of Bioinformatics.

Received: April 20, 2015
Accepted: September 1, 2015
Published: October 1, 2015

Web Resources

The URLs for data presented herein are as follows:

Altrans, http://sourceforge.net/projects/altrans/
Vital-IT, http://www.vital-it.ch

References

1. Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of *Drosophila melanogaster*. Nature *471*, 473–479.
2. Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69.
3. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh,

W., et al.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

4. Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science *302*, 2141–2144.
5. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.
6. Wang, Y.C., Wei, L.J., Liu, J.T., Li, S.X., and Wang, Q.S. (2012). Comparison of cancer incidence between China and the USA. Cancer Biol. Med. *9*, 128–132.
7. Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.
8. Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.
9. Ongen, H., Andersen, C.L., Bramsen, J.B., Oster, B., Rasmussen, M.H., Ferreira, P.G., Sandoval, J., Vidal, E., Whiffin, N., Planchon, A., et al. (2014). Putative cis-regulatory drivers in colorectal cancer. Nature *512*, 87–90.
10. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.
11. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.
12. Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. *28*, 503–510.
13. Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015.
14. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
15. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
16. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691–1692.
17. Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. *40*, 10073–10083.
18. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.
19. Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nat. Methods *9*, 1185–1188.
20. Ongen, H., Buil, A., Brown, A., Dermitzakis, E., and Delaneau, O. (2015). Fast and efficient QTL mapper for thousands of molecular phenotypes. bioRxiv. doi: http://dx.doi.org/10.1101/022301.
21. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA *100*, 9440–9445.
22. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
23. Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.
24. Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., et al.; 1000 Genomes Project Consortium (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. Science *342*, 1235587.
25. Nica, A.C., Ongen, H., Irminger, J.C., Bosco, D., Berney, T., Antonarakis, S.E., Halban, P.A., and Dermitzakis, E.T. (2013). Cell-type, allelic, and genetic signatures in the human pancreatic beta cell transcriptome. Genome Res. *23*, 1554–1562.
26. Schneider, G.F., and Dekker, C. (2012). DNA sequencing with nanopores. Nat. Biotechnol. *30*, 326–328.