Generalized genetical genomics
Li, Yang

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2010

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 7

Global genetic robustness of the alternative splicing machinery in C. elegans

ABSTRACT

Alternative splicing is considered a major mechanism for creating multicellular diversity from a limited repertoire of genes. Different isoforms can be produced at the same time in the same cell type and their ratios can be the same or different between divergent genotypes. Here, we performed the first study of genetic variation controlling alternative splicing patterns by comprehensively identifying quantitative trait loci affecting the differential expression of transcript isoforms in a large recombinant inbred population of C. elegans, using a new generation of whole-genome very-high-density oligonucleotide microarrays. These arrays provide 3 million measured intensity values for each sample, which allowed us to detect heritable differences in gene expression with exquisite sensitivity and resolution. Using 60 experimental lines, we were able to detect 435 genes with substantial heritable variation for at least one exon, of which 36% were regulated at a distance (in trans). Nonetheless, we find only a very small number of examples of heritable variation in alternative splicing (22 transcripts), and most of these genes co-localize with the associated genomic loci. Our findings suggest that the regulatory mechanism of alternative splicing in C. elegans is robust towards genetic variation at the genome-wide scale. This is in striking contrast to earlier observations in humans, which showed much less genetic robustness.

7.1 Introduction

Alternative splicing of pre-mRNAs is part of gene regulation and a major mechanism for increasing the protein repertoire and the resulting phenotypic diversity. Recently, in individual cases variations in number and ratio of splice variants have also been found in C. elegans in different developmental stages (Barberan-Soler and Zahler 2008b), tissues (Kuroyanagi et al. 2007) and genotypes (Fischer et al. 2008). However, the much smaller number of alternative splicing (Kim et al. 2007), and the strong evolutionary conservation of splicing events in C. elegans (Barberan-
Soler and Zahler (2008a) have been interpreted as signifying a fundamental difference in the way that worms and vertebrates generate diversity from their genetic information. The relative rarity of splicing and high degree of stabilizing selection are seen as having parallels in the limited cellular complexity and highly conserved, rigid developmental programs (Zhao et al. 2008) in worms compared to humans. If this is a general trend, and not restricted to just individual cases of splicing, the conservation of splicing patterns should be reflected at the whole-genome level.

In this chapter we explore this question by extending the genetical genomics strategy (Jansen and Nap 2001) to the characterization of the genetic factors contributing to variations in alternative splicing in 60 recombinant inbred (RI) C. elegans strains. This powerful new strategy, also known as expression genetics (Schadt et al. 2003), has emerged in recent years as a versatile tool to study the genetic basis of gene expression by integrating transcriptomics and classical quantitative genetics (Mackay et al. 2009). In this approach, molecular profiling on a large population of densely genotyped individuals is used to map genomic loci that modulate gene expression. This leads to the identification of expression quantitative trait loci (eQTLs), i.e. polymorphic genetic loci that cause heritable differences in mRNA concentration. eQTLs that are found in close vicinity of the transcript-encoding gene are called local or cis eQTL, while influential loci elsewhere in the genome are known as distant or trans eQTL. Using high-resolution tiling microarray we were able to extend this concept to the detection of genetic determinants of alternative splicing, so-called asQTLs, and to the detailed quantification of the genetic robustness of the alternative splicing machinery in C. elegans on a genome-wide scale.

### 7.2 Results

Here, we performed the first genome-wide analysis of genetic variation of alternative splicing in C. elegans using a comprehensive tiling microarray. We used 60 recombinant inbred lines of a cross between two very diverse strains, Bristol (N2) and Hawaii (CB4856), which have been genotyped using 121 markers (Li et al. 2006b). By using tiling array data, with multiple probes targeting every exon of each gene, we obtained a more comprehensive and sensitive picture of heritable variation of gene expression than possible with previous technologies. It also allows us to dissect the genetic component for differences in isoform-specific gene expression. Thus we can detect alternative splicing quantitative trait loci (asQTL), the genome region controlling variation in isoform-specific expression. Two categories of asQTLs can be
Figure 7.1: Mapping location (a) and type (b) of heritable variation in gene expression.

(a) Each dot in the figure represents a single transcript. The physical position of each transcript is indicated on the y-axis, and the position of the locus that is most strongly associated with variation of the corresponding transcript level is shown on the x-axis. Transcripts on and off the diagonal are locally- and distantly-regulated, respectively. The different symbols/colors discriminate the parental allele of the eQTL that caused a higher expression (N2 is indicated by a red cross and CB4856 by a blue dot). Transcripts that overlap with another gene on the genome according to the WormBase genome annotation are shown in pink (N2>CB4856) and light blue (N2<CB4856).

(b) The same eQTL with a different coloring scheme: colors discriminate consistent eQTL (brown) and asQTL (purple). The relative rarity of the latter category is clearly visible. The few cases that are observed are mostly restricted to the cis diagonal, i.e. they are caused by local variation of the gene sequence close to the affected splice site. Variation in trans-acting splicing factors seems to be extremely rare. Transcripts with revised gene definition are indicated in green. Genes with complex and heterogeneous eQTL patterns are shown in black.

distinguished, i.e. those that map in close vicinity to the gene itself (local) and those that map elsewhere in the genome (distant). Local-activity can be explained, for example, by altered functional motifs in exonic splicing enhancers that will affect the splicing activity. The mechanism of distant regulation is often more complicated and can possibly be explained by a polymorphism in an auxiliary splicing factor (e.g. SR protein) that modulate the activity of the spliceosome. In this case we would ex-
pect to see a genetic master regulator at the locus of the splicing factor controlling isoform ratios for large groups of transcripts.

Figure 7.2: Classification of genes showing heritable expression variation (eQTL)
The 435 transcripts were classified into different groups according to their eQTL pattern: Consistent eQTLs (brown) lead to consistent expression differences in all exons of a gene, non-consistent ones are indicating the need for revised gene annotations (green; 8.7%) or potential heritable differences in splicing (i.e. transcripts with alternative splicing QTL)(Kwan et al. 2008). The latter (purple) are subdivided in three classes according to the position of the alternatively spliced exon; they comprise a total of 5% of all cases, compared to 55% out of total 324 transcripts with significant eQTL that showed heritable isoform changes in humans (Kwan et al. 2008). Complex cases (black) contain indications for multiple event types, e.g. various exons with different patterns of heritable difference. Some cases (10.6%) show very heterogeneous eQTL patterns across probes and exons.

Using a nonparametric effect size testing, corrected for genotype imbalance (Materials and Methods) and corresponding to a $p$-value of 0.001 (Wilcoxon’s test), we detected 435 genes with substantial heritable variation for at least one exon. The comparison of gene position and associated polymorphisms shows that most eQTLs
map in close proximity to the affected gene (local eQTL; 277 genes or 64%; Figure 7.1). There are 158 eQTLs mapping to another chromosome (distant eQTL). 267 genes show higher expression in carriers of the N2 allele than in CB4856 carriers, including 53 cases of known gene deletions in the CB4856 strain (Maydan et al. 2007).

A large majority of eQTLs (319, or 70.4%) lead to a consistent differential expression across all exons of the affected gene. Interestingly, the genetic effects (eQTL size) of these consistent eQTLs shows a strong correlation (Spearman’s $\rho = 0.78$) with a previous experiment using cDNA micorarrays (Li et al. 2006b). As shown in Figure 7.2, 8.7% of cases show evidence for a necessary refinement of existing gene definitions, predominantly by expanding known exons (plotted results for all genes are available at www.wormplot.org for detailed examination). In contrast, we find only 22 genes that show evidence for genetic variation of alternative splicing, i.e. an exon-specific asQTL (Figure 7.3). This genome-wide evidence for the genetic robustness of the alternative splicing machinery is consistent with the earlier indication that individual alternative splicing events in C. elegans are highly conserved and hardly tolerate genetic variation. Note, however, that variation in alternative splicing events restricted to a specific cell or tissue type can be diluted in measurements on whole-worm mRNA. In addition, 77% of asQTL were found to be locally regulated. This agrees with recent findings that alternative splicing can be regulated without involvement of an auxiliary splicing factor, by cis-acting RNA sequences that can function as splicing silencer (Yu et al. 2008).

Our results show that most of the reported asQTLs have strong genetic effects (qualitative on-off patterns). We found only few cases of subtle quantitative effects on alternative splicing. Despite the large population used in this study, technical noise and biological variation might limit our ability to detect subtle shifts in isoform proportions. In order to detect more quantitative effects (Figure 7.4), more precise technology such as deep-sequencing would be required. Even then, reliable detection of changes in isoform proportions will depend on extremely large read numbers.

### 7.3 Conclusions

Our genome-wide study provides the first genome-wide evidence supporting earlier hypotheses that in C. elegans the alternative splicing machinery exhibits a general genetic robustness, and only a minor fraction of genes shows heritable variation in splicing forms and relative abundance. This observation points to a profound difference in the regulation of the alternative splicing machinery compared to humans,
7. Global genetic robustness of *C. elegans* alternative splicing machinery

Figure 7.3: Expression intensity and eQTL effect per probe along the genome for selected genes

(a) Detecting consistent heritable differences in gene expression with high resolution. Nearly 300 probes cover the area of this gene, Y87G2A.5. Exon probes show consistently high expression (median intensity=9.64), compared to intron probes. However, there is huge variation between probes, which makes the clear delimitation of exon boundaries challenging. In contrast, there is a clear and highly consistent differential expression between carriers of different alleles (N2 and CB4856). This so-called eQTL effect, indicated by red bars, is highly reproducible across all exon probes within the gene. In this example, the average expression difference between the two alleles is approximately 2.4-fold. In total, there are 306 genes with similar consistent expression differences. It should be noted that a majority of genes show consistently lower intensity (and thus lower eQTL effect) in the 3’ untranslated region (UTR) indicating the decaying end of transcript (Kolmogorov-Smirnov test, \( p \text{ value} < 2.2 \times 10^{-16} \)).

(b) Refining existing genome annotations. The exon probes within gene T21E8.1 show consistently higher expression for individuals carrying the N2 allele than the CB4856 allele. Additionally, several adjacent probes within the sixth intron show the same differential expression pattern, suggesting that this intron contains a previously unannotated additional exon. This would not have been detectable based on the absolute expression levels, due to the high interprobe variability. We find a total of 41 genes which require refined annotation according to the eQTL pattern, mostly extensions of known exons and redefinitions of the transcript start and end sites.
Figure 7.3: (c) and (d) Detecting heritable variation in alternative splicing. These genes do not show heritable expression differences in general, but individual exons show consistently lower signal for carriers of the CB4856 allele. This suggests that these exons are specifically removed by alternative splicing in one of the two alleles. In both cases, this alternative splicing variation is determined by a local sequence variation (QTL mapping in cis). The first example (Y69H2.3) has been confirmed experimentally (Barberan-Soler and Zahler 2008a). We find 22 comparable instances of heritable differences in splicing patterns.

which parallels the differences in cellular diversity and developmental flexibility in the two species and has important consequences for interpreting future studies using *C. elegans* as a model organism for metazoan splicing.

### 7.4 Materials and methods

#### 7.4.1 Worm samples, genotyping and Affymetrix GeneChips

*C. elegans* recombinant inbred lines were generated and genotyped as described in (Li *et al.* 2006b). mRNA was isolated from 60 RILs reared under standard condition and hybridized to Affymetrix 1.0 *C. elegans* tiling arrays. The hybridization was done by ServiceXS (Leiden, The Netherlands). Since polymorphisms in the probe region can lead to spurious local eQTLs (Alberts *et al.* 2007), 80903 probes (out of $2.9 \times 10^6$ probes on each array) with known SNP (including predicted SNP) were removed for the subsequent analysis. Each probe is annotated as exonic, intronic, or intergenic, when the entire probe of 25bp falls in one of the three regions, respectively. Probes spanning exon-intron boundaries are labeled as boundary probes.

#### 7.4.2 Data analysis

Preprocessing of raw data

The raw gene expression data from 60 microarrays (one RIL per array) were log transformed and quantile normalized. Subsequently, the normalized intensity data were corrected for batch effects using the following linear model: $y_i = \mu + B_i + e_i$, where $y_i$ is the genes intensity on the $i$th microarray ($i = 1, \ldots, 60$); $\mu$ is the mean; $B_i$ is the batch effect defined as the date of hybridization and measurement and treated as a categorical variable; and $e_i$ is the residual error.
Figure 7.4: Schematic illustration (a) and power of detection (b) for quantitative changes in alternative splicing.

(a) We consider a transcript with two alternative splicing forms: the second exon is included in isoform 1 but excluded in isoform 2 (cassette exon). Under allele A, x% of the entire transcript amount are of isoform 1, while isoform 2 is expressed at (1-x)%. Similarly, under allele B, the isoform 1 is expressed at y%, and isoform 2 at (1-y)%. Without loss of generality, we assume that the total transcript amount is 1, thus the detected signal for the 2nd exon is x and y under allele 1 and 2, respectively. The difference between these signals (x-y) will be detected as our asQTL effect.

(b) The asQTL effect size changes for different combinations of x and y. The white dotted line corresponds to our QTL threshold; changes in transcript isoform ratios outside the dotted line are reliably detectable for the population size used.

Differential expression between genotypes (eQTL)

We used a robust and powerful statistical approach to associate microarray probe intensity and genotype data in the face of widely different hybridization properties of individual probes. Instead of computing significance of a statistical test, we evaluated a non-parametric effect size (Cliff’s delta(Cliff 1996)) for all 3 million probes at each genomic marker. To increase the robustness of the procedure, the median effect size of probes within each exon was taken as expression QTL effect size of this exon.
7.4 Materials and methods

The raw gene expression data were quantile normalized and corrected for batch effects using a linear model. For each probe on the array we compute the eQTL effect size using Cliff’s non-parametric Delta statistic:

$$\Delta = \frac{\#(X_{i1} > X_{i2}) - \#(X_{i1} < X_{i2})}{n_1 \times n_2}$$

(7.1)

where $n_1$ and $n_2$ are the number of carriers of the N2 and CB4856 allele, and $\#(X_{i1} > X_{i2})$ is the number of possible pair-wise comparisons where the expression level of gene $i$ in an N2 carrier is larger than in a CB4856 carrier. The genotype information of the 60 RILs was previously described (Li et al. 2006b). For an individual probe, a value of $\Delta = 0.45$ corresponds to a $p$-value $= 0.001$ in a Wilcoxon rank sum test.

As several positions in the genome show a strongly imbalanced genotype ratio (i.e. the number of RILs carrying the N2 allele is far larger than the number of RILs carrying the CB4856 allele at a particular locus), the corresponding threshold ($p$-value) for each marker at significance level $= 0.001$ was obtained first, taking the locus-specific imbalance into account. The thresholds of distorted genome regions are expected to be larger than those of balanced marker positions. Subsequently these marker-dependent thresholds were applied to identify significant eQTL.

**Summarizing the eQTL effect for exons**

To increase the robustness of the procedure, the median effect size of probes within each exon was taken as representing the expression QTL effect size of this exon for each genomic marker. Subsequently, the eQTL profile at the marker with maximal summarized eQTL effect was obtained. To achieve a reliable estimate of eQTL effect size, only exons covered by more than 3 probes were considered here. Transcripts with a summarized eQTL effect larger than the threshold for at least one exon were declared as having a significant eQTL and were used for further analysis.

**7.4.3 Classification of eQTLs pattern**

There are 435 transcripts with a significant eQTL for at least one exon were examined in greater details and manually classified as shown in Figure 7.1. By visualizing the intensity level and eQTL size of the entire transcript, we firstly classified transcripts as consistent eQTL if there is an association of majority of probes significant at threshold of $\Delta = 0.45$. Consistent eQTL lead to consistent expression differences in all exons of a gene, non-consistent ones are indicating the need for revised gene annotations or potential heritable differences in splicing (i.e. transcripts with
alternative splicing QTL) (Kwan et al. 2008). The former are subdivided into five subcategories: new exons, new introns, intron inclusions, exon extensions and intron extensions. The latter are subdivided in three classes according to the position of the alternatively spliced exon: cassette exon, alternative initiation and alternative termination. Transcripts showing evidence for multiple types of variation, e.g. having various exons with different patterns of heritable difference, were classified as complex cases. Heterogeneous cases contain transcripts showing very diverse eQTL patterns across probes and exons and belonging to none of the above-mentioned categories.

To validate the classification procedure, all classifications were performed independently by two researchers, and inconsistent cases checked in more detail. A complete list of classifications and the corresponding plots is available in the Supplementary Table 1 and the corresponding plots for all genes are available at www.wormplot.org.

7.4.4 Permutation

A permutation approach was used to estimate the empirical false discovery rates for the detection of genetically regulated alternative splicing. We permute sample labels in the genotype matrix and keep the correlation structure between traits and the correlation structure between markers; this makes this empirical procedure perfectly suited to a non-biased estimation of the significance under the multiple dependence properties of the data (Breitling et al. 2008a). The permuted data were reanalyzed for all genes at chromosome IV to keep the computational burden within reasonable limits: we repeated the QTL detection and classification as we did for the real data. Based on a total of 67,000 permuted instances of genes, we estimated the false discovery rate for genetically regulated alternative splicing case being <1%.

7.4.5 Deleted genes

We validated our ability to detect heritable expression differences by examining published gene deletions in CB4856 worms (Maydan et al. 2007). These genes should show consistently variable expression according to the local genotype. Of 531 CB4856-deleted genes, about 10% (53 genes) are detected as expressed in our experiment. All of these genes show consistent eQTL across all probes with larger expression in N2 allele, well above our threshold. This confirms the sensitivity of our approach.
7.4.6 Comparison with previous experiment

As a further validation step, we compared the detected eQTL to those observed in an earlier study using cDNA microarrays (Li et al. 2006b). Nearly half of the top-500 highly expressed genes (231 genes) are shared in the two experiments. The eQTL effect size also shows strong correlation (locally regulated QTL: \( r = 0.72 \), distantly regulated: \( r = 0.48 \)). Several strong distant eQTL were found in both experiments including ZK488.6, F10D2.9 (fat-7), 56H6.5 (gmd-2), C38D9.2, T21E8.1 (pgp-6), C05A9.1 (pgp-5), F15D4.5.

7.4.7 Power to detect quantitative changes in alternative splicing

Generally, the genetic effect on the abundance of transcript isoforms can be quantitative rather qualitative (shifts in isoform ratios, rather than on-off effects). We calculated the expected effect size for all possible shifts of isoform ratio, assuming that two isoforms differ only by the presence or absence of one exon, and that there is no overall expression difference (Figure 7.2). It turns out that the difference in abundance of transcript isoforms should be at least about 1.86-fold to be picked up in our study. This means that our method has sufficient power to identify quantitative changes in isoform ratio like 90:10 (allele 1) \( \rightarrow \) 20:80 (allele 2) or 60:40 (allele 1) \( \rightarrow \) 12:88 (allele 2).

7.4.8 Supporting information


7.5 Acknowledgments

This work was supported by EU FP7 PANACEA 222936 and the Netherlands Organization for Scientific Research, NWO-86504001.