# Chapter 8
# Mapping of Expression Quantitative Trait Loci Using RNA-seq Data

**Wei Sun and Yijuan Hu**

**Abstract** RNA sequencing (RNA-seq) is replacing expression microarrays for genome-wide assessment of gene expression abundance. Many sophisticated statistical methods have been developed to map gene expression quantitative trait loci (eQTL) using microarray data. These methods can potentially be applied to RNA-seq data with minor modifications. However, they fail to exploit two types of novel information that are available from RNA-seq but not from microarrays: the allele-specific expression (ASE) and the isoform-specific expression (ISE). This chapter gives an overview of the statistical methods that are specifically designed for eQTL mapping using RNA-seq data, as well as the challenges and some future directions.

## 8.1 Introduction

In most living organisms, the DNA information stored in a cell is transcribed into messenger RNA (mRNA) and then translated into protein, which is the working force of the cell. The amount of mRNA produced by a gene is generally referred to as gene expression. Since mid 1990s, gene expression microarrays have been widely employed to assess mRNA abundance genome-wide. The huge amount of data produced by expression microarrays have not only greatly improved our understanding of cell biology, but also provided invaluable resources to guide the diagnosis and treatment of human diseases. For example, gene expression profiles have been used to dissect cancer subtypes [45] and to predict drug sensitivities [20].

W. Sun (✉)
Department of Biostatistics, UNC Chapel Hill, Chapel Hill, NC, USA
e-mail: weisun@email.unc.edu

Y. Hu
Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA
e-mail: yijuan.hu@emory.edu

S. Datta and D. Nettleton (eds.), *Statistical Analysis of Next Generation Sequencing Data*, Frontiers in Probability and the Statistical Sciences, DOI 10.1007/978-3-319-07212-8_8,

The mRNA abundance of a gene may be associated with the genotype of one or more genetic loci, which are referred to as expression quantitative trait loci (eQTL). In most eQTL studies, genome-wide gene expression data and DNA genotype data of genetic markers such as single nucleotide polymorphisms (SNPs) are collected in a common set of samples. Then eQTLs are identified by linkage/association analysis in which the expression of each gene is treated as a quantitative trait. We refer the readers to [10, 51] for reviews on eQTL studies and their potential impacts on understanding the genomic basis of human complex traits, and to [33, 68] for reviews on statistical methods and computational tools for eQTL studies using gene expression from microarrays.

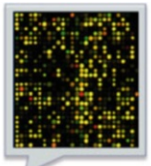
In this chapter, we will focus on eQTL mapping using RNA-seq data. RNA-seq, i.e., high-throughput RNA sequencing, is replacing expression microarrays for transcriptome studies. To explain the motivations of designing statistical methods specifically for RNA-seq data, it is helpful to first describe the differences between the microarray and RNA-seq platforms. In microarray experiments, the abundance of gene expression is measured by fluorescent signals on a set of probes, where each probe contains a specific short piece of DNA sequence (e.g., 25 base pairs for most Affymetrix arrays). The amount of information that can be obtained is limited by the design of the microarray:

- The quantification of gene expression is confined to the regions where the probes are placed. The probes are pre-selected to cover known genes, and in most array platforms, the probes are located at the 3' ends of the transcripts instead of being uniformly distributed across exonic regions. Therefore, previously unknown transcripts cannot be measured for expression and the measurements at known transcripts may be biased by the signals at the 3' ends.
- The same probe sequences are used for all samples and do not accommodate the genetic differences across samples or the differences between the paternal and maternal alleles of a sample. Therefore, the gene expression from the paternal and maternal alleles cannot be distinguished.

In RNA-seq experiments, the expression of a gene is measured by the number of sequence reads mapped to that gene [18, 42]. RNA-seq overcomes the two limitations of microarrays. First, RNA-seq objectively quantifies the genome-wide transcript abundance without relying on pre-selected probes. Second, an RNA-seq read delivers allele-specific information if it overlaps with at least one heterozygous SNP/indel (i.e., a SNP or an insertion or deletion that is heterozygous between the paternal and maternal alleles).

Figure 8.1 illustrates the data generated by the two platforms. In particular, microarray data take continuous values and RNA-seq data are discrete counts. If that is all the difference between the two platforms, then there is no need to develop novel statistical methods for RNA-seq data because one can simply replace the linear regression model for continuous microarray data with the generalized linear regression model (with Poisson or negative binomial distribution assumption) for count data. In fact, the raw sequence data from RNA-seq contain much more information than a single count as shown in Fig. 8.1. First, in a diploid genome such

**a**



| | Sample 1 | Sample 2 | ... | Sample n |
|---|---|---|---|---|
| Gene 1 | 10.45 | 10.52 | | 10.40 |
| Gene 2 | 4.63 | 4.76 | | 4.70 |
| Gene 3 | 8.80 | 8.96 | | 8.82 |
| ... | | | | |
| ... | | | | |
| Gene p | 8.30 | 8.44 | | 8.27 |

**b**

```
@HWI-EAS134:1:1:0:1446#0/1
GGGACTTAATCAACGCAAGCTTATGACCCGCACTT
+
BB<2A81(66B<2>BB@>?B=BA6=8@BCBB@?3/
```

| | Sample 1 | Sample 2 | ... | Sample n |
|---|---|---|---|---|
| Gene 1 | 512 | 339 | | 286 |
| Gene 2 | 1043 | 1212 | | 888 |
| Gene 3 | 20 | 12 | | 10 |
| ... | | | | |
| ... | | | | |
| Gene p | 78 | 65 | | 42 |

**Fig. 8.1** **(a)** Gene expression data from a microarray. Each sample is measured by an array with tens of thousands of pre-selected probes. The expression of one gene is estimated by combining the fluorescent signals of multiple probes. **(b)** Gene expression data from RNA-seq. The data of each sample is stored in a text file, usually in the FASTQ format. An FASTQ file contains millions of records and each record corresponds to an RNA-seq read with four lines: the sequence identifier, the actual DNA sequence, a separator, and the sequencing quality scores for every base pair of the sequence

as the genome of human or mouse, there are two sets of chromosomes, one from the father and one from the mother. Thus most genes (e.g., autosomal genes and X-linked genes in females) have two copies and each copy is called an allele of this gene. The expression of each allele of a gene, i.e., **allele-specific expression** (ASE), can be extracted from the raw RNA-seq data. Second, in a higher organism such as a human or mouse, one gene often comprises of several exons and the exons can be grouped in different ways to produce different proteins or non-coding RNA molecules. Each combination of the exons of a gene is called a transcript or an RNA isoform. The expression of each isoform, i.e., **isoform-specific expression** (ISE), can also be inferred from the raw RNA-seq data. In summary, the RNA-seq platform delivers much more information than the microarrays and thus warrants the development of novel statistical methods to fully exploit the new features.

The remainder of this chapter is organized as follows. Sections 8.2 and 8.3 will introduce eQTL mapping using ASE and ISE, respectively. Section 8.4 will discuss some challenges and future directions.

## 8.2 eQTL Mapping Using ASE

We will first describe the quantification of ASE and show how the ASE enables the detection of *cis*-regulatory eQTLs. Then we will introduce statistical methods for eQTL mapping using ASE under two scenarios, namely, with and without known haplotypes between the candidate eQTL and the gene of interest.

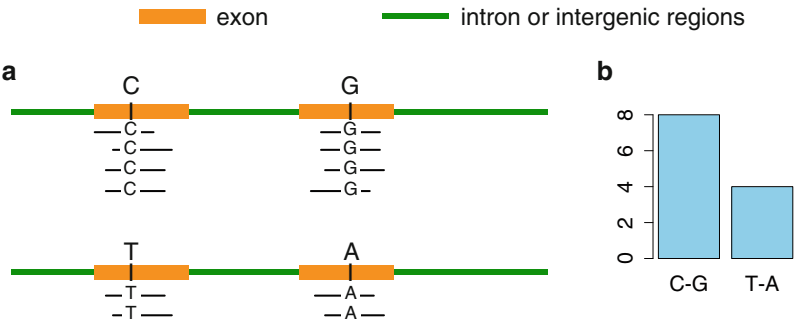



**Fig. 8.2** An example of ASE abundance quantification using RNA-seq, for a hypothetical gene with two exons and one heterozygous SNP within each exon. **(a)** Two haplotypes of this gene. **(b)** The number of allele-specific reads from these two haplotypes

### *8.2.1 Quantification of ASE Using RNA-seq*

ASE can be measured by the number of RNA-seq reads that are mapped to the gene and overlapped with at least one SNP or indel with heterozygous genotype. Figure 8.2 illustrates the quantification of ASE for a hypothetical gene with two exons. There are two SNPs with heterozygous genotypes on the exonic regions of this gene, one SNP for each exon. Given the genotype at each SNP, allele-specific read count (**ASReC**) can be obtained by counting the number of reads harboring a particular SNP allele. For example, there are 6 reads overlapping with the first SNP with genotype CT, and the ASReCs are 4 and 2 for SNP alleles C and T, respectively. Then, the ASE of this gene can be estimated by combining ASReCs across multiple SNPs if the haplotype information is available. In the example shown in Fig. 8.2a, the genotypes of the two SNPs are CT and GA and the possible haplotype pairs are (C-G, T-A) and (C-A, T-G). If we knew that the underlying haplotype pair is (C-G, T-A), we could obtain the gene-level ASReCs as shown in Fig. 8.2b.

Next we discuss a few issues related to ASE quantification: haplotype phasing, sequence mapping bias, and expected ASReC.

#### 8.2.1.1 Haplotype Phasing

Many algorithms (e.g., [8,12,36]) have been developed to infer the haplotype phases from the genotypes of unrelated individuals. It is well known that the phasing accuracy deteriorates as the length of the haplotype increases. However, it is still reasonable to assume that the phasing is accurate within the exonic regions of a gene because those regions are relatively short (∼90 % of the annotated genes are shorter than 100 kb [16]) and tend to undergo less recombination [62]. In addition, the switch errors (i.e., mistaken swapping from one haplotype to the other) in exonic regions can be captured and corrected by RNA-seq reads (either single or

paired-end reads) that overlap with two or more heterozygous SNPs (i.e., SNPs with heterozygous genotypes) and thus provide direct information on the haplotype phase. Some reads may even span over non-adjacent exons due to alternative splicing and thus provide information on long-range phase.

#### 8.2.1.2 Sequence Mapping Bias

A common practice in RNA-seq studies is to map the reads of all samples against the same reference genome. This may induce mapping bias because the reads harboring reference alleles tend to be mapped more accurately than those harboring alternative alleles. There are several solutions to this problem.

1. Identify and remove SNPs that may cause mapping bias by mapping simulated reads to the reference genome [46].
2. Employ an allele-aware sequence aligner [70] that uses both the reference genome and alternate alleles to map reads.
3. Construct the two haploid genomes for each diploid individual and map the reads against the two genomes separately [26, 30].

The third approach is the most unbiased and most comprehensive one, although it requires more information, i.e., the complete haploid genomes, and more computational time. Such an effort can be well justified for certain diploid samples with two very different haploid genomes, e.g., F1 mice from a cross of two inbred mouse strains with different genome backgrounds.

#### 8.2.1.3 Expected ASReC

What proportion of RNA-seq reads are allele-specific? The answer depends on two factors, the density of DNA polymorphisms (usually SNPs or indels) with heterozygous genotypes and the read length. Clearly, the more different are the two haploid genomes, the more reads are allele-specific; the longer the reads are, the more likely they overlap with heterozygous DNA polymorphisms. The expected proportion of allele-specific reads can vary from 0.5 % in a human study with short reads [46, 55] (Fig. 8.3a) to 35 % in an F1 mouse study with longer reads [11] (Fig. 8.3b). To be specific, the human study [46,55] adopted an RNA-seq experiment with 35 bp single-end reads and used ~1.4 million HapMap SNPs to extract allele-specific reads. The number of heterozygous SNPs for an individual ranges from 392,800 to 415,500 with a median of 409,100. In another on-going study involving 550 breast cancer patients from The Cancer Genome Atlas (TCGA) using 2×50 bp paired-end reads and ~30 million 1000G SNPs, we identified 3.4 % reads as allele-specific. The number of heterozygous SNPs across these TCGA samples ranges from 1.91 million to 2.02 million with a median of 1.97 million. The increase of the proportion of allele-specific reads from 0.5 % to 3.4 % in the two human studies can be attributed to both the longer reads and the larger number of heterozygous

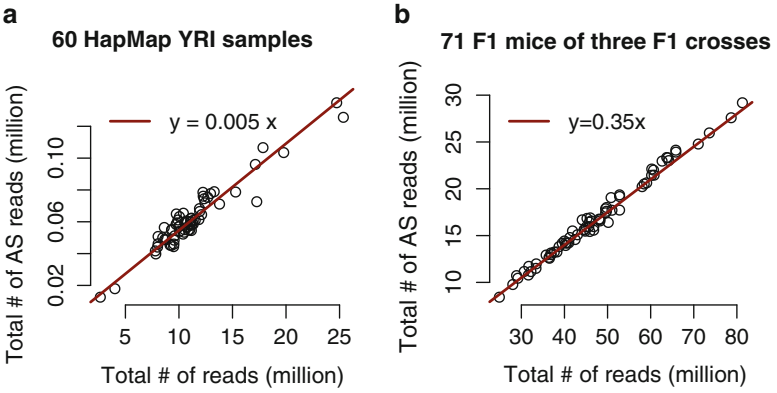



**Fig. 8.3** Scatter plot of the total number of RNA-seq reads versus the total number of allele-specific reads for all the samples in (**a**) a human study of unrelated individuals of African population (HapMap YRI samples) [55] and (**b**) a mouse study of three reciprocal F1 crosses of three mouse inbred strains (CAST/EiJ, PWK/PhJ and WSB/EiJ) representative of three subspecies within the *Mus musculus* species group (*M. m. castaneus*, *M. m. musculus* and *M. m. domesticus*, respectively)

SNPs. By contrast, the mouse study [11] collected $2\times100$ bp paired-end RNA-seq reads from F1 mice with around 17.5 million heterozygous SNPs/indels per sample, making it possible to harvest 35 % of RNA-seq reads as allele-specific.

### *8.2.2 ASE for cis-eQTL Mapping*

Given ASE, we can assess whether there is allelic imbalance of gene expression. In some publications, the terms ASE and allelic imbalance are used exchangeably. In this book chapter, however, ASE indicates the expression measurement from a particular allele. ASE is available for a gene if it has exonic SNPs/indels with heterozygous genotypes, and thus having ASE does not imply allelic balance. A number of pioneering studies have shown that allelic imbalance in gene expression exists and may be associated with disease susceptibility [17, 27, 35, 40, 60, 73]. For example, the reduction in the expression of one allele at the TGFBR1 gene in blood cells (germline) leads to an elevated risk of colorectal cancer [60]. In addition, effective treatments can be developed by silencing the disease allele while sparing the expression of the wild-type allele [41]. Here, we focus on mapping the DNA polymorphism that leads to allelic imbalance of gene expression, which is called a *cis*-eQTL and is a main mechanism of allelic imbalance.

To better understand *cis*-eQTLs, it is helpful to introduce the concept of *trans*-eQTL and clarify their differences. *Cis*-eQTL and *trans*-eQTL have been widely used to refer to eQTLs that are close to the associated genes and eQTLs that are distant, respectively. An arbitrary distance, such as 200 kb or 1 Mb, is often used

to distinguish local and distant eQTLs. It has been pointed out before [51] and is worthwhile to be emphasized again: it is misleading to refer to a local or distant eQTL as a *cis*- or *trans*-eQTL as the latter have their own biological meanings.

The Latin words *cis* and *trans* mean "on the same side" and "across", respectively. A *cis*-eQTL is located on the same chromosome as its target gene and influences the gene expression in an allele-specific manner. Specifically, a mutation in the maternal allele only changes the gene expression from the maternal allele but does not affect the expression from the paternal allele (Fig. 8.4a). A plausible scenario is that a *cis*-eQTL is located at the transcriptional factor binding site of a gene and thus interferes with the transcriptional factor binding in the allele-specific manner. A *cis*-eQTL is likely to be a local eQTL, though this is not always true. By contrast, a *trans*-eQTL of a gene can be located anywhere in the genome and it influences the gene expression of both alleles to the same extent. One possible mechanism is that a *trans*-eQTL modifies the activity or abundance of a protein that regulates the gene and such regulation does not distinguish the two alleles of the gene [67] (Fig. 8.4b). Therefore, *cis*- and *trans*-eQTLs should be distinguished by ASE (Fig. 8.4a, b) [14, 52] rather than their physical distance to the target gene. Note that *cis*- and *trans*-eQTLs cannot be distinguished by the total expression of the gene, which shows the same pattern at the population level (Fig. 8.4c, d).

From the above discussions, it is clear that ASE is informative for *cis*-eQTL mapping. Figure 8.5a–d shows a hypothetical example of *cis*-eQTL mapping using ASE. Assume that the gene of interest has two exons with one SNP for each. We wish to test whether a candidate eQTL, displayed on the left of the gene in Fig. 8.5a, *cis*-regulates the gene expression. First, we count the number of allele-specific reads. As mentioned in Sect. 8.2.1, an RNA-seq read is allele-specific if it can be assigned to one of the two alleles of the gene without ambiguity. As illustrated in Fig. 8.5a, individuals (i) and (ii) have heterozygous genotypes for at least one exonic SNP, and thus their ASE can be measured by the number of reads that overlap with the heterozygous SNPs. Haplotype information is required to combine ASE measured at individual exonic SNPs into the gene-level ASE. For example, for individual (i), we count the number of allele-specific reads mapped to the haplotypes A-A and T-G. Next, we associate ASE with the candidate eQTL. For individual (i) in Fig. 8.5a, given the longer haplotypes C-A-A and T-T-G that span over the gene as well as the candidate eQTL, we can link ASE of the A-A and T-G haplotypes of the gene to the C and T alleles of the candidate eQTL, respectively (Fig. 8.5c). The association testing seeks to answer the question whether one allele of the candidate eQTL is associated with a higher or lower ASE of the gene. If the answer is yes (and assuming there is no other factor inducing the allelic imbalance), then we expect allelic imbalanced expression when the genotype of the candidate eQTL is heterozygous and allelic balanced expression when the genotype is homozygous; in other words, the candidate eQTL is a *cis*-eQTL. For example, individual (i) has a heterozygous genotype C/T at the candidate eQTL and has a higher ASE corresponding to the C allele than the T allele (Fig. 8.5c). Individual (ii) has a homozygous genotype C/C at the candidate eQTL, each C allele corresponding to the same ASE (Fig. 8.5d). A real data example of 65 HapMap samples is shown in Fig. 8.5f.

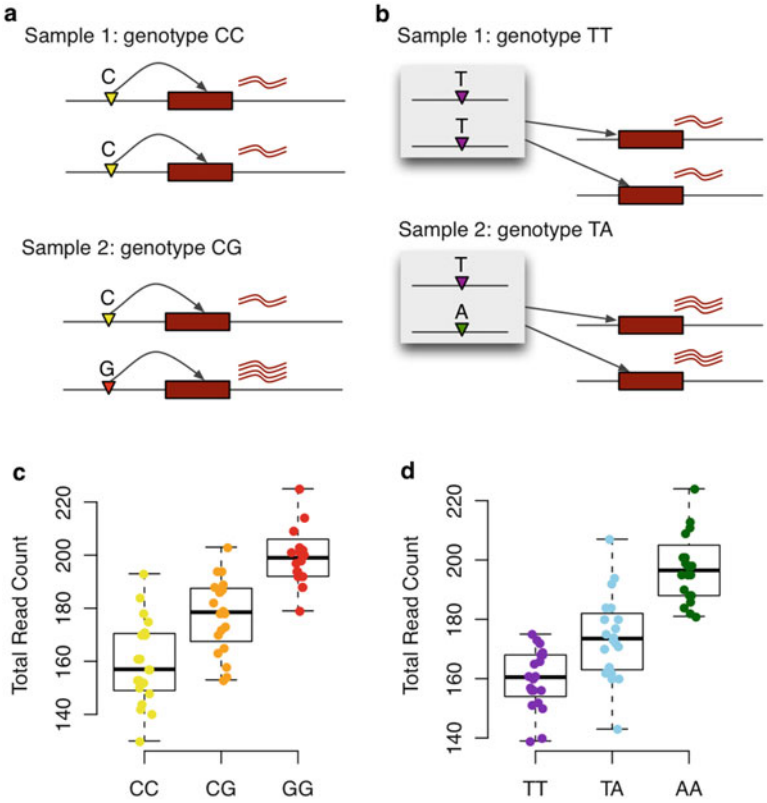



**Fig. 8.4** **(a)** An example of a *cis*-eQTL in two samples. In sample 2 where the candidate eQTL (the SNP for which we test association) has a heterozygous genotype CG, the expression of the two alleles are different. **(b)** An example of a *trans*-eQTL in two samples. In sample 2 where the candidate eQTL has a heterozygous genotype TA, the expression of the two alleles are the same. **(c)** A simulated data for a *cis*-eQTL across 60 samples with 20 samples within each genotype class. **(d)** A simulated data for a *trans*-eQTL across 60 samples with 20 samples within each genotype class. This figure is adapted from Fig. 1 in our earlier paper Sun and Hu (2013) [56]

The total read count (**TReC**) is also informative for *cis*-eQTL mapping, which is similar to the traditional eQTL mapping using gene expression measured by microarrays. While ASE provides information at the allele level, TReC contributes at the individual level and in a way that is consistent with the allele level. In Fig. 8.5a–d, the C allele of the candidate eQTL is associated with a higher ASE, which is manifested at the allele level (Fig. 8.5c, d) and at the individual level (Fig. 8.5b). In general, the TReC of a gene is much greater than the sum of the two ASReCs in that TReC includes many reads that do not overlap with any heterozygous SNPs/indels.

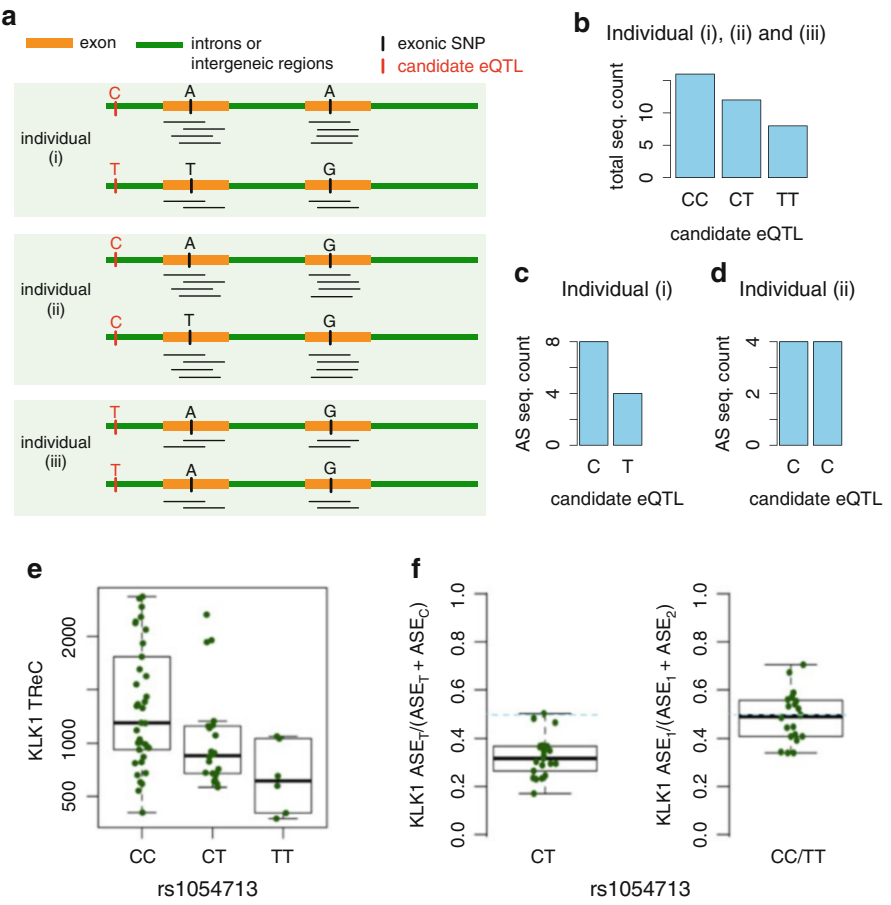



**Fig. 8.5** **(a)–(d)** A hypothetical example of *cis*-eQTL mapping. **(a)** RNA-seq measurements of a gene with two exons in three individuals. **(b)** TReC (total read count) for the three individuals. **(c–d)** ASE for individual (i) and (ii). **(e)–(f)** A real data example of *cis*-eQTL mapping between gene KLK1 and SNP rs1054713. **(e)** Association between the genotypes and TReC. The y-axis is the total number of reads mapped to the gene KLK1 and each point corresponds to one of the 65 samples. **(f)** Association between the genotypes and ASE. When the genotype of rs1054713 is heterozygous, the ASE of the two alleles of this gene can be associated with the two alleles of rs1054713. $ASE_T$ and $ASE_C$ denote the ASReC corresponding to the T and C allele of rs1054713, respectively. When the genotype of rs1054713 is homozygous, we denote the ASReC of the two alleles of this gene by $ASE_1$ and $ASE_2$, respectively. This figure is a modified version of Figs. 2 and 4 of the earlier paper by Sun and Hu (2013) [56]

### *8.2.3 eQTL Mapping Using ASE with Known Haplotypes*

While the haplotypes across the exonic regions of a gene can be accurately phased, those extending from the gene to a candidate eQTL may not be reliably phased

because the candidate eQTL may be far away from the gene. In this section, we assume that the extended haplotypes are known and defer the scenario with unknown haplotypes to the next section.

Our statistical model is for a particular gene of interest. To simplify the notation, we skip the index for gene. The model was originally proposed by Sun (2012) [55] and reviewed by Sun and Hu (2013) [56]. We use the following notation.

- Let $H = (h_1, h_2)$ denote the haplotype pair consisting of haplotypes $h_1$ and $h_2$ across the exonic SNPs. Let $\tilde{H} = (\tilde{h}_1, \tilde{h}_2)$ denote the extended haplotype pair consisting of both the exonic SNPs and the candidate eQTL. Here the order of the two haplotypes is arbitrary and thus $(h_1, h_2)$ is the same as $(h_2, h_1)$ and $(\tilde{h}_1, \tilde{h}_2)$ is the same as $(\tilde{h}_2, \tilde{h}_1)$. We assume that both $H$ and $\tilde{H}$ are known here.
- Let $T$ be the total read count (TReC). Note that a paired-end sequence read is counted as one read.
- Let $N_1$, $N_2$ and $N$ denote the allele-specific read count (ASReC) from haplotypes $h_1$ and $h_2$ and the total ASReC, respectively. Naturally, $N = N_1 + N_2$.
- Let $G$ be the genotype of the candidate eQTL, which has two alleles A and B. Under the additive genetic effect, $G = 0$, 1, and 2 for genotypes AA, AB and BB, respectively. Dominant, recessive, and co-dominant effects can also be modeled using appropriate coding for genotypes.
- Let $\mathbf{X}$ be the relevant covariates including an intercept. Typically, $\mathbf{X}$ include the log form of the total read count per sample reflecting the read depth.

We model the probability of $T$ given $G$ and $\mathbf{X}$ by a negative binomial distribution indexed by parameters $(\boldsymbol{\gamma}, \beta_T, \phi)$, which is denoted by $P_{\mathtt{TReC}}(T|G, \mathbf{X}; \boldsymbol{\gamma}, \beta_T, \phi)$. A negative binomial distribution can be considered as an infinite gamma mixture of Poisson distributions. It allows over-dispersion in the read counts, a phenomenon that is often observed in sequencing data across biological replicates. Thus the negative binomial distribution has been commonly used for RNA-seq data analysis [5]. In particular, we assume that $T$ follows the negative binomial distribution with mean $\mu$ and a dispersion parameter $\phi$:

$$P_{\mathtt{TReC}}(T|G, \mathbf{X}; \boldsymbol{\gamma}, \beta_T, \phi) = \frac{\Gamma(T + 1/\phi)}{T!\Gamma(1/\phi)} \left(\frac{1}{1+\phi\mu}\right)^{1/\phi} \left(\frac{\phi\mu}{1+\phi\mu}\right)^{T},$$

where

$$\log(\mu) = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X} + w(G, \beta_T),$$

and

$$w(G, \beta_T) = \begin{cases} 0 & \text{if } G = 0 \\ \log[1 + \exp(\beta_T)] - \log 2 & \text{if } G = 1 \\ \beta_T & \text{if } G = 2. \end{cases}$$

The functional form of $w(G,\beta_T)$ reflects the additive genetic effect. To see this, we write the means of $T$ given $\mathbf{X}$ and $G=0,1,2$ by $\mu_{AA,\mathbf{X}}$, $\mu_{AB,\mathbf{X}}$ and $\mu_{BB,\mathbf{X}}$, respectively, where

$$\begin{aligned}\mu_{AA,\mathbf{X}} &= \exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}),\\ \mu_{AB,\mathbf{X}} &= \exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}+\log[1+\exp(\beta_T)]-\log 2)\\ \mu_{BB,\mathbf{X}} &= \exp(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}+\beta_T).\end{aligned}$$

We can see that $\beta_T$ characterizes the difference between $\log(\mu_{AA,\mathbf{X}})$ and $\log(\mu_{BB,\mathbf{X}})$ and $\mu_{AB,\mathbf{X}}$ is at the mid point between $\mu_{AA\mathbf{X}}$ and $\mu_{BB,\mathbf{X}}$, i.e., $\mu_{AB,\mathbf{X}} = (\mu_{AA,\mathbf{X}} + \mu_{BB,\mathbf{X}})/2$.

We model the probability of $N_1$ given $N$, $\tilde{H}$ and $\mathbf{X}$ assuming that $N_1$ follows a beta-binomial distribution indexed by parameters $(\beta_A,\psi)$ and denote the model by $P_{\texttt{ASReC}}(N_1|N,\tilde{H},\mathbf{X};\beta_A,\psi)$. A beta-binomial distribution extends a binomial distribution to allow over-dispersion. In particular, we assume that $N_1$ follows a beta-binomial distribution with mean $p$ and a dispersion parameter $\psi$:

$$P_{\texttt{ASReC}}(N_1|N,\tilde{H},\mathbf{X};\beta_A,\psi) = \binom{N}{N_1}\frac{\prod_{k=0}^{N_1-1}(p+k\psi)\prod_{k=0}^{N-N_1-1}(1-p+k\psi)}{\prod_{k=1}^{N-1}(1+k\psi)},$$

where

$$p=\begin{cases}0.5 & \text{if the candidate eQTL has a homozygous genotype AA or BB,}\\ q & \text{if } \tilde{H} \text{ indicates haplotype configuration B-}h_1 \text{ and A-}h_2\text{, respectively,}\\ 1-q & \text{if } \tilde{H} \text{ indicates haplotype configuration A-}h_1 \text{ and B-}h_2\text{, respectively.}\end{cases}$$

Thus $q$ characterizes the proportion of ASReC corresponding to the B allele among the total ASReC corresponding to the heterozygous genotype AB. We further express $q$ as $e^{\beta_A}/(1+e^{\beta_A})$. Note that the covariate effects are ignored here because they are expected to be the same on the two alleles of a gene within an individual. When the candidate eQTL *cis*-regulates the expression of the gene, we have $\beta_A = \beta_T$. To see this, we first define $\mu_A$ and $\mu_B$ as the mean ASReC corresponding to the A and B alleles, respectively, at the baseline of $\mathbf{X}$. Then, $\beta_A = \log[q/(1-q)] = \log(\mu_B/\mu_A)$. On the other hand, $\beta_T = \log(\mu_{BB,\mathbf{X}}/\mu_{AA,\mathbf{X}}) = \log\{(2\mu_B)/(2\mu_A)\}$, where the second equation follows from the additive genetic effect and from canceling out the individual-specific covariate effects. By contrast, when the candidate eQTL *trans*-regulates the gene expression, we have $\beta_T \neq 0$ but $\beta_A = 0$.

The likelihood based on the TReC and ASReC data of $n$ unrelated individuals takes the form

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} P_{\texttt{TReC}}(T_i|G_i,\mathbf{X}_i;\boldsymbol{\gamma},\beta_T,\phi)P_{\texttt{ASReC}}(N_{i1}|N_i,\tilde{H}_i,\mathbf{X}_i;\beta_A,\psi), \tag{8.1}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\gamma}, \beta_T, \phi, \beta_A, \psi)$. We refer to (8.1) as the **TReCASE** model, which is the novel model for *cis*-eQTL mapping using RNA-seq data. For *trans*-eQTL mapping, since ASE data are uninformative, the likelihood is only based on the TReC data: $L(\boldsymbol{\gamma}, \beta_T, \phi) = \prod_{i=1}^{n} P_{\texttt{TReC}}(T_i|G_i, \mathbf{X}_i; \boldsymbol{\gamma}, \beta_T, \phi)$. A hypothesis testing method has been developed to distinguish whether an eQTL is *cis*- or *trans*- by testing $H_0$: $\beta_T = \beta_A$ [55].

### 8.2.4 eQTL Mapping Using ASE with Unknown Haplotypes

When the haplotypes connecting the candidate eQTL and the gene of interest are unknown, we consider all possible haplotype pairs $(\tilde{h}_k, \tilde{h}_l)$ that are compatible with the known haplotypes in the gene body ($H$) and the genotype at the candidate eQTL ($G$). We denote these haplotype pairs as $(\tilde{h}_k, \tilde{h}_l) \sim (G, H)$. Then the likelihood function is a weighted summation of the probabilities, each corresponding to a possible haplotype pair and given by (8.1), i.e.,

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} P_{\texttt{TReC}}(T_i|G_i, \mathbf{X}_i; \boldsymbol{\gamma}, \beta_T, \phi) \times \sum_{(\tilde{h}_k, \tilde{h}_l) \sim (G_i, H_i)} P_{\texttt{ASReC}}(N_{i1}|N_i, \tilde{h}_k, \tilde{h}_l, \mathbf{X}_i; \beta_A, \psi) P(\tilde{h}_k, \tilde{h}_l; \boldsymbol{\pi}) f_{kl}(\mathbf{X}_i), \quad (8.2)$$

where $\boldsymbol{\Theta} = (\boldsymbol{\gamma}, \beta_T, \phi, \beta_A, \psi, \boldsymbol{\pi}, \{f_{kl}(.)\}_{k,l})$. We explain the terms that are not in (8.1) as follows.

Suppose there are $K$ possible haplotypes across the exonic SNPs and the candidate eQTL. Write the frequency of the $k$th haplotype by $\pi_k = \Pr(\tilde{h} = \tilde{h}_k)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. We denote the model for the probability of $\tilde{H} = (\tilde{h}_k, \tilde{h}_l)$ indexed by $\boldsymbol{\pi}$ by $P(\tilde{h}_k, \tilde{h}_l; \boldsymbol{\pi})$. Under the assumption of Hardy-Weinberg equilibrium, $P(\tilde{h}_k, \tilde{h}_l; \boldsymbol{\pi}) = \pi_k \pi_l$.

The density function of $\mathbf{X}$ given $\tilde{H} = (\tilde{h}_k, \tilde{h}_l)$ is denoted by $f_{kl}(\mathbf{X})$. Under the assumption of gene-environment independence, $f_{kl}(\mathbf{X})$ reduces to the marginal density function of $\mathbf{X}$ and will drop out from (8.2). In some applications, $\tilde{H}$ and $\mathbf{X}$ are correlated. One important example is when $\mathbf{X}$ represent the principal components for ancestry. Another example is when the gene influences both the environmental exposure (e.g., cigarette smoking) and the disease occurrence (e.g., lung cancer) [3]. In such cases, $f_{kl}(\mathbf{X})$ can be specified using a generalized odds-ratio function [28].

## 8.3 Isoform-Specific eQTL Mapping

More than 90 % of human multi-exon genes can be alternatively spliced, resulting in RNA isoforms [44, 64]. Alternative splicing may directly cause a disease or

modify certain disease susceptibility [19, 61, 63]. Although several methods have been proposed for detecting the event of alternative splicing and estimating the RNA-isoform abundance [2, 4, 21, 23, 31, 34, 38, 39, 50, 59, 65], only a few have been developed for testing the differential RNA-isoform usage between two groups of samples (e.g., cases vs. controls) [22, 54, 59]. Differential isoform usage refers to the changes of RNA-isoform expression relative to the total expression of the corresponding gene. The purpose of isoform-specific eQTL mapping is to dissect the genetic basis of the differential isoform usage. There are a few points worth mentioning from the statistical perspective on isoform-specific eQTL mapping.

- Because the isoform structure or abundance cannot be directly measured, transcriptome reconstruction and abundance estimation are necessary steps of isoform-specific eQTL mapping. The uncertainty of the transcriptome reconstruction and the abundance estimation should be incorporated into isoform-specific eQTL mapping.
- In most eQTL studies or genome-wide association studies, SNP genotype effects are assumed to be additive. Thus the SNP genotype is essentially a quantitative covariate. However, most existing methods assess the differential isoform usage between two groups of samples (e.g., cases vs. controls) and few methods can test the association between the isoform usage and a quantitative covariate.
- One gene may be differentially expressed with respect to a covariate, both in terms of the total expression and the isoform usage. It will be useful to jointly test for differential expression and differential isoform usage.

### *8.3.1 Transcriptome Reconstruction and Isoform Abundance Estimation*

A gene usually occupies a consecutive segment of the DNA sequence and it is often composed of several exons that are separated by introns. A subset of the exons may be employed by the cell to construct alternatively spliced messenger RNAs (mRNAs). These mRNAs may be translated to different proteins. Each RNA isoform is often referred to as a transcript and thus each gene can be considered as a transcript cluster. In some organism such as a human or a mouse, there are existing annotations on the kinds of transcripts a gene may encode. Such annotations are often incomplete or inaccurate, for example, some transcripts may be express in a particular tissue and/or developmental stage. In some other organisms, such as those without complete reference genomes, such transcriptome annotations are not available at all. Therefore, one may need to reconstruct the transcriptome from the observed RNA-seq data. This task can be achieved with or without a reference genome [18]. The reference genome-guided reconstruction is often more accurate and computationally more efficient than the de novo transcriptome construction without a reference genome. Thus the former approach is more popular for organisms that have reference genomes. Given the transcriptome annotation, the

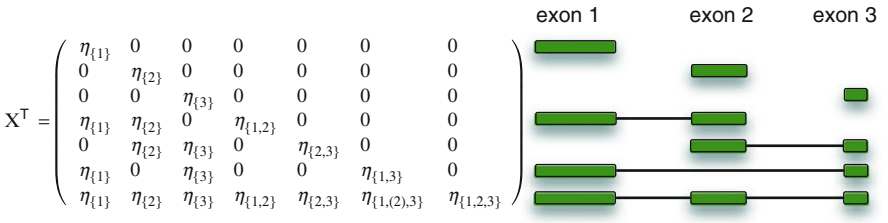



**Fig. 8.6** All possible isoforms of a gene with three exons and the corresponding design matrix $\mathbf{X}^T$

abundance of each transcript can be estimated by the number of RNA-seq reads aligned to that transcript. However, most RNA-seq fragments cannot be uniquely assigned to a specific transcript. To estimate transcript abundance in the presence of such alignment ambiguity is the focus of many existing works [31, 32, 37, 43, 48, 49, 53, 59, 72]. Penalized regression methods have been developed to simultaneously reconstruct transcriptome and estimate transcript/isoform abundance [6, 38, 39, 71]. The method we will describe next is an example of such penalized regression methods.

### *8.3.2 Isoform-Specific eQTL Mapping*

The method presented here is based on Sun et al. (2013) [58]. We first illustrate the statistical model by a hypothetical gene with three exons (Fig. 8.6). An RNA-seq read may overlap with one or more exons. Thus we count the number of RNA-seq reads per exon set. For this simple gene, there are seven possible exon sets, denoted by $\{1\}$, $\{2\}$, $\{3\}$, $\{1,2\}$, $\{2,3\}$, $\{1,3\}$, and $\{1,2,3\}$. Note that each RNA-seq read is only counted once. For example, if an RNA-seq read overlaps with both exon 1 and 2, it will be counted for exon set $\{1,2\}$ instead of exon set $\{1\}$ or $\{2\}$. There are seven possible isoforms (right panel of Fig. 8.6). We code each isoform as a covariate, which corresponds to one row of the design matrix $\mathbf{X}^T$ (left panel), where $^T$ denotes matrix transpose. The seven columns of matrix $\mathbf{X}^T$ correspond to exon sets $\{1\}$, $\{2\}$, $\{3\}$, $\{1,2\}$, $\{2,3\}$, $\{1,3\}$, and $\{1,2,3\}$. Each element in $\mathbf{X}^T$ is the effective length of the column-specific exon set within the row-specific isoform. Intuitively, the effective length of an exon set $A$, denoted by $\eta_A$, is the number of unique locations within $A$, where a randomly selected sequence fragment can be sampled. We defer the details of effective length calculation to the next section, but would like to point out that there are special exon sets that consist of non-contiguous exons in the specific isoform. For example, the exons in set $\{1,3\}$ is non-contiguous with respective to isoform 1-2-3 and the effective length of $\{1,3\}$ is denoted by $\eta_{\{1,(2),3\}}$. Our effective length calculation accurately reflects the fact that sequence reads of exon set $\{1,3\}$ are more likely from isoform 1-3 rather than isoform 1-2-3.

In this example, the gene expression in the $i$th sample is denoted by a vector: $\mathbf{y}_i = (y_{i\{1\}}, y_{i\{2\}}, y_{i\{3\}}, y_{i\{1,2\}}, y_{i\{2,3\}}, y_{i\{1,3\}}, y_{i\{1,2,3\}})^\mathsf{T}$, where $y_{iA}$ indicates the TReC at the exon set $A$. As in Sect. 8.2.3, we model the probability of a TReC via a negative binomial distribution. Let $f_{NB}(\mu, \phi)$ be a negative binomial distribution with mean $\mu$ and a dispersion parameter $\phi$. We assume that $y_{iA} \sim f_{NB}(\mu_{iA}, \phi)$. Assuming independence of $y_{iA}$'s given the underlying RNA isoforms, then $\mathbf{y}_i \sim f_{NB}(\boldsymbol{\mu}_i, \phi) \equiv \prod_A f_{NB}(\mu_{iA}, \phi)$ where $\boldsymbol{\mu}_i = (\mu_{i\{1\}}, \mu_{i\{2\}}, \ldots, \mu_{i\{1,2,3\}})^\mathsf{T}$. By the definition of the design matrix $\mathbf{X}$, we transform the problem of isoform deconvolution to a regression problem: $\mathbf{y}_i \sim f_{NB}(\boldsymbol{\mu}_i, \phi)$, $\boldsymbol{\mu}_i = T_i \mathbf{X}\boldsymbol{\gamma} = T_i \sum_{u=1}^{7} \mathbf{x}_u b_u$, where $T_i$ is TReC of this gene in sample $i$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_7)$, $\boldsymbol{\gamma} = (b_1, \ldots, b_7)^\mathsf{T}$, and $b_u \geq 0$ is the expression rate of the $u$th isoform. Note that $b_u$ quantifies the relative expression abundance with respect to the total expression $T_i$.

Next, we present the general method. Suppose that we study the isoform-specific expression of a gene with $m$ exon sets and $p$ possible isoforms across $n$ individuals, and we are particularly interested in whether a covariate $G$ has an influence on the isoform-specific expression of this gene. We assess this hypothesis by a likelihood ratio test. Under the null hypothesis, we solve the problems of isoform selection and abundance estimation by assuming that the isoform usage is the same for all samples. Thus we use a negative binomial regression with the link function $\boldsymbol{\mu}_i = T_i \mathbf{X}\boldsymbol{\gamma}$. Note that a linear link function instead of commonly used log link function is used to reflect the fact that the total number of reads is the summation of the number of reads from all the isoforms. Under the alternative, we model the effect of $G$ as follows. Let $g_i$ be the value of $G$ in the $i$th sample. Without loss of generality, we restrict the range of $g_i$ to be [0,1]. For example, if $G$ is genotype of a SNP, we set $g_i$ = 0, 1/2, and 1 for genotypes AA, AB, and BB, respectively. Provided $\boldsymbol{\mu}_i = T_i \mathbf{X}\boldsymbol{\gamma}$, we model the influence of $G$ on $b_u$ $(1 \leq u \leq p)$ by a linear model: $b_u = \gamma_u(1 - g_i) + \gamma_{u+p} g_i$, where $\gamma_j \geq 0$ for $1 \leq j \leq 2p$. Therefore, we have two negative binomial problems, with $p$ and $2p$ covariates, under null and alternative, respectively.

The major difficulty of this problem comes from the high dimensionality of the possible isoforms [25]. We address this difficulty by two sequential steps. First we identify the candidate isoforms for a gene using a modified connectivity graph approach [23, 38]. Next we select among the candidate isoforms using a penalized negative binomial regression problem. For example, under the alternative, the objective function becomes $f(\boldsymbol{\gamma}, \phi) = \sum_{i=1}^{n} \log[f_{NB}(\boldsymbol{\mu}_i, \phi)] - \sum_{j=1}^{2p} \lambda \log(\gamma_j + \tau)$, where $\lambda$ and $\tau$ are two tuning parameters that can be selected by BIC or extended BIC [57]. We use the log penalty $\lambda \log(\gamma_j + \tau)$ because of its superior theoretical and empirical advantages over other penalties [9, 15, 57]. Given $\lambda$ and $\tau$, the parameters $\boldsymbol{\gamma}$ and $\phi$ can be estimated by a coordinate descent algorithm [57]. The above model is formulated when the isoform usage is associated with one quantitative covariate; it is straightforward to extend it to include multiple quantitative covariates. For a categorical covariate (e.g., under the dominant or recessive effect of a SNP), we can simply code it as a number of dummy variables, which can be treated as multiple quantitative covariates.

Due to the variable selection (i.e., selecting expressed RNA isoforms) under both the null and the alternative hypotheses, the asymptotic distribution of the likelihood ratio statistic is unknown. Thus we estimate the null distribution of the statistic by parametric bootstrap. Specifically, we generate the $v$th bootstrap sample, denoted by $\tilde{\mathbf{y}}^{(v)}$ (a vector of length $nm$), by sampling from a negative binomial distribution with mean $\hat{\boldsymbol{\mu}}_0$ and a dispersion parameter $\hat{\phi}_0$, where $\hat{\boldsymbol{\mu}}_0$ (a vector of length $nm$) and $\hat{\phi}_0$ are estimated under the null. Then using this bootstrap sample, we apply the penalized regression approach under the null and the alternative to obtain a likelihood ratio statistic $LR_v$. Repeat the parametric bootstrap for a large number of times (e.g. 10,000 times) and pool the $LR_v$'s, we obtain the null distribution for the observed statistic $LR$. The final p-value is the proportion of $LR_v$'s that are equal to or larger than the likelihood ratio statistic from original data.

The above solution only tests differential isoform usage, which is the difference of relative abundance of an isoform with respect to the total expression of the gene for different values of $G$. If we are interested in testing both the differential expression and the differential isoform usage of a gene, the original link function $\boldsymbol{\mu}_i = T_i\mathbf{X}\boldsymbol{\gamma}$ can be changed to be $\boldsymbol{\mu}_i = R_i\mathbf{X}\boldsymbol{\gamma}$, where $R_i$ is the total number of RNA-seq reads of the $i$th sample across all genes. The reason is as follows. The original link function can be written as $\boldsymbol{\mu}_i = T_i\mathbf{X}\boldsymbol{\gamma} = R_i(T_i/R_i)\mathbf{X}\boldsymbol{\gamma}$, where $(T_i/R_i)$ measures the total expression of the gene in the $i$th sample. Then skipping the ratio $(T_i/R_i)$ in the original link function leads to the new link function, which is equivalent to assuming this gene has a constant expression rate across samples.

### 8.3.3 Calculation of Effective Length

An RNA-seq fragment is a segment of RNA to be sequenced. Usually only part of an RNA-seq fragment is sequenced: one end or both ends, hence single-end sequencing or paired-end sequencing. All the discussions in this section are for paired-end reads, though the extension to single-end reads is straightforward. The minimum fragment size is the read length, denoted by $d$. This happens when the two reads of a fragment completely overlap. We impose an upper bound for the fragment length based on prior knowledge of the experimental procedure and denote the upper bound by $l_M$. Then the fragment length $l$ satisfies $d \le l \le l_M$. We denote the distribution of the fragment length for sample $i$ by $\varphi_i(l)$, which can be calculated using observed read alignment information. The fragment length distribution is incorporated in our model to allow across-sample variations due to the differences in fragment length distribution.

For the $i$th sample, the effective length of exon $j$ of $r_j$ base pairs (bps) is

$$\eta_{i,\{j\}} = f(r_j, d, l_M, \varphi_i) = \begin{cases} 0 & \text{if } r_j < d \\ \sum_{l=d}^{\min(r_j, l_M)} \varphi_i(l)(r_j + 1 - l) & \text{if } r_j \ge d \end{cases}.$$

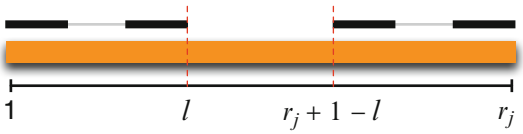



**Fig. 8.7** An illustration of effective length calculation for an exon of $r_j$ bps and RNA-seq fragment of $l$ bps. The *orange box* indicates the exon, and the *black lines* above the *orange box* indicate two RNA-seq fragments, while each RNA-seq fragment is sequenced by a paired-end read. There are $r_j + 1 - l$ distinct choices to select an RNA-seq fragment of $l$ bps from this exon, and thus the effective length is $r_j + 1 - l$

If $r_j < d$, the exon is shorter than the shortest fragment length, and thus the effective length of this exon is 0. In other words, no RNA-seq fragment is expected to overlap and only overlap with this exon. If $r_j \geq d$, the effective length is $r_j + 1 - l$, i.e., there are $r_j + 1 - l$ distinct RNA-seq fragments that can be sequenced from this exon (Fig. 8.7). Then $\sum_{l=d}^{\min(r_j, l_M)} \varphi_i(l)(r_j + 1 - l)$ is summation across all likely fragment lengths, weighted by the probability of having fragment length $l$.

In the following discussions, to simplify the notation, we skip the subscript of $i$. For two exons $j$ and $k$ ($j < k$) of lengths $r_j$ and $r_k$, which are adjacent in the transcript, the effective length for the fragments that cover both exons is

$$\eta_{\{j,k\}} = f(r_j + r_k, d, l_M, \varphi) - \eta_{\{j\}} - \eta_{\{k\}}. \tag{8.3}$$

For three exons $j$, $h$, and $k$ ($j < h < k$) of lengths $r_j$, $r_h$ and $r_k$, which are adjacent in the transcript, the effective length for the fragments that cover all three exons is

$$\eta_{\{j,h,k\}} = f(r_j + r_h + r_k, d, l_M) - \eta_{\{j,h\}} - \eta_{\{h,k\}} - \eta_{\{j,(h),k\}} - \eta_{\{j\}} - \eta_{\{h\}} - \eta_{\{k\}},$$

where $\eta_{\{j,(h),k\}}$ is the effective length in the scenario that the transcript covers consecutive exons $j$, $h$, and $k$, whereas the observed paired-end read only covers exons $j$ and $k$.

$$\eta_{\{j,(h),k\}} = \begin{cases} 0 & \text{if } (r_j, r_h, r_k) \in R_1 \\ \sum_{l=2d+r_h}^{\min(r_j+r_h+r_k, l_M)} \varphi(l)\delta_l & \text{otherwise} \end{cases}$$

where $R_1 = \{(r_j, r_h, r_k) : \ r_j < d \text{ or } r_k < d \text{ or } r_h + 2d > l_M\}$, and $\delta_l = \min(r_j, l - r_h - d) - \max(d, l - r_h - r_k) + 1$. The above formula is derived by the following arguments. Let $l_j$ and $l_k$ be the lengths of the parts of the fragment that overlaps with exon $j$ and $k$, respectively. Given $l$, the restriction of $l_j$ and $l_k$ are $l = l_j + l_k + r_h$, $d \leq l_j \leq r_j$, and $d \leq l_k \leq r_k$, and thus the range of $l_j$ is $\max(d, l - r_h - r_k) \leq l_j \leq \min(r_j, l - r_h - d)$. For more than three consecutive exons, the effective lengths can be calculated using recursive calls to the above equations.

In practice, a few sequence fragments may be observed even when the effective length is zero, which may be due to sequencing errors. To improve the robustness of our method, we modify the design matrix $\mathbf{X}$ by adding a pre-determined constant `eLenMin` to each element of $\mathbf{X}$.

## 8.4 Discussion

We conclude this chapter by a few discussion points.

### *8.4.1 eQTL Mapping Using Both ASE and ISE*

We have introduced statistical methods of using ASE or ISE for eQTL mapping. A natural extension is to use both ASE and ISE for eQTL mapping. The likelihood can be similar to the one for eQTL mapping using ASE, but using count data from exon sets intend of genes. Such a model can explain more subtle changes in the gene expression data. For example, one isoform is used in one allele, but not in the other allele, i.e., allele-specific isoform usage. A major challenge would be computational feasibility. Thus a more computationally efficient implementation is needed for such an effort.

### *8.4.2 cis-eQTL and Imprinting*

Allelic imbalance of gene expressions may be due to factors other than *cis*-eQTL. Arguably, the second most likely factor causing allelic imbalance, after *cis*-eQTL, is imprinting. Imprinted genes are differentially expressed on maternal and paternal alleles. Thus imprinting is also referred to as the parent-of-origin effect [47]. An important lesson we learned from our recent study of ASE in F1 mice [11] is that "imprinting is incomplete for most genes and *cis*-acting mutations can modify the strength of imprinting". Usually imprinting effect is much more subtle than *cis*-eQTL effects. Therefore, to obtain more sensitive and more accurate estimates of imprinting effects, it is crucial to jointly study imprinting and *cis*-eQTL.

### *8.4.3 Quality Control and Possible Non-genetic Factors*

Quality control (QC) is a necessary step for eQTL mapping using RNA-seq data. Low quality samples may be detected by checking the sequencing quality scores, mapping quality, percentage of uniquely mapped reads, percentage of reads

mapped to exonic regions, percentage of rRNA reads, and the distribution of insert size for paired-end reads [1, 13, 66]. Sample identity check is a very important step in genome-wide genomic studies. Between sample contamination may be detected by the percentage of heterozygous SNPs, sex-mismatch (recorded sex from demographic information vs. sex inferred from genomic data), or the D-statistic that measures the median correlation of gene expression between one sample versus each of the other samples [1, 69]. Sample swap will seriously reduce the power of eQTL analysis. Fortunately, checking for sample swap is relatively easy using RNA-seq data than using microarray data [29]. A QC step that is crucial for ASE data is the mapping bias toward reference alleles, which has been discussed at Sect. 8.2. For ISE data, checking the coverage of the whole gene body is important because there may be a trend of increasing read depth towards the 3' end of a gene. The method described in Sect. 8.3 assumes a uniform distribution of read depth, though the hypothesis testing method is not sensitive to this assumption due to the resampling nature of the test [58].

The effect of non-genetic factors can be accounted for by including them (or an appropriate transformation of them) as covariates in eQTL mapping. First, the overall read depth per sample is one factor that should always be included. In addition, GC content and dinucleotide frequencies may influence gene expression in a sample-specific manner. For example, gene expression and GC content may be positively correlated in some samples, but negatively correlated in other samples [74]. A conditional quantile normalization method has been proposed to model such sample-specific effects from sequence contents within the framework of generalized linear regression models [24]. This approach can be employed in the eQTL-mapping framework described in this book chapter.

### *8.4.4 The Genetic Architecture of Gene Expression*

Figure 8.8 shows the results of two genome-wide eQTL studies: a yeast study of ∼6,000 genes and ∼1,000 SNPs in 112 yeast segregants (offspring) (Fig. 8.8a) and a human study of ∼18,000 genes and ∼1,000,000 SNPs (germline genotype) in 550 breast cancer patients. Gene expression abundance was measured by microarrays in the yeast study and by RNA-seq in the human study. The difference in the genetic architecture of gene expression between the two studies is remarkable. In both studies, the eQTL plots have a diagonal pattern, which corresponds to a large number of local eQTLs. In the yeast study, there are several vertical bands, each corresponding to an eQTL hotspot, i.e., a genetic locus that is eQTL of many genes. In contrast, there is no such eQTL hotspot in the human study. The two studies are representative for experimental cross and human studies. In experimental cross, usually two strains with very different genetic backgrounds are crossed and thus some loci may have large and broad effects on many genes. For example, in the yeast study, several eQTL hotspots arise because one strain has several genes deleted. In human studies, the genetic differences across humans are much smaller than

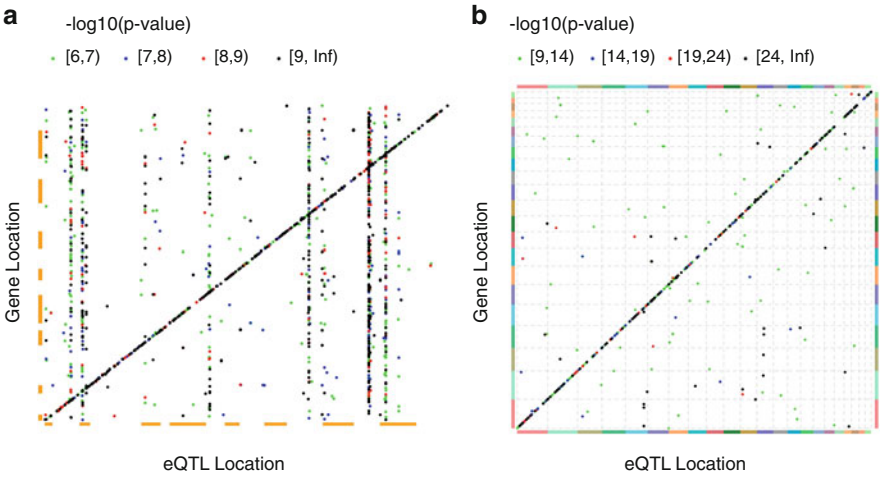



**Fig. 8.8** The results of eQTL studies in **(a)** 112 yeast sergeants of two yeast strains [7] and **(b)** 550 breast cancer patients of an on-going study. Each point represents a genome-wide significant association. The *color* indicates certain range of the p-value. More liberal p-values are used for the yeast study because there is a smaller number of genes and SNPs and hence less burden of multiple testing correction

in experimental crosses and generally no single locus can substantially alter the expression of many genes. We have reported similar findings in a recent human eQTL studies with 2,494 twins and a validation data set of 1,895 independent subjects [69]. The conclusion is that, for human studies, the vast majority of genetic effects on gene expression are through local eQTL and most of the local eQTL are likely to be *cis*-eQTL [55]. This implies that the identification of distant eQTLs may be as difficult as or even more difficult than genome-wide association studies for complex traits.

## References

[1] A C't Hoen, P., Friedländer, M.R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F., Buermans, H.P., Karlberg, O., Brännvall, M., et al.: Reproducibility of high-throughput mrna and small rna sequencing across laboratories. Nat. Biotechnol. **31**, 1015–1022 (2013)

[2] Ameur, A., Wetterbom, A., Feuk, L., Gyllensten, U.: Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol. **11**(3), R34 (2010)

[3] Amos, C.I., Wu, X., Broderick, P., Gorlov, I.P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., et al.: Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25. 1. Nature Genet. **40**(5), 616–622 (2008)

[4] Au, K., Jiang, H., Lin, L., Xing, Y., Wong, W.: Detection of splice junctions from paired-end RNA-seq data by splicemap. Nucleic Acids Res. **38**(14), 4570–4578 (2010)

[5] Auer, P.L., Doerge, R.: Statistical design and analysis of rna sequencing data. Genetics **185**(2), 405–416 (2010)

[6] Bohnert, R., Rätsch, G.: rquant. web: a tool for RNA-seq-based transcript quantitation. Nucleic Acids Res. **38**(Suppl 2), W348–W351 (2010)
[7] Brem, R.B., Storey, J.D., Whittle, J., Kruglyak, L.: Genetic interactions between polymorphisms that affect gene expression in yeast. Nature **436**(7051), 701–703 (2005)
[8] Browning, S., Browning, B.: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. **81**(5), 1084–1097 (2007)
[9] Chen, T., Sun, W., Fine, J.: Designing penalty functions in high dimensional problems: the role of tuning parameters. Technical Report, UNC Chapel Hill (2011)
[10] Cookson, W., Liang, L., Abecasis, G., Moffatt, M., Lathrop, M.: Mapping complex disease traits with global gene expression. Nat. Rev. Genet. **10**(3), 184–194 (2009)
[11] Crowley, J.J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I.K., Kim, Y., Wang, J.R., Morgan, A.P., Calaway, J.D., Aylor, D.L., Yun, Z., Bell, T.A., Buus, R.J., Calaway, M.E., Didion, J.P., Gooch, T.J., Hansen, S.D., Robinson, N.N., Shaw, G.D., Spence, J.S., Quackenbush, C.R., Barrick, C.J., Xie, Y., Valdar, W., Lenarcic, A.B., Wang, W., Welsh, C.E., Fu, C.P., Zhang, Z., Holt, J., Guo, Z., Threadgill, D.W., Tarantino, L.M., Miller, D., R., Zou, F., McMillan, L., Sullivan, P.F., Pardo-Manuel de Villena, F.: Pervasive allelic imbalance revealed by allele-specific gene expression in highly divergent mouse crosses. Nat. Genet. (2013, in revision)
[12] Delaneau, O., Zagury, J., Marchini, J., et al.: Improved whole-chromosome phasing for disease and population genetic studies. Nat. Meth. **10**(1), 5–6 (2013)
[13] DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.D., Williams, C., Reich, M., Winckler, W., Getz, G.: Rna-seqc: Rna-seq metrics for quality control and process optimization. Bioinformatics **28**(11), 1530–1532 (2012)
[14] Doss, S., Schadt, E., Drake, T., Lusis, A.: Cis-acting expression quantitative trait loci in mice. Genome Res. **15**(5), 681 (2005)
[15] Fan, J., Lv, J.: Non-concave penalized likelihood with np-dimensionality. IEEE Trans. Inf. Theory **57**(8), 5468–5484 (2011)
[16] Flicek, P., Amode, M., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al.: Ensembl 2011. Nucleic Acids Res. **39**(Suppl 1), D800 (2011)
[17] Fogarty, M., Xiao, R., Prokunina-Olsson, L., Scott, L., Mohlke, K.: Allelic expression imbalance at high-density lipoprotein cholesterol locus mmab-mvk. Hum. Mol. Genet. **19**(10), 1921–1929 (2010)
[18] Garber, M., Grabherr, M., Guttman, M., Trapnell, C.: Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Meth. **8**(6), 469–477 (2011)
[19] Garcia-Blanco, M., Baraniak, A., Lasda, E.: Alternative splicing in disease and therapy. Nat. Biotechnol. **22**(5), 535–546 (2004)
[20] Garnett, M., Edelman, E., Heidorn, S., Greenman, C., Dastur, A., Lau, K., Greninger, P., Thompson, I., Luo, X., Soares, J., et al.: Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature **483**(7391), 570–575 (2012)
[21] Grabherr, M., Haas, B., Yassour, M., Levin, J., Thompson, D., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al.: Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. **29**(7), 644–652 (2011)
[22] Griffith, M., Griffith, O., Mwenifumbo, J., Goya, R., Morrissy, A., Morin, R., Corbett, R., Tang, M., Hou, Y., Pugh, T., et al.: Alternative expression analysis by RNA sequencing. Nat. Meth. **7**(10), 843–847 (2010)
[23] Guttman, M., Garber, M., Levin, J., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M., Gnirke, A., Nusbaum, C., et al.: Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. **28**(5), 503–510 (2010)
[24] Hansen, K.D., Irizarry, R.A., Zhijin, W.: Removing technical variability in rna-seq data using conditional quantile normalization. Biostatistics **13**(2), 204–216 (2012)
[25] Hiller, D., Jiang, H., Xu, W., Wong, W.: Identifiability of isoform deconvolution from junction arrays and RNA-Seq. Bioinformatics **25**(23), 3056 (2009)

[26] Holt, J., Huang, S., McMillan, L., Wang, W.: Read annotation pipeline for high-throughput sequencing data. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, p. 605. ACM (2013)
[27] Hosokawa, Y., Arnold, A.: Mechanism of cyclin d1 (ccnd1, prad1) overexpression in human cancer cells: analysis of allele-specific expression. Genes Chrom. Cancer **22**(1), 66–71 (1998)
[28] Hu, Y., Lin, D., Zeng, D.: A general framework for studying genetic effects and gene–environment interactions with missing data. Biostatistics **11**(4), 583–598 (2010)
[29] Huang, J., Chen, J., Lathrop, M., Liang, L.: A tool for rna sequencing sample identity check. Bioinformatics **29**(11), 1463–1464 (2013)
[30] Huang, S., Kao, C.Y., McMillan, L., Wang, W.: Transforming genomes using mod files with applications. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, p. 595. ACM (2013)
[31] Jiang, H., Wong, W.: Statistical inferences for isoform expression in RNA-seq. Bioinformatics **25**(8), 1026 (2009)
[32] Katz, Y., Wang, E., Airoldi, E., Burge, C.: Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Meth. **7**(12), 1009–1015 (2010)
[33] Kendziorski, C., Wang, P.: A review of statistical methods for expression quantitative trait loci mapping. Mamm. Genome **17**(6), 509–517 (2006)
[34] Li, B., Dewey, C.: Rsem: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinform. **12**(1), 323 (2011)
[35] Li, Y., Grupe, A., Rowland, C., Nowotny, P., Kauwe, J., Smemo, S., Hinrichs, A., Tacey, K., Toombs, T., Kwok, S., et al.: Dapk1 variants are associated with Alzheimer's disease and allele-specific expression. Hum. Mol. Genet. **15**(17), 2560–2568 (2006)
[36] Li, Y., Willer, C., Ding, J., Scheet, P., Abecasis, G.: Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiol. **34**(8), 816–834 (2010)
[37] Li, B., Ruotti, V., Stewart, R., Thomson, J., Dewey, C.: RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics **26**(4), 493–500 (2010)
[38] Li, W., Feng, J., Jiang, T.: IsoLasso: a LASSO regression approach to RNA-seq based transcriptome assembly. J. Comput. Biol. **18**(11), 1693–1707 (2011)
[39] Li, J., Jiang, C., Hu, Y., Brown, B., Huang, H., Bickel, P.: Sparse linear modeling of RNA-seq data for isoform discovery and abundance estimation. Proceedings of the National Academy of Sciences **108**(50), 19867–19872 (2011)
[40] Meyer, K., Maia, A., O'Reilly, M., Teschendorff, A., Chin, S., Caldas, C., Ponder, B.: Allele-specific up-regulation of fgfr2 increases susceptibility to breast cancer. PLoS Biol. **6**(5), e108 (2008)
[41] Miller, V., Xia, H., Marrs, G., Gouvion, C., Lee, G., Davidson, B., Paulson, H.: Allele-specific silencing of dominant disease genes. Proc. Natl. Acad. Sci. U.S.A. **100**(12), 7195 (2003)
[42] Ozsolak, F., Milos, P.: RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. **12**(2), 87–98 (2010)
[43] Pachter, L.: Models for transcript quantification from RNA-seq. Arxiv preprint arXiv:1104.3889 (2011)
[44] Pan, Q., Shai, O., Lee, L., Frey, B., Blencowe, B.: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. **40**(12), 1413–1415 (2008)
[45] Perou, C., Sørlie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., et al.: Molecular portraits of human breast tumours. Nature **406**(6797), 747–752 (2000)
[46] Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J., Stephens, M., Gilad, Y., Pritchard, J.: Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature **464**(7289), 768–772 (2010)
[47] Reik, W., Walter, J., et al.: Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. **2**(1), 21–32 (2001)

[48] Richard, H., Schulz, M., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., et al.: Prediction of alternative isoforms from exon expression levels in RNA-seq experiments. Nucleic Acids Res. **38**(10), e112 (2010)
[49] Roberts, A., Trapnell, C., Donaghey, J., Rinn, J., Pachter, L., et al.: Improving RNA-seq expression estimates by correcting for fragment bias. Genome Biol. **12**(3), R22 (2011)
[50] Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S., Mungall, K., Lee, S., Okada, H., Qian, J., et al.: De novo assembly and analysis of RNA-seq data. Nat. Meth. **7**(11), 909–912 (2010)
[51] Rockman, M.V., Kruglyak, L.: Genetics of global gene expression. Nat. Rev. Genet. **7**(11), 862–872 (2006)
[52] Ronald, J., Brem, R., Whittle, J., Kruglyak, L.: Local regulatory variation in Saccharomyces cerevisiae. PLoS Genet. **1**(2), e25 (2005)
[53] Salzman, J., Jiang, H., Wong, W.: Statistical modeling of RNA-seq data. Stat. Sci. **26**(1), 62–83 (2011)
[54] Singh, D., Orellana, C., Hu, Y., Jones, C., Liu, Y., Chiang, D., Liu, J., Prins, J.: Fdm: a graph-based statistical method to detect differential transcription using RNA-seq data. Bioinformatics **27**(19), 2633–2640 (2011)
[55] Sun, W.: A statistical framework for eQTL mapping using RNA-seq data. Biometrics **68**(1), 1–11 (2012)
[56] Sun, W., Hu, Y.: eQTL mapping using RNA-seq data. Stat. Biosci. **5**(1), 198–219 (2013)
[57] Sun, W., Ibrahim, J., Zou, F.: Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. Genetics **185**(1), 349 (2010)
[58] Sun, W., Liu, Y., Crowley, J.J., Chen, T.H., Zhou, H., Chu, H., Huang, S., Kuan, P.F., Li, Y., Miller, D., Shaw, G., Wu, Y., Zhabotynsky, V., McMillan, L., Zou, F., Sullivan, P.F., Pardo-Manuel de Villena, F.: IsoDOT detects differential RNA-isoform usage with respect to a categorical or continuous covariate with high sensitivity and specificity. arXiv preprint arXiv:1402.0136 (2014)
[59] Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**(5), 511–515 (2010)
[60] Valle, L., Serena-Acedo, T., Liyanarachchi, S., Hampel, H., Comeras, I., Li, Z., Zeng, Q., Zhang, H., Pennison, M., Sadim, M., et al.: Germline allele-specific expression of tgfbr1 confers an increased risk of colorectal cancer. Science **321**(5894), 1361 (2008)
[61] Venables, J.: Aberrant and alternative splicing in cancer. Cancer Res. **64**(21), 7647 (2004)
[62] Wahls, W.P., Davidson, M.K.: Dna sequence-mediated, evolutionarily rapid redistribution of meiotic recombination hotspots commentary on genetics 182: 459–469 and genetics 187: 385–396. Genetics **189**(3), 685–694 (2011)
[63] Wang, G., Cooper, T.: Splicing in disease: disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. **8**(10), 749–761 (2007)
[64] Wang, E., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S., Schroth, G., Burge, C.: Alternative isoform regulation in human tissue transcriptomes. Nature **456**(7221), 470–476 (2008)
[65] Wang, K., Singh, D., Zeng, Z., Coleman, S., Huang, Y., Savich, G., He, X., Mieczkowski, P., Grimm, S., Perou, C., et al.: Mapsplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. **38**(18), e178 (2010)
[66] Wang, L., Wang, S., Li, W.: Rseqc: quality control of rna-seq experiments. Bioinformatics **28**(16), 2184–2185 (2012)
[67] Wittkopp, P., Haerum, B., Clark, A.: Evolutionary changes in cis and trans gene regulation. Nature **430**(6995), 85–88 (2004)
[68] Wright, F.A., Shabalin, A.A., Rusyn, I.: Computational tools for discovery and interpretation of expression quantitative trait loci. Pharmacogenomics **13**(3), 343–352 (2012)

[69] Wright, F., Sullivan, P., Brooks, A., Zou, F., Sun, W., Xia, K., Madar, V., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T., W., C., et al.: Heritability and genomics of gene expression in peripheral blood. Nature Genet. **46**(5), 430–437 (2014)
[70] Wu, T.D., Nacu, S.: Fast and snp-tolerant detection of complex variants and splicing in short reads. Bioinformatics **26**(7), 873–881 (2010)
[71] Xia, Z., Wen, J., Chang, C., Zhou, X.: Nsmap: A method for spliced isoforms identification and quantification from RNA-seq. BMC Bioinform. **12**(1), 162 (2011)
[72] Xing, Y., Yu, T., Wu, Y., Roy, M., Kim, J., Lee, C.: An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. Nucleic Acids Res. **34**(10), 3150 (2006)
[73] Zhao, Q., Kirkness, E., Caballero, O., Galante, P., Parmigiani, R., Edsall, L., Kuan, S., Ye, Z., Levy, S., Vasconcelos, A., et al.: Systematic detection of putative tumor suppressor genes through the combined use of exome and transcriptome sequencing. Genome Biol. **11**(11), R114 (2010)
[74] Zheng, W., Chung, L.M., Zhao, H.: Bias detection and correction in rna-sequencing data. BMC Bioinform. **12**(1), 290 (2011)