Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

# Statistical Inference of Allelic Imbalance from Transcriptome Data

Michael Nothnagel,[1]* Andreas Wolf,[1] Alexander Herrmann,[2] Karol Szafranski,[3] Inga Vater,[4] Mario Brosch,[2] Klaus Huse,[3] Reiner Siebert,[4] Matthias Platzer,[3] Jochen Hampe,[2] and Michael Krawczak[1]

[1]*Institute of Medical Informatics and Statistics, Christian-Albrechts University, Kiel, Germany;* [2]*Department of Internal Medicine I, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany;* [3]*Leibniz Institute for Age Research, Fritz Lipmann Institute, Jena, Germany;* [4]*Institute of Human Genetics, Christian-Albrechts University and University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany*

Communicated by Ulf Landegren

Received 16 July 2010; accepted revised manuscript 19 October 2010.
Published online 30 November 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/humu.21396

**ABSTRACT:** Next-generation sequencing and the availability of high-density genotyping arrays have facilitated an analysis of somatic and meiotic mutations at unprecedented level, but drawing sensible conclusions about the functional relevance of the detected variants still remains a formidable challenge. In this context, the study of allelic imbalance in intermediate RNA phenotypes may prove a useful means to elucidate the likely effects of DNA variants of unknown significance. We developed a statistical framework for the assessment of allelic imbalance in next-generation transcriptome sequencing (RNA-seq) data that requires neither an expression reference nor the underlying nuclear genotype(s), and that allows for allele miscalls. Using extensive simulation as well as publicly available whole-transcriptome data from European-descent individuals in HapMap, we explored the power of our approach in terms of both genotype inference and allelic imbalance assessment under a wide range of practically relevant scenarios. In so doing, we verified a superior performance of our methodology, particularly at low sequencing coverage, compared to the more simplistic approach of completely ignoring allele miscalls. Because the proposed framework can be used to assess somatic mutations and allelic imbalance in one and the same set of RNA-seq data, it will be particularly useful for the analysis of somatic genetic variation in cancer studies.
Hum Mutat 32:98–106, 2011. © 2010 Wiley-Liss, Inc.

**KEY WORDS:** sequencing; gene expression; transcription; maximum likelihood; HapMap

## Introduction

Next-generation sequencing and the availability of high-density single-nucleotide polymorphism (SNP) genotyping arrays have facilitated an assessment of human genetic variation, and of its association with clinical or intermediate phenotypes, at unprecedented level. However, drawing sensible conclusions about the functional relevance of the large number of somatic and meiotic variants that are currently uncovered still remains a formidable challenge. Available technologies simply do not allow functional annotation to proceed at a similar pace as genotyping, thereby rendering the biological interpretation of the observed genotype–phenotype relationships difficult. One of the auxiliary approaches to this problem is the analysis of intermediate RNA phenotypes, such as allelic imbalance [Yan et al., 2002], which is defined as a substantial departure of autosomal transcription activity from parity. A multitude of mechanisms may underlie allelic imbalance, including mutations/polymorphisms in *cis*-acting regulatory elements that affect transcription [Ge et al., 2009], splicing [Caux-Moncoutier et al., 2009], RNA stability, nonsense-mediated decay, and methylation status [Wang et al., 2008]. Random mono-allelic expression, where occasionally only one or the other of the two alleles present in a given cell is transcribed, has also been found to be common for autosomal genes [Gimelbrant et al., 2007; Pastinen et al., 2004; Pollard et al., 2008; Wang et al., 2007]. Allelic imbalance analysis has been featured successfully in studies of alternative splicing [Caux-Moncoutier et al., 2009], micro-RNA directed gene repression [Kim and Bartel, 2009] and cancer transcriptome specificity [Nakanishi et al., 2009], and it bears the potential to advance greatly the functional interpretation of variants of unknown significance [Caux-Moncoutier et al., 2009; Domchek and Weber, 2008]. Technically, most approaches to allelic imbalance analysis in the past were based upon targeted chemistries, such as pyrosequencing [Kim and Bartel, 2009] and SNaPshot assays [Caux-Moncoutier et al., 2009]. At the same time, genotyping microarrays have been adapted for use in RNA-based protocols as well, moving from low-resolution panels [Nakanishi et al., 2009; Pant et al., 2006] to increasingly denser marker sets [Ge et al., 2009; Gimelbrant et al., 2007; Liu et al., 2010]. These developments have now facilitated an assessment of allelic imbalance at the genome-wide level.

Despite the great attention paid to allelic imbalance in scientific practice, a formal framework for its statistical analysis, particularly in genome-wide settings, is only beginning to emerge. A straightforward approach in this direction has been the pin-pointing of SNPs with clearly disparate genotypes [Coenen et al., 2008; Nakanishi et al., 2009] or copy numbers [Lamy et al., 2007] in different samples, such as cancer versus normal tissue or blood versus urine. A similar all-or-nothing approach [Gimelbrant et al., 2007] was based upon contrasting heterozygous nuclear genotypes with homozygous transcriptome-derived genotypes. Other proposals involved the definition of “normal” (i.e., balanced) expression and required expression level ratios of mutant versus

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Michael Nothnagel, Institute of Medical Informatics and Statistics, Christian-Albrechts University, Brunswiker Strasse 10, 24105 Kiel, Germany. E-mail: nothnagel@medinfo.uni-kiel.de wild-type allele to fall into a two standard deviation range [Caux-Moncoutier et al., 2009] or 99% confidence interval [Palacios et al., 2009] around the mean taken in a reference population, or into a fixed ($\pm$7%) range around parity [Loeuillet et al., 2007]. More sophisticated methods involved one-sided Binomial tests [Degner et al., 2009; Kim and Bartel, 2009], *t*-tests [Yamamoto et al., 2007], regression-based tests [Chen et al., 2008], chi-square tests [Heap et al., 2010], or segment analysis to identify closely linked SNPs with similar allele-specific expression levels [Staaf et al., 2008]. In addition, the joint use of genome-wide expression and nuclear genotype data has been proposed as a means to augment haplotype-based eQTL identification [Montgomery et al., 2010; Pickrell et al., 2010], but these methods cannot be applied directly to study allelic imbalance at single variants.

Here, we present a novel statistical framework for the assessment of allelic imbalance using RNA-seq data. Our method is applicable both with and without knowledge of the underlying nuclear genotype(s), does not require an expression reference, and allows for sequencing errors. We evaluated this framework using extensive simulation as well as RNA-seq data from HapMap individuals in order to explore its power and limits for a wide range of practically relevant scenarios.

## Materials and Methods

### Statistical Inference of Allelic Imbalance from Transcriptome Data

#### *Likelihood ratio test for allelic imbalance when the nuclear genotype is known*

For every heterozygous single-nucleotide substitution (henceforth simply referred to as “substitution”) in a given individual, there are two *correct* types of call from aligned transcriptome sequencing reads, namely, those of the two constituent alleles A1≠A2. Let $f_{A1}$ and $f_{A2=}\,1-f_{A1}$ denote the *actual* frequencies of A1 and A2 among all transcripts, respectively. Erroneous calls yield one of the two remaining nucleotides M1 ≠ M2. Let $c_X$ denote the number of reads for which nucleotide X has been called and let $c = c_{A1}+c_{A2}+c_{M1}+c_{M2}$ be the overall number of reads.

If there are no erroneous calls, then the number of calls of either nuclear allele of a given substitution should follow a binomial distribution, $\beta(c,\ 1/2)$, under the null hypothesis of balanced transcription, $f_{A1\,=}f_{A2\,=}\,1/2$. In this case, the optimal statistical test for allelic imbalance would be a binomial test. In practice, however, sequencing data will almost always contain calling errors, which we accommodate in our framework by introducing two types of error parameter. First, let $\pi$ denote the probability that a call from a given read is wrong, and we assume that this probability is independent of the correct call. Second, we introduce conditional probabilities $\pi_{X,Y}$ of calling nucleotide X, given that a miscall has occurred and that the correct call would be Y. *Valid* calls for a heterozygous SNP occur either when A1 or A2 are called correctly, or when one allele is erroneously called from the other. *Invalid* calls are always miscalls. This implies that the log-likelihood of $f = f_{A1}$, given numbers of calls $\{c_X\}$, equals

$$\begin{aligned}\ln\ \mathrm{L(f)} = \mathrm{const} &+ c_{A1}\cdot\ln(f\cdot(1-\pi)+(1-f)\cdot\pi\cdot\pi_{A1,A2})\\ &+ c_{A2}\cdot\ln((1-f)^{*}(1-\pi)+f^{*}\pi^{*}\pi_{A2,A1})\\ &+ c_{M1}\cdot\ln(\pi\cdot[f\cdot\pi_{M1,A1}+(1-f)\cdot\pi_{M1,A2}])\\ &+ c_{M2}\cdot\ln(\pi\cdot[f\cdot\pi_{M2},_{A1}+(1-f)\cdot\pi_{M2,A2}])+ \qquad (1)\end{aligned}$$

For the time being, we assume that proper estimates of the miscalling probabilities are available. If the nuclear genotype underlying the transcriptome data is known, then the identity of A1 and A2 is also known, and a straightforward likelihood ratio test for allelic imbalance is given by

$$G(c_{A1,}\ c_{A2,}\ c_{M1,}\ c_{A2,}) = -2\ \ln\frac{\mathrm{L}(\frac{1}{2})}{\max\limits_{f}\mathrm{L}(f)}. \qquad (2)$$

Under $H_0$, statistic $G$ follows a chi-square distribution with 1 degree of freedom. Maximization of L($f$) in (2) can be done numerically using, for example, the optimize function of the stats package of the R statistical software [R Development Core Team, 2010] (http://www.r-project.org/).

#### *Inference of the nuclear genotype from transcriptome data*

Transcriptome sequencing can be used to infer allelic imbalance even if the underlying nuclear genotype is unknown. Such an endeavor may seem paradox at first glance because, by definition, extreme allelic imbalance cannot be distinguished from homozygosity using transcriptome data alone. However, with less than extreme allelic imbalance, an underlying heterozygous genotype may still be sufficiently more probable than other genotypes, assuming that there is no allelic imbalance. Therefore, it appeared worthwhile exploring the power of a two-stage approach whereby the most likely nuclear genotype is first inferred from the transcriptome sequence data using Bayes’ theorem under the assumption of balanced transcription, followed by a test for allelic imbalance in the case of a sufficiently well supported heterozygous genotype (posterior probability $>0.5$).

Because the true nuclear genotype is assumed to be unknown here, all possible genotypes are assigned an equal prior probability, which implies that their posterior probability gets proportional to the corresponding likelihood for the transcriptome data. For a presumed heterozygous genotype A1A2, this likelihood is calculated from formula (1) using miscalling probabilities that were estimated from empirical data as described below. For a homozygous nuclear genotype A1A1, there is only one correct allele (namely, A1), whereas the other three nucleotides (M1, M2, and M3) represent miscalls. In this case, the log-likelihood equals

$$\begin{aligned}&\mathrm{const}+c_{A1}\cdot\ln(1-\pi)+c_{M1}\cdot\ln(\pi\cdot\pi_{M1,A1})+c_{M2}\cdot\ln(\pi\cdot\pi_{M2,A1})\\ &\quad+c_{M3}\cdot\ln(\pi\cdot\pi_{M3,A1})\end{aligned}$$

and the posterior probability of A1A1 gets proportional to the antilog of this expression.

### Estimation of Miscalling Probabilities

#### *Cell lines, genotyping, and sequencing*

Transcriptome sequence data were generated for the publicly available lymphoblastoid cell lines GM10847, GM12760, GM12864, GM12870, and GM12871 (Coriell Institute for Medical Research, Camden, NJ) as follows. Cells were cultivated as suggested by the supplier. DNA was extracted from harvested cells using the Qiagen DNA extraction kit (Cat.No. 13343, Qiagen, Hilden, Germany). Genotyping of single-nucleotide polymorphisms (SNPs) using Affymetrix Chip 6.0 and the Illumina Omni chip (San Diego, CA) was performed according to the manufacturers’ protocols. The genomic positions of observed nucleotide substitutions were retrieved from the UCSC database (http://genome.ucsc.edu/; NCBI Build 36.1; genome freeze hg18) [Kent et al., 2002; Pruitt et al., 2005]. RNA was extracted from $1 \times 10^8$ cells using the RNEasy kit (Qiagen). The mRNA-Seq libraries for Illumina/Solexa GAII sequencing were prepared according to the manufacturer's instructions, starting with 5 μg RNA. Libraries were sequenced on two lanes per sample, generating between 27 and 33 million 76-nt reads for each cell line. An overview of the generated sequence data is provided in Supp. Table S1.

### *Alignment and sequence analysis*

FASTQ-formatted Illumina/Solexa sequences were mapped onto the human genome (hg18) [Rhead et al., 2010] and a collection of splice junctions using the novoalign software package v2.05.13 (http://www.novocraft.com/). The mapping procedure was run in SE mode (parameters -R 30 -r All -e 5) and included adapter trimming (option -a). A comprehensive list of known exons was compiled from UCSC *knowngenes*. Only splice junctions from adjacent exons located within a gene region were considered. Each of the generated splice junctions was 140 bp in length, containing the last 70 bases of the upstream exon and the first 70 bp of the downstream exon. Use of 70 bp of flanking sequence ensured that all junction-mapped reads matched a minimum of 5 bp on either side of the junction. Only reads that mapped to a unique position in the genome were considered for further analysis (between 66% and 75% of reads per cell line). Positions with low sequencing quality (phred quality $<10$) were excluded.

Only those SNPs were used for the estimation of miscalling probabilities for which the nuclear genotype (Affymetrix 6.0 or Illumina Omni) and transcriptome sequence data with at least 20-fold coverage were available (see below). SNPs known to be located in the same region as imprinted genes were excluded. The sequencing data from all five cell lines (Supp. Table S1) were pooled, providing a total of 66,943 SNPs for analysis. Although some of these data represented multiple (up to five) counts of one and the same SNP, pooling was deemed reasonable since miscalling probability estimates obtained from individual cell lines were not found to be significantly different (data not shown).

## Maximum-Likelihood Estimation of Miscalling Probabilities

The test for allelic imbalance defined in formulae (1) and (2) requires knowledge of the unconditional and conditional miscalling probabilities $\pi$ and $\{\pi_{X,Y}\}$, respectively. If nuclear genotypes for a sufficiently large number of SNPs are available to complement RNA-seq data, then these error parameters can be estimated empirically. Thus, any additional alleles called for a homozygous genotype, say YY, must be due to an error so that the relative proportions of the different miscalls $X \neq Y$ provide reasonable estimates of $\pi_{X,Y}$. Obviously, this approach assumes the absence of errors in the assignment of the nuclear genotype and, in fact, includes these errors into $\pi$.

An assessment of allelic imbalance is only meaningful for heterozygotes, and because the miscalling probability $\pi$ may vary depending upon whether the allele under study is present in homozygous or heterozygous state, only heterozygous nuclear genotypes were used for the maximum-likelihood estimation of $\pi$ as well. *Lege artis* estimation of $\pi$ would have involved maximization of the global log-likelihood function according to (1), taking into account SNP-specific values of $f$ and a single genome-wide value of $\pi$. In this case, however, the global log-likelihood would have depended upon $f$ values for thousands of SNPs, thereby rendering maximization with respect to $\pi$ computationally intractable. We therefore performed SNP-wise maximization of (1) with respect to $f$ and $\pi$, and used the average of the SNP-specific estimates of $\pi$ in subsequent analyses. Maximization of (1) was again carried out numerically, using the optim function of the R stats package with the "L-BFGS-B" option set in order to restrict the possible parameter space to [0,1] [R Development Core Team, 2010].

## Framework Evaluation by Simulation

We carried out simulations to evaluate the power of the statistical framework described above for correctly inferring nuclear genotypes and for detecting allelic imbalance. Simulations and statistical analyses were implemented in R v2.10.1 [R Development Core Team, 2010] and Perl.

### *Genotype simulation*

An idealized genome comprising 500,000 nucleotide triplets was generated picking triplets randomly from their frequency distribution in the human genome (http://www.kazusa.or.jp/codon/). For 125,000 of these triplet (25%), the central triplet position was deemed to carry a single nucleotide substitution, and the alternative triplet center allele was chosen at random in accordance with known nearest-neighbor-dependent mutation rates in the human genome [Krawczak et al., 1998]. All substitution-carrying triplets were henceforth assumed to be heterozygous in the simulated genome, whereas the remaining 375,000 triplets were deemed to be homozygous.

### *Transcript simulation*

Various sets of sequencing data were generated assuming coverage levels of 5, 10, 20, 50, and 100 reads, respectively, for each triplet. For heterozygous genotypes, each coverage level was combined with one of six levels of allelic imbalance, defined by a minor transcript frequency of 0.5, 0.4, 0.3, 0.2, 0.1, or 0.05. Errors were added to the reads by randomly changing the call at the central triplet position in accordance with empirical estimates of the miscalling probabilities $\pi$ and $\pi_{X,Y}$. To emulate noisy RNA-seq data, we also performed simulations adopting a miscalling probability of $\pi = 0.05$.

### *Genotype inference and assessment of allelic imbalance*

We evaluated our statistical framework separately in terms of its genotype inference and allelic imbalance detection capabilities. This means that only simulated data from correctly inferred heterozygotes were tested for allelic imbalance. Joint power estimates for both steps were taken as equal to the product of the two individual power estimates. Real transcriptome sequencing data usually comprise variants with different read coverage and different allelic imbalance. Simulations assuming a single coverage level may therefore lead to unrealistic power estimates. To avoid such a mistake, we first pooled all simulated data sets with the same minor transcript frequency, but different coverage level, and randomly split the two pools of homozygotes and heterozygotes into 10 equally sized subsets each. Then, following a Latin-Square-like crossvalidation design, five subsets of homozygous and five subsets of heterozygous genotypes were used for parameter estimation in the "training stage" of each validation round. For the remaining subsets comprising the "validation stage," the underlying genotype was inferred as described above, using the miscalling probability estimates obtained in the training stage. The power of correct genotype inference was then estimated by the proportion of correctly inferred genotypes, separately for homo- and heterozygotes. Next, allelic imbalance was tested statistically only for those heterozygous genotypes that were inferred correctly in the validation stage. The power of allelic imbalance detection was estimated as the proportion of *P*-values below either 0.05 or a threshold that was Bonferroni-adjusted to the number of tested genotypes. The validation procedure comprised 10 rounds so that each subset was used five times for training and five times for validation. Substitution-specific *P*-values from the allelic imbalance tests were averaged over those substitutions for which the heterozygous nuclear genotype was inferred correctly in all five validation pools.

### *Test for allelic imbalance ignoring allele miscalls*

To assess the relative benefit of taking allele miscalls explicitly into account in our statistical framework, we also subjected both the simulated and the HapMap RNA-seq data to a simple chi-square test for equal frequencies of the genotype-constituting alleles, as proposed by Heap et al. [2010].

### *Descriptive statistical analysis*

The R software v2.10.1 [R Development Core Team, 2010] was used for statistical analysis and for creating graphs. Receiver operator characteristic (ROC) curves were created using 100 equidistant *P*-values per combination of coverage level and allelic imbalance ratio. The corresponding area under the curve (AUC) was calculated by means of linear interpolation.

## Application to Real Data

We evaluated our approach using the publicly available RNA-seq data from 60 HapMap individuals of European descent (CEPH Utah residents with ancestry from northern and western Europe; CEU) [Montgomery et al., 2010]. Alignment files (all_sam_data.tar) and derived genotypes (RNASEQ60_snps.full.txt.gz) were downloaded from a dedicated Web site (http://jungle.unige.ch/rnaseq_CEU60/). We only considered reads mapping to chromosomes 1 to 22, according to the UCSC database (http://genome.ucsc.edu/; NCBI Build 36.1; genome freeze hg18) [Kent et al., 2002; Pruitt et al., 2005]. For each genotype, the sequence read information was extracted from the alignment file using SAMtools [Li et al., 2009]. Information on the site-specific reads in all individuals was then merged into a single data set. Where information on one and the same SNP was available for more than one individual, the respective genotypes were considered independent observations. For comparison, we again applied a chi-square test for equal allele frequencies in addition to our proposed likelihood ratio test.

# Results

## Estimates of Miscalling Probabilities

From the analysis of the cell line transcriptome data for SNPs with known nuclear heterozygous genotype, we estimated $\pi$ to be $2.17 \times 10^{-3}$. The conditional miscalling probabilities $\pi_{X,Y}$ as estimated from homozygous SNPs were found to be substantially skewed (Table 1). Thus, although G was found to be preferentially ($>50\%$) miscalled as A, and C as T, both A and T were miscalled as either G or C with nearly equal probability (~45%). The remaining base (i.e., T or A) was approximately four times less likely to represent the respective miscall.

**Table 1. Estimates of the Conditional Miscalling Probabilities $\pi_{X,Y}$ Obtained from the Pooled Homozygous SNPs of Five Human Cell Lines (see Supp. Table S1)**

| | | Allele Call (X) | | | |
|---|---|---|---|---|---|
| | | A | C | G | T |
| Nuclear allele (Y) | A | – | 0.40210 | 0.48006 | 0.11784 |
| | C | 0.22214 | – | 0.25456 | 0.52330 |
| | G | 0.53012 | 0.26919 | – | 0.20070 |
| | T | 0.13354 | 0.44217 | 0.42429 | – |

**Table 2. Percentage of Simulated Substitutions ($n = 125{,}000$) with Correctly Inferred Heterozygous Genotype**

| | Sequencing coverage | | | | |
|---|---|---|---|---|---|
| Allelic imbalance | 5 × | 10 × | 20 × | 50 × | 100 × |
| 50:50 | 93.5 | 98.9 | 100.0 | 100.0 | 100.0 |
| 60:40 | 90.9 | 97.9 | 99.8 | 100.0 | 100.0 |
| 70:30 | 82.8 | 93.0 | 98.3 | 100.0 | 100.0 |
| 80:20 | 67.2 | 79.7 | 88.3 | 97.3 | 99.6 |
| 90:10 | 40.9 | 51.7 | 51.4 | 52.6 | 52.8 |
| 95:5 | 22.8 | 29.5 | 20.1 | 9.9 | 3.6 |

## Genotype Inference

In many RNA-seq studies, the nuclear genotypes of the investigated substitutions will be unknown, and the available transcript data will have to be used to infer them. Our simulations revealed that a heterozygous genotype can often be inferred reliably at low to moderate levels of allelic imbalance (up to 80:20), even at a coverage as low as 20 reads (Table 2). At a coverage of five reads, heterozygous genotypes were still identified correctly in $>80\%$ of cases if the allelic imbalance was less than 70:30. At 50-fold or higher coverage, the proportion of correctly identified heterozygous genotypes was found to exceed 97%. On the other hand, extreme allelic imbalance (95:5) could scarcely be distinguished from homozygosity, particularly at high coverage. Strong allelic imbalance (90:10) appeared to represent a change point at which a heterozygous genotype was inferred correctly in 50 to 60% of the replications, except for extremely low coverage (i.e., 5 reads). Homozygous genotypes were inferred correctly in almost all cases (from 98.9% at 10-fold or lower coverage to 100% at 50-fold or higher coverage). For virtually all simulated genotypes ($>99.999\%$), the maximum posterior probability exceeded 50%, thereby leading to the inference of a (correct or incorrect) genotype with sufficient confidence.

## Assessment of Allelic Imbalance

For correctly inferred (or known) heterozygous genotypes, the proposed likelihood ratio test was capable of discriminating well between the presence and absence of allelic imbalance (Fig. 1 and Table 3). Thus, the AUC at balanced transcription (50:50) was approximately 0.50 as expected, with some random variation

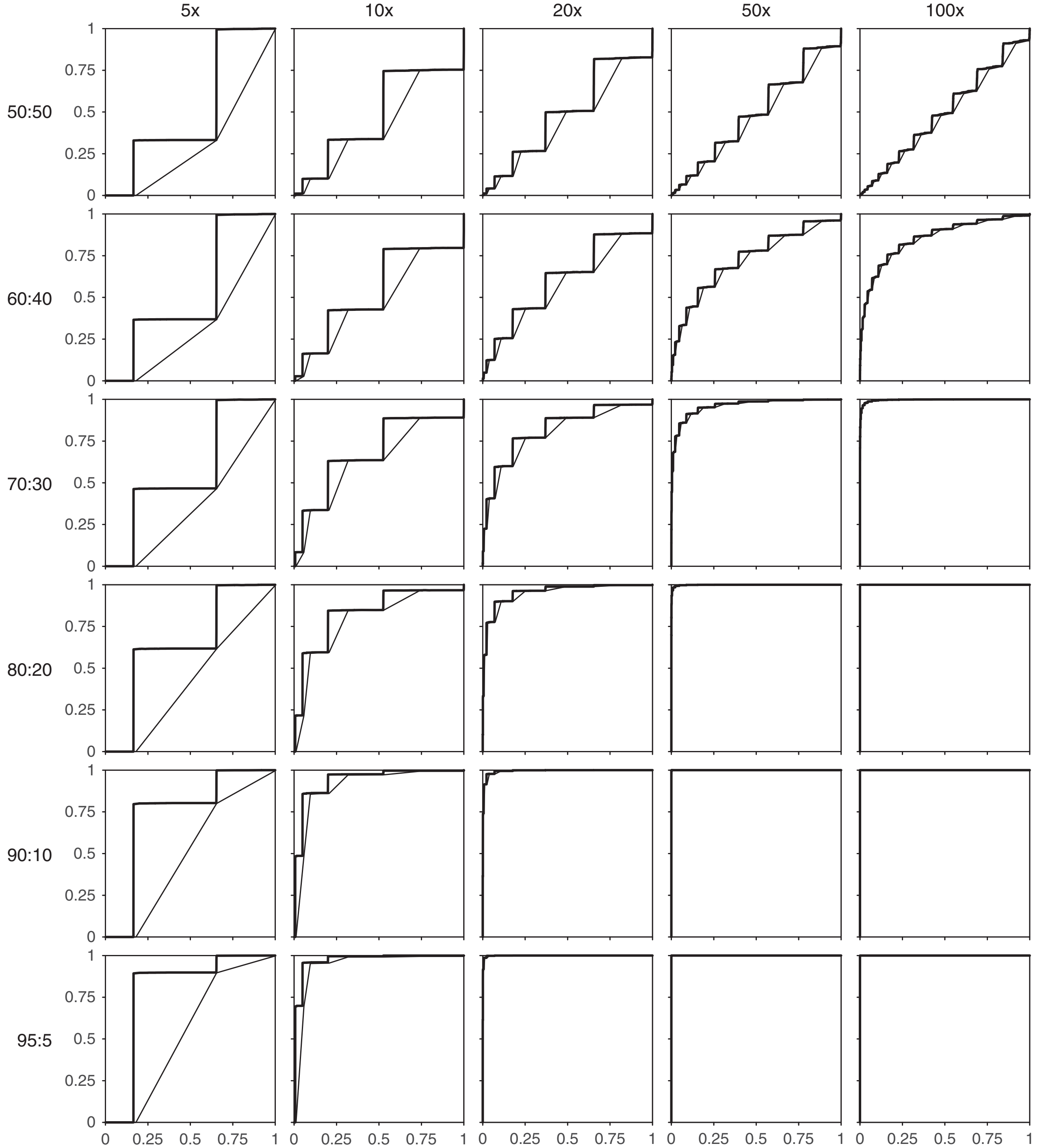

**Figure 1.** Receiver operator characteristic (ROC) curves for the detection of allelic imbalance using either a likelihood ratio test (thick line) or a chi-square test (thin line). Columns correspond to different levels of sequencing coverage, lines correspond to different levels of allelic imbalance (125,000 simulated substitutions; only substitutions with a correctly inferred heterozygous genotype were considered). For details, see text.

observed at low coverage (5- to 50-fold). Slight allelic imbalance (60:40) required at least 100-fold coverage to yield satisfactory AUC values ($>0.85$). The AUC exceeded 0.95 for moderate imbalance (70:30) at 50-fold coverage, and for strong imbalance (80:20) already at 20-fold coverage. For stronger imbalance or higher coverage, the AUC approached 1.0. The AUC values of the chi-square test were usually smaller than those of the likelihood ratio test, partially because the limited number of reads per SNP allowed only a small number of different $p$ values to be obtained by the former.

At a nominal significance level of 0.05, the statistical test for allelic imbalance defined in formulae (1) and (2) showed high power ($>97\%$) to detect strong imbalance (90:10) at 20-fold coverage, and still had $>85\%$ power at 10-fold coverage (Table 4). Higher coverage was required to detect more moderate imbalance with reasonable power. Coverage of only five reads resulted in a

**Table 3. Area Under Curve (AUC) for the Inference of Allelic Imbalance Using Either a Likelihood-Ratio Test or, in Parentheses, a Chi-square Test**

| | Sequencing coverage | | | | |
|---|---|---|---|---|---|
| Allelic imbalance | 5 × | 10 × | 20 × | 50 × | 100 × |
| 50:50 | 0.508 (0.308) | 0.481 (0.419) | 0.494 (0.446) | 0.497 (0.464) | 0.500 (0.476) |
| 60:40 | 0.526 (0.323) | 0.541 (0.481) | 0.607 (0.564) | 0.732 (0.711) | 0.856 (0.847) |
| 70:30 | 0.573 (0.363) | 0.683 (0.627) | 0.824 (0.796) | 0.962 (0.957) | 0.996 (0.996) |
| 80:20 | 0.647 (0.424) | 0.833 (0.789) | 0.957 (0.947) | 0.999 (0.998) | 1.000 (1.000) |
| 90:10 | 0.738 (0.500) | 0.941 (0.904) | 0.995 (0.992) | 1.000 (1.000) | 1.000 (1.000) |
| 95:5 | 0.785 (0.540) | 0.973 (0.939) | 0.999 (0.998) | 1.000 (1.000) | 1.000 (1.000) |

The analysis was based upon variable subsets of the original 125,000 simulated substitutions; each subset comprised only those substitutions for which the heterozygous genotype was inferred correctly.

**Table 4. Power of Two Statistical Tests to Detect Allelic Imbalance at the 5% Significance Level and, in Parentheses, after Bonferroni Correction for the Number of Substitutions Analyzed**

| | Sequencing Coverage | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 5 × | | 10 × | | 20 × | | 50 × | | 100 × | |
| Allelic imbalance | LRT | Chi-square test | LRT | Chi-square test | LRT | Chi-square test | LRT | Chi-square test | LRT | Chi-square test |
| 50:50 | 0.000 (0.000) | 0.000 (0.000) | 0.099 (0.000) | 0.011 (0.000) | 0.042 (0.000) | 0.042 (0.000) | 0.063 (0.000) | 0.063 (0.000) | 0.055 (0.000) | 0.055 (0.000) |
| 60:40 | 0.000 (0.000) | 0.000 (0.000) | 0.162 (0.000) | 0.027 (0.000) | 0.125 (0.000) | 0.125 (0.000) | 0.329 (0.000) | 0.329 (0.000) | 0.538 (0.001) | 0.537 (0.001) |
| 70:30 | 0.000 (0.000) | 0.000 (0.000) | 0.333 (0.000) | 0.084 (0.000) | 0.407 (0.000) | 0.407 (0.000) | 0.856 (0.007) | 0.856 (0.007) | 0.987 (0.156) | 0.987 (0.113) |
| 80:20 | 0.000 (0.000) | 0.000 (0.000) | 0.590 (0.000) | 0.216 (0.000) | 0.777 (0.000) | 0.777 (0.000) | 0.997 (0.167) | 0.997 (0.160) | 1.000 (0.907) | 1.000 (0.866) |
| 90:10 | 0.000 (0.000) | 0.000 (0.000) | 0.859 (0.000) | 0.485 (0.000) | 0.977 (0.000) | 0.977 (0.000) | 1.000 (0.882) | 1.000 (0.760) | 1.000 (1.000) | 1.000 (1.000) |
| 95:5 | 0.000 (0.000) | 0.000 (0.000) | 0.957 (0.000) | 0.697 (0.000) | 0.998 (0.000) | 0.998 (0.000) | 1.000 (0.998) | 1.000 (0.992) | 1.000 (1.000) | 1.000 (1.000) |

Estimates were based upon 125,000 simulated substitutions; only substitutions with a correctly inferred heterozygous genotype were considered. LRT, likelihood ratio test.

nearly complete lack of power to detect any allelic imbalance. With extremely high coverage (500-fold; data not shown), the test would achieve >99% power to detect even weak imbalance (60:40). The likelihood ratio test showed only minor inflation of the type I error in the simulations, with a maximum of 0.099 observed at 10-fold coverage (see "50:50" line in Table 4).

The chi-square test showed similar power as the likelihood ratio test and also the same inflation of the type I error at 20-fold or higher coverage (Table 4). For 10-fold coverage, however, the power of the chi-square test was substantially lower than that of the likelihood ratio test. Moreover, with a notably increased miscalling probability of 5%, where both tests were found to perform poorly (Supp. Table S2), the likelihood ratio test still retained at least modest power at low coverage (10 reads), whereas the chi-square test failed to provide any power.

Simultaneously investigating hundreds of thousands of SNPs for allelic imbalance may represent a serious multiple-testing problem. Therefore, we also quantified the power of the two tests using Bonferroni correction for the number of correctly inferred genotypes (Table 3). As was to be expected, the power dropped dramatically upon Bonferroni correction, in particular at 20-fold or lower coverage for which the power to detect allelic imbalance approached zero (Table 4). At least 50-fold coverage was required to detect strong imbalance (90:10) with >85% power, whereas 100-fold coverage was required for moderate imbalance (80:20). Again, the power of the $\chi^2$ test was found to be substantially lower than that of the likelihood ratio test for many combinations of coverage and allelic imbalance (Table 4). Moreover, an increased allele miscalling probability of 5% required at least 20-fold coverage for the chi-square test to provide any power to detect imbalance whereas the likelihood ratio test yielded satisfactory power already at 10-fold coverage (Supp. Table S2).

## Assessment of Allelic Imbalance in Real Transcriptome Sequence Data

We applied both the likelihood ratio test and the chi-square test to real autosome-wide RNA-seq data [Montgomery et al., 2010] for which auxiliary nuclear genotype information was available in HapMap. Given the lack of power of both tests at low coverage, we restricted our analysis to SNPs with at least five reads. We also limited the maximum coverage to 100 reads per SNP because of the rarity of SNPs with even higher coverage. When counting each SNP multiple times according to the number of individuals analysed, a total of 434,509 SNPs met the above criteria. Of these, a total of 139,535 (32.1%) were found to be heterozygous. Our genotype inference framework correctly inferred 82.2% of the heterozygous SNPs, whereas the remainder were deemed homozygous. Both the correctly and the incorrectly inferred heterozygous SNPs were subsequently tested for allelic imbalance.

We found the coverage per SNP in the analyzed RNA-seq data to be highly skewed, following an almost exponential-like distribution (Fig. 2). Nearly half (49.3%) of the heterozygous SNPs had at most 10 reads and 75.5% had at most 20 reads. Thus, the vast majority of SNPs were characterised by comparatively low coverage.

Because the true level of allelic imbalance was unknown for the HapMap RNA-seq data, their analysis cannot serve as a gold standard for assessing the power of an allelic imbalance test. This notwithstanding, the likelihood ratio test classified a substantially higher proportion of SNPs with auxiliary genotype information as showing allelic imbalance than the chi-square test, in particular, for small coverage of up to 20 reads (Fig. 3, continuous line). Upon closer inspection, this excess was also found to be due partially to the limited number of different

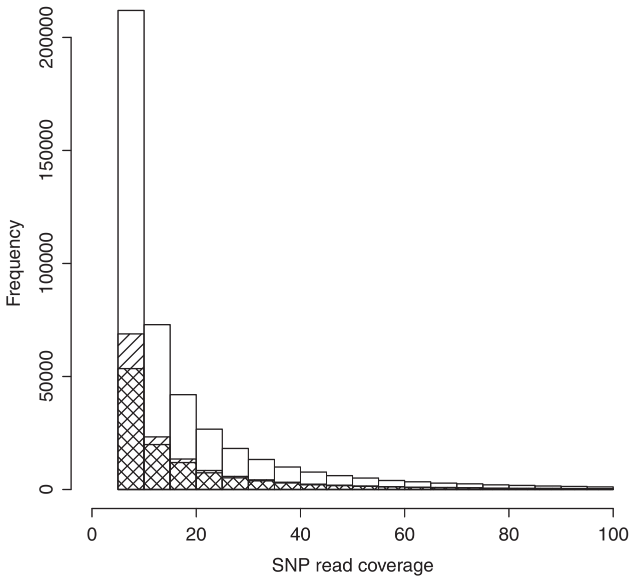

**Figure 2.** Distribution of the coverage level per SNP in the RNA-seq data from HapMap [Montgomery et al., 2010]. The range of coverage levels was restricted to 5 to 100 reads. Hatched bars: heterozygous SNPs according to HapMap [Montgomery et al., 2010]; crosshatched bars: heterozygous SNPs that were inferred correctly by the proposed genotype inference approach. For details, see text.

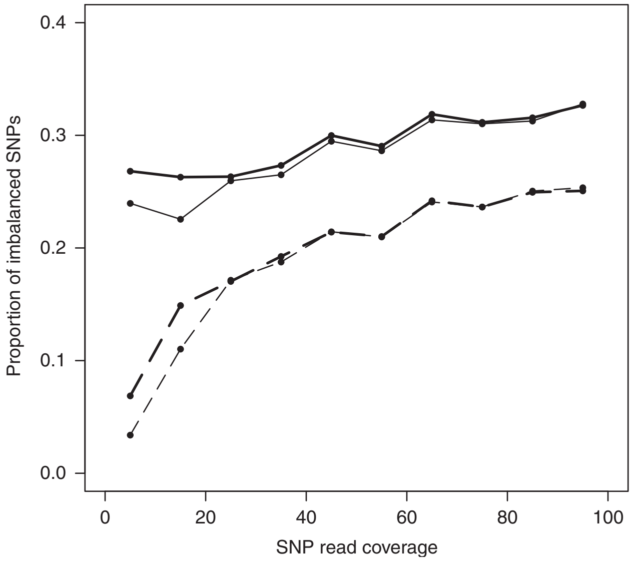

**Figure 3.** Proportion of heterozygous SNPs in the RNA-seq data from HapMap [Montgomery et al., 2010] that were found to show significant allelic imbalance by the likelihood ratio test (thick line) or the chi-square test (thin line). The analysis was confined to SNPs with fivefold to 10-fold coverage. Dashed lines refer to the subset of SNPs for which the genotype was inferred correctly by the proposed genotype inference approach. SNPs were grouped according to coverage into bins of size 10.

*p*-values that are technically possible with a chi-square test. If auxiliary genotype information is not available and the nuclear genotype has to be inferred from the allelic transcripts instead, both tests have decreased power to detect imbalance (Fig. 3, dashed line), also reinforcing the notion that extreme allelic imbalance is indistinguishable from homozygosity. However, for up to 20-fold coverage, the likelihood ratio test would again confer a substantial power gain compared to the chi-square test.

## Discussion

Transcriptome sequencing (RNA-seq) allows the detection of splice variants and the assessment of allelic imbalance in a single experiment. We proposed a coherent statistical framework for the analysis of RNA-seq data that does not only address extreme (near all-or-nothing) allelic imbalance, but that also provides sufficient power to detect moderate and even weak allelic imbalance if sufficient sequencing coverage is provided. Most importantly, our method allows for calling errors, thereby being more realistic than other approaches about the idiosyncrasies of real sequencing data, including technology- and species-specific error signatures. Moreover, by explicitly taking the different probabilities of different allele miscalls into account, the proposed likelihood ratio test for allelic imbalance may provide substantially more power than a simplistic chi-square test that ignores such information. Our framework can easily be adapted to different sequencing platforms, genome compositions and other factors that might impact upon the output of RNA-seq experiments. Furthermore, in the course of defining and evaluating our proposed likelihood ratio test, we derived empirical estimates of the conditional miscalling probabilities of such experiments that corresponded well to earlier observations, and that likely reflect the signature of DNA polymerase infidelities [Dohm et al., 2008].

The statistical framework presented herein can also be applied to infer the unknown nuclear genotype underlying error-prone RNA-seq data as long as allelic imbalance is at most moderate (up to a ratio of 80:20). This means that it would still be possible to analyze allelic imbalance even if prior information about nuclear genotypes is lacking or difficult to obtain. The resolution of allelic imbalance achievable in such instances roughly coincides with that resulting from a restriction of the analysis to variants with a minor allelic transcript frequency $>15\%$ [Heap et al., 2010]. Potential applications of RNA-based nuclear genotype inference include the profiling of somatic cells, particularly in cancer, which often show an accumulation of somatic mutations [Pleasance et al., 2010]. It should be noted, however, that extreme allelic imbalance is inherently indistinguishable from homozygosity if nuclear genotype information is missing. Moreover, RNA editing may hamper the inference of nuclear genotypes from RNA-seq data even further. Another point of concern may be the fact that, because the prior probability of a given genotype depends upon many factors including the species and population under study, the genomic region of interest, etc., we chose to assign equal prior probabilities to all possible genotypes in our simulations and subsequent analyses. If the true underlying distribution is substantially skewed in a given research context, this could, of course, be accounted for in more specific evaluations outside the scope of our manuscript. Furthermore, at least as regards power considerations, the effects of a certain genotype being rare and therefore inferred incorrectly more often than others owing to equal priors are likely to average out. Finally, we are, of course, aware that alternative algorithms have been proposed to infer genotypes from sequence reads using, for example, probabilistic graphical models and a binomial distribution [Goya et al., 2010]. However, because our main focus was the assessment of allelic imbalance, not genotype inference, we refrained from comparing our framework with these other methods.

Our simulations clearly demonstrated that, once the underlying nuclear genotype is known, the proposed likelihood ratio test has high power to detect strong allelic imbalance even at moderate coverage, and is still well powered to detect moderate imbalance if higher coverage is provided. The expected level of allelic imbalance should therefore determine the envisaged sequencing coverage of RNA-seq experiments. If Bonferroni correction for multiple testing is deemed necessary, for example, in genome-wide experiments, then at least 100-fold coverage seems mandatory for detecting allelic imbalance weaker than 90:10. Particularly at low coverage or when Bonferroni correction is applied, the proposed likelihood ratio test provides substantially more power to detect allelic imbalance than a chi-square test that ignores miscalling altogether. This power gain became even more evident with "noisy" RNA-seq data that would result from an increased allele miscalling probability. The observed drop in power of both tests to detect allelic imbalance when the nuclear genotype has to be inferred first is not surprising because only variants that were balanced enough to be called heterozygous were tested in the validation stage. Finally, although our simulations revealed a minor inflation of the type I error for both tests, especially at high coverage, we wish to emphasize that this lack of conservativeness should be of minor practical relevance. Testing for allelic imbalance will most often serve the purpose of (biological) hypothesis generation so that a higher type II error rate, as has been noted for the chi-square test at low coverage, would be of much greater concern.

If both genotype inference and allelic imbalance assessment are to be performed on the same data set, our method will work best for moderate to strong imbalance (70:30 to 80:20) with at least moderate coverage (50-fold or more). However, many research and clinical applications of transcriptome sequencing will be limited in coverage and will rarely exceed 20-fold. In view of the coverage distribution seen in the whole-transcriptome data from HapMap individuals analyzed here [Montgomery et al., 2010], and given the high costs still arising from high-throughput sequencing, low coverage is therefore likely to be the rule rather than the exception for the majority of variants investigated in RNA-seq studies in the foreseeable future. Although such studies will generally have only little power to detect moderate imbalance (up to 70:30), it is exactly this type of data for which our statistical approach provides reasonable power to detect higher levels of allelic imbalance.

In summary, we presented a likelihood-based statistical framework that takes allele miscalls into account and that allows the joint detection of somatic variants and imbalanced allelic expression. In providing a means, for example, to rank somatic mutations according to their likely functional relevance, we therefore expect our approach to be especially useful in RNA-seq studies of human cancer.

## Acknowledgments

We thank Ivonne Görlich, Claudia Becher, and Reina Zühlke for expert technical assistance. This work was supported by the German National Genome Research Network (NGFNplus; Grant 01GS08182) through the Colon Cancer Network (CCN), by the German Research Foundation (DFG) Cluster of Excellence "Inflammation at Interfaces" and by the German Federal Ministry of Education and Research (BMBF) through the Services@MediGRID Project. None of the funding organizations had any influence on the design, conduct, or conclusions of the study. All R scripts and data files used in the present study are available from the authors upon request.

## References

Caux-Moncoutier V, Pages-Berhouet S, Michaux D, Asselain B, Castera L, De Pauw A, Buecher B, Gauthier-Villars M, Stoppa-Lyonnet D, Houdayer C. 2009. Impact of BRCA1 and BRCA2 variants on splicing: clues from an allelic imbalance study. Eur J Hum Genet 17:1471–1480.

Chen X, Weaver J, Bove BA, Vanderveer LA, Weil SC, Miron A, Daly MB, Godwin AK. 2008. Allelic imbalance in BRCA1 and BRCA2 gene expression is associated with an increased breast cancer risk. Hum Mol Genet 17:1336–1348.

Coenen MJ, Ploeg M, Schijvenaars MM, Cornel EB, Karthaus HF, Scheffer H, Witjes JA, Franke B, Kiemeney LA. 2008. Allelic imbalance analysis using a single-nucleotide polymorphism microarray for the detection of bladder cancer recurrence. Clin Cancer Res 14:8198–8204.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25:3207–3212.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36:e105.

Domchek S, Weber BL. 2008. Genetic variants of uncertain significance: flies in the ointment. J Clin Oncol 26:16–17.

Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KC, Gagne V, Dias J, Hoberman R, Montpetit A, Joly MM, Harvey EJ, Sinnett D, Beaulieu P, Hamon R, Graziani A, Dewar K, Harmsen E, Majewski J, Goring HH, Naumova AK, Blanchette M, Gunderson KL, Pastinen T. 2009. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat Genet 41:1216–1222.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. Science 318:1136–1140.

Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP. 2010. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26:730–736.

Heap GA, Yang JH, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, Albert TJ, Rodesch MJ, Clayton DG, Todd JA, van Heel DA, Plagnol V. 2010. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Hum Mol Genet 19:122–134.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12:996–1006.

Kim J, Bartel DP. 2009. Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression. Nat Biotechnol 27:472–477.

Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63:474–488.

Lamy P, Andersen CL, Dyrskjot L, Torring N, Wiuf C. 2007. A Hidden Markov Model to estimate population mixture and allelic copy-numbers in cancers using Affymetrix SNP arrays. BMC Bioinformatics 8:434.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Liu Z, Li A, Schulz V, Chen M, Tuck D. 2010. MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. PLoS One 5:e10909.

Loeuillet C, Weale M, Deutsch S, Rotger M, Soranzo N, Wyniger J, Lettre G, Dupre Y, Thuillard D, Beckmann JS, Antonarakis SE, Goldstein DB, Telenti A. 2007. Promoter polymorphisms and allelic imbalance in ABCB1 expression. Pharmacogenet Genomics 17:951–959.

Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET. 2010. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464:773–777.

Nakanishi H, Matsumoto S, Iwakawa R, Kohno T, Suzuki K, Tsuta K, Matsuno Y, Noguchi M, Shimizu E, Yokota J. 2009. Whole genome comparison of allelic imbalance between noninvasive and invasive small-sized lung adenocarcinomas. Cancer Res 69:1615–1623.

Palacios R, Gazave E, Goni J, Piedrafita G, Fernando O, Navarro A, Villoslada P. 2009. Allele-specific gene expression is widespread across the genome and biological processes. PLoS One 4:e4150.

Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. 2006. Analysis of allelic differential expression in human white blood cells. Genome Res 16:331–339.

Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brandstrom H, Beck A, Verner A, Kingsley J, Harmsen E, Labuda D, Morgan K, Vohl MC, Naumova AK, Sinnett D, Hudson TJ. 2004. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics 16:184–193.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464:768–772.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, Mudie LJ, Ning Z, Royce T, Schulz-Trieglaff OB, Spiridou A, Stebbings LA, Szajkowski L, Teague J, Williamson D, Chin L, Ross MT, Campbell PJ, Bentley DR, Futreal PA, Stratton MR. 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191–196.

Pollard KS, Serre D, Wang X, Tao H, Grundberg E, Hudson TJ, Clark AG, Frazer K. 2008. A genome-wide approach to identifying novel-imprinted genes. Hum Genet 122:625–634.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ. 2010. The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38:D613–D619.

Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Goransson H, Juliusson G, Rosenquist R, Hoglund M, Borg A, Ringner M. 2008. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol 9:R136.

Wang J, Valo Z, Smith D, Singer-Sam J. 2007. Monoallelic expression of multiple genes in the CNS. PLoS One 2:e1293.

Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, Clark AG. 2008. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. PLoS One 3:e3839.

Yamamoto G, Nannya Y, Kato M, Sanada M, Levine RL, Kawamata N, Hangaishi A, Kurokawa M, Chiba S, Gilliland DG, Koeffler HP, Ogawa S. 2007. Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. Am J Hum Genet 81:114–126.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. 2002. Allelic variation in human gene expression. Science 297:1143.