Quantifying RNA allelic ratios by microfluidic multiplex PCR and sequencing

Rui Zhang1, Xin Li2, Gokul Ramaswami1, Kevin S Smith2, Gustavo Turecki3, Stephen B Montgomery1,2 & Jin Billy Li1

We developed a targeted RNA sequencing method that couples microfluidics-based multiplex PCR and deep sequencing (mmPCR-seq) to uniformly and simultaneously amplify up to 960 loci in 48 samples independently of their gene expression levels and to accurately and cost-effectively measure allelic ratios even for low-quantity or low-quality RNA samples. We applied mmPCR-seq to RNA editing and allele-specific expression studies. mmPCR-seq complements RNA-seq for studying allelic variations in the transcriptome.

RNA allelic ratios, including those generated by RNA editing and allele-specific expression (ASE), are quantitative traits. Adenosine-to-inosine (A-to-I) editing, the most common type of RNA editing in metazoans1, is tightly controlled2,3. The editing level of specific sites is critical, as aberrant editing has been linked with various diseases4. ASE is a phenomenon in which two alleles of a gene within an individual exhibit unequal expression. It is largely considered to reflect the effects of functional cis-acting genomic sequence variants5. The ability to accurately measure RNA allelic ratios is critical to the study of RNA editing and ASE and the functional role of genetic regulatory variation.

RNA-seq has been used to quantify RNA editing (editotyping) and ASE (allelotyping)6–9. However, RNA-seq is intrinsically limited by the large dynamic range of RNA expression, which leads to inaccurate quantification of allelic ratios for genes with low-to-moderate expression levels. This limitation cannot be overcome by conventional targeted resequencing technologies (ASE, for which levels are mostly around 50%). To capture the failure rate in conventional single-plex primer designs17. Of the 220 successful amplicons, 201 (~91%) were covered with reads within a 16-fold difference (24, from 210 to 214, Fig. 1b). Notably, mmPCR-seq amplicon coverage was independent of gene expression levels, in contrast to RNA-seq (Fig. 1c).

Received 7 June; accepted 18 October; published Online 24 November 2013; DOI:10.1038/NMETH.2736

1Department of Genetics, Stanford University, Stanford, California, USA. 2Department of Pathology, Stanford University, Stanford, California, USA. 3McGill Group for Suicide Studies, Douglas Mental Health University Institute, McGill University, Montreal, Quebec, Canada. Correspondence should be addressed to J.B.L. (jin.billy.li@stanford.edu) or S.B.M. (smontgom@stanford.edu).
BRIEF COMMUNICATIONS

We reasoned that the accuracy of allelic-ratio quantification using mmPCR-seq may depend on the amount of cDNA input because of sampling bottlenecks for sites located in low-expression genes (Supplementary Note 2). To assess this, we performed technical replicates using different amounts (100, 200, 500 and 1,000 ng) of cDNA. As expected, the reproducibility of measurements increased with more input template (Fig. 1d and Supplementary Figs. 3 and 4). We estimate that 1,000 ng of input cDNA corresponds to RNA from ~200 cells used in each one of the 48 PCR reactions, thus demonstrating accurate measurement even for genes with few copies per cell (Supplementary Note 2).

To confirm that our reproducible measurements were also accurate, we compared our results with editing levels measured by RNA-seq (a combination of 300 million reads of the HBRR...
subsamplings of mmPCR-seq data. The enrichment score was calculated as the number of detected A-to-G or T-to-C mismatches per 10 kb in RNA samples divided by the counterpart in DNA samples. The average value of two samples is shown. Sites with ≥1,000 reads were used. (b) Pairwise comparison of the editing level of each novel site and the nearest known site. (c) Cumulative distribution of RNA-editing levels for different groups of sites. For each site, the highest editing level among eight samples is shown. (d) Nucleotide composition in positions immediately upstream and downstream of the edited sites. The control is all A sites that are covered by mmPCR-seq reads and not edited in any samples tested.

We further examined seven postmortem brain Brodmann area 44 samples. In total, we identified 914 novel sites, most of which are edited at extremely low levels (Fig. 2c). Nevertheless, 109 novel sites were edited at levels ≥10%, which suggests that mmPCR-seq can also complement RNA-seq by identifying additional moderately and highly edited sites. Of all the novel sites, 136 are nonrepetitive nonsynonymous events; as such, the repertoire of recording targets is greatly expanded (Supplementary Data 4). As expected18, increased editing levels are associated with under- and overrepresentation of guanosines immediately 5′ and 3′ of the edited adenosine (Fig. 2d and Supplementary Fig. 7) as well as the higher TAG and AAG triplet fraction (Supplementary Fig. 8). In contrast to our observation that a large number of sites were near moderately and highly edited events, there was a deficit of lowly edited sites at loci that lack known RNA-editing sites. This is consistent with the coupling hypothesis19, whereas it does not fully support the continuous probing hypothesis20 (Supplementary Note 4).

We next reasoned that mmPCR-seq would provide high-resolution ASE measurements. We examined 960 single-nucleotide polymorphisms (SNPs) (Online Methods) in lymphoblastoid cell lines (LCLs) from 16 individuals within a three-generation family

**Figure 2** | Characterization of novel RNA-editing sites identified by mmPCR-seq. (a) Relationship between the enrichment score and the minimum frequency of a variant nucleotide. The enrichment score was calculated as the number of detected A-to-G or T-to-C mismatches per 10 kb in RNA samples divided by the counterpart in DNA samples. The average value of two samples is shown. Sites with ≥1,000 reads were used. (b) Pairwise comparison of the editing level of each novel site and the nearest known site. (c) Cumulative distribution of RNA-editing levels for different groups of sites. For each site, the highest editing level among eight samples is shown. (d) Nucleotide composition in positions immediately upstream and downstream of the edited sites. The control is all A sites that are covered by mmPCR-seq reads and not edited in any samples tested.

**Figure 3** | ASE analysis of mmPCR-seq data. (a) Proportion of sites with ASE among all heterozygous sites using mmPCR-seq or RNA-seq for genes with various expression levels. The matched sites obtained from mmPCR-seq and RNA-seq data were used. FPKM, fragments per kilobase of transcript per million mapped reads. (b) Correlation of ASE of identical-by-descent siblings calculated from subsamplings of mmPCR-seq data. $R^2$, Pearson correlation coefficient. Error bars, s.d. (c) The percentage and number of cis-acting expression quantitative trait loci (cis-eQTL) genes identified using different permutation $p$ value thresholds. Only genes with mmPCR-seq sites were used in the analysis. Three models were used for the mapping. TRec, total read count, an association model using gene expression information only; TRecASE (RNA-seq), a joint model of TRec and ASE measured by RNA-seq; TRecASE (mmPCR-seq), a joint model of TRec and ASE measured by mmPCR-seq.
(Supplementary Fig. 9), with additional genome and RNA sequencing data available (X.L., A. Battle, K. Karczewski, K.S.S., K. Kukurba et al., unpublished data). We designed 48 pools of 20-plex primers that allowed us to examine all 960 SNPs on a single chip.

We carried out mmPCR-seq for 16 LCLs in technical triplicates. By sequencing all 48 samples in one Illumina HiSeq lane with 101-bp single-end reads, we obtained an average of ~1.6 million mapped reads per sample (Supplementary Table 4), and an average of 770 sites had ≥100 reads per LCL. At 100 reads, this provides the ability to detect allelic effects of 1.56-fold with a binomial probability of 0.05. We confirmed the accuracy of allelic assays (Supplementary Note 5, Supplementary Figs. 10–13 and Supplementary Table 5). Because personal genome sequencers can be more amenable to targeted studies, we sequenced the same samples on one Illumina MiSeq run with 150-bp single-end reads for comparison. Although ten times fewer reads were obtained, an average of 647 sites had ≥100 reads per LCL (Supplementary Note 5). ASE measurements were also highly similar between HiSeq and MiSeq measurements (Supplementary Fig. 14). These data show the feasibility of using MiSeq to quantify ASE of hundreds of SNPs and support the possibility of using HiSeq to examine more SNPs and/or samples.

We next assessed the ability of mmPCR-seq to detect ASE. Of all the heterozygous SNPs we investigated, 14–22% showed ASE in genes expressed at any level. In contrast, RNA-seq detected a substantially smaller fraction of ASE, especially in low- or moderate-expression genes, when we used the matched sites of mmPCR-seq (Fig. 3a). The same conclusion was obtained when we combined all heterozygous SNPs in the same gene to call ASE (Supplementary Note 6 and Supplementary Figs. 15 and 16). As the accumulation of regulatory variation—and, therefore, the expectation of ASE discovery—could potentially differ for high- versus low-expression genes, we further tested the heritability of the discovered effects across expression levels. We discovered better capture of heritable effects for mmPCR-seq than for RNA-seq across all expression levels (Supplementary Fig. 17). These results highlight the utility of mmPCR-seq to detect ASE for genes regardless of overall expression level. We further determined the optimal read depth for capturing heritable ASE. By taking random subsets of reads from mmPCR-seq, we observed that heritability estimates become saturated for sites with 400 reads (Fig. 3b), which suggests the requirement of high coverage to capture ASE more accurately.

By combining genotypic and allelic effects on expression, it is possible to jointly detect cis-acting expression quantitative trait loci (cis-eQTL)22. However, using RNA-seq data for this approach has limited success because of insufficient sequencing coverage for accurate ASE quantification21. When we compared the number of cis-eQTL genes identified using (i) gene expression information only, (ii) gene expression information and ASE measured by RNA-seq, and (iii) gene expression information and ASE measured by mmPCR-seq, we found that a combination of RNA-seq and ASE measured by mmPCR-seq nearly doubled the power to detect cis-eQTL (Fig. 3c).

We have demonstrated that mmPCR-seq provides a flexible high-throughput methodology to measure RNA allelic ratios. mmPCR-seq is cost-effective and efficient (<8 min per sample; Supplementary Table 6). These features, along with its utility for low-quantity and low-quality RNA samples, will enable future large-scale editotyping and allelotyping studies. At its current throughput, mmPCR-seq enables measuring RNA-editing levels for all known nonrepetitive recording sites in the human genome, thus providing a single standardized solution. For ASE studies, by focusing on a customized set of SNPs, mmPCR-seq may help identify regulatory effects linked to variants identified in genome-wide association studies22, map causal regulatory variation23 or identify epistatic interactions24, pinpoint genetic interactions that define the variable penetrance of coding variants25, and provide more complete insight into the genetics of gene expression.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**Accession codes** NCBI Sequence Read Archive: SRP029341.

**Note:** Any Supplementary Information and Source Data files are available in the online version of the paper.

**ACKNOWLEDGMENTS**

We thank M. Snyder for access to the Fluidigm Access Array system and W. Sun for advice on TReCASE analysis. R.Z. was partially supported by a Dean’s fellowship from Stanford University School of Medicine. G.R. was supported by a Stanford Graduate Fellowship. This work was supported by the US National Institutes of Health (GM102484); Ellison Medical Foundation and United States–Israel Binational Science Foundation (to J.B.L.); and Edward Mallinckrodt, Jr. Foundation (to S.B.M.).

**AUTHOR CONTRIBUTIONS**

R.Z. developed and optimized the mmPCR-seq method with the help from G.R., K.S.S., S.B.M. and J.B.L. R.Z. and X.L. performed computational analyses with help from S.B.M. and J.B.L. G.T. provided the brain samples. R.Z., X.L., S.B.M. and J.B.L. wrote the paper.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

**References**

1. Nishikura, K. Annu. Rev. Biochem. 79, 321–349 (2010).
2. Wahlstedt, H., Daniel, C., Ensterö, M. & Öhman, M. Genome Res. 19, 978–986 (2009).
3. Li, J.B. et al. Science 324, 1210–1213 (2009).
4. Maas, S., Kawahara, Y., Tamburro, K.M. & Nishikura, K. RNA Biol. 3, 1–9 (2006).
5. Pastinen, T. & Hudson, T.J. Science 306, 647–650 (2004).
6. Montgomery, S.B. et al. Nature 464, 773–777 (2010).
7. Pickrell, J.K. et al. Nature 464, 768–772 (2010).
8. Ramaswami, G. et al. Nat. Methods 9, 579–581 (2012).
9. Ramaswami, G. et al. Nat. Methods 10, 128–132 (2013).
10. Ng, S.B. et al. Nature 461, 272–276 (2009).
11. Levin, J.Z. et al. Genome Biol. 10, R115 (2009).
12. Mercer, T.R. et al. Nat. Biotechnol. 30, 99–104 (2012).
13. Zhang, K. et al. Nat. Methods 6, 613–618 (2009).
14. Eran, A. et al. Mol. Psychiatry 18, 1041–1048 (2013).
15. Main, B.J. et al. BMC Genomics 10, 422 (2009).
16. Zhang, K. et al. Nat. Genet. 38, 382–387 (2006).
17. Andersson, R., Möls, T. & Remm, M. Nucleic Acids Res. 36, e66 (2008).
18. Polson, A.G. & Bass, B.L. EMBO J. 13, 5701–5711 (1994).
19. Ensterö, M., Daniel, C., Wahlstedt, H., Major, F. & Öhman, M. Nucleic Acids Res. 37, 6916–6926 (2009).
20. Gommans, W.M., Mullen, S.P. & Maas, S. Bioessays 31, 1137–1145 (2009).
21. Sun, W. Biometrics 68, 1–11 (2012).
22. Hindorff, L.A. et al. Proc. Natl. Acad. Sci. USA 106, 9362–9367 (2009).
23. Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E.T. PLoS Genet. 7, e1002144 (2011).
24. Lappalainen, T., Montgomery, S.B., Nica, A.C. & Dermitzakis, E.T. Am. J. Hum. Genet. 89, 459–463 (2011).
25. MacArthur, D.G. et al. Science 335, 823–828 (2012).
ONLINE METHODS

Multiplex PCR primer design. To design multiplex PCR primers, we modified the yamPCR program, which designs multiplex PCR primers on the basis of genome sequence information. For a given group of loci of interest, yamPCR involves four steps to design primers: (i) cutting flanking sequences of given loci and identifying candidate primers using a modified Primer 3 program, (ii) searching all partial and complete matches of the candidate primers on the genome by BLAST, (iii) predicting the amplification products of all possible primer combinations via a thermodynamic model, and (iv) deducing a group of compatible primers that are specific to the loci of interest. We made three modifications to this program to design multiplex primers according to transcriptome information. First, for a given site, we mapped the site to a specific transcript on the basis of gene annotation information and cut flanking cDNA sequences from given genome file. Notably, for RNA-editing site primer design, because editing occurs before splicing, the editing level of a specific site should be the same for different isoforms of a gene. Therefore, the selection of different transcripts in this step does not affect the quantification of editing levels. For ASE-site primer design, the incorporation of transcript selection in this step can be used to examine ASE in different isoforms of a selected gene. Second, we made a database that contains cDNA sequences of all human genes and used this database instead of genome sequence for BLAST search. For each gene, we merged all of its exons to generate a representation of editing levels. For ASE-site primer design, the incorporation of transcript selection in this step can be used to examine ASE in different isoforms of a selected gene. Second, we made a database that contains cDNA sequences of all human genes and used this database instead of genome sequence for BLAST search. For each gene, we merged all of its exons to generate a representation of transcript selection. Third, the editing or ASE sites were designed to be within 100 bp from the 5′ end of one primer. To ensure that the designed primers were not located in regions with variants, which may potentially affect the allelic ratio quantification, we used the genome that had been masked using known SNPs and editing sites to design primers. A Perl script used to design the primers is in Supplementary Software and also available at http://lilab.stanford.edu/mmPCR/.

Editing-site collection. We collected exonic A-to-I editing sites from two resources. One was a subset (88) of 400 nonrepetitive editing sites identified in a genome-wide editing-site scanning. Our previous study identified 400 nonrepetitive editing sites in the human genome. To obtain high-confidence sites, we selected 88 of them that are edited in the brain RNA-seq data sets we collected (Supplementary Table 7). The second resource was 49 newly identified nonrepetitive nonsynonymous editing sites and 176 UTR sites located in nonrepetitive regions.

We designed 24 10-plex primers that cover 287 editing sites (Supplementary Data 1) within 240 loci. The sizes of amplicons range from 150 to 350 bp. Primer sequences are listed in Supplementary Data 2.

Allele-specific expression site collection. We selected two sets of sites from 16 LCL samples. Group 1 contains 617 sites that are known eQTL genes. Group 2 contains 755 non-eQTL sites. All selected SNPs are within expressed genes and heterozygous in at least three individuals within the family. We designed 48 20-plex primers that cover 960 sites, including 410 and 550 Group 1 and 2 sites, respectively (Supplementary Data 1). The sizes of amplicons range from 150 to 350 bp. Primer sequences are listed in Supplementary Data 2.

RNA and cDNA preparation. Total RNAs were extracted with the RNaseasy kit (Qiagen) or Trizol (Invitrogen). After DNase I treatment, 2–10 µg of total RNA was used to synthesize the cDNA with iScript Advanced cDNA Synthesis Kit (Bio-Rad). cDNA was purified with MinElute PCR Purification Kit (Qiagen) and concentrated using SpeedVac if needed.

Preamplification of cDNA samples. To preamplify cDNA samples before the microfluidic multiplex PCR, 1 µl of cDNA sample was added to 9 µl of pre-sample mix containing 5 µl of KAPA2G Fast Multiplex Mix (2×) and 2.4 µl of primer pool (104 nM per primer). We used the following PCR program for preamplification: 95 °C for 10 min, 15 cycles of 95 °C for 15 s, 55 °C for 30 s and 72 °C for 60 s; and 72 °C for 3 min. Following preamplification, amplified product was purified with MinElute PCR Purification Kit.

Preparation of mixtures with known allelic ratio. We randomly chose six SNPs from the set of 960 ASE sites with one requirement: we are able to find an individual from the 16-person family with homozygous AA genotype and another with homozygous BB genotype. For each of the six loci, we carried out PCR to obtain amplicons with either A or B allele, and we mixed them at different allelic frequencies (A/(A + B) = 1%, 2%, 5%, 10%, 20%, 30% and 40%). In addition to testing different allelic frequencies, we also titrated the concentration of different templates in the Fluidigm reactions to test the effect of different amounts of template-molecule copies on the accuracy of allelic-ratio measurements.

Target amplification on the Fluidigm Access Array microfluidic system. We loaded 4 µl of individual 5-, 10- or 20-plex primer pools (1 µM per primer) into the primer inlets of the 48.48 Access Array IFC (Fluidigm). To prepare the cDNA templates, we added 2.25 µl of each cDNA sample to 2.75 µl of pre-sample mix containing 2.5 µl of KAPA2G 2× Fast Multiplex Mix (Kapa Biosystems) and 0.25 µl of 20× Access Array sample loading buffer (Fluidigm). After the loading of both samples and primers via IFC Controller AX (Fluidigm), the IFC was subject to thermal cycling using FC1 Cycler (Fluidigm) with the following program: 50 °C for 2 min; 70 °C for 20 min; 95 °C for 10 min; five, 10 or 15 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 60 s; two cycles of 95 °C for 15 s, 80 °C for 30 s, 60 °C for 30 s and 72 °C for 60 s; eight cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 60 s; two cycles of 95 °C for 15 s, 80 °C for 30 s, 60 °C for 30 s and 72 °C for 60 s; eight cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 60 s; five cycles of 95 °C for 15 s, 80 °C for 30 s, 60 °C for 30 s and 72 °C for 60 s; and 72 °C for 3 min.

Sequencing adaptor and barcode addition. For each sample, 0.5 µl of the 100-fold–diluted PCR products was added to 9.5 µl of pre-sample mix containing 5 µl of KAPA2G 2× Fast Multiplex Mix, 2 µl of primer mix (a universal forward primer (2×M)) and a reverse primer with different barcode sequence (2×M) and 2.5 µl of water. Primer sequences are listed in Supplementary Table 8. We used the following PCR program: 95 °C for 10 min; four cycles of 95 °C for 30 s, 55 °C for 30 s and 72 °C 1 min; ten cycles of 95 °C for 30 s and 72 °C for 1 min; and 72 °C for 5 min. All pools were then combined at equal volume and purified via QIAquick PCR purification kit (Qiagen).
**Fluidigm library sequencing data analysis.** Libraries were pooled and sequenced in Illumina HiSeq with 101-bp single-end reads. We used FASTX Toolkit to demultiplex the raw reads. We used BWA to align reads to a combination of the reference genome and exonic sequences surrounding known splicing junctions from gene models annotated in RefSeq and Gencode V12. We chose the length of the splicing junction regions to be slightly shorter than the reads to prevent redundant hits. For allelic-ratio count, we took the bases with a minimum quality score of 20. For read-depth count, we took the coverage of the representative sites in each amplicon.

To call novel RNA-editing sites, we required variants to be supported by at least two mismatch reads with base quality score \( \geq 20 \) and mapping quality score \( \geq 20 \). We also removed all known SNPs present in dbSNP (except SNPs of molecular type “cDNA”; database version 135; [http://www.ncbi.nlm.nih.gov/SNP/](http://www.ncbi.nlm.nih.gov/SNP/)), the 1000 Genomes Project or the University of Washington Exome Sequencing Project ([http://evs.gs.washington.edu/EVS/](http://evs.gs.washington.edu/EVS/)).

**RNA-seq data analysis.** The Human Brain Reference RNA sample (HBRR, Ambion, catalog #6050) consists of total RNA extracted from several regions of the brains from 23 adult donors. We obtained three Illumina RNA-seq data sets for this sample from the NCBI Sequence Read Archive ([http://www.ncbi.nlm.nih.gov/sra](http://www.ncbi.nlm.nih.gov/sra)). A list of data sets is shown in Supplementary Table 9. We mapped RNA-seq reads to the human genome as previously described. To obtain gene expression levels, we used a pipeline comprising TopHat and Cufflinks.

**ASE analysis.** Whole-genome sequencing data of the family were obtained from Complete Genomics. Samples were sequenced to an average genome-wide coverage of 80×. SNPs are called by Complete Genomics Analysis Pipeline (version 2.0.0).

We performed a binomial test to obtain the statistical \( P \) values of deviations from 0.5 on the raw allelic counts. \( q \) values were then used to estimate the proportion of non-nulls. \( q \) values were calculated using the qvalue function from R package. A \( q \) value \( \leq 0.05 \) was used as the cutoff of statistical significance. We also required that the magnitude of allelic drift be larger than 0.1 (allelic frequency <0.4 or >0.6) to be considered allele specific. We required at least five reads for both reference and alternative alleles with allelic frequency >0.01 and <0.99 to avoid wrongly assigned homozygous SNPs due to potential genotyping or sequencing errors.

To perform cis-eQTL mapping using both genetic data and RNA-seq or by also including ASE information, we employed asSeq R package ([http://www.bios.unc.edu/~weisun/software/asSeq.htm](http://www.bios.unc.edu/~weisun/software/asSeq.htm)). We examined all SNPs within the gene body or outside the gene body (within 200 kb of the transcription start or end sites) to identify cis-eQTLs. All 16 individuals were treated as unrelated for mapping. We considered the most significant cis-eQTL for each gene and calculated a permutation \( p \) value for each gene using the trecaseP function in the asSeq package with the following parameters: \( \text{min.AS.reads} = 20 \), \( \text{min.AS.sample} = 5 \), \( \text{min.n.het} = 5 \), \( \text{local.only} = \text{TRUE} \), \( \text{local.distance} = 200000 \), \( \text{np.max} = 500 \), \( \text{np} = c(20, 100) \), \( \text{aim.p} = c(0.5, 0.2) \). For a fair comparison between mmPCR-seq and RNA-seq, we used the same set of ASE sites (selected in the mmPCR-seq assay) for analysis.

We investigated the heritability of ASE within the family by comparing the ASE between identical-by-descent (IBD) siblings. IBD sharing block are defined by recombination positions in the family. Haplotype blocks are inferred by method PedIBD. Pearson correlation coefficient \( R^2 \) was used to reflect degree of correlation between ASE among IBD siblings.

**Statistical analysis.** All statistical analyses were performed with either R packages or Matlab.

26. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).
27. Trapnell, C., Pachter, L. & Salzberg, S.L. *Bioinformatics* **25**, 1105–1111 (2009).
28. Trapnell, C. *et al.* *Nat. Biotechnol.* **28**, 511–515 (2010).
29. Li, X., Yin, X. & Li, J. *Bioinformatics* **26**, 1191–1198 (2010).