Chapter 6

Identification of Imprinted Loci by Transcriptome Sequencing

Tomas Babak

Abstract

Enabled by high-throughput technologies that are capable of generating millions of sequencing reads, transcriptome sequencing is emerging as an important approach for mapping allelic imbalance (AI), where transcription is biased toward one allele in a diploid system. AI is identified by counting sequencing reads that map to genomic regions containing heterozygous SNPs, where the base identity of the SNP is used to distinguish allelic origin. Genomic imprinting is a special case of AI where bias is toward parental sex and can be identified by transcriptome sequencing of systems that represent reciprocally inherited loci. The focus of this protocol is on experimental design, analysis, and interpretation of genomic imprint discovery using whole transcriptome sequencing.

Key words: Imprinting, Transcriptome sequencing, Allelic imbalance

1. Introduction

While a comprehensive map of imprinted loci in all cell/tissue types of all mammals facilitates the evolutionary characterization of genomic imprinting, inherent challenges of traditional discovery approaches have mostly limited their application to developing mouse stages. Classical genetic screens based on uniparental disomies and reciprocal translocations (1, 2) and genetic mapping of parent-of-origin phenotypes in humans (e.g., ref. 3) revealed the initial imprinted loci. Most imprinted genes emerged from fine mapping of these initial regions, typically by RFLP analysis of cDNA. Microarray profiling of embryos with uniparental disomies or entire genomes (4, 5) extended the map, but embryonic lethality induced by these genetic perturbations has limited their discovery potential. A two-dimensional RFLP approach (6) and genotyping microarrays (7) have been applied to imprint discovery in adult/wild-type tissues but require extensive analysis to rule out signal

Nora Engel (ed.), *Genomic Imprinting: Methods and Protocols*, Methods in Molecular Biology, vol. 925,
DOI 10.1007/978-1-62703-011-3_6, © Springer Science+Business Media, LLC 2012

from noise. Mapping imprinting by transcriptome sequencing, which has only recently become possible, is advantageous in that it does not require a priori knowledge of the expected genomic localization or phenotype, and can be applied to any diploid progeny of genetically diverged parents.

NextGen sequencing (NGS) has rapidly changed the landscape of nucleic acid-driven research. In addition to standard sequence determination, the ability to generate millions of sequencing reads has also enabled quantification of input abundance by simply counting the reads. RNA-Seq (8, 9), the most commonly utilized form of transcriptome sequencing, is based on random priming of polyA + purified RNA to generate cDNA, which is again random primed to generate double-stranded DNA that is then sequenced. Biological inference typically involves mapping of the reads to a reference genome and quantifying events (e.g., splicing or gene expression) by counting. Similarly, when a mapped sequencing read overlaps a heterozygous SNP, allelic origin of that read can be discerned using the identity of the polymorphic base. AI can be inferred by comparing the two sums of allelic reads that map to that SNP. Since AI is widespread in mammalian cells (10) and genomic imprinting is likely to be causal for only a small proportion (11), demonstrating AI in a reciprocally inherited fashion is imperative for imprinting discovery. The most straightforward experimental design consists of two crosses of inbred strains, where each parental sex is represented by both genetic backgrounds. In this simple system each progeny is heterozygous at all SNPs and both maternal and paternal copies of each allele are represented (Fig. 1). RNA-Seq is then applied to both crosses and genomic imprinting distinguished as AI that is consistently biased toward parental sex (vs. AI biased toward same parental strain which is much more common). Although the classical definition of genomic imprinting is based on monoallelic expression, it has been expanded

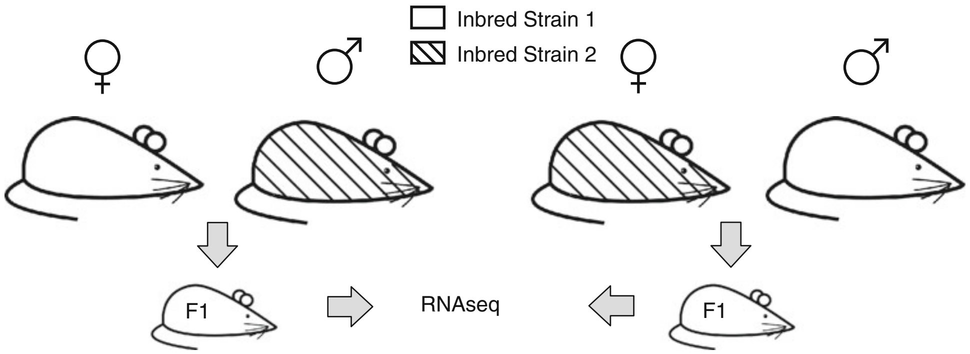

Fig. 1. Schematic of reciprocal cross. Inbred parental strains are crossed in reciprocal to produce F1 progeny that are sequenced and analyzed in pairs.

to include incomplete parent-of-origin expression biases (11) and enables identification of cell/tissue-specific imprinting that is diluted by tissue heterogeneity.

Several groups have successfully used transcriptome sequencing for imprint discovery (12–14) and the approach has outstanding promise, but unresolved challenges exist. For example, all published studies using RNA-Seq to map AI have thus far not addressed systematic bias. It has been shown that AI reproduces very well across technical and biological replicates when considering the same SNP (15) but little has been done to assess concordance across different SNPs. Systematic priming biases induced by SNPs could lead to incorrect AI calls. One would expect strong agreement on AI among two SNPs within the same exon, for example, and these could be used to empirically estimate systematic bias. Until AI is robustly modeled, mock reciprocal crosses (i.e., biological replicates) can be used to gauge false-discovery and I expand on this below. A second challenge is application outside the somewhat artificial setting of inbred mouse crosses. Transcriptome sequencing has not yet been applied for imprint discovery in species where inbred strains or controlled crosses are not available (such as humans). Two experimental designs could be employed to make this feasible. The first is a controlled screen in a family where parents and offspring are genotyped; AI is determined at all heterozygous SNPs in offspring, and parent-of-origin inheritance determined from phased haplotypes. A large family with distantly related parents would be ideal. The second approach is to screen an outbred population and identify imprinting as AI with no sequence dependence (i.e., AI is always observed but biased for either allele with equal likelihood) since consistent bias toward one allele suggests a genetic mechanism and this is generally the case (10).

The focus of this protocol is on identification of genomic imprinting in reciprocally crossed F1 mouse strains using standard RNA-Seq with emphasis on analysis practices. The assumption is that the reader has access to total RNA from reciprocally crossed mouse tissues and a reasonable (>5 million) map of SNPs for these strains. The scope of the approach could be expanded to any diploid species and additional recommendations for application outside inbred mouse strains are also included.

2. Materials

2.1. Constructed RNA-Seq Libraries

Construction of RNA-Seq libraries was first described in yeast and mouse (8, 16) and is now available in kit format from several manufacturers. In my experience the standard RNA-Seq kit sold by Illumina works very well and the end result is a library of high complexity (measured by the proportion of sequencing reads that align to unique genomic locations). The TruSeq RNA kit (Illumina), which is designed for higher sample throughput, works well and I have seen great data from as little as 100 ng total RNA input. A few practical changes involving deoxy-uridine triphosphate (dUTP) and uracil-*N*-glycosylase (UNG) (17, 18) effectively introduce strand specificity and are recommended since imprinted antisense transcription is known to exist. Library complexity and even coverage are essential for measuring AI and some of the low-input kits suffer in this regard because they have multiple series of amplification. NSR-seq (not-so-random primer sequencing) (19) is one of the earlier approaches that was applied for identifying AI (12), and has the advantage of capturing non-polyadenylated transcripts and is also strand specific, but personal experience and a recent evaluation (17) have revealed undesirable evenness of coverage (i.e., coverage is "spiky"). In summary, any approach that quantitatively captures input transcript abundance and yields a library of high complexity will work and I have seen excellent data suitable for mapping AI from libraries made with mRNAseq/TruSeq RNA kits purchased from Illumina modified with dUTP/UNG treatment (18) to achieve strand specificity.

2.2. NextGen Sequencing Capacity

454, SOLiD, and Illumina are currently the major suppliers of NGS sequencers. Any of these platforms and likely many other emerging platforms will work, although Illumina and SOLiD are currently the only commercially available RNA-Seq platforms for generating tens to hundreds of millions of reads. Overall sequencing depth is dependent on the length and number of sequencing reads and the heterozygous SNP density of the system. Methods exist to estimate the minimum required sequencing (20) and more will always improve sensitivity. In practice, 4 Gb of single-end RNA-Seq data from reciprocally crossed C57BlxCAST samples (i.e., 8 Gb total data, 4 Gb from each cross) is sufficient to confidently identify >90% of previously validated imprints in that tissue. 2 Gb will result in slightly lower performance (70–80% sensitivity at the same detection threshold) and even 1 Gb will yield acceptable results (~60% sensitivity). The ideal read length is a trade-off between molecular complexity (long reads and PE reads limit the number of molecules represented in the library) and sequencing of SNPs. The ideal read length would on average capture 1 SNP/read and can be estimated using a published model (20). Considering practical challenges I recommend using single-end 75–100 bp reads. Paired-end (PE) data improves mapping performance but only marginally. With a mean RNA-Seq insert size of ~200 bp, the 3′ ends of pairs can overlap which leads to diminishing returns. Reads shorter than 50 bp are not recommended since this will lead to significant mapping bias (see Note 9).

2.3. Computational Resources

1. Access to Linux/Unix working environment with at least 6 Gb RAM.
2. Installation of Novoalign v2.07.11 or newer (21) or equivalent short read sequence aligner. V2.07.11 has capability of reporting mismatches to masked bases (Ns).
3. Reference genome. Most sequenced mammalian genomes can be downloaded from UCSC (22). If the genome for the species is not available, reads could be assembled into transcript models that then serve as a reference, but this is beyond the scope of this protocol.
4. Map of SNPs. 15 mouse strains were recently sequenced by the Sanger Institute and SNP maps are available for download (23). If working with a system that has a reference genome but where SNPs are unknown, genotyping arrays or genome sequencing can be used to map heterozygous SNPs. In humans additional SNPs can be imputed and phased using MaCH (24) to improve sensitivity of AI mapping. SNPs can also be inferred from the RNA-Seq data. This does not work well for mapping AI since discovery favors SNPs biased in expression toward the non-reference allele and thus the resulting AI profile becomes artificially skewed toward non-reference alleles. However, the approach can be effective for imprint discovery by calling SNPs on pooled (in equal amount) sequencing data from the reciprocal crosses where imprinted SNPs are supported by near 50:50 proportions. SAMtools (25), GATK (26), or soapSNP (27) can all be used to identify SNPs from mapped RNA-Seq data and are comparable in performance.
5. Perl/Python installation or equivalent for custom manipulation of data.
6. Matlab, R, Excel, or equivalent for visualizing results.

3. Methods

1. Generate a masked version of genome where known SNPs are replaced with Ns (see Note 1).
2. Index genome by running Novoindex (default options and -k 14 -s 3) on a single fasta file of masked genome (see Note 2).
3. Align fastq-formatted raw reads files (paired-end or single-end) against masked genome using Novoalign (default options and -a -o IUBMatch -r None, and -i 0 1000 if aligning PE reads) (see Note 3). A summary of the alignment approach is shown in Fig. 2.

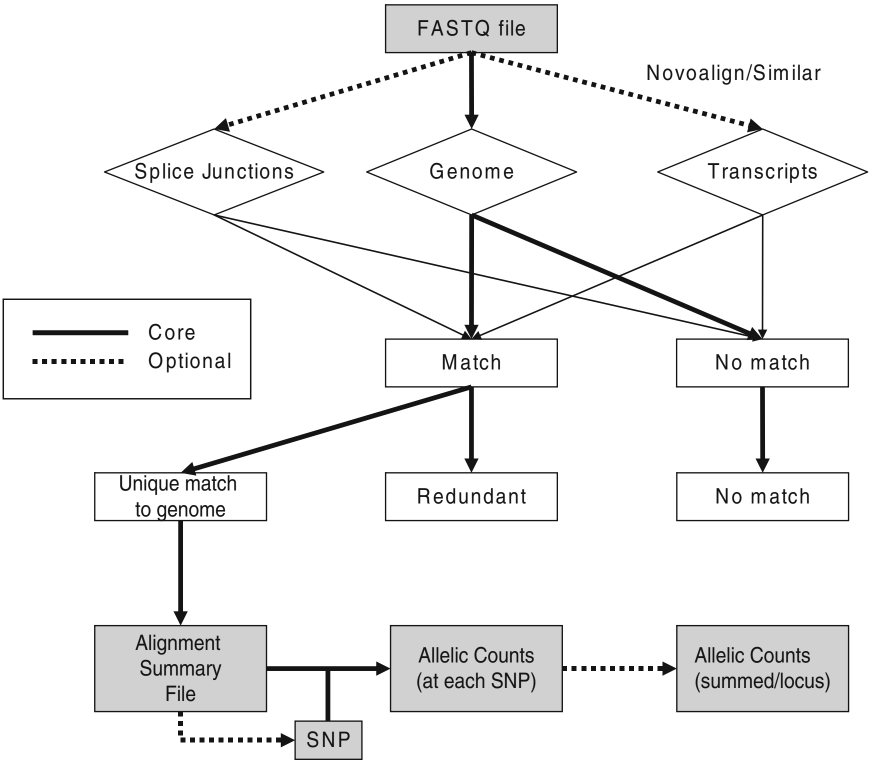

Fig. 2. Alignment, SNP-identification, AI-quantification pipeline. Alignment is accomplished with an independent algorithm (e.g., Novoalign (21)) against the genome, and optionally splice junctions and full-length transcripts. Unique matches (in the genome) are retained and used for SNP prediction and quantification of ASE.

4. If improved alignment sensitivity is desired, also align reads to splice junctions and full-length transcripts (especially useful if aligning paired-end reads) using Novoalign (same settings as above except -r All 50) (see Note 4). Convert transcript alignment coordinates to genomic coordinates using the transcript genomic coordinates (see Note 5).
5. Discard redundant alignments (where reads map to more than one genomic location) and generate report files that store alignment coordinates and genomic mismatches for each read (see Note 6).
6. For each SNP, tally the number of reads that support the reference and alternate bases (see Note 6).
7. At each SNP, let A represent the number of reference-specific reads and B the number of alternate allele-specific reads. Quantify the degree of AI as $A/(A+B)$. The probability of AI can be estimated using the cumulative binomial distribution. This can also be done in Excel where binomial-p (probability of no AI) = binomdist(min(A, B), $A+B$, 0.5, 1). In Matlab the binomcdf function from the statistics package can be called.

The same principles can also be used on all allele-specific reads summed across a transcript (i.e., sum over all SNPs within the transcript) (see Note 7).

8. Genomic imprinting requires AI to be measured in tissue-matched reciprocally crossed samples. If s1 = sample 1 and s2 = reciprocal sample, Genomic imprinting may exist if ($AI_{s1} > 0.5$ and $AI_{s2} < 0.5$) or if ($AI_{s1} < 0.5$ and $AI_{s2} > 0.5$), i.e., reciprocal bias exists. The probability of imprinting can be estimated as the less significant binomial estimate of the two samples (see Note 8).
9. Select a suitable threshold of significance for calling imprinting by using a mock reciprocal cross as a negative control (see Note 9).

4. Notes

1. SNP maps can be downloaded from Sanger (23) and masking greatly reduces alignment biases (28).
2. This step creates an .idx file that is used as input for genome alignment. Junction and transcript indices (step 4) can be made with the same settings.
3. Over a dozen short-read alignment algorithms are currently available. BWA (29), SOAP2 (30), and Bowtie (31) are based on the Burrows–Wheeler Transform (BWT) algorithm and are by far the fastest aligners with sensitivity and specificity comparable or better than most. However, in testing these and eight other popular aligners on simulated single-end and paired-end data (with imputed mismatches representative of quality scores and expected variation), Novoalign (21) attained the highest sensitivity (7–8% higher sensitivity than BWT approaches with avg. alignment rate of 87% vs. 79–80%) and comparable specificity (<0.1% erroneous alignments) to all aligners. Not surprisingly, the AI profile was less biased toward reference alleles than for other approaches tested owing to a better ability to align over SNPs. -a will trim adapter sequences, -o IUBMatch will report N > (ACGT) base changes, -r None will not report reads that align in more than one genomic regions, and -i 0 1000 will allow pairs to match up to 1,000 bp apart.
4. Extensive custom scripting will be required to perform this step and there is more than one way to compile a reference transcriptome. I made a splice junction coordinate file from all possible exon skipping events (up to two exons skipped) from RefSeq, ENSEMBL, UCSC known gene, and Genbank mRNA BED files downloaded from UCSC (22). I then sorted to

remove duplicate entries (sort -k 6,6 -k 1,1 -k 2,2n -k 3,3n -u unsorted_with_6_columns.bed > sorted_unique_junctions.bed) and retrieved the fasta equivalent from UCSC Table Browser and indexed using Novoindex (-k 14 -s 3). Aligning paired-end reads to transcripts will considerably improve the number of reads that align as pairs. I again recommend RefSeq, ENSEMBL, UCSC KG, and Genbank mRNAs as a comprehensive transcript set. It is important to allow reporting of all matches since -r None will ignore matches to multiple isoforms which will be most of them (i.e., use -r All 50). Redundant filtering (step 5) done in genomic space removes truly redundant matches. BED files for junctions and transcripts can be used to convert alignments back into genomic coordinates.

5. If mapping paired-end reads, a paired match takes precedence over single matches (i.e., if maps as a pair once take that alignment and disregard all others). At this point reads that do not contain N > (ACGT) changes can be discarded if further SNP discovery will not be done.

6. Reads mapping to opposite strands should be tallied independently if strand-specific RNA-Seq was used (i.e., each SNP may have up to two sets of counts).

7. The cumulative binomial distribution models the maximum number of successes in a sequence of independent binary events, each of which yields success with some probability. For example, the chance of getting three or fewer heads when flipping a fair coin ten times is 17.2%. Summing reads across SNPs violates the binomial assumption when a single sequencing read spans more than one SNP since it expects all counts to be independent. Ideally, a read (whether single-end or paired-end) should only be counted once. An ad hoc approach to ensure that this is the case is to only consider SNPs that are further apart than the read length (fragment length if using paired-reads). In practice, the extent of systematic error in measuring AI with RNA-Seq contributes significantly more uncertainty in the binomial calculation than violating counting independence as described. Negative controls are imperative for estimating false-discovery (see step 9).

8. A suitable threshold for making an imprint call depends on the extent of acceptable false-discovery (i.e., proportion of calls that are not truly imprinted; see step 9) and will vary from sample to sample and with the selected RNA-Seq protocol. In practice, a binomial *p*-value of 0.001 results in a false-discovery rate (FDR) of ~10% using standard Illumina mRNAseq.

9. For this control to be valid, samples need to be prepared completely in parallel, they must be sequenced to equivalent depths, and all must pass quality control criteria. The FDR can be

estimated by plotting the number of imprinted sites as a function of binomial-p cutoff from data generated from biological replicates. Since there is no genuine reciprocal inheritance of any allele in this scheme, all calls are false-positives and their rate will translate to a genuine cross if all samples are sequenced to an equal depth. Random removal of reads should be done to ensure that all samples have the same number of input reads. A similar plot for a genuine reciprocal cross can be used to estimate sensitivity (number of known imprinted sites detected) and by combining the data into a plot of FDR vs. sensitivity a useful threshold for making imprinting calls can be selected. The same criteria can be applied to AI inferred from reads summed across SNPs in the same transcript.

References

1. Cattanach BM, Kirk M (1985) Differential activity of maternally and paternally derived chromosome regions in mice. Nature 315: 496–498
2. Surani MA, Reik W, Allen ND (1988) Transgenes as molecular probes for genomic imprinting. Trends Genet 4:59–62
3. Nicholls RD, Knoll JH, Butler MG, Karam S, Lalande M (1989) Genetic imprinting suggested by maternal heterodisomy in nondeletion Prader-Willi syndrome. Nature 342: 281–285
4. Choi JD, Underkoffler LA, Collins JN, Marchegiani SM, Terry NA, Beechey CV, Oakey RJ (2001) Microarray expression profiling of tissues from mice with uniparental duplications of chromosomes 7 and 11 to identify imprinted genes. Mamm Genome 12: 758–764
5. Mizuno Y, Sotomaru Y, Katsuzawa Y, Kono T, Meguro M, Oshimura M, Kawai J, Tomaru Y, Kiyosawa H, Nikaido I, Amanuma H, Hayashizaki Y, Okazaki Y (2002) Asb4, Ata3, and Dcn are novel imprinted genes identified by high-throughput screening using RIKEN cDNA microarray. Biochem Biophys Res Commun 290:1499–1505
6. Plass C, Shibata H, Kalcheva I, Mullins L, Kotelevtseva N, Mullins J, Kato R, Sasaki H, Hirotsune S, Okazaki Y, Held WA, Hayashizaki Y, Chapman VM (1996) Identification of Grf1 on mouse chromosome 9 as an imprinted gene by RLGS-M. Nat Genet 14:106–109
7. Morcos L, Ge B, Koka V, Lam KC, Pokholok DK, Gunderson KL, Montpetit A, Verlaan DJ, Pastinen T (2011) Genome-wide assessment of imprinted expression in human cells. Genome Biol 12:R25
8. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628
9. http://www.illumina.com
10. 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
11. Morison IM, Ramsay JP, Spencer HG (2005) A census of mammalian imprinting. Trends Genet 21:457–465
12. Babak T, Deveale B, Armour C, Raymond C, Cleary MA, van der Kooy D, Johnson JM, Lim LP (2008) Global survey of genomic imprinting by transcriptome sequencing. Curr Biol 18:1735–1741
13. Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science (New York, NY) 329:643–648
14. 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
15. Babak T, Garrett-Engele P, Armour CD, Raymond CK, Keller MP, Chen R, Rohl CA, Johnson JM, Attie AD, Fraser HB, Schadt EE (2010) Genetic validation of whole-transcriptome sequencing for mapping expression affected by cis-regulatory variation. BMC Genomics 11:473
16. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome

defined by RNA sequencing. Science (New York, NY) 320:1344–1349

17. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 7:709–715
18. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 37:e123
19. Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, Rohl CA, Johnson JM, Raymond CK (2009) Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat Methods 6:647–649
20. Fontanillas P, Landry CR, Wittkopp PJ, Russ C, Gruber JD, Nusbaum C, Hartl DL (2010) Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. Mol Ecol 19(Suppl 1):212–227
21. http://www.novocraft.com/main/index.php
22. 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. Nucleic Acids Res 38:D613–D619
23. http://www.sanger.ac.uk/resources/mouse/genomes/
24. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834
25. 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 (Oxford, England) 25:2078–2079
26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303
27. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19:1124–1132
28. 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 (Oxford, England) 25: 3207–3212
29. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25:1754–1760
30. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics (Oxford, England) 25:1966–1967
31. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25