APPLICATIONS OF NEXT-GENERATION SEQUENCING

Genome-wide allele-specific analysis: insights into regulatory variation

Tomi Pastinen

Abstract | Functional genomics is rapidly progressing towards the elucidation of elements that are crucial for the *cis*-regulatory control of gene expression, and population-based studies of disease and gene expression traits are yielding widespread evidence of the influence of non-coding variants on trait variance. Recently, genome-wide allele-specific approaches that harness high-throughput sequencing technology have started to allow direct evaluation of how these *cis*-regulatory polymorphisms control gene expression and affect chromatin states. The emerging data is providing exciting opportunities for comprehensive characterization of the allele-specific events that govern human gene regulation.

Gene expression is a complex trait that is influenced by *cis*- and *trans*-acting genetic and epigenetic variation and also by environmental factors. To date, characterization of the human genetic variation that affects gene expression has largely focused on expression quantitative trait loci (eQTL) mapping[1]. However, direct assessment of *cis*-regulatory variation requires allele-specific approaches.

Allele-specific analyses of genome function hinge on the intrinsic power of using a within-sample control — the other allele — to sensitively and specifically query the effects of genetic and epigenetic differences on *cis*-regulatory control in diploid genomes. Early allelic analyses of gene expression[2] or transcriptional activity[3] indicated the power of allelic-specific approaches but were restricted to individual loci. Recent advances in genomic technologies are now making allele-specific analyses of expression, DNA–protein interactions and chromatin states (including DNA methylation or modified histone occupancy) possible at a genome-wide scale across the human genome.

So what are the advantages of using methods with allelic discrimination to focus on *cis* effects in functional assays? In principle, elimination of environmental or *trans*-acting influences that alter gene expression or DNA–protein interactions should provide higher sensitivity for uncovering the direct influence of sequence and epigenetic variation in *cis*-regulatory elements. This advantage can be illustrated in a simple scenario: expression of a transcript is under strict negative-feedback control (that is, transcription is repressed by an increasing amount of the transcriptional product) and therefore the total expression of the transcript varies little across samples with different *cis*-regulatory genotypes. However, in heterozygotes, allele-specific studies can readily detect differences in expression between the alleles and therefore reveal the effect of *cis*-regulatory genetic variation (FIG. 1).

Early results indicate that allele-specific differences among transcripts within an individual can affect up to 30% of loci and, at the population level, ~30% of expressed genes show evidence that their *cis* regulation is influenced by common alleles[4]. In population studies, an even larger proportion of genes showed allelic variation in expression that could not be mapped, which could be due to rare genetic variants or epigenetic effects. Therefore, allele-specific genetic and epigenetic effects on gene regulation are likely to be widespread. Furthermore, surveys of allelic expression and allele-specific DNA–protein interactions have so far been limited to single cell types and small sample sizes. Consequently, only a subset of loci has yielded data at sufficient depth to assess allele specificity. Expansion of such efforts is timely given the number of common disease associations to non-coding DNA sequences[5].

In this Progress article, I discuss the techniques, results and challenges of the current transition of human functional genomics from diploid or allele-insensitive analyses to haploid or allele-specific interrogation of our genome. Although further technical improvements would be advantageous, allele-specific analyses show great promise.

Genome-wide allele-specific analysis

The extraction of functional data with allelic resolution can be directed to variable sites (polymorphisms) that have been identified within the genome or, alternatively, next-generation sequencing (NGS) technologies can be used for global functional genomics assays. In this section, I discuss these two general approaches and their relative advantages.

Polymorphism-directed approaches. The rapidly advancing characterization of common variants in the human genome provides an opportunity to query individual variant sites for allele-specific function. The advantages of this approach are twofold: only sites likely to be informative (heterozygous) for allelic analyses are observed, which increases the information density of the genomic data; and these same sites can be targeted in genomic DNA control samples with equal allelic content, which can control for the technical biases inherent in quantitative analysis of allele ratios (FIG. 2a).

Genome-wide genotyping arrays provide a convenient and relatively low-cost approach for assessing allele specificity of polymorphic sites in a range of contexts, including: expressed transcripts using genomic DNA and RNA (converted to cDNA) in parallel[4,6,7]; regulatory element DNA prepared by chromatin immunoprecipitation (ChIP)[8]; and methylated DNA, prepared by methylation-sensitive restriction enzyme digestion of genomic DNA[9]. However, coverage of allelic differences in regulatory elements can be low because current standard SNP arrays contain only a small subset of polymorphic regulatory elements (<5%). In the case of

DNA methylation, coverage is also limited because methylation differences are only detected at sites that are recognized by the restriction enzyme used in DNA preparation. However, multi-exon transcripts span large genomic regions and include intronic SNPs. Therefore, the use of RNA that contains unspliced primary transcripts in allele-specific expression analysis[6] provides information about a larger proportion of genes than if only coding SNPs are assayed. Combined with ongoing improvements in genome-wide genotyping arrays[10], this will allow allelic expression studies of nearly all human genes[4].

An alternative approach to assay allelic expression is the use of custom padlock probes to capture known exonic polymorphisms on a large scale[11,12]. This method allows targeted analysis of tens of thousands of sites in the genome. The digital quantification of the alleles in captured sequences from a genomic DNA control and corresponding RNA (cDNA) is carried out by NGS followed by allele counting of the short reads (using the same approach as used previously for quantifying alleles in PCR or RT-PCR products)[13]. Given the highly selective capture achieved by this method, over one-third of expressed SNPs yield greater than 50-fold coverage using a single Illumina Genome Analyser 2 sequencing flow cell. This approach was recently extended to understand allelic variation in DNA methylation by padlock capture of CpG islands in bisulphite-treated DNA[14], which shows that functional elements involved in gene regulation can be targeted for allelic analysis.

The parallel analysis of control (untreated or genomic) DNA and test samples (RNA, protein-bound DNA or methylated DNA) helps to limit artificial allelic biases in these polymorphism-directed techniques and also confirms the heterozygosity of the control sample at each informative SNP. Nevertheless, the control and test samples are present at variable concentrations, so additional normalization and combination of data across multiple measured polymorphisms are required to yield quantitative estimates of allelic biases[4].

Global approaches. The rapid progress in sequencing technologies is transforming functional genomics and allowing unbiased views of transcriptomes or functional DNA elements. The single-base resolution and digital nature of the data provide information on the abundance and the allelic biases in transcripts or regulatory DNA, which could not be achieved using hybridization-based microarrays.

The first assessments of allelic expression in the human genome have used NGS of the transcriptome (RNA-seq)[15–17] (FIG. 2b), and these studies present converging views on the current potential and limitations of this technology. As RNA-seq collects short sequence reads relatively uniformly across expressed transcripts and these reads correlate with the abundance of the RNA species, genes that have abundant genetic variation in their mRNA and are robustly expressed in the studied tissue can yield high coverage for expressed polymorphic sites. RNA-seq reads (with filtering of clonal reads) at the polymorphic sites provide digital estimates of allelic abundance, and simple statistical approaches (for example, binomial tests) allow the detection of sites that show potentially skewed expression. The power of these statistical tests depends on the read counts at the polymorphic sites. Owing to the unequal representation of different RNA species (that is, highly expressed transcripts are sampled at much higher coverage) and limited genetic variation in the mRNA in some cases, only a minority of expressed genes yield useful information on allelic expression at the transcriptome-sequencing coverage depths presented to date (10–20 million 36 bp reads per sample). Notably, RNA-seq is the only approach that provides concurrent allelic and total expression data, which is attractive for the within-study validation of *cis*-regulatory changes detected by total expression[16,17]. This technique also provides the future possibility of quantifying the relative contribution of *cis*- versus *trans*-regulatory changes, as shown earlier for model organisms using independent assays[18].

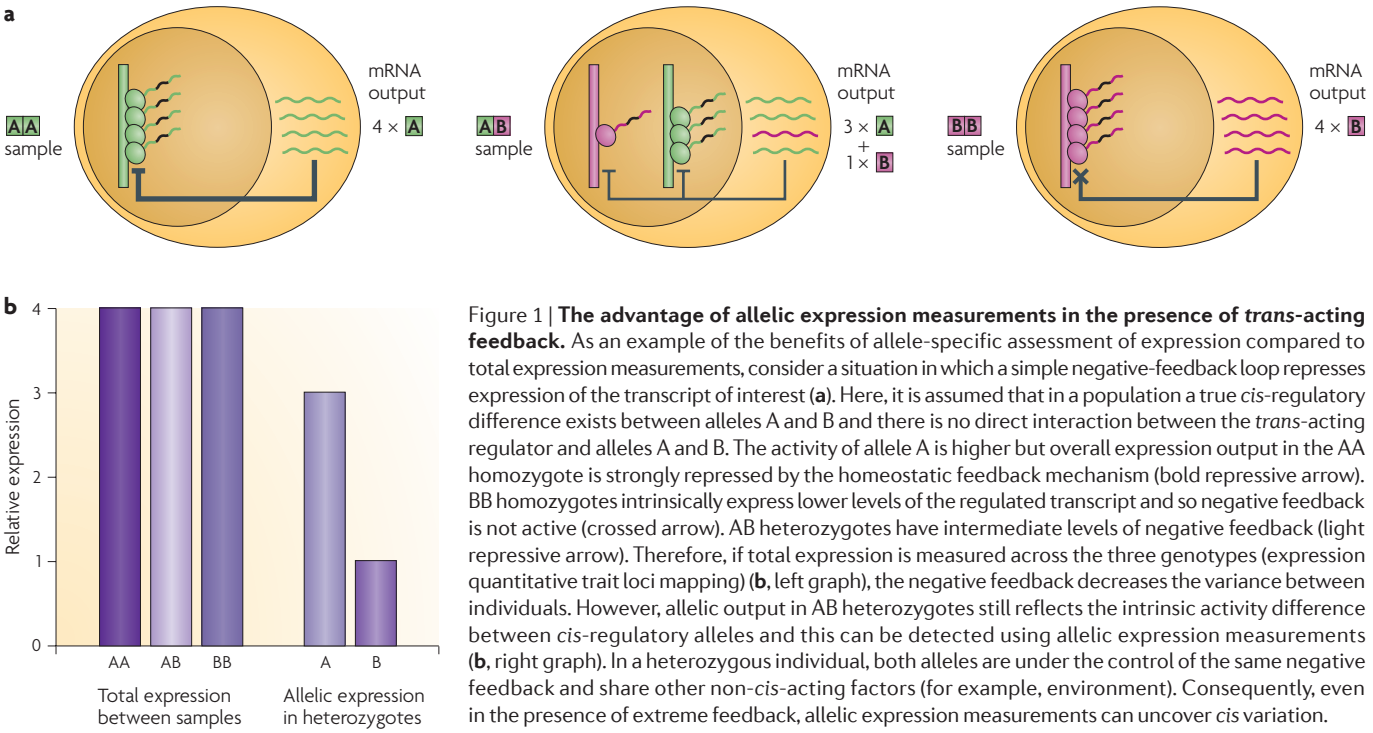


Figure 1 | **The advantage of allelic expression measurements in the presence of *trans*-acting feedback.** As an example of the benefits of allele-specific assessment of expression compared to total expression measurements, consider a situation in which a simple negative-feedback loop represses expression of the transcript of interest (**a**). Here, it is assumed that in a population a true *cis*-regulatory difference exists between alleles A and B and there is no direct interaction between the *trans*-acting regulator and alleles A and B. The activity of allele A is higher but overall expression output in the AA homozygote is strongly repressed by the homeostatic feedback mechanism (bold repressive arrow). BB homozygotes intrinsically express lower levels of the regulated transcript and so negative feedback is not active (crossed arrow). AB heterozygotes have intermediate levels of negative feedback (light repressive arrow). Therefore, if total expression is measured across the three genotypes (expression quantitative trait loci mapping) (**b**, left graph), the negative feedback decreases the variance between individuals. However, allelic output in AB heterozygotes still reflects the intrinsic activity difference between *cis*-regulatory alleles and this can be detected using allelic expression measurements (**b**, right graph). In a heterozygous individual, both alleles are under the control of the same negative feedback and share other non-*cis*-acting factors (for example, environment). Consequently, even in the presence of extreme feedback, allelic expression measurements can uncover *cis* variation.

Allele specificity in specific transcription factor–DNA interactions in living cells can be found systematically by mapping transcription factor-binding sites by ChIP–chip followed by assaying the immunoprecipitated DNA by quantitative genotyping[19] (FIG. 2b). ChIP-seq and related approaches are now widely being applied to study transcription factor binding, histone modifications and DNase I hypersensitivity. ChIP-seq allows the detection of total binding at specific sequences and also of allele-specific chromatin activity in cases in which heterozygous sites overlap ChIP-seq peaks. For example, a recent report extended global allele-specific analysis by NGS[20] across individuals to active chromatin (DNase I hypersensitivity) and DNA–protein binding (ChIP-seq for CTCF).

This early phase of studies of expressed transcripts and regulatory DNA based on counting NGS reads from individual alleles has faced a common challenge: currently, all alignment methods for short sequence reads are biased towards the alleles that are represented in the reference genome[15,20,21]. Therefore, careful adjustment for the resulting artificial allelic biases is required. The lack of a parallel reference sample (which would need to be high-coverage, full genomic DNA sequence generated by the same method) has limited most studies to already sequenced and/or deeply genotyped genomes or resulted in the exclusion of extreme allelic deviations that may either be sequence errors or true polymorphic sites that have strong biased expression of one allele[15].

A second difficulty, alluded to above, is the relatively low coverage of polymorphic sites in all global NGS studies to date, which leaves only a small subset of potentially informative sites amenable to analysis. Indeed, a recent simulation study[22] suggested that, for genomes with relatively low levels of heterozygosity (such as human), higher average coverage than is normally achieved in RNA-seq studies may be required to provide the power to detect allelic biases. For example, hundreds of reads per base pair may be required for subtle variants (those with <1.5-fold difference between transcripts)[22]. Therefore, none of the studies published to date have come close to the coverage required to include the full range of allelic biases in the genome. Decreasing sequencing cost and improved analytical methods should alleviate the concerns of limited coverage and alignment biases, respectively. An alternative approach was suggested by Heap *et al.*[15], in which the transcript region of interest was enriched by sequence capture and then subjected to high-coverage NGS analysis. This allowed detection of both exonic and intronic (in primary transcripts) expressed SNPs at increased coverage. However, the capture process introduces another variable that has potential for spurious allelic biases.

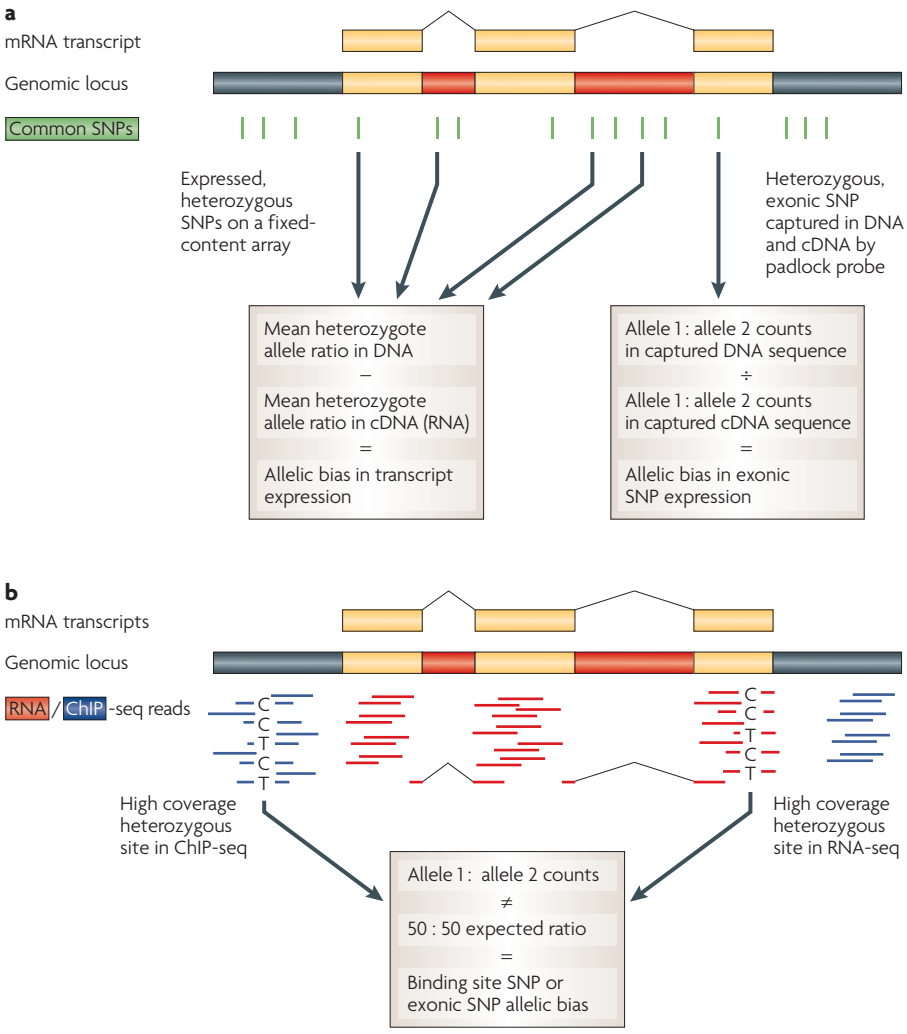


Figure 2 | **Targeted and global approaches to study allele-specific function. a** | Polymorphism-directed approaches can integrate data from SNPs (green bars) in unspliced primary transcripts[4] or target specific exons (yellow)[11,12].When primary transcripts are studied (left side), quantitative measurements are taken for multiple phased polymorphisms in the same transcript (mean allele ratio in cDNA) and normalized to genomic DNA from the same sample (mean allele ratio in DNA). In the targeted method (right side), highly informative sites (such as common variants in exons) are captured and digital quantification by next-generation sequencing (NGS) of the polymorphic bases (allele 1 and allele 2) at heterozygous sites is carried out in parallel for cDNA and genomic DNA. Biased allelic expression is defined by differences in allele counts (using the chi-square test). The advantages of polymorphism-directed approaches are that the data provide the maximum information about allelic effects and allelic biases detected in RNA are normalized to genomic DNA. **b** | NGS-based functional assays such as RNA-seq[15–17] and ChIP-seq[20] can be used to interrogate allelic effects when reads overlap a site that provides a high-quality heterozygote call (that is, a site with a polymorphism confirmed by a high coverage of reads). The ratio of reads from each allele (allele 1: allele 2) is calculated. Allelic bias at such sites is determined if this ratio deviates from the expected 50: 50, based on binomial statistics. Notably, given the average heterozygosity in the human genome, only a small proportion of the data generated will be informative for allelic biases. As polymorphic sites are not specifically targeted, the depth of sequencing required to obtain allelic resolution may be high. Experimental biases (such as technical causes for unequal allele counts) require careful consideration as the genomic DNA sample is not assayed in parallel with the cDNA and data cannot be normalized to correct for these biases.

Understanding allelic biases

Prevalence of allelic biases. Differential allelic expression has been widely studied. On a per-sample basis, similar estimates of the proportion of SNPs (at approximately 10% false-discovery rate based on biological replicates) or loci that show differential allelic expression are emerging from SNP-targeted studies: 20% of sites show 1.5-fold[12] and 30% of measured transcripts show 1.2-fold[4] difference between alleles. The reported per-sample allelic bias rates from RNA-seq[15] or ChIP-seq data[20] are presently lower, and range from 7 to 11% of sites measured. One likely reason for the discrepancy is the relatively low power of the current non-targeted NGS studies to detect subtle allelic differences, as discussed above. However, undetected technical biases in targeted studies might also contribute. Overall, the data to date suggests that at least 20% of transcripts show measureable allelic differences in regulation within one cell type, but much of the variation is subtle and therefore it may be difficult to ascertain biologically significant phenotypic differences, even at a cellular level. At the population level, in 53 unrelated immortalized lymphoblastoid cell lines, up to half of the expressed loci show allelic differences that were detected in over 10% of samples studied[4].

Determinants of allelic biases. Twin studies[23] and the much higher correlation of allelic expression within independent cell lines from the same individual[11,12] compared to allelic expression between individuals suggest a genetic basis for allelic expression differences. Furthermore, a positive correlation between eQTL alleles detected by RNA-seq and allelic biases in corresponding informative polymorphisms are observable both in lymphoblasts from Yoruban African[17] and Caucasian (CEU)[16] individuals. The latter study noted that many of the significant differences in allelic expression were observed in genes that were not detected to have eQTLs. It was suggested that in some cases this allelic variation may be explained by rare variants because the lengths of haplotypes shared by these loci were larger than expected[16]. In phased CEU lymphoblasts, when differential allelic expression was used as a quantitative trait and mapped to local common variants (those with a minor allele frequency of 10% or greater), up to one-third of expressed transcripts showed significant association with the local common variants[4]. Furthermore, 90% of these mapped associations showed Mendelian segregation of allelic expression in independent three-generation pedigrees, which provides strong evidence for heritability of differences in allelic expression. Finally, ChIP-seq data showed that there was preferable transmission of allelic biases in CTCF binding from parents to offspring[20]. Therefore, these studies highlight *cis*-regulatory sequence variation as the most common tractable mechanism for functional allelic biases.

Allelic differences in expression that depend on the parent of origin of the allele — which suggest imprinting — accounted for less than 1% of population variation in *cis* regulation in lymphoblasts[4]. Also, in a study of human lymphoblasts, features consistent with imprinting were found in <2% of sites that showed allelic biases in chromatin activity[20]. It should be noted that random monoallelic expression[6] has been detected in individually derived clonal cell lines and suggested to be a confounder for all allelic expression phenomena in immortalized cells[24]. However, attempts to control for this 'epigenetic noise' did not substantially alter the results of mapping common *cis*-regulatory SNPs (rSNPs) by allelic expression[4]. Nevertheless, an excess of extreme allelic expression is observable in immortalized cells versus primary cells[4], suggesting that some stochastic phenomena induced in cell culture may account for a small fraction of detected allelic biases.

Tissue specificity of allelic expression. The percentage of *cis*-eQTLs that show tissue specificity has been suggested to be 50–90%[25,26]. By contrast, when cells from the same individual are reprogrammed using induced pluripotent stem cell technology and differences in *cis* regulation are directly monitored by targeted allelic expression tests at different stages of reprogramming, only 10% of loci substantially alter their allelic expression[11]. Similarly, heritable allelic expression traits detected in lymphoblasts predicted 40–70% of the variance in allelic expression when the same loci were analysed in primary cells from the mesenchymal lineage[4]. Overall, the higher estimates of the tissue independence of *cis*-regulatory variation that have been observed by monitoring allelic expression may indicate that *cis*-eQTL mapping is more sensitive than allelic methods to broader changes — such as differences in overall expression levels between tissues — that reflect environmental influences. Indeed, if less stringent cut offs for *cis*-eQTL significance are used[24], the estimates of tissue-independent heritable variation obtained from the two methods converge[4] and at least half of *cis*-regulatory SNPs show effects in multiple cell types.

Further characterization of tissue-specific genetic variation in the context of differences in chromatin structure and the overall regulatory patterns of the genes involved are now needed. This work will be important to understand how to optimally correlate disease-associated genomic regions with functional genomic data.

Allelic chromatin states and expression. Chromatin states predict gene expression patterns and, consequently, allelic differences in chromatin would be expected to yield changes in allelic expression status. This general assumption seems to hold true for sequence-specific changes in DNA–protein interactions[8] and for sequence-specific alterations in DNA methylation[9]. In a genome-wide study, SNPs were recently identified that correlate with population variation in DNA–protein interactions (assessed by ChIP-seq), which also showed

Glossary

Clonal read
NGS produces, in principle, independent reads from each molecule in the input sample. However, in some cases, the amplification of molecules yields copies of the same short read, which can potentially bias allelic read counts.

DNase I hypersensitivity
The susceptibility of a genomic region to digestion by DNase I. Promoter, enhancer and other active regulatory DNA sequences are more easily digested than inactive non-coding sequences.

Expression quantitative trait locus
A locus at which gene expression variance in a population —typically measured by microarrays or by RNA-seq — correlates significantly with genotype. This locus can be near the measured gene (*cis*) or elsewhere in the genome (*trans*).

Histone modification
Regulatory elements in actively transcribed versus repressed loci have differences in post-translational modifications (for example, methylation or acetylation of lysines) of histones that can be identified by ChIP using modification-specific antibodies.

Padlock probe
An oligonucleotide with 5′ and 3′ sequences that are specific for target regions of the genome and are separated by generic sequence. After binding to its target, the probe can be circularized by ligase, and the generic sequence portion is used to amplify or capture the probes.

Standing genetic variation
Allelic variation that is currently segregating within a population; as opposed to alleles that appear by new mutation events.

that *cis* variation influences gene expression and regulatory DNA activity across the genome[27]. Allelic expression traits that map to functional DNA elements have also revealed allelic differences in chromatin activity (assayed by nuclease sensitivity) and DNA–protein binding[28]. More comprehensive combination of allele-specific tools is now required to establish how often and what types of allelic chromatin states predict differential expression of alleles.

Allelic differences and disease SNPs

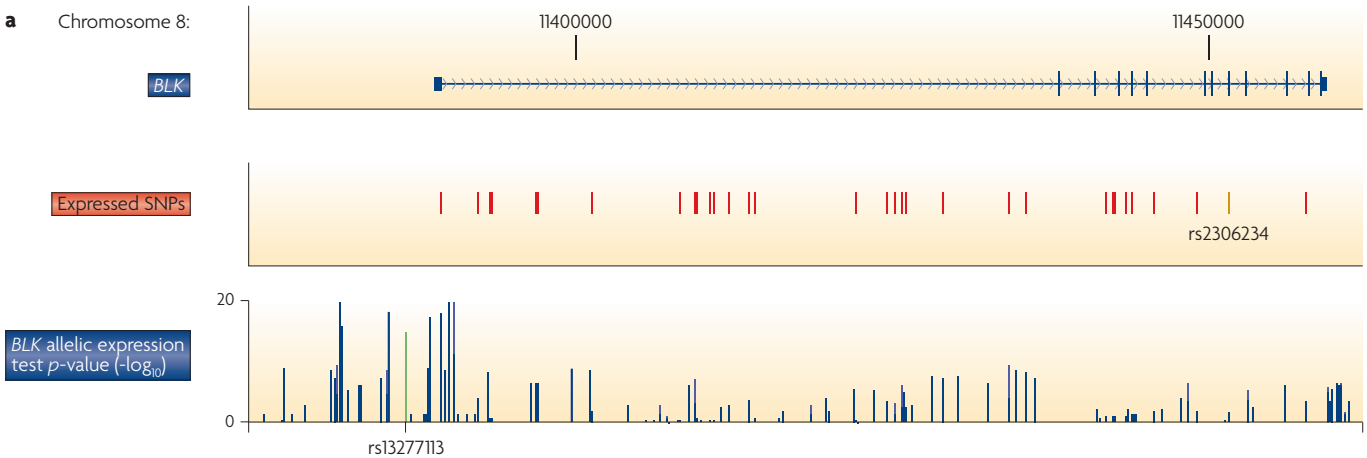
Disease alleles that map to non-coding DNA are often assumed to alter gene regulation in *cis*. Analysing the intersection of independent allelic expression-mapping and disease association data (FIG. 3a) allows estimation of how well the functional data and disease traits are correlated[4,28]. Furthermore, this approach can establish whether there is global enrichment of disease alleles among *cis*-regulatory variants[4], which can be used to evaluate the appropriateness of a tissue for studying a specific disease. Cell-based functional traits also have higher effect sizes than disease traits so cell-based traits might allow more robust isolation (functional fine mapping) of putative causal alleles[28]. Alternatively, links between disease-associated SNPs and their allele-specific function can be made by hypothesis-driven tests for expression[29,30] or chromatin activity[31] (FIG. 3b).

The obvious advantages of gathering independent mapping data for functional and disease traits may be offset by the need to have phased genotype data from the cells used in allelic expression tests[32]. However, alternative methods to establish statistical correlations between allelic function and SNPs in unphased allelic expression data are emerging[29]. At present, the data are scarce for estimating the global use of allelic functional assays in the dissection of disease, although efforts so far have shown promise.

Conclusions and future directions

Functional genomic assays based on NGS technologies that allow read-outs at single nucleotide resolution are now routinely yielding allele-specific data. Consequently, in the short-term, the benefits of these internally controlled assays for understanding human genome function will be widely evaluated. The current data suggest that functional allelic effects are frequent, interrelated and could allow disease associations to be translated to molecular mechanisms.

There are several considerations for the optimization of future study designs. First, the depth of sequencing required for allelic analyses is much higher than that required to observe total expression or DNA–protein-binding events, and alignment methodologies need to be unbiased by the reference genome. Therefore, high-coverage parallel genomic sequence data will eventually be needed to ensure the capture of extreme deviations in allelic biases and also to allow their fine mapping in populations. Second, different cellular lineages should be sampled, for example, by taking advantage of adult cell-reprogramming methods to conclusively establish the effect of the epigenetic environment on genetic *cis* regulation. In addition, collection of data from family-based samples will allow the straightforward



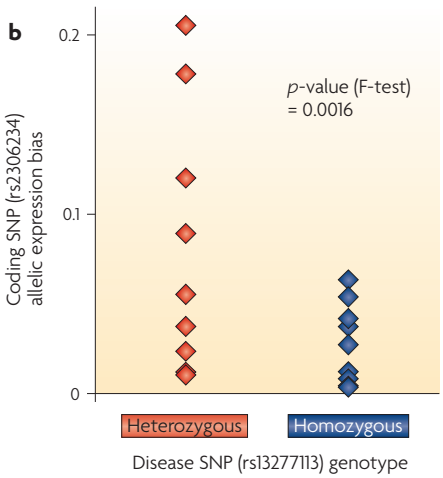


Figure 3 | ***Cis*-regulatory SNP mapping by allelic expression. a** | In a population of CEU (Caucasian) lymphoblastoid cells, allelic expression was mapped across the B lymphoid tyrosine kinase (*BLK*) gene by using the average allelic bias detected at expressed intronic and exonic SNPs (red track) (data from Reference[4]). Allelic expression in phased population samples was used as a quantitative trait and was correlated with local SNP genotypes using a regression test. This yielded very strong ($p = 1 \times 10^{-20}$) association of primary transcript allelic variation to the 5′ proximal region of *BLK* (blue track). The region of strongest *cis*-(regulatory SNP) rSNP association also contains the top association for systemic lupus erythematosus susceptibility[33] (rs13277113 highlighted in green). **b** | Fogarty *et al.*[29] provided several tests to apply in cases in which individual SNPs in coding regions are tested for allelic expression and the goal is to understand whether disease-linked SNPs alter *cis* regulation. Applying the F-test[29] using coding SNP allelic expression data provided by Ge *et al.*[4] in a distal exon in the *BLK* gene (highlighted in orange in part **a**), the heterozygotes for the disease SNP showed significantly more allelic expression variance compared with homozygotes. This type of allelic expression test does not require high-density allelic expression data nor the ability to accurately phase genotypes but, compared to a transcript-mapping approach[4], its power to detect differences is lower as the variance of single allelic expression measurements can be high owing to non-biological reasons and only a subset of samples are informative for allelic expression. Part **a** modified from *Nature Genetics* REF. 4 © (2009) Macmillan Publishers Ltd.

correlation of *cis*-rSNPs with allelic functional data. The parallel collection of allelic data at multiple levels of genome function is also required. This should include the analysis of sequence-specific DNA–protein interactions, general chromatin activity (such as histone modifications and DNase I hypersensitivity) and gene expression to allow the simultaneous interpretation of the mechanistic basis and consequences of allelic differences. In addition, higher-order regulatory organization (for example, chromatin conformation) could, in future, be queried at allelic resolution. The development of longer NGS read-lengths would allow more comprehensive assessment of relative transcript isoform differences, which require the simultaneous identification of a polymorphic base and altered mRNA structure. Finally, larger sample sizes would allow examination beyond common variants and maximize the use of standing genetic variation in understanding genome function. Undoubtedly, progress towards these goals will play an important part in understanding the 'genetic code' of non-coding DNA and its role in phenotypic variation among humans.

Departments of Human and Medical Genetics, McGill University and Génome Québec Innovation Centre, 740 Dr. Penfield Avenue, Room 6202, Montréal, Québec, Canada, H3A 1A4.
e-mail: tomi.pastinen@mcgill.ca

doi:10.1038/nrg2815
Published online 22 June 2010

1. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nature Rev. Genet.* **10**, 184–194 (2009).
2. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297**, 1143 (2002).
3. Knight, J. C., Keating, B. J., Rockett, K. A. & Kwiatkowski, D. P. *In vivo* characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nature Genet.* **33**, 469–475 (2003).
4. Ge, B. *et al.* Global patterns of *cis* variation in human cells revealed by high-density allelic expression analysis. *Nature Genet.* **41**, 1216–1222 (2009).
5. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
6. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136–1140 (2007).
7. Milani, L. *et al.* Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome Res.* **19**, 1–11 (2009).
8. Maynard, N. D., Chen, J., Stuart, R. K., Fan, J. B. & Ren, B. Genome-wide mapping of allele-specific protein–DNA interactions in human cells. *Nature Methods* **5**, 307–309 (2008).
9. Kerkel, K. *et al.* Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nature Genet.* **40**, 904–908 (2008).
10. Gunderson, K. L. Whole-genome genotyping on bead arrays. *Methods Mol. Biol.* **529**, 197–213 (2009).
11. Lee, J. H. *et al.* A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet.* **5**, e1000718 (2009).
12. Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nature Methods* **6**, 613–618 (2009).
13. Verlaan, D. J. *et al.* Targeted screening of *cis*-regulatory variation in human haplotypes. *Genome Res.* **19**, 118–127 (2009).
14. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* 23 Apr 2010 (doi:10.1101/gr.104695.109).
15. Heap, G. A. *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Genet.* **19**, 122–134 (2010).
16. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
17. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
18. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in *cis* and *trans* gene regulation. *Nature* **430**, 85–88 (2004).
19. Ameur, A., Rada-Iglesias, A., Komorowski, J. & Wadelius, C. Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. *Nucleic Acids Res.* **37**, e85 (2009).
20. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239 (2010).
21. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
22. Fontanillas, P. *et al.* Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Mol. Ecol.* **19** (Suppl. 1), 212–227 (2010).
23. Cheung, V. G. *et al.* Monozygotic twins reveal germline contribution to allelic expression differences. *Am. J. Hum. Genet.* **82**, 1357–1360 (2008).
24. Plagnol, V. *et al.* Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS ONE* **3**, e2966 (2008).
25. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250 (2009).
26. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
27. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
28. Verlaan, D. J. *et al.* Allele-specific chromatin remodeling in the *ZPBP2/GSDMB/ORMDL3* locus associated with the risk of asthma and autoimmune disease. *Am. J. Hum. Genet.* **85**, 377–393 (2009).
29. Fogarty, M. P., Xiao, R., Prokunina-Olsson, L., Scott, L. J. & Mohlke, K. L. Allelic expression imbalance at high-density lipoprotein cholesterol locus *MMAB–MVK*. *Hum. Mol. Genet.* **19**, 1921–1929 (2010).
30. McCarroll, S. A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nature Genet.* **40**, 1107–1112 (2008).
31. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nature Genet.* **42**, 255–259 (2010).
32. Pastinen, T. *et al.* Mapping common regulatory variants to human haplotypes. *Hum. Mol. Genet.* **14**, 3963–3971 (2005).
33. Hom, G. *et al.* Association of systemic lupus erythematosus with *C8orf13–BLK* and *ITGAM–ITGAX*. *N. Engl. J. Med.* **358**, 900–909 (2008).

Acknowledgements
I thank T. Kwan for critical reading of the manuscript. T.P. holds a Canada Research Chair and is supported by grants from Genome Canada, Genome Quebec and the Canadian Institutes of Health Research.

Competing interests statement
The author declares no competing financial interests.