Review

Advances in the identification and analysis of allele-specific expression

Christopher G Bell and Stephan Beck

Address: Medical Genomics, University College London Cancer Institute, Huntley Street, London WC1E 6BT, UK.

Correspondence: Stephan Beck. Email: s.beck@ucl.ac.uk

Published: 29 May 2009

Genome Medicine 2009, **1**:56 (doi:10.1186/gm56)

The electronic version of this article is the complete one and can be found online at http://genomemedicine.com/content/1/5/56

Abstract

Allele-specific expression (ASE) is essential for normal development and many cellular processes but, if impaired, can result in disease. ASE is a feature of organisms with genomes consisting of more than one set of homologous chromosomes. The higher the number of chromosome sets (ploidy) per cell, the higher the potential complexity of ASE. Humans, for instance, are diploid (except germ cells, which are haploid), resulting in multiple possible expression states in time and space for each set of alleles. ASE is invoked and modulated by both genetic and epigenetic changes, affecting the underlying DNA sequence or chromatin of each allele, respectively. Although numerous methods have been developed to assay ASE, they usually require RNA to be available and are dependent upon genetic polymorphisms (such as single nucleotide polymorphisms (SNPs)) to differentiate between allelic transcripts. The rapid convergence to second-generation sequencing as the method of choice to examine genomic, epigenomic and transcriptomic data enables an integrated and more general approach to define and predict ASE, independent of SNPs. This 'Omni-Seq' approach has the potential to advance our understanding of the biology and pathophysiology of ASE-mediated processes by elucidating subtle combinatorial effects, leading to the accurate delineation of sub-phenotypes with consequential benefit for improved insight into disease etiology.

Allele-specific expression

The interrelationship between the haploid fractions of diploid (or polyploid) genomes and how they control a co-ordinated regulation of gene expression is still poorly understood. This is despite the fact that the contribution of expression variation to phenotypic diversity, adaptive evolution and disease susceptibility is well recognized [1]. It has been challenging, however, to identify the underlying mechanisms. For instance, only a small minority of single nucleotide polymorphisms (SNPs) identified from the recent plethora of genome-wide disease association studies involved protein-coding sequence changes. Most of the disease-associated SNPs were found within non-coding intronic or intergenic regions from where they are thought to operate on gene expression through *cis*-acting mechanisms [2].

Here, we will only consider diploidy, which is the situation found in all nucleated, somatic cells in humans, giving rise to five possible expression states for each set of alleles (Table 1). Expression states 1 and 2 refer to the situations where expression is either 'on' or 'off' for both alleles and, therefore, do not result in allele-specific expression (ASE).

Expression states 3 and 4 refer to the extreme ends of the ASE spectrum, leading to monoallelic expression caused by different mechanisms. The first of these is autosomal

Table 1

Possible expression states of allele sets in diploid genome

	Expression states				
	Non-ASE		$ASE_{i/xi}$		ASE_{Δ}
Allele	1	2	3	4	5
A	On	Off	On	Off	Δ
a	On	Off	Off	On	Δ

Expression states 1 and 2 refer to alleles that do not display allele-specific expression (ASE). Expression states 3 and 4 refer to alleles that do display ASE, for example, due to imprinting (ASE_i) and X-inactivation (ASE_{xi}). Expression state 5 refers to alleles that display differential ASE (ASE_{Δ}), due to currently unknown mechanism(s).

imprinting: this is a parent-of-origin specific action, where either the paternal or maternal allele has complete expression output, either within the entire body, specific tissue/cell types, specific developmental stages or only for a particular isoform [3]. Computational prediction suggests that our current knowledge of imprinted genes is an underestimate [4]. A second mechanism is X-inactivation, the random assignment and maintenance of a clonal lineage, whereby functional hemizygosity of the homogametic female genome is invoked for dosage critical genes on one X chromosome [5]. The third is an as yet unknown mechanism, resulting in wide-spread monoallelic expression of autosomal genes [6]. This involves the apparently stochastic choice of either allele to be expressed and was first recognized in a subset of immune and neurological genes, including those encoding odorant, T cell, and natural killer cell receptors, as well as immunoglobulin and interleukin genes. In a subsequent study, almost 10% of the 3,939 genes assessed were found to have one allele switched off [6]. In these cases, and contrary to X–inactivation, ASE was not stable within a clone lineage, as the allele that was expressed could alternate, and the choice of expression was made independently for each gene, not for a chromosome in its entirety. Although genes of diverse functions were involved, those encoding cell-surface proteins were over-represented, as well as those undergoing lineage-specific accelerated evolution. Due to the small sample size, however, some of these genes may in fact display differential rather than monoallelic expression [7].

Finally, expression state 5 refers to differential expression between the two alleles (ASE_{Δ}), which, arguably, is the most common ASE state and is discussed in more detail in the next section.

Methods for the identification of allele-specific expression

Although the monoallelic mechanism of imprinting was first identified in 1984 [8], quantitative variance in expression of the two different alleles was only first acknowledged in 2002, in a small study where 6 of only 13 genes investigated showed allelic differences [9]. Since these early studies drew attention to possible *cis*-regulatory effects causing ASE, additional individual loci were queried by PCR-based methods, such as real-time quantitative PCR or discrimination of PCR products by differing primer extension [10]. However, in order to identify and characterize this variation on a more genome-wide level, PCR techniques were coupled with microarray technology. Initially performed by Lo *et al*. [11] using an early Affymetrix HuSNP array with approximately 1,000 exonic SNPs in 602 genes, a surprisingly high estimate of >50% of genes showed some ASE pattern and the majority of these were not known to be imprinted.

Thereafter, many further studies have used this approach dependent upon heterozygous SNPs residing within the gene's coding region and, subsequently, compared ratios of copy DNA (cDNA) from RNA to quantify differences [12,13]. The direct measurement of both alleles within the same system removes the possible confounding influence of *trans*-acting environmental factors. This can identify plausible imprinted genes, which familial studies can verify. However, they may also be developmental-, time- or tissue-specific [13] or display ASE that will show co-segregation through a pedigree.

Two widely used commercially developed techniques for ASE analysis include the BeadArray platform and the Oligo Pool All (OPA) method (Illumina Inc.). In the BeadArray method, genomic and converted RNA are assessed for the ratio of each allele by primer extension assays with fluorescence-labeled allele-specific primers. The resolution allows a 1.5-fold ASE change to be detected robustly from experimental noise. This method was used by Serre *et al*. to estimate that approximately 20% of human genes display ASE [14]. The OPA method is based on the Golden Gate assay [15]. By excluding any SNP within within 45 base pairs (bp) of the start or end of exons, in order to ensure that there was an equivalent chance of working between genomic DNA and converted DNA, this method was used to investigate the unrelated 210 individuals within the HapMap population [16]. By exploring the interaction between non-synonymous SNPs and *cis*-regulatory features, this study estimated ASE to be approximately 18%.

The major issues with these aforementioned techniques are threefold, and various adaptations have been developed to overcome them. Firstly, the influence of bias in PCR amplification in these ASE examinations has been acknowledged and a custom ASE array has been developed that removes this possible confounding factor [17]. Secondly, issues of cross-hybridization were also reduced in the custom ASE array study by the use of longer probes (39 to 49 bp) and a new probe design. The use of shorter probes may have also contributed to a possible overestimation of

ASE in earlier studies. Probes were designed for the mismatch base to have a balanced T_m on either side, thus placing it at the most thermodynamically disruptive location. Thirdly, the necessity of a SNP to reside within the transcript limits the number of genes for which there are informative haplotypes. However, when multiple SNPs occur, robust results can be elucidated with consistent ratio differences, which is considerably aided if these SNPs are in strong linkage disequilibrium (LD) with each other. Verlaan *et al.* [18] have modified their method to investigate unspliced primary transcripts, thereby including the intronic regions and thus greatly increasing the number of polymorphisms that can be used to delineate allele calls. These results estimated that >10% of genes expressed in a lymphoblastoid cell line exhibited ASE. An alternative approach, independent of the transcription of coding region SNPs, is to use a marker of transcriptional activity (such as phosphorylated RNA polymerase II (Pol II)), as developed in the haploChIP method [19]. By immunoprecipitation of phosphorylated Pol II cross-linked to chromatin, the relative DNA fragment amounts of these protein-DNA interactions are differentiated by the use of any SNP within the location of interest.

Epigenetically driven allele-specific expression

The significant role of DNA methylation as a common driving factor in ASE has been substantiated by a number of studies, including a recent investigation of a cohort of pediatric patients with acute lymphoblastic leukaemia. In this study, DNA and RNA from both blood and bone marrow were analyzed and 16% were found to display ASE [7]. A direct quantitative correlation between ASE and CpG site methylation was observed within individual samples. The unequal epigenetic state of each haploid genome is clearly a major factor in the asymmetrical expression of the two alleles. Genome-wide methods to investigate the epigenome, including DNA methylation with methylated DNA immunoprecipitation (MeDIP) [20] and chromatin structure with chromatin-dependent immunoprecipitation (ChIP) [21] have enabled a primary view of these features.

The concept of specific epigenetic haplotypes or 'hepitypes' that may add additional power to phenotype-related studies was proposed by Murrell *et al.* [22] and initial support for this concept has come from the work of Kerkel *et al.* [23]. Although allele-specific methylation of CpG dinucleotides is a characteristic feature of imprinted loci, a genomic examination by methylation-sensitive restriction enzymes and subsequent analysis on microarray revealed regions where adjacent SNPs were influential [23]. The sequence, as opposed to parental origin (as would be expected in imprinted loci), was associated with methylation state in a dozen regions. Thus, a comparison of common haplotypes within linkage disequilibrium blocks of disease-association SNPs for their methylation status may well reveal whether the epigenetic state of these hepitypes links SNPs with expression and possible phenotype and/or disease susceptibility. However, the necessity for functional examination of *cis*-regulation in the appropriate tissue has also been highlighted by the observation that in approximately 50% of differing tissues the same haplotype has differing effects [18]. Furthermore, Zhang *et al.* [24] found that while DNA methylation levels were usually very comparable in different cells, ASE was not, reinforcing the possibility of environment-specific hepitype interactions.

Integrated approach for the identification of allele-specific expression

The recent combination and integration of more than one method of investigation has opened a novel route into ASE-related research [25]. Exploring the discriminatory power of different epigenetic states, Wen *et al.* [25] investigated if ASE (and imprinting) could be analyzed independently of SNPs. For that, they assayed the two parentally derived genomes for DNA methylation using MeDIP and the histone methylation mark H3 lysine-4 dimethylation (H3K4Me2) using ChIP. While DNA methylation (particularly at promoters) is associated with heterochromatic and transcriptionally inactive regions, the presence of H3K4Me2 is associated with euchromatic and actively transcribed regions of the genome. The authors hypothesized that the regions where DNA methylation and H3K4Me2 co-localized ('double hits') should be enriched for ASE due to imprinting (ASE_i) if the two marks were to map to separate parental chromosomes. Using custom tiling arrays enriched for imprinted and non-imprinted genes, respectively, they found that imprinted genes were enriched (more than fivefold) for 'double hits', which were frequently located at transcriptional start sites near antisense or alternative transcripts. If a third mark, CTCF binding sites (an insulator protein often associated with imprinted genes [26]), was assayed as well, known imprinted genes were enriched (>75-fold) by 'triple hits'. As expected, the 'double hits' (DNA methylation and H3K4Me2) were mapped to opposite alleles.

The obvious limiting factor of this approach is the dependence upon preselected regions present on the microarrays, which results in only a small fraction of the genome being evaluated. The coupling of the immuno-capturing techniques of MeDIP and ChIP with the power of massive parallel second-generation sequencing for MeDIP-Seq [27] and ChIP-Seq [28] is swiftly leading to the replacement of microarrays as the platform of choice for epigenome analysis, resulting in far improved resolution and coverage. The revolutionary possibilities of this technique have been rapidly seized upon with an exponential expansion in whole-genome investigations, including transcriptome analysis. Using RNA-Seq [29], all species of RNA transcripts, including coding RNAs and non-coding RNAs, such as microRNAs, Piwi-interacting RNAs, short interfering RNAs

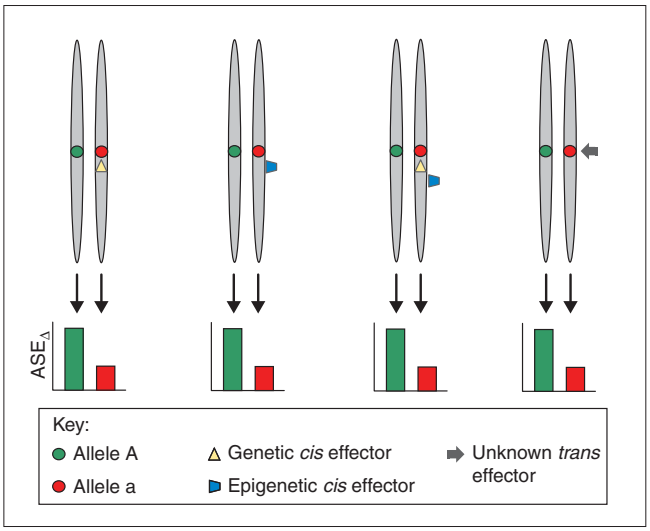

Figure 1
Mechanisms capable of invoking differential allele-specific expression (ASE$_\Delta$): genetically driven by *cis*-acting polymorphisms; epigenetically driven by *cis*-acting epigenetic effectors; hepitype driven by a combination of genetic and epigenetic effectors; and driven by as yet unknown *trans* effectors.

or large intervening non-coding RNAs, can be analyzed at single base pair resolution and a wide dynamic range of expression.

All these different but complementary techniques (MeDIP-Seq, ChIP-Seq and RNA-Seq) can now be conducted on one sequencing platform, facilitating a coordinated and integrated 'Omni-Seq' approach for genomic, epigenomic and expression analyses to be run in parallel. Such combinatorial analyses have the potential and power to delineate subtle modifications that may not be detectable by one technique alone. The challenges for bioinformatics are obviously daunting, with mammoth amounts of data currently being generated using just one method. However, this concern is being tackled and experience tells us that it is not insurmountable [30]. Looking back, solutions were eventually found for the exponential growth of data in the past decade. We predict that such a harmonized approach will be able to tease out the most informative combinatorial causes of ASE and will become the integrated method of biological examination, with as yet unforeseen benefits for human health prediction, prognostication and diagnosis.

Concluding remarks and outlook

ASE is clearly a common and highly complex phenomenon. Simplistically, ASE and ASE$_\Delta$ in particular, can be invoked by three *cis*-acting mechanisms (Figure 1): genetically driven, for example, via sequence variation; epigenetically driven, for example, via DNA methylation and/or chromatin modification; or (epi)genetically driven, for example, via hepitypes. In addition, there is evidence for as yet unknown *trans*-acting mechanisms effecting ASE$_\Delta$. Emerging technologies such as '4C' [31], which use chromosome confirmation capture in conjunction with microarrays or second-generation sequencing platforms, increasingly allow the ability to assay the three-dimensional nuclear organization for unbiased interaction of genomic sequences in *cis* as well as *trans*.

There can be little doubt that the advances in ASE discussed here will translate into clinical benefits in the not too distant future. In the area of diagnostics, for instance, the exploitation of ASE is already well underway for tumor type classification, evaluation of differential gene expressivity or penetrance in monogenic hereditary conditions and the discovery of currently unknown imprinted genes with respect to their roles in developmental syndromes. In addition, ASE can be expected to shed more light on the critical pathophysiology and epistatic factors in complex diseases. The exploitation of ASE for therapeutics may take a little longer, but certainly has tremendous potential, particularly when coupled with targeted RNA-based therapy [32]. In this context, ASE may enable the identification of critical disease-causing haplotypes, epitypes and hepitypes for down-regulation with targeted RNA interference-based therapies.

Thus, accurate delineation of an allele's genetic and epigenetic state, linked with knowledge of its transcriptional output, will undoubtedly improve our understanding of ASE and its wide-ranging implications for genome biology and medicine.

Abbreviations

ASE, allele specific expression; bp, base pair; ChIP, chromatin dependent immunoprecipitation; H3K4Me2, H3 lysine-4 dimethylation; LD, linkage disequilibrium; MeDIP, methylated DNA immunoprecipitation; OPA, oligo pool all method; Pol II, RNA polymerase II; SNP, single nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CB and SB wrote the review jointly.

Authors' information

CB is a Postdoctoral Researcher and Genetic Pathologist in Professor Beck's Medical Genomics group at the University College London Cancer Institute, and his research is focused on the Epigenomics of Common Diseases. SB is Professor of

Medical Genomics at University College London and is interested in the genomics and epigenomics of phenotypic plasticity in health and disease [33].

Acknowledgements

CB and SB were supported by the Wellcome Trust.

References

1. Knight JC: **Allele-specific gene expression uncovered.** *Trends Genet* 2004, **20:**113-116.
2. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
3. Morison IM, Ramsay JP, Spencer HG: **A census of mammalian imprinting.** *Trends Genet* 2005, **21:**457-465.
4. Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, Hartemink AJ: **Computational and experimental identification of novel human imprinted genes.** *Genome Res* 2007, **17:**1723-1730.
5. Chow JC, Yen Z, Ziesche SM, Brown CJ: **Silencing of the mammalian X chromosome.** *Annu Rev Genomics Hum Genet* 2005, **6:**69-92.
6. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A: **Widespread monoallelic expression on human autosomes.** *Science* 2007, **318:**1136-1140.
7. Milani L, Lundmark A, Nordlund J, Kiialainen A, Flaegstad T, Jonmundsson G, Kanerva J, Schmiegelow K, Gunderson KL, Lonnerholm G, Syvanen AC: **Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation.** *Genome Res* 2009, **19:**1-11.
8. McGrath J, Solter D: **Completion of mouse embryogenesis requires both the maternal and paternal genomes.** *Cell* 1984, **37:**179-183.
9. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW: **Allelic variation in human gene expression.** *Science* 2002, **297:**1143.
10. Bray NJ, Buckland PR, Owen MJ, O'Donovan MC: **Cis-acting variation in the expression of a high proportion of genes in human brain.** *Hum Genet* 2003, **113:**149-153.
11. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP: **Allelic variation in gene expression is common in the human genome.** *Genome Res* 2003, **13:**1855-1862.
12. Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA: **Analysis of allelic differential expression in human white blood cells.** *Genome Res* 2006, **16:**331-339.
13. Pollard KS, Serre D, Wang X, Tao H, Grundberg E, Hudson TJ, Clark AG, Frazer K: **A genome-wide approach to identifying novel-imprinted genes.** *Hum Genet* 2008, **122:**625-634.
14. Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T, Fan JB, Hudson TJ: **Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression.** *PLoS Genet* 2008, **4:**e1000006.
15. Fan JB, Chee MS, Gunderson KL: **Highly parallel genomic assays.** *Nat Rev Genet* 2006, **7:**632-644.
16. Dimas AS, Stranger BE, Beazley C, Finn RD, Ingle CE, Forrest MS, Ritchie ME, Deloukas P, Tavare S, Dermitzakis ET: **Modifier effects between regulatory and protein-coding variation.** *PLoS Genet* 2008, **4:**e1000244.
17. Bjornsson HT, Albert TJ, Ladd-Acosta CM, Green RD, Rongione MA, Middle CM, Irizarry RA, Broman KW, Feinberg AP: **SNP-specific array-based allele-specific expression analysis.** *Genome Res* 2008, **18:**771-779.
18. Verlaan DJ, Ge B, Grundberg E, Hoberman R, Lam KC, Koka V, Dias J, Gurd S, Martin NW, Mallmin H, Nilsson O, Harmsen E, Dewar K, Kwan T, Pastinen T: **Targeted screening of cis-regulatory variation in human haplotypes.** *Genome Res* 2009, **19:**118-127.
19. Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP: **In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading.** *Nat Genet* 2003, **33:**469-475.
20. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D: **Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.** *Nat Genet* 2005, **37:**853-862.
21. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290:**2306-2309.
22. Murrell A, Rakyan VK, Beck S: **From genome to epigenome.** *Hum Mol Genet* 2005, **14:**R3-R10.
23. Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, Li K, Murty VV, Schupf N, Vilain E, Morris M, Haghighi F, Tycko B: **Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation.** *Nat Genet* 2008, **40:**904-908.
24. Zhang Y, Rohde C, Tierling S, Jurkowski TP, Bock C, Santacruz D, Ragozin S, Reinhardt R, Groth M, Walter J, Jeltsch A: **DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution.** *PLoS Genet* 2009, **5:**e1000438.
25. Wen B, Wu H, Bjornsson H, Green RD, Irizarry R, Feinberg AP: **Overlapping euchromatin/heterochromatin- associated marks are enriched in imprinted gene regions and predict allele-specific modification.** *Genome Res* 2008, **18:**1806-1813.
26. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B: **Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.** *Cell* 2007, **128:**1231-1245.
27. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Backdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavare S, Beck S: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.** *Nat Biotechnol* 2008, **26:**779-785.
28. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448:**553-560.
29. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57-63.
30. Doctorow C: **Big data: welcome to the petacentre.** *Nature* 2008, **455:**16-21.
31. Ohlsson R, Gondor A: **The 4C technique: the 'Rosetta stone' for genome biology in 3D?** *Curr Opin Cell Biol* 2007, **19:**321-325.
32. Wood M, Yin H, McClorey G: **Modulating the expression of disease genes with RNA-based therapy.** *PLoS Genet* 2007, **3:**e109.
33. **Medical Genomics at University College London** [http://www.ucl.ac.uk/cancer/research-groups/medical-genomics/].