Human Molecular Genetics, 2010, Vol. 19, No. 1 **122–134**
doi:10.1093/hmg/ddp473
Advance Access published on October 13, 2009

Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing

Graham A. Heap[1], Jennie H.M. Yang[2], Kate Downes[2], Barry C. Healy[2], Karen A. Hunt[1], Nicholas Bockett[1], Lude Franke[1], Patrick C. Dubois[1], Charles A. Mein[3], Richard J. Dobson[3], Thomas J. Albert[4], Matthew J. Rodesch[4], David G. Clayton[2], John A. Todd[2], David A. van Heel[1,†] and Vincent Plagnol[2,*,†]

[1]Centre for Digestive Diseases, Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK, [2]Department of Medical Genetics, Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK, [3]Genome Centre, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK and [4]Roche NimbleGen, 504 S. Rosa Rd. Madison, WI 35393

Received May 14, 2009; Revised September 17, 2009; Accepted October 9, 2009

Many disease-associated variants identified by genome-wide association (GWA) studies are expected to regulate gene expression. Allele-specific expression (ASE) quantifies transcription from both haplotypes using individuals heterozygous at tested SNPs. We performed deep human transcriptome-wide resequencing (RNA-seq) for ASE analysis and expression quantitative trait locus discovery. We resequenced double poly(A)-selected RNA from primary CD4$^+$ T cells ($n = 4$ individuals, both activated and untreated conditions) and developed tools for paired-end RNA-seq alignment and ASE analysis. We generated an average of 20 million uniquely mapping 45 base reads per sample. We obtained sufficient read depth to test 1371 unique transcripts for ASE. Multiple biases inflate the false discovery rate which we estimate to be ~50% for random SNPs. However, after controlling for these biases and considering the subset of SNPs that pass HapMap QC, 4.6% of heterozygous SNP-sample pairs show evidence of imbalance ($P < 0.001$). We validated four findings by both bacterial cloning and Sanger sequencing assays. We also found convincing evidence for allelic imbalance at multiple reporter exonic SNPs in *CD6* for two samples heterozygous at the multiple sclerosis-associated variant rs17824933, linking GWA findings with variation in gene expression. Finally, we show in CD4$^+$ T cells from a further individual that high-throughput sequencing of genomic DNA and RNA-seq following enrichment for targeted gene sequences by sequence capture methods offers an unbiased means to increase the read depth for transcripts of interest, and therefore a method to investigate the regulatory role of many disease-associated genetic variants.

INTRODUCTION

Genome-wide association (GWA) studies using single nucleotide polymorphism (SNP) maps have revolutionized the mapping of common genetic loci determining susceptibility to a wide range of common, multifactorial disorders (1), in particular autoimmune diseases (2). The next steps to follow up on these findings are the identification of particular

*To whom correspondence should be addressed. Tel: +44 1223762107; Fax: +44 1223762102; Email: vincent.plagnol@cimr.cam.ac.uk
†These authors contributed equally to this work.

candidate variants and haplotypes, and the investigation of the molecular effects of these genetic variants. Because current evidence suggests that only a small fraction of the causal loci consists of variants (non-synonymous SNPs, copy-number variants or indels) directly affecting the protein amino-acid sequence, we expect a large fraction of the loci to have a regulatory role on gene expression via effects on transcription, message stability and splicing. To investigate the potential effects of candidate causal variants and haplotypes on gene regulation researchers have been correlating SNPs with inherited gene expression, known as expression quantitative trait loci (eQTLs). The combination of genome-wide genotyping with quantification of mRNA transcripts using microarray technology in sufficiently large cohorts has already demonstrated the widespread presence of eQTLs in the human genome (3–7). Most of these studies (3–5), however, used lymphoblastoid cell lines immortalized using Epstein Barr Virus and relied on observing differences between individuals despite the large inter-individual variability of gene expression measurements that is not explained by *cis* genetic variation, in addition to the limited accuracy of hybridization-based gene expression assays. This high variability generated by environmental factors and additional non-measured genetic or epigenetic variability significantly reduces the statistical power for eQTL discovery. Therefore, measurement of expression levels across multiple individuals may be so noisy that reliable correlations between SNP alleles and gene expression levels cannot always be demonstrated when the difference of expression between haplotypes is small (less than 1.3 fold). Moreover, cell lines may not be representative of *in vivo* biology and may introduce even greater variability (8–10), and, therefore, gene expression analyses using purified primary cell populations are urgently required (6–7,11).

An alternative experimental design well suited to address these limitations is allele-specific expression (ASE) analysis. This approach quantifies (un)equal transcription (or splicing) from the two alleles or haplotypes using RNA samples from individuals who are heterozygous at the eQTL SNP of interest. The elegant ASE approach has the major advantage of assessing expression within an individual rather than across subjects thereby avoiding major sources of error and variation.

In parallel, recent advances in high-throughput resequencing technologies have enabled highly quantitative sequencing-based analysis of human transcriptomes [RNA-seq (10,12,13)]. Because these techniques separately resequence both haplotypes, they have the potential to be used for the quantification of allelic imbalance, provided that a heterozygous SNP which can be used as a marker for each haplotype exists in the transcript of interest. The potential of this method for ASE analysis has been demonstrated in pooled cDNA samples and human cell lines (14). Here, we extend this approach to eight independently sequenced human poly(A)-selected transcriptomes obtained from primary cells from healthy donors using high-throughput paired-end (PE) resequencing. In the context of the recently identified shared pathways between multiple autoimmune disorders (2,15) that motivated this study, many of the most relevant genes in regions identified by GWA studies are immune genes that are highly expressed in $CD4^+$ T cells. This observation suggested the use of primary $CD4^+$ T cells in the current ASE study, thus illustrating the potential of this approach to identify regulatory effects in purified primary cell subsets.

RESULTS

Data description

We used Illumina Genome Analyzer II (GAII) high-throughput resequencing of cDNA libraries obtained from poly(A)-purified mRNA from four individuals analysed under T-cell activation (stimulated) or unstimulated conditions (see Materials and Methods), resulting in a total of eight samples. We obtained 45 bp reads, the majority of them are paired end (i.e. containing reads from both the 3′ and 5′ end of a ~250 bp fragment, see Table 1), that were mapped to a transcriptome reference sequence set specifically constructed for PE RNA-seq (see Materials and Methods). This reference set includes a spliced transcript for each annotated gene (Ensembl CCDS), as well as additional sequences for introns and non-standard splice junctions. A full version of the reference gDNA genome with annotated gene regions masked was added to enable: capture of transcribed, but not annotated, chromosome regions; detect gDNA contamination of the mRNA preparation and importantly to allow assessment of repetitive sequence. Only reads mapping with high confidence to a unique location in our reference sequence set (defined as quality reads, see Materials and Methods) were included in this study.

Owing to the complex nature of the RNA-seq reference genome, taking advantage of PE sequence reads relies on the ability of the mapping algorithm to map 'chimeric' fragments: for example, the first read of a pair may map to a non-standard exon–exon junction sequence and the second read to the main spliced transcript. An algorithm implementing this feature was provided by the novoalign (www.novocraft.com) software package, which we used to align resequencing reads to our reference set. Another useful feature provided by novoalign is the ability to set a lower penalty for alignment when a 'chimeric' paired read maps to two sequences that are part of the same gene.

Transcript coverage

In the absence of experimental biases, the ability to detect an allelic imbalance using ASE depends on two parameters: the strength of the allelic imbalance and the read depth at the reporter heterozygous SNP. We analytically computed the read depth required to demonstrate allelic imbalance for different allelic ratios. Power calculations (Fig. 1) show that for a read depth of 50 and a 67:33 allelic imbalance, which corresponds to an average of one cycle difference in a qPCR experiment between individuals homozygous at both alleles (a two fold difference), the probability to observe a *P*-value more significant than 0.001 is 19% (Fig. 1). Therefore, to remove SNPs providing almost no power to detect allelic imbalance, we only tested for ASE SNPs with read depth ≥ 50.

While this approach is not limited to previously known SNPs, we first tested 589 673 dbSNPs for ASE (obtained from Ensembl release 52) and located in annotated spliced

Table 1. Number of 45 bp quality reads (after filtering out low mapping score and clonal reads, see Materials and Methods), heterozygous SNPs and number of heterozygous and imbalanced SNPs (at $P < 0.001$) for each sample

	Number of quality 45 bp reads			Number of loci with depth >50 at any dbSNP		Number of loci with depth >50 at a heterozygous SNP		Number of imbalanced loci ($P < 0.001$) with read depth >50	
	Total	In pairs	Single	*N* genes	*N* SNPs	*N* genes	*N* SNPs	*N* genes	*N* SNPs
Individual 1, stimulated	15 151 170	15 151 170	0	2245	11 976	379	559	42	51
Individual 1, unstimulated	9 498 415	9 498 415	0	1089	6514	176	260	14	15
Individual 2, stimulated	35 964 198	24 244 978	11 719 220	4091	24 239	831	1348	80	114
Individual 2, unstimulated	20 722 639	20 722 639	0	2432	10 574	383	579	41	67
Individual 3, stimulated	16 336 921	16 336 921	0	1952	11 113	308	465	15	18
Individual 3, unstimulated	18 172 107	13 704 083	4 468 024	2030	12 491	379	619	28	28
Individual 4, stimulated	16 400 802	11 736 947	4 663 855	2067	12 178	351	536	37	39
Individual 4, unstimulated	15 240 509	10 688 635	4 551 874	2043	11 561	350	563	32	38

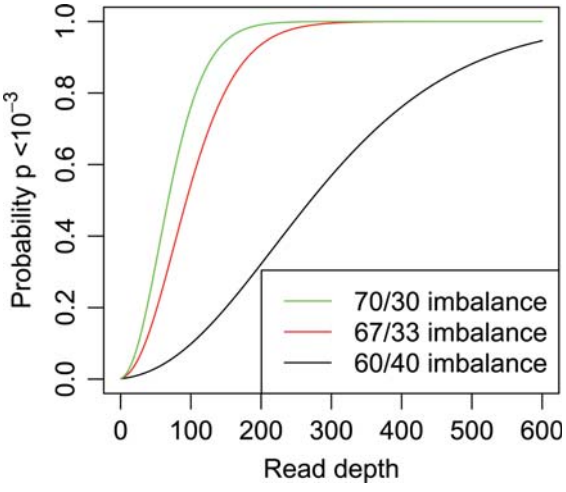

Figure 1. Probability to detect an allelic imbalance at $P < 0.001$ as a function of the read depth at a single SNP. We considered three levels of allelic imbalance: 60/40, 67/33 and 70/30. The 67/33 scenario corresponds to a two-fold difference in expression level, i.e. an average of one cycle difference between individuals homozygous at both alleles in a qPCR experiment.

transcripts (including 5′/3′ untranslated regions) in order to limit biases associated with simultaneous SNP discovery and ASE testing. We also tested for ASE 4 282 208 intronic dbSNPs but owing to the use of poly(A)-selected mRNA the vast majority of the SNPs with sufficient read depth were located in spliced transcripts (96.4%). Heterozygous SNPs with sufficient read depth were located in highly expressed transcripts and we obtained sufficient read depth to test 4929 pairs of heterozygous SNPs/samples for allelic imbalance. Grouping these SNPs by transcripts for each of the eight samples provided 3107 pairs transcripts/samples with sufficient coverage for ASE analysis. When counting each SNP uniquely, we had sufficient data to test 1371 transcripts and 2701 SNPs for ASE. In each individual, the number of testable transcripts (i.e. containing at least one heterozygous SNP with sufficient read depth) was proportional to the number of mapped reads (Fig. 2, $R^2 = 0.91$).

Overall distribution of allelic imbalance

When considering the overall distribution of the test statistic, we found evidence of widespread departure from 1:1 allelic ratio. Summing over the eight samples, we tested a total of 4929 pairs of heterozygous SNPs/samples with sufficient read depth (>50), and 370 SNPs (7.5%) showed evidence of allelic imbalance at $P < 0.001$ (Table 1). At this significance level, the estimated theoretical false discovery rate (FDR) is ∼1%. However, we have identified several biases inflating the false positive rate (see below for a description). Since SNPs passing HapMap quality filters are less likely to be affected by some of these biases (for example, indels close to an SNP will influence primer/probe hybridization), the ASE rate for HapMap SNPs is a more reliable estimate of the true rate of ASE. Indeed, when we restricted the analysis to the subset of 87 796 dbSNPs in spliced transcripts that passed HapMap quality filters, the ASE rate was significantly lower: 4.6% of heterozygous HapMap SNPs/sample pairs

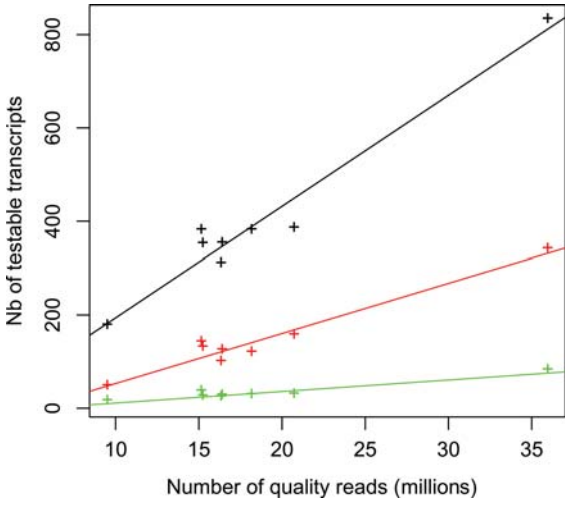

Figure 2. Number of transcripts containing at least one heterozygous dbSNP with read depth 50 (black), 100 (red) and 250 (green) as a function of the number of quality 45 bp reads. Each point represents one condition/individual sample. To provide a more intuitive reference, a mean 1× read-depth across the human genomic DNA requires 65 million 45 bp reads.

were ASE positive (115 out of 2478 heterozygous HapMap SNPs with sufficient read depth, $P = 2.47 \times 10^{-14}$). However, our statistical power to detect allelic imbalance is limited by read depth and, therefore, this estimate of 4.6% is a lower bound. Indeed, when restricting this analysis to HapMap SNPs with read depth >100, we found a higher ASE rate of 7.5% (66 of 878). A complete list of all the heterozygous dbSNPs with sufficient read depth for ASE testing is provided in Supplementary material, Table S1.

Genotyping data and epigenetic effects

Our approach consisting of identifying heterozygous SNPs using solely the RNA-seq data restricts our ASE analysis to SNPs for which the observed imbalance level does not exceed 15–85%, the threshold beyond which SNPs are called homozygous. With the double purpose of confirming RNA-seq calls and identifying heterozygous SNPs completely imbalanced in the RNA-seq data, we therefore genotyped the four individuals using the Illumina Quad660W BeadChip. We lowered to 20 the minimum read depth required to call SNPs in the RNA-seq data and identified 9727 pairs of SNP/samples shared between RNA-seq and Illumina Quad660W genotyping chip. Out of these 9727 calls, 6886 are homozygous based on the genotyping chip data. Out of these 6886 calls, only one call is inconsistently called heterozygous in RNA-seq (rs7484182 in individual 2) and visual inspection of the genotyping chip clustering plots showed poor quality data consistent with a genotyping error.

Conversely, out of the 2841 genotyping chip heterozygous calls, 14 are homozygous based on the RNA-seq data. Four of these 14 inconsistent calls were located in the transcript *SNRPN*, a known parentally imprinted gene, thus explaining its monoallelic expression. Three calls were located in *ERAP2*, a gene with known complete *cis*-acting differential allelic control [but not parentally imprinted, see (16)]. After removing these, 7 of 2841 calls (0.24%) located in seven different transcripts remain inconsistent. For four of these calls (located in *IQGAP1*, *TMBIM4*, *CALHM2*, *AL031281.6*), the read depth was <30 and we observed low levels of expression of the alternate allele. This is consistent with either highly skewed allelic expression or simply the result of random sampling of both alleles at low read depth. The remaining three calls (located in *CD44*, *VAMP1* and *NAPRT1*) may be real, or alternatively the result of unexplained artefacts either in the RNA-seq or the genotyping data.

Single locus validation

In order to confirm that some of our findings are not the consequence of technical biases, we selected four pairs of HapMap SNPs/individuals and validated them using two different locus-specific assays [clone-based allele-specific expression (C-BASE) (17) and PeakPicker (18)]. The C-BASE ASE assay relies on amplification by PCR of cDNA or genomic DNA fragment containing the reporter heterozygous SNP, followed by ligation into a vector and transformation into bacterial competent cells. Bacterial colonies are genotyped and both SNP alleles are counted, generating allelic counts reflecting the relative proportion of the two alleles *in vivo*. The PeakPicker ASE method uses sequencing trace data to quantify the relative fluorescent intensity for both alleles at heterozygous SNPs. The sequencing trace fluorescent intensities are summarized by the ratio of peak heights for both alleles.

We selected four combinations of HapMap SNPs/samples showing convincing evidence of ASE in the transcriptome resequencing data for replication using the locus-specific approach. All four initial RNA-seq results replicated and PeakPicker/C-BASE provided consistent results (Table 2 and Supplementary material, Table S2 for details of replication assay). Note that because we have observed for C-BASE assay allelic bias even using genomic DNA level as the PCR template (see, for example, rs1060819 in Table 2), the C-BASE ASE test uses a 1df goodness-of-fit χ^2 test for equal cDNA/gDNA allelic ratio, thus using the gDNA allelic distribution as control. Using gDNA allelic ratio as an internal control for mRNA imbalance is a more robust design, but obtaining this gDNA control for high-throughput sequencing data is, currently, prohibitively expensive owing to the very large amount of high-throughput resequencing needed to obtain high read depth for gDNA using whole genomes. This limitation points to future modifications of the approach by using sequence capture in order to enrich the cDNA for target sequences and to increase read depth for genomic DNA and RNA-seq (see below).

ASE analysis of disease-associated genes

The subset of primary cells used in this study ($CD4^+$ T cells) was chosen for its relevance to autoimmunity. Therefore, we reviewed the literature and identified 79 genes previously associated with autoimmune disorders (2,19). To maximize the number of testable heterozygous SNPs, we included in this analysis new heterozygous SNPs not listed in the Ensembl database but discovered by analysing our transcriptome

Table 2. Comparison of allelic imbalance between mRNA transcriptome resequencing (RNA-seq) and two validation assays: locus-specific bacterial cloning (C-BASE) and PeakPicker. Both validation assays tested total RNA and genomic DNA. Allele 1 indicates the allele in the reference genome. RNA-seq *P*-values use a χ^2 goodness-of-fit test for a 50−50 allelic ratio. For the C-BASE validation assay *P*-values test for equal allelic ratio in mRNA and gDNA using a 1df χ^2 goodness-of-fit test.

Sample	Gene	SNP	mRNA resequencing (RNA-seq)			Validation: C-BASE					Validation: PeakPicker	
			mRNA Counts			Total RNA Counts		gDNA counts			Ratio of normalized peak heights, allele 1 by allele 2	
			Allele 1	Allele 2	*P*	Allele 1	Allele 2	Allele 1	Allele 2	*P* (gDNA by total RNA)	Total RNA	gDNA
Individual 1, stimulated	*FAM118A*	rs2064068 (G/A)	82	23	8.5×10^{-9}	229	101	186	150	2.5×10^{-4}	1.91	0.95
Individual 4, stimulated	*FAM118A*	rs2064068 (G/A)	73	25	1.2×10^{-6}	211	95	168	166	2.4×10^{-6}	1.76	0.97
Individual 4, stimulated	*CALM3*	rs10405893 (A/G)	149	86	4×10^{-5}	202	117	176	167	2×10^{-3}	1.54	1.08
Individual 4, stimulated	*ATHL1*	rs1060819 (T/C)	82	223	6.8×10^{-16}	95	232	196	145	1.3×10^{-13}	0.43	0.86

resequencing data. In these 79 genes (Supplementary material, Table S3), we found 61 heterozygous SNPs with read depth >50 and eight of them were not listed in dbSNP (13.1%). These 61 heterozygous SNPs were found in 22 genes, i.e. 28% of the targeted genes, an elevated rate consistent with a higher expression level of these immune genes in $CD4^+$ T cells. Counting separately each pair of SNP/individual, we tested a total of 127 pairs and 13 were imbalanced at $P < 0.001$. After controlling for the elevated read depth, the frequency of ASE positive pairs of SNP/sample in these 79 transcripts was consistent with genome-wide estimates ($P > 0.05$).

Firstly, 8 of 13 imbalanced SNPs were located in the multiple sclerosis-associated gene *CD6* (20). In individual 2 (stimulated and unstimulated), two heterozygous exonic SNPs (HapMap SNP rs2074225 and rs11230562, separated by 97 bp in the spliced transcript) showed strong level of imbalance (allelic counts 253/142 at rs11230562 and 237/170 at rs2074225). Both SNPs were sufficiently close to obtain phase information from the RNA-seq 45 bp PE reads (using novoPhase, see Materials and Methods) and these data confirmed that the same haplotype was over-represented at both SNPs. Note that a third SNP (rs1050922) in individual 2 had 1:1 allelic ratio but also unexpectedly high read depth (two fold increase over rs2074225/rs11230562). This could be the consequence of reads from another transcript mapping to a homologous region in the CD6 transcript and we believe that the data at this SNP are artefactual. We also found three *CD6* heterozygous SNPs in individual 3 and these SNPs were sufficiently close to be phased using the RNA-seq paired reads. For all three SNPs, the pattern was consistent after phasing, the same haplotype being over-represented at all three SNPs (counts 176/155 at rs61899223, 95/56 at rs11230562 and 83/64 at ENSSNP10443844, where the first number relates to the same haplotype). For individual 4, we found four heterozygous SNPs (including two adjacent SNPs) but after phasing the data were inconsistent: different haplotypes were over-represented at different SNPs, suggesting the possibility of unidentified biases [e.g. alternative splicing at CD6 (21)]. Given the convincing evidence of association for individuals 2 and 3, we genotyped all individuals for the SNP currently strongly associated with multiple sclerosis in *CD6*, rs17824933 (20). Both individuals 2 and 3 (but not individuals 1 and 4) were heterozygous at this SNP. Taken together, these data suggest that the alteration of CD6 expression in $CD4^+$ T cells is a potential mechanism for the effect of the multiple sclerosis-associated variant in the *CD6* gene region. SNP phasing obtained from linkage disequilibrium information at HapMap SNPs rs2074225 and rs17824933 shows that the minor allele G of rs17824933, which is more common in multiple sclerosis patients than controls, is associated with lower CD6 expression. However, the biological interpretation is complex as resting CD6 expression inhibits T-cell activation, but CD166 ligand engagement of CD6 enhances T-cell activation (22).

Secondly, we observed an intriguing pattern at *ICOSLG*, in which five heterozygous SNPs, all not listed in dbSNPs, were identified in individual 2 and a strong skew was observed for four of them. Moreover, after phasing the data, we identified three distinct haplotypes instead of two, suggesting that the observed skew is artefactual. We scanned the database of genomic variants and found that, according to two independent studies (23,24), the entire gene is located in a copy number variable region. Our data suggest that this copy number variation is a duplication and that the duplicated gene is expressed: the additional observed haplotype in individual 2 probably originates from this duplicated copy. This additional copy is probably expressed at a lower level than the original transcript which would explain the imbalance in the RNA-seq data.

Lastly, we identified six heterozygous SNPs in the *IL21R* gene (25) for individual 1, two of them imbalanced at $P < 0.001$. We could phase three of six heterozygous SNPs (rs2285452, rs3093387 and rs2239928) using the RNA-seq data, and the trend was consistent with 1.3:1 allelic ratio. This finding is consistent with a previous report of a *IL21R* eQTL using mRNA from the peripheral blood of 110 celiac cases (7).

Controlling for sequencing chemistry biases

The Illumina GAII sequencing chemistry induces biases that can increase the rate of false positive ASE. A first issue is PCR biases over-amplifying identical cDNA fragments, the final allelic ratio being eventually affected by the presence of a large number of over-represented identical read clones. To control for these biases, we identified clonal molecules in the PE data using as a signature the mapping location for both reads in a pair (or if one read did not map uniquely, the first 10 bp of sequence was used as a signature). Identical mapping location for both 45 bp reads is unlikely to occur randomly, and for each set of clonal read pairs we only included a single read pair in the analysis. This filtering process removed on average 10% of the reads from the final analysis (option rmDup in novoPile). To ensure our ASE results are not affected by clonality, especially for reads mapping as single end (SE) data for which the filter described above is not applicable, we also verified that among the reads covering a heterozygous SNP the observed allele was independent of the position of the SNP within each 45 bp read (see Materials and Methods).

Secondly, we also observed biases that are specific to forward/reverse read direction. Therefore, we verified that for heterozygous SNPs the allelic ratio distribution was consistent for read counts obtained for the reverse and forward strands. Using a $P = 0.01$ threshold for both tests, we removed 18% of heterozygous dbSNPs from ASE testing.

Controlling for *in silico* mapping biases

In silico artefacts in the mapping procedure can also bias the allelic ratio distribution and increase the false positive rate. A general difficulty is the mapping of short reads to a standard reference sequence that does not match perfectly the genetic sequence of the sample. In particular, for heterozygous SNPs, the allele concordant with the reference is more likely to be mapped, resulting in a bias towards the reference allele and an inflated rate of false positive ASE.

In silico mapping procedures relying on a conservative threshold (no more than two mismatches) are heavily affected by this bias (Fig. 3). We solved these issues by incorporating two additions to our mapping algorithm. Firstly, for all known dbSNPs, we replaced the reference allele with the corresponding genetic ambiguity code coding both possible alleles at this SNP. Secondly, we relaxed the threshold on the number of mismatches for allowing reads to be mapped in order to limit the impact of a small number of mismatched SNPs. Our mapping procedure uses a probability-based alignment score that typically allows up to five mismatches (option -t150 in novoalign), and reads with more than five mismatches were discarded at later stages (see Materials and Methods). These two additions corrected most of the reference allele bias (Fig. 3).

Additional differences, such as other heterozygous SNPs or indels, further contribute to differential mapping of both haplotypes. The novoalign algorithm provides calls for small indels (up to 13 bp). Using these calls, we found an increased proportion of imbalanced SNPs within 45 bp of indels. Before applying any quality control filter (see above), 42% of heterozygous SNPs within 45 bp of a called indel were imbalanced at $P < 0.001$ (233 out of 545 calls). This proportion was 11.7% when no indel was called within 45 bp (737 out of 6311 calls). After applying the quality filters, these proportions became 24% for SNPs within 45 bases of an indel, and 7.5% otherwise. In both cases the difference is highly significant ($P << 10^{-10}$). These results suggest that the quality filters capture only a fraction of the artefacts caused by indels. Consequently, we excluded from our ASE analysis heterozygous SNPs within 45 bp of a called indel in the same individual.

Additional sources of false positive ASE

However, it is likely that even after applying these quality filters copy number variation and mapping biases still account for a number of false positive ASE results. In particular, intermediate size indels larger than $\sim$13 bp cannot be called by novoalign with the settings we used for read mapping. We used the software Pindel (26) to re-analyse the RNA-seq data with the purpose of identifying new larger indels (this tool can detect 1–10 kb indels). However, we could find new indels within 300 bp of the heterozygous SNP for only 11 positive ASE results.

In addition, low-complexity or repeat sequence surrounding heterozygous SNPs is associated with an elevated ASE positive rate. We used as criteria for low-complexity/repeat sequences the fact that more than 25% of the 90 bp surrounding sequence (45 bp on each side) was masked by RepeatMasker. We found that roughly 12% of the heterozygous SNPs in this study were located in low-complexity regions. For SNPs in such low-complexity regions, the ASE positive rate was 23%. Overall, 28% of ASE positive SNPs are located in low-complexity DNA regions. The ASE rate was also elevated for intronic SNPs (33/208=15.8% ASE positive SNPs), which are more likely to contain small indels.

Lastly, the ASE positive rate was lower for heterozygous SNPs that passed HapMap quality filters: 115 ASE positive SNPs out of 2478 = 4.6%, compared with 7.5% for the non-filtered set (370/4929 heterozygous SNPs). This difference between HapMap SNPs and non-HapMap SNPs is significant ($P = 2.47 \times 10^{-14}$) and consistent with the fact that SNPs passing HapMap quality filters are not likely to be located in repeat sequences or near indels. A similar reduction of ASE positive rate was observed for SNPs on the Illumina Quad660W genotyping chip (32/797 = 4.01%).

To obtain a more accurate assessment of potential biases, we attempted to validate additional ASE results. We used individual 1 data and selected 22 transcripts with a unique imbalanced heterozygous SNP in each transcript (Supplementary material, Table S4). We first resequenced the gDNA to identify potential indels and obtained usable sequence data for 19 of 22 SNPs. We found only one previously undetected 1 bp indel within 45 bp of the heterozygous SNP, providing a likely explanation for the observed imbalance at this SNP. However, resequencing data showed that 5 of 19 SNPs are in fact homozygous. None of these five SNPs passed HapMap QC, and this discrepancy is most likely the consequence of RNA-seq mapping errors with homologous DNA regions.

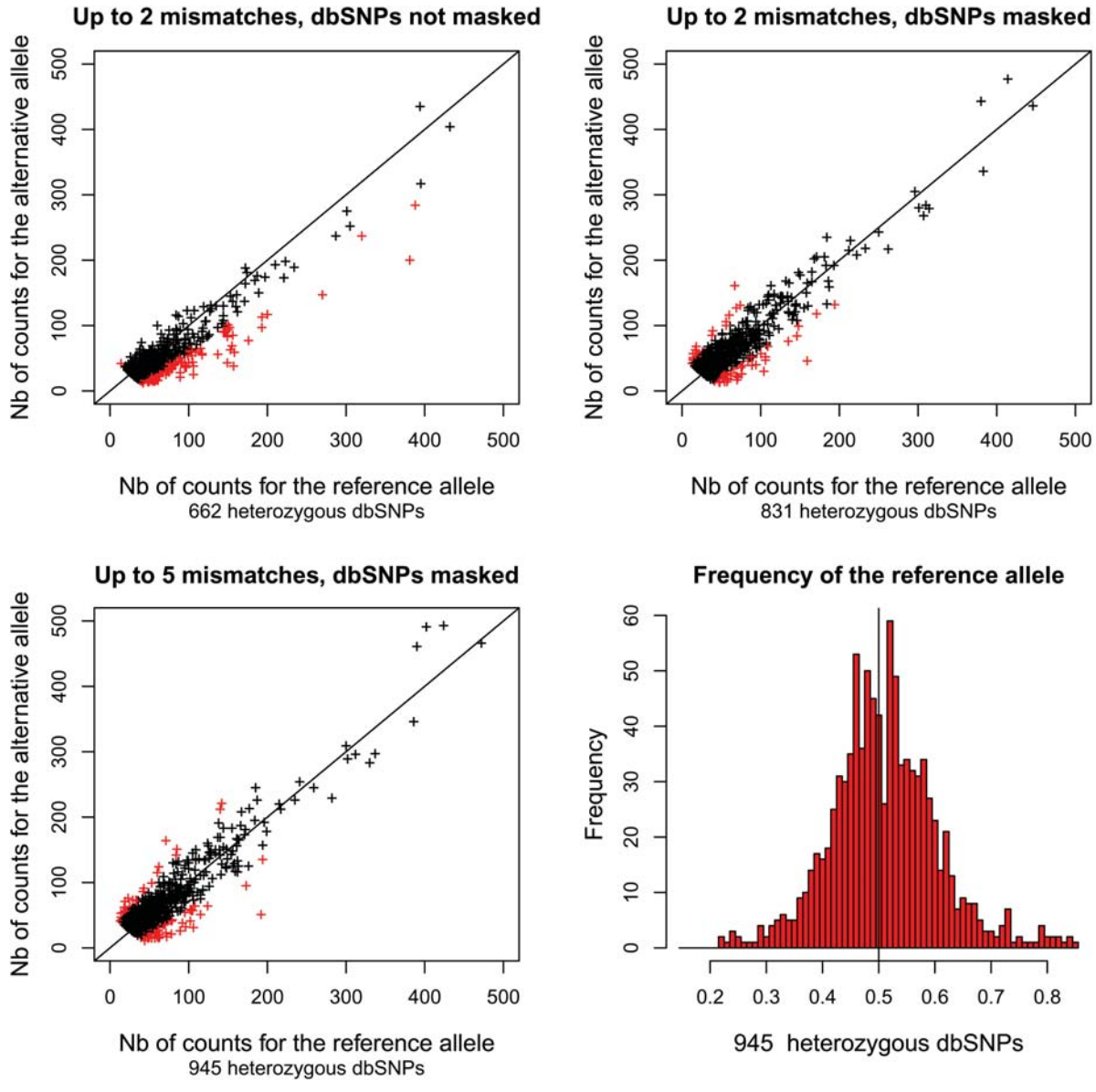

Figure 3. Effect of mapping parameters (number of mismatches allowed for read mapping, and/or masking dbSNPs from the reference sequence set) on the allelic ratio distribution using the PE data for individual 2 stimulated at heterozygous dbSNPs passing quality checks and with read depth >50. Red crosses indicate significant allelic imbalance ($P < 0.001$). The bottom-right histogram has been generated allowing up to five mismatches and mapping reads to the masked reference set. Nb, number.

For the remaining 13 SNPs, we designed short-range primers to resequence both gDNA and total RNA (Supplementary material, Table S4). Sequence data were usable for 9 of 13 SNPs for which we used the PeakPicker method (18) to quantify potential imbalance. Three of nine SNPs convincingly validated (cDNA by gDNA allelic ratio >1.5). For two out of six remaining SNPs, the estimated imbalance from RNA-seq was low (no more than 60:40) and for these SNPs the limited sensitivity of the PeakPicker validation assay when the allelic imbalance is <1.5-fold can potentially explain the lack of validation. For the remaining four SNPs, the PeakPicker validation data showed no evidence of allelic imbalance in spite of a strong bias observed in the RNA-seq data and we found no obvious explanation to account for this discrepancy. In addition to limited sensitivity, the interpretation is also complicated by the fact that our validation data used total RNA but the RNA-seq findings used poly(A)-selected RNA. We did not apply the more sensitive counting ASE method C-BASE owing to time constraints and the laborious nature of this assay.

Taken together, these data suggest that several biases, some of them unknown, increase the FDR. Even if one assumes that the ASE FDR is relatively small for HapMap SNPs, the difference in ASE positive rate with non-HapMap SNPs suggests that the false positive rate among non-HapMap SNPs is ~50%. However, this rate could potentially be higher, especially for heterozygous SNPs surrounded by repeat sequences or small indels. Thus 'unproven quality' SNPs

should only be used for ASE estimation with great caution and the presence of multiple imbalanced SNPs is required to provide convincing evidence of ASE. In addition, the corresponding ASE for gDNA is a desirable control.

PE versus SE reads

To compare PE and SE mapping, we used the PE data for individual 2 stimulated (as shown in Fig. 3) which we mapped either as PE or SE reads. Using the frequency of the reference allele as a diagnostic for *in silico* mapping issues, the distributions obtained using PE and SE mapping were indistinguishable (data not shown). However, using the same quality threshold, the PE approach mapped 19% more quality reads than the SE approach. This difference indicates that the PE approach is more cost-effective than the SE approach for this ASE design, in addition to other advantages for RNA-seq experiments including splice junction mapping and *de novo* assembly.

Sequence capture

The main limitation of the full poly(A)-selected transcriptome resequencing is the limited read depth at genes that are insufficiently expressed. To address these issues, we investigated the potential of the cDNA enrichment with sequences of interest using recently developed sequence capture methods (27). As a pilot study, we designed a sequence capture Nimblegen hybridization array targeting 25 genes located in immune disorders-associated chromosome regions. We also included, as a positive control, the *FAM118A* transcript, a gene identified in this study and another (4) as an eQTL transcript. In total we targeted 26 genes and a total sequence length of 1.7 Mb.

Based on available *FAM118A* eQTL data (4) we used the Cambridge BioResource [www.cambridgebioresource.org.uk (28)] to recruit for a blood donation individual 5, selected to be heterozygous at the *FAM118A* eQTL. The Cambridge BioResource is a collection of over 5000 healthy volunteers in the Cambridge area that can be recalled by genotypes. Using the Nimblegen capture array, we captured genomic DNA and cDNA obtained from poly(A)-selected RNA for this individual. We then resequenced both samples using Illumina GAII sequencer and 76 bp PE reads (9.35 M pairs of reads for gDNA and 20 M pairs of reads for RNA).

Sequence capture successfully enriched both resulting products with target sequences. In the poly(A)-selected data, 61.58% of the total number of reads and 87.46% of the uniquely mapping reads mapped within 76 bp of the 1.7 Mb target sequence. For the gDNA data, these numbers were 76.61% of the total number of reads and 67.27% of the number of uniquely mapping reads. In contrast, in the absence of sequence capture for individuals 1–4, we found that 0.2–0.4% of the full poly(A)-selected transcriptome reads mapped to the 1.7 Mb regions targeted by the capture array, indicating that capture enriched for sequences of interest by an average 200-fold.

However, following sequence capture, we discarded 86.7% of the RNA reads and 52.96% of the genomic DNA reads because of PCR clonality issues. Such clonal reads were identified on the basis of having identical mapping locations for both reads in a pair. These biases most likely result from our use of PCR amplification both before and after hybridization on the capture array and are particularly problematic when dealing with limited levels of input DNA/cDNA, either prior to array capture or following array capture at the sequencing step. We are currently considering alternative library preparations and sequence capture protocols to circumvent these issues.

Nevertheless, after excluding duplicate clonal reads, we obtained sufficient data to perform ASE analysis for gDNA and RNA at the positive control gene *FAM118A* (Fig. 4). After applying the quality control procedures described in this study, we found 43 heterozygous SNPs with read depth >50 in the gDNA data. Only one SNP was not in dbSNP and 21 of 43 passed HapMap quality filters. In the RNA data, we found 21 heterozygous SNPs with read depth >50. Seven of 21 passed HapMap quality filters and 1 of 21 was not in dbSNP. Of the 21 heterozygous RNA SNPs, 14 also had sufficient read depth in the gDNA and as expected the gDNA data were consistent with a 1:1 allelic ratio ($P>0.001$ at all 43 SNPs). The consistency with 1:1 allelic ratio observed in the gDNA data indicates that no allelic bias was introduced by the sequence capture, a critical requirement for this experimental design.

In contrast, 10 of 21 heterozygous SNPs were significantly imbalanced in the RNA data at this $P < 0.001$ threshold (and 14/21 at $P < 0.01$). Analysis of the RNA data also showed that 77.5% of the heterozygous SNPs with read depth >50 are intronic SNPs, indicating that the sequence capture approach retains sufficient read depth even in introns when capturing cDNA from poly(A)-selected RNA.

DISCUSSION

Using a full poly(A)-selected transcriptome resequencing approach (RNA-seq) applied to eight $CD4^+$ T-cell samples obtained in untreated and activated conditions from four healthy individuals, we performed a transcriptome-wide scan for ASE. We identified several biases that inflate the ASE false positive rate and to control for these biases while maximizing the information obtained from the data, we developed a novel approach for mapping PE sequence reads to the human transcriptome. For pairs of HapMap SNPs/samples, the proportion of ASE positive results was 4.6%, a rate significantly lower than what we observed for the full set of heterozygous SNPs (7.5%). This comparison shows that for randomly selected ASE positive heterozygous SNPs of 'unknown quality', the FDR is probably as high as 50%. However, the fact that, using two independent locus-specific assays, all four attempted ASE validations at HapMap SNPs were successful suggests that this FDR is limited for SNPs that passed HapMap quality filters. In situations where there is uncertainty regarding genetic variation in the sequences surrounding the assayed SNPs, the presence of several imbalanced SNPs in the same transcript is necessary to provide reassurance that an ASE finding is indeed a true positive. Nevertheless, only independent validation can unequivocally

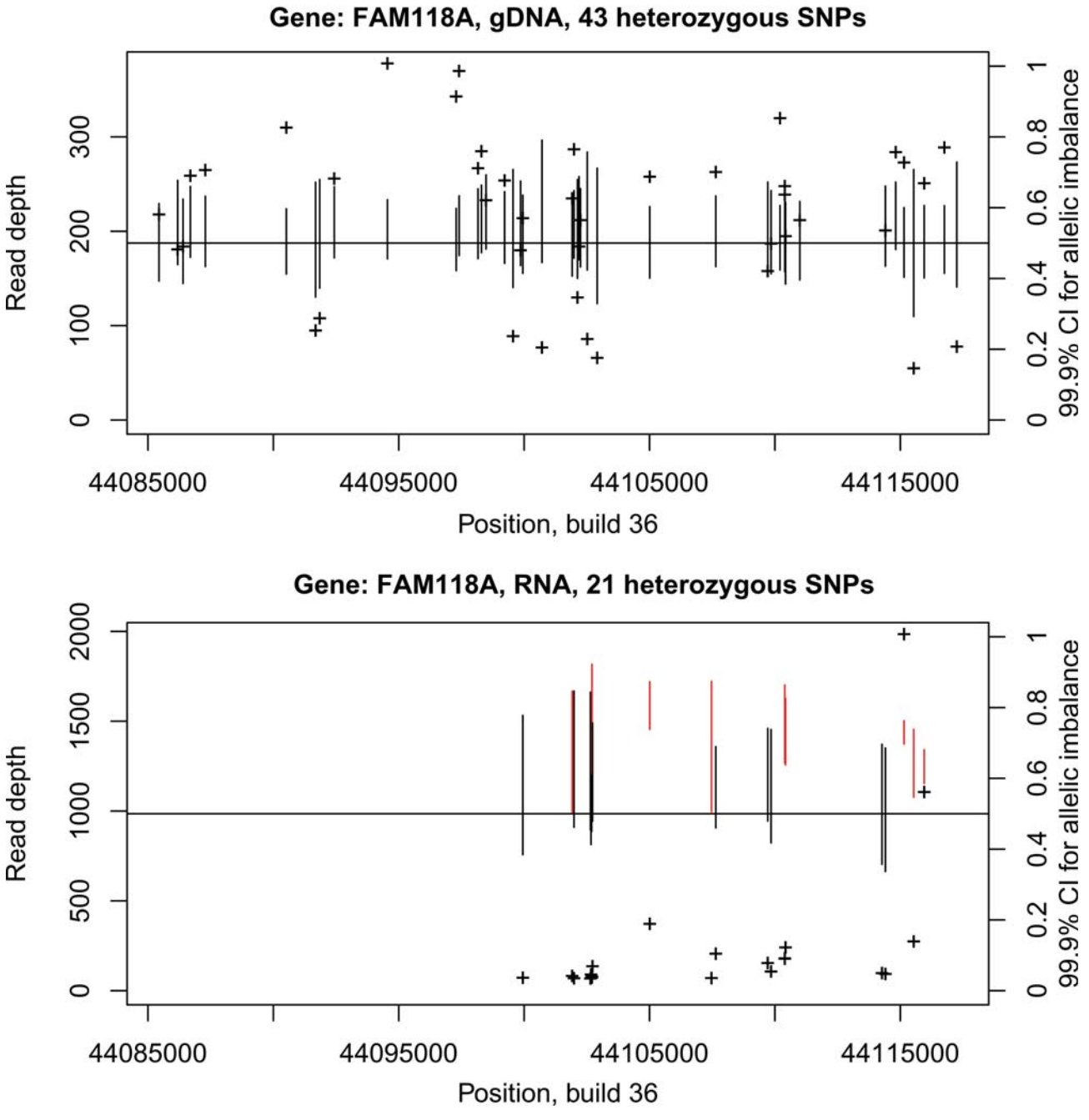

Figure 4. ASE data for individual 5 following sequence capture for genomic DNA (top) and RNA (bottom) at the *FAM118A* positive control known eQTL transcript. Crosses mark the read depth at each heterozygous SNP (only heterozygous SNPs with read depth >50 are shown). Vertical bars show the 99.9% confidence interval for the imbalance computed as the ratio between the most common and the least common of the two alleles for each heterozygous SNPs. Red bars mark heterozygous SNPs showing significant imbalance (at $P < 0.001$). The horizontal bar marks the 1:1 allelic ratio.

confirm findings from the genome-wide ASE screening methodology presented here.

Our data show that the effect of indels on the mapping of short sequencing reads is a major source of false positive for ASE. Longer sequencing reads, improved analytical methods and use of PE reads will facilitate indel calling in the future and longer reads will also improve the reliability of mapping procedures. However, other biases, some yet to be discovered, also contribute to the false positive rate. In particular, a previous study (28) has shown that the RNA fragmentation and the subsequent gel purification step for the sequencing library preparation can also introduce biases differentially affecting both haplotypes. Such biases will not necessarily be solved by the improvement of DNA sequencing technologies.

An advantage of our full transcriptome resequencing approach is the ability to interrogate the entire genome for ASE while avoiding the biases associated with the amplification of specific mRNA/cDNA sequences. The drawback is

the fact that the current ASE approach is restricted to highly expressed transcripts also including a heterozygous SNP in the transcript-encoding sequence. Indeed, read depth is the key determinant of statistical power for this ASE approach and the analysis of rested and activated samples in this study was initially motivated by the need to increase the expression level of specific transcripts. For example *IL2*, a T1D disease-associated gene (www.t1dbase.org), requires activation to be expressed at a sufficient level for ASE analysis. To capture more subtle levels of allelic imbalance, an even greater read depth (as high as 1000) would be beneficial, if practical, both for a genome-wide screen or locus-specific validation assay. The rapidly increasing throughput of DNA/cDNA resequencing technologies will most likely soon provide that level of read depth for a large number of transcripts, either in poly(A)-selected or even total RNA.

Our preliminary data show that the enrichment of cDNA with sequences of interest using sequence capture methods (27) followed by high-throughput sequencing is an effective way to increase the read depth at transcripts of interest. Alternatively, another study (29) has proposed padlock-capture of targeted sequences containing known SNPs to increase read depth. In addition to the increased read depth, the sensitivity to detect introns in preRNA (probably from partially unspliced heteronuclear RNA with a poly(A) tail) provides increased polymorphism levels compared with exonic sequences, and therefore more heterozygous SNPs for ASE testing. Joint sequencing of genomic DNA, also enriched for the regions of interest, enables screening of the heterozygous SNPs for 1:1 ratio in genomic DNA, thus facilitating detection of small indels and repeat sequences that increase the ASE false positive rate. Our preliminary study using Nimblegen sequence capture arrays for T-cell total RNA detect retained introns at high read depth in several genes from target regions. Furthermore, if a particular transcript is of interest, specific samples can be selected from a collection of individual volunteers of known genotype (e.g. www.cambridgebioresource.org.uk) to ensure the presence of heterozygous SNPs.

A successful combination of sequence capture and deep sequencing of cDNA generated from total or poly(A)-purified RNA with sufficient coverage has the potential to simultaneously test for ASE, in specific primary cell subsets, most findings identified by GWA scans, provided that the expression levels of the transcripts are sufficient and heterozygous SNPs are available. This experiment would be useful not only to demonstrate a potential regulatory effect of these disease-associated alleles, but also to fine-map-associated regions by testing for ASE in carefully selected rare recombinant individuals. Last, repeating the experiment for different cell subsets will enable the identification of the cell populations in which this regulatory role exists (28), thus helping researchers link GWA findings with the molecular function of these susceptibility alleles.

MATERIALS AND METHODS

Sample description

Primary $CD4^+$ human cells were extracted with Rosettesep (Stem Cell Technologies, UK) negative selection from leucofilters obtained from the National Blood Service (individuals 2, 3 and 4). Cells from individuals 1 and 5, obtained from venous blood through the Cambridge BioResource (28), were purified from ficoll-isolated PBMC using Dynabeads untouched human $CD4^+$ T cells negative selection kit (Invitrogen Corp., UK). Purity was assessed by flow cytometry (>95% for all samples) while viability was assessed by Trypan blue staining (>93% for all samples). $CD4^+$ T cells were cultured for 3 h in X-Vivo media with 1% human AB serum alone (Lonza, UK) or stimulated for 3 h at 37°C, 5% CO_2 in plates pre-coated with 1 μg anti-CD3 (clone: OKT3, eBioscience, USA) with the addition of 2 μg anti-CD28 (eBioscience).

Illumina library construction

RNA was extracted using the RNeasy kit (QIAGEN, UK) following the manufacturer's instruction. Samples were subjected to additional DNase treatment using Turbo-DNase (Ambion, UK). RNA quantification and quality were assessed using Nanodrop (Nanodrop Technologies, USA) and RNA 6000 Agilent Bioanalyzer chip (Agilent Technologies, USA). A double poly(A) RNA isolation was performed on 10 μg of total RNA (Invitrogen). Poly(A) RNA was fragmented for exactly 5 min at 70°C in fragmentation buffer (Ambion) prior to random hexamer reverse transcription and second strand synthesis as previously described (10). Illumina GAII PE adapters were ligated to the DNA and the library generated according to the standard library generation protocol (Illumina, USA). A 300 bp size range was excised from the library on 2% agarose gel. The resultant library was subjected to 15 cycles of PCR (Phusion* DNA polymerase, Finnzymes, Finland). The library was quantified by Nanodrop (Nanodrop Technologies) and assayed for size using a DNA 7500 Agilent Bioanalyzer chip (Agilent).

Description of mRNA sequencing protocol

Sequencing was performed on GAII (Illumina). Cluster generation and sequencing was performed according to manufacturer's instructions. Forty-five cycles of sequencing were performed on $n = 8$ samples (four individuals) using the SBS sequencing kit v1 (Illumina) and cycle sequencing kit v1/2. For the sequence capture sample (individual 5), 76 bp sequencing reads using SBS sequencing kit v2 were performed. All the sequencing data used in this study have been submitted to the short read archive (accession number SRA008367).

ASE validation by C-BASE

cDNA used for ASE validation was synthesized from the fragmented total RNA using Superscript III RT kit (Invitrogen). Forward and reverse PCR primers were designed against conserved sequences near the SNPs rs2064068, rs10405893 and rs1060819, in *FAM118A*, *CALM3* and *ATHL1*, respectively, to amplify a 200–230 bp product using AmpliTaq Gold (Applied Biosystems, ABI, Warrington, UK). Primer sequences are listed in Supplementary material, Table S2. Both genomic DNA and cDNA, synthesized from fragmented total RNA, of individuals that are heterozygous at the reporter

SNP were amplified. PCR products were excised from a 1.5% agarose gel using a QIAquick gel extraction kit (QIAGEN). Purified PCR product was ligated into pCR4-TOPO Vector and transformed chemically into One Shot Top 10 competent cells (Invitrogen). Transformed cells were spread onto Luria-Broth agar plates containing Ampicillin and incubated at 37°C overnight. Colonies were picked from the agar plates and directly inserted into plates with pre-aliquoted PCR mastermix of AmpliTaq Gold and T3 (5′-ATT AACCCTCACTAAAGGGA-3) and T7 (5′-TAATACGACT CACTATAGGG-3′) primers. Colonies were screened using TaqMan genotyping assays designed by the assay-by-design service from ABI. Alleles were scored on SDS v2.2.2 software (ABI) manually and counted.

ASE validation by sequencing

Purified PCR products, as above, were sequenced using the corresponding forward or reverse primer in duplicate reactions. The sequencing reactions were performed using Applied Biosystems' BigDye (version 3.1) chemistry and resolved using an ABI 3700 Genetic Analyzer. PeakPicker software v2.0 was used to calculate normalized peak height ratios as described in (18).

Read alignment

Our reference sequence set was designed for RNA-seq PE mapping. We used the Ensembl database version 52 to obtain, for each annotated gene, the transcript with the largest number of exons and included this main spliced transcript in our reference set. Second, we added one sequence per intron, extending intron boundaries 40 bp on each side to allow mapping of reads overlapping exon–intron junctions, but ensuring that 45 bp read would still map uniquely. In addition, we included one sequence per non-standard exon–exon junction (up to three skipped exons, i.e. junctions 1–3, 1–4, 1–5 but not 1–6). These junction sequences consist of 40 bp on each side of the exon–exon junction, such that a 45 bp read would have to overlap the junction by at least 5 bp. Last, a version of the reference DNA genome with masked annotated transcripts was included. This reference set was designed such that no region longer than 40 bp is represented twice, thus avoiding multiple mapping issues with 45 bp reads. However, the same SNP can be represented in different sequences and our final processing merged the data when a single SNP mapped to multiple sequences (Perl tools for reference building available at http://www.gene.cimr.cam.ac.uk/todd/).

Reads were mapped using novoalign V2.05.12 PE mode for paired reads, using the group mapping option for all the sequences (main transcript, introns and non-standard exon–exon junctions) originating from the same transcript (parameters -v 20 20 200 [>]([^_]*)_). PE mapping also used adapter trimming (option -a). SE mode was used for SE data (parameters -t 150). Quality reads were defined as uniquely mapped reads with phred-scaled probability score (30) $\geq$20 (i.e. maximum estimated mapping error rate of 1 per 100 reads). On average, 60% of the total number of sequenced 45 bp passed this quality read threshold. We wrote a tool, novoPile (available at http://www.gene.cimr.cam.ac.uk/todd/) to generate a 'pileup' file of bases mapping to individual reference sequence positions directly from the novoalign output. We excluded from this 'pileup' reads with more than five mismatches. We also filtered out duplicate reads that were identified in the PE data (option rmDup).

Identifying heterozygous dbSNPs and indels

To call SNPs heterozygous for ASE, we required that, following the exclusion of low mapping score reads and low-quality score calls, the read depth was $\geq$50 and the frequency of the second most common SNP was at least 15%. Genes located in the HLA region were excluded from this analysis because of the elevated polymorphism level and potential biases this polymorphism might introduce during the mapping process. Indels were called using the output from novoalign. Only indels supported by at least four independent reads were considered to be true calls.

Quality filtering for heterozygous SNPs

We used two statistical tests to filter out unreliable heterozygous SNPs. To verify that the allelic call is independent of the position of the SNP within the 45 base reads, we compared both distributions of positions using the Kolmogorov–Smirnov test. SNPs with *P*-values more significant than 0.01 were excluded. To check for strand-specific biases, we used a goodness-of-fit χ^2 test on the 2 by 2 table of allelic calls by strand. Again, SNPs with *P*-values more significant than 0.01 were excluded. SNPs within 45 bases of a called indel were also excluded from this analysis.

When restricting the analysis to HapMap SNPs, we used HapMap SNPs in the CEPH population with minor allele frequency >1% and call rate >95% (2 239 392 SNPs total and 87 796 of them were located in spliced transcripts).

Testing for allelic imbalance

Testing a single SNP for allelic imbalance uses a χ^2 goodness-of-fit test for even frequencies of both alleles. When more than one heterozygous SNP are present in the same transcript and the phase is known, the counts can be summed across heterozygous SNPs, counting only once reads overlapping multiple SNPs. This phasing information can be obtained directly from the RNA-seq data, provided that at least one read pair overlaps both heterozygous SNPs. Tools for phasing data from short PE reads (novoPhase, available at http://www.gene.cimr.cam.ac.uk/todd/). Alternatively, SNPs in high linkage disequilibrium can be phased *in silico*, provided that this information is available.

Sequence capture and sequencing for individual 5

Sequence capture probes were designed by Nimblegen (Madison, USA) to tile CCDS reference transcript for the following genes: *BACH2*, *CAPZA1*, *CD69*, *CLEC16A*, *CTLA4*, *DEXI*, *FAM118A* (included as a positive control known eQTL), *ICOS*, *IFIH1*, *IL2*, *IL2RA*, *IL15RA*, *IL18RAP*, *IL21*, *KIAA1109*, *LSP1*, *MAPKAPK2*, *PFKFB3*, *PRKCQ*, *PTPN2*,

RBM17, *RGS1*, *SH2B3*, *SOCS1*, and *TAGAP*. Probes were also designed to tile the intronic regions within each transcript, extending 40 bp into the exons on each side to capture retained intronic–exonic spanning reads since poly(A)-selected RNA contains preRNA. Oligonucleotide probes (between 60 and 100 bp) were designed, excluding any sequences mapping to repeat regions or that were not unique in the human genome permitting 80% of the total 1.7 Mb genomic regions to be tiled. Probes were synthesized on 385 000 feature glass slide arrays.

Standard Solexa library preps were conducted according to the manufacturer's instructions (Illumina). For the RNA samples, a poly(A)-selected RNA-seq library was constructed as described above. For the genomic DNA sample, DNA was extracted from whole blood using the Gentra PUREGENE DNA Extraction kit (QIAGEN) according to manufacturer's instructions. Five micrograms of DNA was fragmented by sonication (Bioruptor, Diagenode, USA) before a standard PE library was constructed. After a single gel-extraction step to size select a 350 bp library, the adapter ligated library was amplified with 12 cycles of PCR. One microgram of either genomic DNA or RNA (cDNA) PE Genome Analyzer library was added to 100 μg of Cot-1 DNA (Invitrogen) with 1 μl each of the PE enhancing oligos (1 mM). The libraries were then hybridized to the sequence capture array for 72 h at 42°C according to the manufacturer's instructions. Arrays were stringently washed at 47.5°C, and hybridized DNA eluted with NaOH. At multiple stages of the protocol, DNA size ranges were checked by Agilent Bioanalyzer and concentration by OD260/280 (Nanodrop). Hybridization and elution of libraries was undertaken by Roche Nimblegen in Madison, USA.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *HMG* online.

ACKNOWLEDGEMENTS

We acknowledge the use of the NIHR Biomedical Research Centre Cambridge BioResource and thank Sarah Nutland, Jennifer Sambrook and Willem Ouwehand for contributing to this study. We also thank Colin Hercus at Novocraft Technologies Pty for his help, time and expertise, Bing Ge for providing the PeakPicker software as well as the Barts and the London Genome Centre for data storage, analysis server and GAII sequencer use. Vincent Plagnol is a JDRF advanced post-doctoral fellow.

Conflict of Interest statement. None declared.

FUNDING

The study was funded by the Wellcome Trust (WT084743MA to D.A.v.H.) and the Juvenile Diabetes Research Foundation (JDRF) (JDRF common mechanisms grant 33-2008-402 to D.A.v.H. and J.A.T.). The JDRF, Wellcome Trust and the National Institute for Health Research (NIHR) funded the Diabetes and Inflammation Laboratory. The Cambridge Institute for Medical Research (CIMR) is in receipt of a Wellcome Trust Strategic Award (079895). Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

REFERENCES

1. Wellcome Trust Case Control Consortium (WTCCC) (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
2. Zhernakova, A., van Diemen, C. and Wijmenga, C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.*, **10**, 43–55.
3. Morley, M., Molony, C., Weber, T., Devlin, J., Ewens, K., Spielman, R. and Cheung, V. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
4. Dixon, A., Liang, L., Moffatt, M., Chen, W., Heath, S., Wong, K., Taylor, J., Burnett, E., Gut, I., Farrall, M. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.
5. Stranger, B., Forrest, M., Dunning, M., Ingle, C., Beazley, C., Thorne, N., Redon, R., Bird, C., de Grassi, A., Lee, C. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.
6. Goring, H., Curran, J., Johnson, M., Dyer, T., Charlesworth, J., Cole, S., Jowett, J., Abraham, L., Rainwater, D., Comuzzie, A. *et al.* (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.*, **39**, 1208–1216.
7. Heap, G.A., Trynka, G., Jansen, R.C., Bruinenberg, M., Swertz, M.A., Dinesen, L.C., Hunt, K.A., Wijmenga, C., Vanheel, D.A. and Franke, L. (2009) Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genom.*, **2**:1.
8. Choy, E., Yelensky, R., Bonakdar, S., Plenge, R., Saxena, R., De Jager, P., Shaw, S., Wolfish, C., Slavik, J., Cotsapas, C. *et al.* (2008) Genetic analysis of human traits *in vitro*: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.*, **4**, e1000287.
9. Plagnol, V., Uz, E., Wallace, C., Stevens, H., Clayton, D., Ozcelik, T. and Todd, J. (2008) Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS ONE*, **3**, e2966.
10. Mortazavi, A., Williams, B., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
11. Dimas, A., Deutsch, S., Stranger, B., Montgomery, S., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Arcelus, M., Sekowska, M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.
12. Wang, E., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S., Schroth, G. and Burge, C. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
13. Cloonan, N., Forrest, A., Kolle, G., Gardiner, B., Faulkner, G., Brown, M., Taylor, D., Steptoe, A., Wani, S., Bethel, G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
14. Verlaan, D., Ge, B., Grundberg, E., Hoberman, R., Lam, K., Koka, V., Dias, J., Gurd, S., Martin, N., Mallmin, H. *et al.* (2009) Targeted screening of cis-regulatory variation in human haplotypes. *Genome Res.*, **19**, 118–127.
15. Smyth, D., Plagnol, V., Walker, N., Cooper, J., Downes, K., Yang, J., Howson, J., Stevens, H., McManus, R., Wijmenga, C. *et al.* (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. *New Eng. J. Med.*, **359**, 2837–2838.
16. Bjornsson, H.T., Albert, T.J., Ladd-Acosta, C.M., Green, R.D., Rongione, M.A., Middle, C.M., Irizarry, R.A., Broman, K.W. and Feinberg, A.P. (2008) SNP-specific array-based allele-specific expression analysis. *Genome Res.*, **18**, 771–779.
17. Rainbow, D.B., Esposito, L., Howlett, S.K., Hunter, K.M., Todd, J.A., Peterson, L.B. and Wicker, L.S. (2008) Commonality in the genetic control of Type 1 diabetes in humans and NOD mice: variants of genes in the IL-2 pathway are associated with autoimmune diabetes in both species. *Biochem. Soc. Trans.*, **36**, 312–315.
18. Ge, B., Gurd, S., Gaudin, T., Dore, C., Lepage, P., Harmsen, E., Hudson, T.J. and Pastinen, T. (2005) Survey of allelic expression using EST mining. *Genome Res.*, **15**, 1584–1591.

19. Barrett, J., Clayton, D., Concannon, P., Akolkar, B., Cooper, J., Erlich, H., Julier, C., Morahan, G., Nerup, J., Nierras, C. *et al.* (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.*, **41**, 703–707.
20. De Jager, P., Jia, X., Wang, J., de Bakker, P., Ottoboni, L., Aggarwal, N., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C. *et al.* (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.*, **41**, 776–782.
21. Castro, M., Oliveira, M., Nunes, R., Fabre, S., Barbosa, R., Peixoto, A., Brown, M., Parnes, J., Bismuth, G., Moreira, A. *et al.* (2007) Extracellular isoforms of CD6 generated by alternative splicing regulate targeting of CD6 to the immunological synapse. *J. Immunol.*, **178**, 4351–4361.
22. Hassan, N., Simmonds, S., Clarkson, N., Hanrahan, S., Puklavec, M., Bomb, M., Barclay, N. and Brown, M. (2006) CD6 regulates T-cell responses through activation-dependent recruitment of the positive regulator SLP-76. *Mol. Cell. Biol.*, **26**, 6727–6738.
23. de Smith, A., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R., Bruhn, L. *et al.* (2007) Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.*, **16**, 2783–2794.
24. Jakobsson, M., Scholz, S., Scheet, P., Gibbs, R., Vanliere, J., Fung, H.-C., Szpiech, Z., Degnan, J., Wang, K., Guerreiro, R. *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.
25. Hunt, K., Zhernakova, A., Turner, G., Heap, G., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L., Ryan, A., Panesar, D. *et al.* (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet*, **40**, 395–402.
26. Ye, K., Schulz, M., Long, Q., Apweiler, R. and Ning, Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, in press, doi:10.1093/bioinformatics/btp394.
27. Albert, T., Molla, M., Muzny, D., Nazareth, L., Wheeler, D., Song, X., Richmond, T., Middle, C., Rodesch, M., Packard, C. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, **4**, 903–905.
28. Dendrou, C., Plagnol, V., Fung, E., Yang, J., Downes, K., Cooper, J., Nutland, S., Coleman, G., Himsworth, M., Hardy, M. *et al.* (2009) Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat. Genet.*, **41**, 1011–1015.
29. Zhang, K., Billy Li, J., Gao, Y., Egli, D., Xie, B., Deng, J., Li, Z., Lee, J., Aach, J., Leproust, E. *et al.* (2009) Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Meth.*, **6**, 613–618.
30. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.