RNA-Seq in 296 phased trios provides a high resolution map of genomic imprinting

Running Title: Imprinted gene identification in trios

Bharati Jadhav[1,#], Ramin Monajemi[2,#], Kristina K. Gagalova[3], Harmen H.M. Draisma[2,4], Mark A. van de Wiel[5], Lude Franke[6], Bastiaan T. Heijmans[2], Joyce van Meurs[7], Rick Jansen[8], GoNL Consortium, BIOS Consortium, Peter A.C. 't Hoen[4,9], Andrew J. Sharp[1,#], Szymon M. Kiełbasa[2,#]

[1] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA

[2] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

[3] GenomeScan B.V., Plesmanlaan 1D, 2333 BZ Leiden, the Netherlands

[4] Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands

[5] Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, the Netherlands.

[6] Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands

[7] Department of Internal Medicine, Erasmus MC, Rotterdam, the Netherlands

[8] Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, the Netherlands

[9] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands

[#] These authors contributed equally

Consortium Members BIOS Consortium

Bastiaan T. Heijmans, Peter A.C. ’t Hoen, Joyce van Meurs, Dorret I. Boomsma, René Pool, Jenny van Dongen, Jouke J. Hottenga, Marleen MJ van Greevenbroek, Coen D.A. Stehouwer, Carla J.H. van der Kallen, Casper G. Schalkwijk, Cisca Wijmenga, Sasha Zhernakova, Ettje F. Tigchelaar, P. Eline Slagboom, Marian Beekman, Joris Deelen, Diana van Heemst, Jan H. Veldink, Leonard H. van den Berg, Cornelia M. van Duijn, Bert A. Hofman, André G. Uitterlinden, P. Mila Jhamai, Michael Verbiest, H. Eka D. Suchiman, Marijn Verkerk, Ruud van der Breggen, Jeroen van Rooij, Nico Lakenberg, Hailiang Mei, Jan Bot, Dasha V. Zhernakova, Peter van ’t Hof, Patrick Deelen, Irene Nooren, Matthijs Moed, Martijn Vermaat, René Luijk, Marc Jan Bonder, Maarten van Iterson, Freerk van Dijk, Michiel van Galen, Wibowo Arindrarto, Szymon M. Kiełbasa, Morris A. Swertz, Erik. W van Zwet, Aaron Isaacs, Rick Jansen, Lude Franke.

Consortium Members GoNL Consortium

Francioli LC, Menelaou A, Pulit SL, van Dijk F, Palamara PF, Elbers CC, Neerincx PB, Ye K, Guryev V, Kloosterman WP, Deelen P, Abdellaoui A, van Leeuwen EM, van Oven M, Vermaat M, Li M, Laros JF, Karssen LC, Kanterakis A, Amin N, Hottenga JJ, Lameijer EW, Kattenberg M, Dijkstra M, Byelas H, van Setten J, van Schaik BD, Bot J, Nijman IJ, Renkens I, Marschall T, Schönhuth A, Hehir-Kwa JY, Handsaker RE, Polak P, Sohail M, Vuzman D, Hormozdiari F, van

Enckevort D, Mei H, Koval V, Moed MH, van der Velde KJ, Rivadeneira F, Estrada K, Medina-Gomez C, Isaacs A, McCarroll SA, Beekman M, de Craen AJ, Suchiman HE, Hofman A, Oostra B, Uitterlinden AG, Willemsen G, Study LC, Platteel M, Veldink JH, van den Berg LH, Pitts SJ, Potluri S, Sundar P, Cox DR, Sunyaev SR, den Dunnen JT, Stoneking M, de Knijff P, Kayser M, Li Q, Li Y, Du Y, Chen R, Cao H, Li N, Cao S, Wang J, Bovenberg JA, Pe'er I, Slagboom PE, van Duijn CM, Boomsma DI, van Ommen GJ, de Bakker PI, Swertz MA, Wijmenga C.

Address for correspondence:

Andrew J. Sharp, Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, Hess Center for Science and Medicine, 1470 Madison Avenue, Room 8-116, Box 1498, New York, NY 10029 USA. Telephone: +1-212-824-8942, Email: andrew.sharp@mssm.edu

Szymon M. Kiełbasa, Department of Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, 2333 ZC Leiden, the Netherlands, Telephone: +31-71-5269700, Email: smkielbasa@lumc.nl

Keywords

imprinting, allele-specific expression, Bayesian analysis, parent-of-origin, phased genotypes

Abstract

Combining allelic analysis of RNA-Seq data with phased genotypes in family trios provides a powerful method to detect parent-of-origin biases in gene expression. We report findings in 296 family trios from two large studies: 165 lymphoblastoid cell lines from the 1000 Genomes Project, and 131 blood samples from the Genome of the Netherlands participants (GoNL). Based on parental haplotypes we identified >2.8 million transcribed heterozygous SNVs phased for parental origin, and developed a robust statistical framework for measuring allelic expression. We identified a total of 45 imprinted genes and one imprinted unannotated transcript, 16 of which have not previously been reported as showing parental expression bias. Multiple novel imprinted transcripts showing incomplete parental expression bias were located adjacent to known strongly imprinted genes. For example, *PXDC1*, a gene which lies adjacent to the paternally-expressed gene *FAM50B*, shows a 2:1 paternal expression bias. Other novel imprinted genes had promoter regions that coincide with sites of parentally-biased DNA methylation identified in uniparental disomy samples, thus providing independent validation of our results. Using the stranded nature of the RNA-Seq data in LCLs we identified multiple loci with overlapping sense/antisense transcripts, of which one is expressed paternally and the other maternally. Using a sliding window approach, we searched for imprinted expression across the entire genome, identifying a novel imprinted putative lncRNA in 13q21.2. Our methods and data provide a robust and high resolution map of imprinted gene expression in the human genome.

Introduction

Genomic imprinting is a special case of mono-allelic expression where genes are expressed in a parent-of-origin (PofO) specific manner. This type of mono-allelic expression can be observed in mammals at different developmental stages and is dependent on stage, cell and tissue type. Genomic imprinting plays a vital role in normal development, and errors of imprinting can underlie developmental disorders and contribute to certain cancers (Moore and Oakey 2011). Imprinting significantly influences the development of cell lineages, prenatal growth, normal brain function and metabolism (Bartolomei and Ferguson-Smith 2011). Any disruption to imprinted genes can lead to disturbed gene function and can have a deleterious effect on health. If such disruption happens at imprinted loci, it can result in imprinting disorder such as Beckwith-Wiedemann, Silver-Russell (Azzi et al. 2009), Prader-Willi, Angelman syndromes (Nicholls, Saitoh, and Horsthemke 1998), neonatal diabetes (Mackay et al. 2008) and cancer. Wilm's tumor, colorectal cancer, and hepatoblastoma are few examples of cancer caused due to aberrant imprinting in *IGF2* gene (Steenman et al. 1994; Kaneda and Feinberg 2005).

There are many screening methods developed and applied to discover imprinted genes such as DNA methylation, histone modification, and gene expression assays. RNA sequencing (RNA-Seq) is the most direct and comprehensive way to identify imprinted genes as it allows for quantifying relative expression of the maternal and paternal alleles (allele specific expression or ASE) at all heterozygous sites with sufficient coverage. However, the technology is subject to several technical biases resulting in potential false positives (Piskol, Ramaswami, and Li 2013). The reference bias, caused by additional penalties in the alignment for non-reference alleles is the most prominent of these biases (Castel et al. 2015). Moreover, the availability of additional DNA genotype information is essential because the heterozygous sites may appear as

homozygous in the RNA because of mono-allelic expression of the imprinted genes. Typically, such studies are performed without allelic inheritance information and make use of the bimodal distribution of the expression at heterozygous sites. This type of analyses lacks the ability to identify directionality of parental bias (*i.e.* assessing maternal versus paternal imprinting). Adding PofO information allows determination of maternal vs. paternal allele-specific expression and should have more power, in particular in the case of incomplete imprinting (slight bias towards the paternal or maternal allele), where bimodality in the distribution is difficult to assess. The use of PofO information is straightforward in mouse studies where reciprocal cross design is often used to identify maternal/paternal gene expression and imprinted genes (Gregg et al. 2010; Wang et al. 2008). So far, there are few studies performed in humans where PofO information is available. However, those studies are usually limited to either small number of trios or analysis at specific loci (Baran et al. 2015; Morcos et al. 2011; Metsalu et al. 2014; Apostolidou et al. 2007).

To circumvent these limitations we present a robust genome-wide approach to find PofO specific gene expression and identify the signature of imprinted genes at heterozygous sites using phased DNA genotypes from parent-offspring trios and RNA-Seq data aggregated at gene level. Our method is applied to two large scale studies with a total of 296 trios: 165 trios from the HapMap / 1000 genomes projects with RNA-seq data from lymphoblastoid cell lines (LCLs) (REF), and 131 trios from the Genome-of-the-Netherlands (Genome of the Netherlands Consortium 2014). We focus on the identification of genes and transcripts that are consistently imprinted in the population, detecting both complete imprinting (exclusive expression of the paternal or maternal allele) or incomplete imprinting (bias in expression towards the maternal or paternal allele).

Results

We tested for imprinted gene expression using allele-specific RNA-Seq analysis of 296 parent-offspring trios derived from two independent cohorts: (i) 165 LCLs collected as part of the HapMap project, and (ii) 131 whole blood (WB) samples studied by the GoNL Consortium. In each cohort, we used phased genotypes to compute the relative expression from the maternal and paternal alleles in RNA-Seq reads at expressed heterozygous single nucleotide variants (SNVs). We analyzed 23,003 Gencode genes which had at least one heterozygous SNV with ≥1 overlapping RNA-Seq reads in >10% of the samples, and summed the paternal and maternal counts for all heterozygous SNVs contained in a gene, irrespective of their exonic or intronic nature. The inclusion of intronic SNVs increased the power of our test considerably despite their low individual coverage, as there were generally many more informative intronic than exonic SNVs. We applied two statistical tests to check for consistent parental expression bias of autosomal genes within the populations. The rationale for using two statistical tests, Wilcoxon Signed Rank (WSR) test and ShrinkBayes (SB), is their differences in power and false positive rate in case of low numbers of informative individuals and low expression. More details are given in the Supplementary Note.

Quantile-Quantile plots showed a clear excess of genes with highly significant observed p-values above the null expectation with both statistical tests and cohorts, indicating strong evidence for imprinting. Furthermore, there was no evidence of genomic inflation in our study, with all values of λ between 0.9999 and 1.02 (Figure 1). Overall a total of 78 unique gene fragments (UGFs, see Methods) across the two populations showed significant evidence of imprinting (Supplementary Tables 1 and 2): 66 in LCLs and 43 in WB. However, the presence of overlapping transcripts, some of which were split into multiple separate annotations by our use

of UGFs, created redundancy in this list. After removal of these redundancies, we further manually curated signals to (i) assign signals of imprinted expression to the gene annotation which showed best consistency with the strand and location of data, (ii) remove transcripts where biased expression was driven by outlier samples with extreme read depth, and (iii) at loci containing multiple overlapping gene annotations, to avoid inflating the number of reported genes, we removed anonymous transcripts which appeared to represent partial gene fragments (see comments in Supplementary Table 1). This identified a total of 45 imprinted genes across the two cohorts: 38 in LCLs and 31 in WB, with 23 identified in both populations (Figure 1). The paternal ratios for each of these genes in each individual are plotted in Figure 2. Among the list of 45 genes, 30 genes have been previously reported as imprinted, while 15 are putative novel imprinted genes (Tables 1 and 2).

For each dataset, we classified genes as high confidence if significant in both statistical tests (34 in LCLs and 20 in WB), and low confidence if a gene was identified as significant with only a single statistical test (4 in LCLs and 11 in WB). At 10% FDR using the Paired Sample Wilcoxon Signed Rank (WSR) test, we found 36 and 24 significant genes in LCLs and WB, respectively. With *ShrinkBayes* (SB), we found 37 and 27 significant genes in LCLs and WB, respectively at 10% FDR (Tables 1 and 2).

We compared the 45 imprinted genes in our dataset with those reported as imprinted by the Genotype-Tissue Expression (GTEx) project (Baran et al. 2015). Of the 29 genes previously reported as imprinted in either LCL or WB that were successfully assayed in our analysis, 19 showed significant parental expression bias in our study (Figure 2). In all cases we observed consistent directionality of parental bias between the two studies (Supplementary Table 3).

Using only female samples, we searched for signals of imprinting on the X chromosome. We first estimated X chromosome inactivation ratios (XCIR) in each female, removing those

samples that showed highly biased XCIR (>80% silencing of one X chromosome), and then normalized allelic read counts for X-linked genes in each sample based on their XCIR. Analyses of these data resulted in one gene showing putative significant parental bias in LCLs (*RNA28S5*), and one gene in WB (*ARSD*). However, both were discounted as false positive signals due to clear reference bias in both cases (Supplementary Figure 1).

Novel incompletely imprinted genes occur in clusters

Most previous studies have identified imprinted genes based on the complete silencing of one parental allele. However, our large population sample and the quantitative nature of our assay identified several genes with biallelic expression, but which showed a significant bias for increased expression of one of the two parental alleles (Figure 2). In many cases, these incompletely imprinted genes occurred in close proximity to previously known imprinted genes that show mono-allelic expression. For example, we identified *PXDC1*, which lies ~100kb distal to the known imprinted *FAM50B* at 6p25.2, as showing a 2:1 paternal expression bias (Figure 3). Similarly, *ADAM23*, which lies ~130kb distal to *ZDBF2* at 2q33.3, also exhibits ~2-fold over-expression from the paternal allele. Overall, we identified 11 clusters of imprinted genes (defined here as two or more imprinted genes separated by <500kb), with 25 of the 46 imprinted genes we report located in these clusters. Thus our observations extend the known clustering of imprinted genes in the mammalian genome, showing that effects of genomic imprinting can extend over broad regions, and cause genes to show differing extents of parentally biased expression.

In another example, we identified two anonymous transcripts *RP11-134O21.1* and *GS1-57L11.1* at 8p23.2 as novel imprinted genes showing a ~2:1 preferential expression of the

paternal allele (Figure 4). Our previous studies of patients with uniparental disomy (R. S. Joshi et al. 2016) identified a maternally methylated region located at the bidirectional promoter of these two transcripts, thus providing independent validation of our results.

Strand specific RNA-Seq data reveals overlapping sense/anti-sense genes with opposite imprinting

In LCLs, the availability of strand-specific RNA-Seq data allowed the quantification of maternal and paternal counts from the forward and reverse strands separately. In the majority of cases, results obtained using stranded data were very similar to those obtained when aggregate data from both strands were considered. However, at loci where overlapping genes were transcribed from both forward and reverse strands, the use of unstranded data yielded misleading results. For *KCNQ1*/*KCNQ1OT1*, *RB1*/*LPAR6*, *NAA60*/*ZNF597,* and *PER3*/*RP3-467L1.4* only the use of strand-specific data was able to unambiguously determine the imprinting status of these genes (Figure 5). Notably, the strand-specific data demonstrated that several sense and antisense transcript pairs displayed opposite parental bias: *KCNQ1* is maternally expressed, whereas *KCNQ1OT1* is paternally expressed; *RB1* is maternally expressed, whereas *LPAR6* is paternally expressed (Figure 5 and Table 1).

Genome-wide scan for imprinting outside of known gene annotations

In order to search for novel signatures of imprinting outside of current gene annotations, we utilized a sliding window approach to systematically analyze the entire genome in an unbiased fashion. We chose a window size of 25kb as this was close to the median transcript length, with

a 5kb incremental slide. At each position, we aggregated maternal and paternal read counts for all available heterozygous SNVs within the 25kb window, and calculated the WSR and SB test statistics (Supplementary Table 4). Using this approach, as expected, we identified significant associations at nearly all imprinted genes found using our gene-centric approach. In several cases (e.g. *ZNF331* and *ZDBF2*), significant signals of imprinted expression were observed downstream of annotated genes, likely representing transcriptional read-through beyond annotated 3' boundaries (Supplementary Figure 2). However, we also identified a significant signal of expression outside of known gene annotations on 13q21.1 in the LCL population. Here, a cluster of 35 informative SNVs spread over ~8kb showed a strong paternal bias, with 87% of reads supporting transcription from the paternal allele in 73 informative samples. We propose that this represents a novel paternally imprinted transcript transcribed from the forward strand that apparently shares a bidirectional promoter with *LINC00434* (Figure 6). In support of this, data from the ENCODE Project in cell line GM12878 indicates the presence of an anonymous transcript at this position that is consistent in size and strand with our observations. There was no significant expression from this locus detected in whole blood.

Discussion

Here we report a detailed survey of imprinted gene expression in two human populations. We used a robust pipeline, incorporating the latest methods for allele-specific expression analysis, including rigorous removal of reads with potential mapping bias. Compared to previous methods, the availability of phased genotype information from whole genome sequencing of trios allowed direct assignment of expression levels from the two parental alleles at >2.8 million transcribed SNVs, providing a direct approach to assess imprinting genome-wide. This method

provides a considerable increase in sensitivity compared to approaches where parental origin information is lacking, allowing us to detect much more subtle imprinting effects than have been observed previously.

Further, we developed a robust statistical framework to account for population heterogeneity of imprinting. While many previous studies have called events at the level of individual samples and variants, we studied nearly 300 independent trios, and employed two complementary statistical tests that considered aggregated read counts at the gene level. The paired WSR is a non-parametric test that has the advantage of a low false positive rate, but with reduced power at small sample size and low expression (Supplementary Figure 7). In contrast, SB uses the zero-inflated negative binomial distribution to fit the data, well-suited for zero-inflated count data such as RNA-Seq, providing increased power for genes with low expression. These approaches have the advantage of assessing differences between paternal and maternal RNA-seq counts at multiple heterozygous loci across all individuals simultaneously, thus providing both increased robustness, and considerably greater power to resolve subtle biases in expression from the two parental alleles, when compared to the study of single data points. Consistent with prior studies, we found that utilizing aggregated read counts across all heterozygous sites per gene in each individual, including intronic reads and SNVs covered by only a single read, gave the most power in our analysis (Baran et al. 2015; DeVeale, van der Kooy, and Babak 2012). Finally, we filtered putative imprinted transcripts to remove false signals caused by reference bias, before manually curating each locus to resolve signals from overlapping and antisense transcripts. Importantly, curation to remove reference bias was an important step to avoid false positive imprinting signals: despite the fact that we masked non-unique genomic regions and applied stringent filtering to remove reads with ambiguous mapping, we still identified several genes with significant signals of parental expression bias that

were attributable to reads mapping preferentially to the reference sequence (as assessed by statistical comparison of coverage of the reference and alternative alleles) (Supplementary Figure 1).

Overall, this pipeline led to the identification of 45 imprinted transcripts, 15 of which are novel, in addition to a novel unannotated imprinted locus on 13q21.2. Of the novel imprinted genes identified, two notable examples are *PER3* and *IGF2BP3*. *PER3* [Period, Drosophila, Homolog of, 3; OMIM# 603427] is a member of the Period family of genes and is expressed in a circadian pattern in multiple tissues (Zylka et al. 1998). *PER3* is one of several genes that regulate circadian rhythms, and has been linked to Seasonal Affective Disorder by both human and mouse studies (Delaunay et al. 2000; Zhang et al. 2016). *IGF2BP3* [Insulin-like Growth Factor 2 mRNA-Binding Protein 3; OMIM# 608259] binds to the 5' UTR of the imprinted gene *IGF2*, suggesting it has a role in the regulation of *IGF2* production and is expressed ubiquitously across fetal and adult tissues (Monk et al. 2002; Nielsen et al. 1999). While previous reports have shown that *IGF2BP3* is bi-allelically expressed, we identify a slight bias for increased expression from the paternal allele. This may point at a coordinated PofO-based regulation of *IGF2* signalling cascade.

Classical studies of imprinting typically define imprinted genes as showing monoallelic expression from just one of the two parental alleles. However, recent studies in mouse have identified examples of incomplete, or non-canonical, imprinting (Bonthuis et al. 2015) – such genes are bi-allelically expressed, but show a significant allelic bias, such that the two parental alleles are expressed at different levels. Our study also finds multiple examples of incomplete imprinting in the human genome, and we report nine imprinted genes that each show consistent 2- to 3-fold higher expression from the paternal allele. In several cases, these incompletely imprinted genes occur in close proximity to known imprinted genes that show mono-allelic

expression, consistent with the known clustering of imprinted genes (Edwards and Ferguson-Smith 2007). While it is possible that some of these genes with incomplete imprinting in blood and/or LCLs might be fully imprinted (*i.e.* monoallelically expressed) in other tissues, we note that none were found in a prior survey of imprinting that assayed 34 human tissues (Baran et al. 2015), making this unlikely.

In addition to a gene-centric approach, we also utilized a sliding window analysis to screen for imprinted transcription across the genome, independent of known transcript annotations. This identified a novel imprinted locus at 13q21.2, apparently corresponding to an anonymous lncRNA approximately 8kb in length. This imprinted transcript is antisense to *LINC00434*, with the two genes apparently sharing a bidirectional promoter. Although we did not detect any expression from *LINC00434* in LCLs, given that these two genes are likely transcribed from the same promoter, we hypothesize that *LINC00434* may also be imprinted.

Given a previous report of sex-specific variations in imprinting (Baran et al. 2015), we tested whether age or gender influenced the imprinting status for any of the 46 imprinted transcripts we identified. However, we did not detect any significant effects of these two variables on parental expression bias (Supplementary Note). Furthermore, as studies in mouse (Davies et al. 2005; Raefski and O'Neill 2005) have previously identified a cluster of imprinted genes on the X chromosome, and phenotypic studies in human have led to the suggestion that genes on the human X chromosome may also be subject to imprinting (Skuse et al. 1997), we specifically searched for imprinting on the X chromosome. Although this analysis utilized only female samples, and thus suffered a reduction in power compared to our analysis of the autosomes, we were unable to detect any evidence to support the presence of imprinted genes on the human X chromosome.

We compared the list of genes we detected as imprinted with those found in previous studies of imprinting, and overall found good concordance. However, for ten genes that were reported as imprinted in the GTEx cohort (Baran et al. 2015), we did not observe evidence of imprinting, despite these genes having sufficient informative SNVs to be adequately assessed in our samples (*UTS2*, *MEST*, *UBE3A*, *PLAGL1*, *CPA4*, *MAGI2*, *INPP5F_V2*, *PRSS50*, *THEGL*, *RP11-7F17.7*). We note that of these ten genes, *MEST*, *UBE3A*, *PLAGL1*, *CPA4*, *MAGI2* and *INPP5F_V2* have all been reported as imprinted in other prior studies. While it is possible these may represent false-negatives in our analysis, many apparently show tissue-specific imprinting, with normal biparental expression in blood and LCLs, thus explaining our results (Kosaki et al. 2000; Vu and Hoffman 1997; Valleley, Cordery, and Bonthron 2007; Kayashima et al. 2003). In addition, we note that *UTS2* overlaps and is antisense to *PER3*, a gene which we identify as showing a weak paternal bias in LCLs. Given our improved methodology that utilized strand-specific RNA-seq, we suggest that the previously reported imprinting of *UTS2* instead likely reflects paternally-biased expression of *PER3*.

Our study has some limitations. Primarily, as our approach relies on measuring read depth over transcribed SNVs, we were limited to the study of genes that both contained heterozygous variants, and were expressed at sufficient levels to be analyzed. Thus, genes that were not expressed at detectable levels in a sufficient number of individuals, or which lacked heterozygous variants in our samples, were not assayed. Similarly, we had little discriminatory power to detect imprinting for genes that contained very few SNVs in our cohort, or for those that were expressed at very low levels. Further, as we studied samples of peripheral blood and LCLs, we were unable to detect genes that show imprinting confined to other tissues (Baran et al. 2015). Finally, as the LCLs we studied are immortalized cell lines, it is possible this process may have disrupted epigenetic processes such as imprinting. However, arguing against this

possibility, there was both strong concordance of our results obtained in LCLs with previous studies of imprinting, and several of the novel imprinted genes detected in LCLs were also supported by methylation and/or RNA-seq data from whole blood.

Given that our study assessed the imprinting status of ~41% of human transcripts, and identified 45 that are imprinted, our findings are broadly consistent with previous projections that have suggested that the human genome likely contains approximately 100 genes that are imprinted in somatic tissues (Barlow 1995).

Methods

Strand-specific RNA-Seq in 165 Lymphoblastoid Cell Lines

We generated RNA-Seq data from lymphoblastoid cell lines (LCL) for 57 CEPH (CEU), 58 Yoruba (YRI) and 50 Han Chinese (CHS) samples, all of whom were offspring of multi-generation pedigrees studied as part of The HapMap (http://hapmap.ncbi.nlm.nih.gov/) and/or 1000 Genomes (http://www.1000genomes.org/) Projects. Samples are listed in Supplementary Table 5.

Genotype data processing

For 163 samples, genotype data from the complete mother/father/child trio were available, while for the two samples, genotype data for only one parent was available. We obtained 1000 Genomes and HapMap project data from multiple releases: this included data from The 1000 Genomes Project Phase 1 and Phase 3 generated from low-coverage Illumina whole genome

sequencing, high coverage Complete Genomics whole genome sequencing data, exome sequencing, Illumina Omni 2.5M SNV array data, and HapMap3 project data genotyped on Illumina 1.6M and Affymetrix 6.0 SNV arrays. We included high quality filtered and curated DNA genotype data from the final releases of all these resources and combined into population-specific datasets. We performed quality control on the merged data such as resolving strand inconsistencies, removing multi allelic SNVs, indels, removing SNVs not present in the 1000 genomes data and converting coordinates from hg18 to hg19 where required using PLINK (versions 1.07 and 1.9) (Purcell et al. 2007; Chang et al. 2015), vcftools (version 0.1.15) (Danecek et al. 2011) and Beagle utilities.

Due to the differing genotyping approaches and resulting SNV densities available across different individuals, we performed combined imputation and phasing to increase SNV density and infer the two parental haplotypes in each offspring with Beagle 4.0 (S. R. Browning and Browning 2007). This used family pedigree information with the 1000 Genomes Phase3 reference panel downloaded from Beagle website (http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes_phase3_v5). Using 493 HapMap samples from the CEU, YRI and CHS populations, we created population specific-reference panels to improve imputation accuracy. Since many of the samples in our target panel are also part of 1000 Genomes Project reference panel, for each population group we created subsets of target and reference panel in such a way that there are no overlapping samples in two sets, and imputed and phased each of these subsets of target panel separately. Each chromosome was divided into segments to efficiently perform imputation and phasing, and these segments were subsequently merged together to yield chromosome-wide imputed and phased genotypes. Imputed genotypes were filtered to retain only high-quality genotypes ($R^2 \geq 0.95$). We also removed sites with Mendelian errors in each trio, Hardy-Weinberg

Equilibrium $p<10^{-4}$, and retained only biallelic SNVs with Minor Allele Frequency ≥5% in at least one of the three ethnicities in the cohort. This yielded ~3.9 million high-quality SNVs phased for parental origin.

To reduce phase-switch errors introduced during phasing that would result in incorrect parental origin assignment of SNVs, we used an R script developed in-house (https://github.com/SharpLabMSSM/PofOAssignment). This method utilizes the phased genotypes generated using BEAGLE, as follows: Each offspring's haplotype is compared with the parental haplotypes using a sliding window of 100 SNVs with 50 SNV incremental slide. Within each window we check for perfect matches between each offspring haplotype, and the four possible haplotypes within the parents. Parental origin assignments for each haplotype in the offspring are based on an unambiguous match to a single parental haplotype. This approach allows assignment of parental origin at uninformative sites where all members of the trio are heterozygous, and also provides an error check for phase switching. In the case when offspring's haplotypes do not perfectly match a parental haplotype, the genotypes in the window are set to missing. Subsequently, we then recover any such lost sites using simple rules of Mendelian Inheritance to each individual SNV genotype in the trio. Thus, by using a combined approach leveraging both statistical phasing with rules of Mendelian inheritance, we are able to generate maximally informative assignment for parental origin at heterozygous SNVs, with a minimal error rate.

Sample preparation

Lymphoblastoid cell lines were obtained from the Coriell Institute (Camden, NJ). Cells were grown in RPMI1640 media supplemented with 1mM L-glutamine, 10% FBS and 100u/L each of

penicillin and streptomycin, according to recommended protocols. Total RNA was extracted from frozen cell pellets (5-10 million cells) using TRIZOL, according to manufacturer's instructions (ThermoFisher Scientific). Strand-specific RNA-Seq libraries were prepared using NEBNext Ultra Directional RNA Library Prep Kit from Illumina. 1µg of total RNA was used as input, polyA+ selected, followed by strand synthesis was performed. Libraries were sequenced on an Illumina Hiseq 2500 instrument, with 10 samples pooled per lane, to generate 100bp single-end reads to a median depth of ~16 million reads per sample.

RNA seq data processing

Quality control analysis was performed on RNA-Seq reads using fastqc (version 0.11.2) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Over-represented sequences were removed using trimmomatic (version 0.32) (Bolger, Lohse, and Usadel 2014), and trimmed reads ≥30bp in length were kept. Cleaned reads were mapped to the human reference genome (hg19) with Gencode v16 annotations using the STAR aligner (version 2.3.0) (Dobin et al. 2013), yielding a mean of 79% uniquely mapped reads respectively. Picard (version 1.112) (https://github.com/broadinstitute/picard) was used for intermediate bam file processing such as add read groups, sorting and merging bam files of the same samples. To correct for mapping errors and biases which can result in false-positive allele-specific read-assignments, we used a collection of utilities in the WASP software (version 0.1) (van de Geijn et al. 2015), resulting in the removal of a mean of 36% of reads that overlapped SNVs in each sample, for which unambiguous allelic assignment could not be made. After parental-origin assignment for SNVs in each offspring, heterozygous sites were used to determine allele-specific expression. We first quantified reference and alternate RNA-Seq reads mapped at heterozygous loci using

AlleleCounter (v0.2, https://github.com/secastel/allelecounter) implemented in Python (Castel et al. 2015). Then, reference and alternate allele counts were used with PofO information to assign counts to the maternal and paternal alleles at each heterozygous site. Reads that did not uniquely map, or had base quality ≤10, were discarded. To further reduce mapping errors we applied additional filters, removing heterozygous SNVs that: (i) had a mappability score <1 (based on the "CRG GEM Alignability of 50mers with no more than 2 mismatches" track, downloaded from UCSC genome browser), (ii) overlapped CNVs with MAF ≥5% identified in samples from the 1000 Genomes and HapMap Projects (ftp://ftp.1000genomes.ebi.ac.uk/vol1/withdrawn/phase3/integrated_sv_map/ and common CNVs (Conrad et al. 2010), (iii) Segmental Duplications, and (iv) Simple Repeats (both downloaded from "Variation and Repeats" track group of the UCSC genome browser). These filters resulted in the removal of 21% of heterozygous sites, leaving ~3.1 million sites for downstream analysis.

Unstranded RNA-Seq in 131 Whole Blood Samples

The Genome-of-the-Netherlands (GoNL) project (Genome of the Netherlands Consortium 2014) performed whole genome sequencing of 250 family trios, a subset of which also had whole blood transcriptomes sequenced as part of the BBMRI-NL Biobank-based Integrative Omics Study (BIOS) (Zhernakova et al. 2017; Bonder et al. 2017). From these, we utilized data from 131 children with whole blood RNA-Seq data that passed all quality criteria and had genotypes concordant with those obtained by whole genome sequencing (listed in Supplementary Table 4). The individuals were participants from one of four biobanks: LifeLines-DEEP, The Leiden Longevity Study, Netherlands Twin Registry, and the Rotterdam Study.

Genotype data processing

DNA genotypes of 250 Dutch families were phased and imputed using BEAGLE (B. L. Browning and Yu 2009) and IMPUTE2. An integrated phase panel was constructed using SNV genotype likelihoods from the GATK:UnifiedGenotyper as input for BEAGLE, treating all samples as unrelated. SHAPEIT2 and MVNcall19 were then used along with trio information to phase the complete set of SNVs. Each haplotype transmitted to the offspring, and therefore allelic parental origin, was then obtained from the phased haplotypes (Genome of the Netherlands Consortium 2014).

Sample preparation

Total RNA from whole blood was treated using Ambion's GLOBIN clear kit, and subsequently processed for sequencing using the Illumina Truseq version 2 library preparation kit. Paired-end 50bp reads were generated using an Illumina HiSeq 2000 instrument, pooling 10 samples per lane. Read sets per sample were generated using CASAVA, retaining only reads passing Illumina's Chastity Filter for further processing. Data was generated by the Human Genotyping facility (HugeF) of ErasmusMC (The Netherlands, see URLs). Full details are described in (Zhernakova et al. 2017).

RNA-Seq data processing

Initial quality control was performed using FastQC (v0.10.1). Removal of adaptors was performed using Cutadapt (v1.1) (Martin 2011). Sickle (v1.2) (N. A. Joshi, Fass, and Others 2011) was used to trim low quality ends of the reads (minimum length 25, minimum quality 20). The reads were mapped with the STAR aligner (v2.3.125) (Dobin et al. 2013) to human reference genome hg19 masked at all single nucleotide variants with MAF>0.01 in GoNL samples. Full details are described in (Zhernakova et al. 2017). To reduce the influence of reference bias, we utilized WASP (version 0.1) (van de Geijn et al. 2015) to remove reads that aligned to different genomic positions after substituting the variant site. A summary of the influence of masking SNV positions in the reference and utilizing WASP to remove reads that show ambiguous mapping positions is shown in Supplementary Figure 3.

To obtain the parent-of-origin allelic counts, we first computed RNA-Seq reference and alternative counts using the GATK (v3.6-0-g89b7209) ASEReadCounter tool (McKenna et al. 2010). A script was then used to re-label the reference and alternative counts with parental origin based on the transmitted allele, leaving ~0.9 million heterozygous sites with paternal and maternal read counts for downstream analysis. A summary of the complete analytical pipeline is shown in Supplementary Figure 4.

Statistical analysis to identify imprinted expression

Since overlapping genes are common in the eukaryotic genome (Sanna, Li, and Zhang 2008), care must be taken when assigning reads to specific transcripts. To avoid misassignment of

reads at SNVs located within overlapping transcripts, we compiled all genes from Gencode annotations into a model where we consider overlapping regions of different genes as a separate unit, termed “unique gene fragments” (UGFs) (Supplementary Figure 5). The resulting gene models comprised 79,452 UGFs, and were used for assigning each heterozygous SNV to specific genes.

To maximize statistical power for detecting PofO biased expression, we summed read counts for all SNVs within each UGF. We calculated the paternal allelic ratio (defined as the fraction of reads derived from the paternally-inherited allele) for each individual using aggregated read counts across all informative SNVs within each UGF. We used the paternal allelic ratio of each informative individual to calculate the mean paternal ratio per UGF.

To formally test for parental bias in expression of UGFs, we utilized two complementary statistical approaches. We chose (i) a frequentist non-parametric approach, the Paired Wilcoxon Signed Rank (WSR) test, and (ii) an empirical Bayes approach *ShrinkBayes (van de Wiel et al. 2014)*. ShrinkBayes computes a Bayesian False Discovery Rate (BFDR), and we applied Benjamini–Hochberg False Discovery Rate (FDR) correction to the results of the WSR test, considering those UGFs with FDR q<0.1 (10% FDR) as showing significant evidence of imprinting. In each cohort, we only considered results for those genes in which at least 10% of individuals had ≥1 read informative for parental origin. Based on the results of these two tests, we classified predicted imprinted genes into those with high confidence (identified as significant by both tests) and low confidence (significant by one of the two tests). WSR test is a paired difference non-parametric test. It assigns ranks to the paternal/maternal differences with H_0: mean difference in pairs is symmetric around 0. The test is robust against outliers and has no distributional assumption. ShrinkBayes is an advanced statistical method specifically designed to handle zero-inflated count data allowing multi-parameter inference and modeling of random

effects in a Bayesian setting. It relies on INLA (Rue, Martino, and Chopin 2009) for the parameter estimation per gene, while borrowing information across genes by empirical Bayes type shrinkage of parameters. It allows a spike-and-slab prior for the parameter of interest (*patmat: mean difference in pairs*) to test H_0. Per UGF, we use a simple model with a single predictor parameter for imprinting (patmat) and a random effect parameter (indiv) to account for within individual variability.

$$y \sim 1 + patmat + f(indiv)$$

To assess the performance of the test procedures ShrinkBayes and WSR we developed a simulation scheme. ShrinkBayes is superior to WSR in terms of statistical power (Supplementary Figures 6 and 7) at a cost of increased computational resources. Using the two tests together reduces the false positive rate (Supplementary Figures 7), which motivates our definition of high-confidence genes.

Following statistical testing, we manually curated the UGF level results based on visual inspection of data plots, considering both gene annotations and strand-specific data in LCLs. Here we removed redundancies, and in the case of overlapping transcripts assigned imprinted expression to the correct gene. At several loci where we detected imprinted expression, gene annotations included transcripts with anonymous clone IDs. An example of this is the *L3MBTL1*/*SGK2* locus on chromosome 20. Here Gencode annotations include a transcript *RP1-138B7.5*, which is almost identical to an isoform of *SGK2*. In such cases, even though the transcript *RP1-138B7.5* was included in our initial list of significant imprinted genes, to avoid artificially inflating the number of imprinted transcripts we report, where these anonymous clone IDs likely corresponded to other annotated genes, we did not report them in our final curated list (Tables 1 and 2). Furthermore, although we filtered reads for potential mapping bias using WASP, we performed an additional check of UGF-level data for reference bias. We aggregated

reference and alternate allele read counts at the UGF level, and applied a two-sided WSR test to check whether the distribution of reference and alternate read counts were significantly different after multiple testing corrections (5% FDR), removing genes that showed significant reference bias.

Data Access

The raw and processed RNA seq data for 165 LCL samples have been deposited in the NCBI GEO database under the accession number GSE92521. The 131 WB STAR aligned BAM files (freeze 2) are submitted to European Genome-phenome Archive (EGA) under the study EGAS00001001077 and dataset accession number EGAD00001003937. The phased/imputed SNV data are part of the The Genome of the Netherlands (GoNL) Project with EGA accession number EGAS00001000644.

Acknowledgements

We would like to thank Tuuli Lappalainen and Stephane Castel for facilitating the collaboration underlying this study. This work was supported by NIH grant HG006696 to AJS. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. Work on the WB samples was performed within the framework of the Biobank-Based Integrative Omics

Studies (BIOS) Consortium and the GoNL Project which are funded by BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007).

Author contributions

AJS and PACH conceived and planned the study. BJ, RM, KKG, SMK, HHMD, and MAW performed bioinformatic analyses. BJ, RM, PACH, AJS and SMK prepared the manuscript.

Disclosure Declarations

The authors do not have any conflict of interests to declare.

References

Apostolidou, S., S. Abu-Amero, K. O'Donoghue, J. Frost, O. Olafsdottir, K. M. Chavele, J. C. Whittaker, P. Loughna, P. Stanier, and G. E. Moore. 2007. "Elevated Placental Expression of the Imprinted PHLDA2 Gene Is Associated with Low Birth Weight." *Journal of Molecular Medicine* 85 (4):379–87.

Azzi, Salah, Sylvie Rossignol, Virginie Steunou, Theo Sas, Nathalie Thibaud, Fabienne Danton, Maryline Le Jule, et al. 2009. "Multilocus Methylation Analysis in a Large Cohort of 11p15-Related Foetal Growth Disorders (Russell Silver and Beckwith Wiedemann Syndromes) Reveals Simultaneous Loss of Methylation at Paternal and Maternal Imprinted

Loci." *Human Molecular Genetics* 18 (24). Oxford Univ Press:4724–33.

Baran, Yael, Meena Subramaniam, Anne Biton, Taru Tukiainen, Emily K. Tsang, Manuel A. Rivas, Matti Pirinen, et al. 2015. "The Landscape of Genomic Imprinting across Diverse Adult Human Tissues." *Genome Research* 25 (7):927–36.

Barlow, D. P. 1995. "Gametic Imprinting in Mammals." *Science* 270 (5242):1610–13.

Bartolomei, Marisa S., and Anne C. Ferguson-Smith. 2011. "Mammalian Genomic Imprinting." *Cold Spring Harbor Perspectives in Biology* 3 (7). https://doi.org/10.1101/cshperspect.a002592.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15):2114–20.

Bonder, Marc Jan, René Luijk, Daria V. Zhernakova, Matthijs Moed, Patrick Deelen, Martijn Vermaat, Maarten van Iterson, et al. 2017. "Disease Variants Alter Transcription Factor Levels and Methylation of Their Binding Sites." *Nature Genetics* 49 (1):131–38.

Bonthuis, Paul J., Wei-Chao Huang, Cornelia N. Stacher Hörndli, Elliott Ferris, Tong Cheng, and Christopher Gregg. 2015. "Noncanonical Genomic Imprinting Effects in Offspring." *Cell Reports* 12 (6):979–91.

Browning, Brian L., and Zhaoxia Yu. 2009. "Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-Wide Association Studies." *American Journal of Human Genetics* 85 (6):847–61.

Browning, Sharon R., and Brian L. Browning. 2007. "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering." *American Journal of Human Genetics* 81 (5):1084–97.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. "Tools and Best Practices for Data Processing in Allelic Expression

Analysis." *Genome Biology* 16 (September):195.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (February):7.

Conrad, Donald F., Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, et al. 2010. "Origins and Functional Impact of Copy Number Variation in the Human Genome." *Nature* 464 (7289):704–12.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15):2156–58.

Davies, William, Anthony Isles, Rachel Smith, Delicia Karunadasa, Doreen Burrmann, Trevor Humby, Obah Ojarikre, et al. 2005. "Xlr3b Is a New Imprinted Candidate for X-Linked Parent-of-Origin Effects on Cognitive Function in Mice." *Nature Genetics* 37 (6):625–29.

Delaunay, F., C. Thisse, O. Marchand, V. Laudet, and B. Thisse. 2000. "An Inherited Functional Circadian Clock in Zebrafish Embryos." *Science* 289 (5477):297–300.

DeVeale, Brian, Derek van der Kooy, and Tomas Babak. 2012. "Critical Evaluation of Imprinted Gene Expression by RNA–Seq: A New Perspective." *PLoS Genetics* 8 (3). Public Library of Science:e1002600.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1):15–21.

Edwards, Carol A., and Anne C. Ferguson-Smith. 2007. "Mechanisms Regulating Imprinted Genes in Clusters." *Current Opinion in Cell Biology* 19 (3):281–89.

Geijn, Bryce van de, Graham McVicker, Yoav Gilad, and Jonathan K. Pritchard. 2015. "WASP:

Allele-Specific Software for Robust Molecular Quantitative Trait Locus Discovery." *Nature Methods* 12 (11):1061–63.

Genome of the Netherlands Consortium. 2014. "Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population." *Nature Genetics* 46 (8):818–25.

Gregg, Christopher, Jiangwen Zhang, Brandon Weissbourd, Shujun Luo, Gary P. Schroth, David Haig, and Catherine Dulac. 2010. "High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain." *Science* 329 (5992):643–48.

Joshi, N. A., J. N. Fass, and Others. 2011. "Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files (Version 1.33)[Software]."

Joshi, Ricky S., Paras Garg, Noah Zaitlen, Tuuli Lappalainen, Corey T. Watson, Nidha Azam, Daniel Ho, et al. 2016. "DNA Methylation Profiling of Uniparental Disomy Subjects Provides a Map of Parental Epigenetic Bias in the Human Genome." *American Journal of Human Genetics* 99 (3):555–66.

Kaneda, Atsushi, and Andrew P. Feinberg. 2005. "Loss of Imprinting of IGF2: A Common Epigenetic Modifier of Intestinal Tumor Risk." *Cancer Research* 65 (24). AACR:11236–40.

Kayashima, Tomohiko, Kentaro Yamasaki, Takahiro Yamada, Hideki Sakai, Nobutomo Miwa, Tohru Ohta, Koh-Ichiro Yoshiura, et al. 2003. "The Novel Imprinted Carboxypeptidase A4 Gene (CPA4) in the 7q32 Imprinting Domain." *Human Genetics* 112 (3). Springer:220–26.

Kosaki, K., R. Kosaki, W. J. Craigen, and N. Matsuo. 2000. "Isoform-Specific Imprinting of the Human PEG1/MEST Gene." *American Journal of Human Genetics* 66 (1):309–12.

Mackay, Deborah J. G., Jonathan L. A. Callaway, Sophie M. Marks, Helen E. White, Carlo L. Acerini, Susanne E. Boonen, Pinar Dayanikli, et al. 2008. "Hypomethylation of Multiple Imprinted Loci in Individuals with Transient Neonatal Diabetes Is Associated with Mutations

in ZFP57." *Nature Genetics* 40 (8). nature.com:949–51.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal* 17 (1):10–12.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9):1297–1303.

Metsalu, Tauno, Triin Viltrop, Airi Tiirats, Balaji Rajashekar, Ene Reimann, Sulev Kõks, Kristiina Rull, et al. 2014. "Using RNA Sequencing for Identifying Gene Imprinting and Random Monoallelic Expression in Human Placenta." *Epigenetics: Official Journal of the DNA Methylation Society* 9 (10):1397–1409.

Monk, D., L. Bentley, C. Beechey, M. Hitchins, J. Peters, M. A. Preece, P. Stanier, and G. E. Moore. 2002. "Characterisation of the Growth Regulating Gene IMP3, a Candidate for Silver-Russell Syndrome." *Journal of Medical Genetics* 39 (8):575–81.

Moore, Gudrun E., and Rebecca Oakey. 2011. "The Role of Imprinted Genes in Humans." *Genome Biology* 12 (3):106.

Morcos, Lisanne, Bing Ge, Vonda Koka, Kevin C. L. Lam, Dmitry K. Pokholok, Kevin L. Gunderson, Alexandre Montpetit, Dominique J. Verlaan, and Tomi Pastinen. 2011. "Genome-Wide Assessment of Imprinted Expression in Human Cells." *Genome Biology* 12 (3):R25.

Nicholls, Robert D., Shinji Saitoh, and Bernhard Horsthemke. 1998. "Imprinting in Prader–Willi and Angelman Syndromes." *Trends in Genetics: TIG* 14 (5). Elsevier:194–200.

Nielsen, J., J. Christiansen, J. Lykke-Andersen, A. H. Johnsen, U. M. Wewer, and F. C. Nielsen. 1999. "A Family of Insulin-like Growth Factor II mRNA-Binding Proteins Represses

Translation in Late Development." *Molecular and Cellular Biology* 19 (2):1262–70.

Piskol, Robert, Gokul Ramaswami, and Jin Billy Li. 2013. "Reliable Identification of Genomic Variants from RNA-Seq Data." *American Journal of Human Genetics* 93 (4):641–51.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3):559–75.

Raefski, Adam S., and Michael J. O'Neill. 2005. "Identification of a Cluster of X-Linked Imprinted Genes in Mice." *Nature Genetics* 37 (6):620–24.

Rue, Håvard, Sara Martino, and Nicolas Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 71 (2). [Royal Statistical Society, Wiley]:319–92.

Sanna, Chaitanya R., Wen-Hsiung Li, and Liqing Zhang. 2008. "Overlapping Genes in the Human and Mouse Genomes." *BMC Genomics* 9 (April):169.

Skuse, D. H., R. S. James, D. V. Bishop, B. Coppin, P. Dalton, G. Aamodt-Leeper, M. Bacarese-Hamilton, C. Creswell, R. McGurk, and P. A. Jacobs. 1997. "Evidence from Turner's Syndrome of an Imprinted X-Linked Locus Affecting Cognitive Function." *Nature* 387 (6634):705–8.

Steenman, M. J., S. Rainier, C. J. Dobry, P. Grundy, I. L. Horon, and A. P. Feinberg. 1994. "Loss of Imprinting of IGF2 Is Linked to Reduced Expression and Abnormal Methylation of H19 in Wilms' Tumour." *Nature Genetics* 7 (3):433–39.

Valleley, Elizabeth M., Sarah F. Cordery, and David T. Bonthron. 2007. "Tissue-Specific Imprinting of the ZAC/PLAGL1 Tumour Suppressor Gene Results from Variable Utilization of Monoallelic and Biallelic Promoters." *Human Molecular Genetics* 16 (8):972–81.

Vu, T. H., and A. R. Hoffman. 1997. “Imprinting of the Angelman Syndrome Gene, UBE3A, Is Restricted to Brain.” *Nature Genetics* 17 (1):12–13.

Wang, Xu, Qi Sun, Sean D. McGrath, Elaine R. Mardis, Paul D. Soloway, and Andrew G. Clark. 2008. “Transcriptome-Wide Identification of Novel Imprinted Genes in Neonatal Mouse Brain.” *PloS One* 3 (12):e3839.

Wiel, Mark A. van de, Maarten Neerincx, Tineke E. Buffart, Daoud Sie, and Henk M. W. Verheul. 2014. “ShrinkBayes: A Versatile R-Package for Analysis of Count-Based Sequencing Data in Complex Study Designs.” *BMC Bioinformatics* 15 (April):116.

Zhang, Luoying, Arisa Hirano, Pei-Ken Hsu, Christopher R. Jones, Noriaki Sakai, Masashi Okuro, Thomas McMahon, et al. 2016. “A PERIOD3 Variant Causes a Circadian Phenotype and Is Associated with a Seasonal Mood Trait.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (11):E1536–44.

Zhernakova, Daria V., Patrick Deelen, Martijn Vermaat, Maarten van Iterson, Michiel van Galen, Wibowo Arindrarto, Peter van ’t Hof, et al. 2017. “Identification of Context-Dependent Expression Quantitative Trait Loci in Whole Blood.” *Nature Genetics* 49 (1):139–45.

Zylka, M. J., L. P. Shearman, D. R. Weaver, and S. M. Reppert. 1998. “Three Period Homologs in Mammals: Differential Light Responses in the Suprachiasmatic Circadian Clock and Oscillating Transcripts outside of Brain.” *Neuron* 20 (6):1103–10.

Tables

Table 1. High confidence imprinted genes identified in LCLs and whole blood.

	Gene name	Chr	Start (hg19)	End (hg19)	Cytogenetic Band	Strand	Pat Ratio (LCL_S/LCL_U/WB)	Preferentially expressed allele	Confidence (LCL/WB)
1	*PER3*	1	7844380	7905237	p36.23	+	0.63/0.65/0.65	Paternal	HC/LC
2	*RP3-467L1.4*	1	7870302	7887402	p36.23	-	0.81/0.69/0.62	Paternal	HC/-
3	*PPIEL*	1	39997510	40024379	p34.3	-	0.76/0.74/0.92	Paternal	HC/HC
4	*ZDBF2*	2	207139387	207179148	q33.3	+	0.94/0.94/0.95	Paternal	HC/HC
5	*ADAM23*	2	207308263	207485851	q33.3	+	0.71/0.70/0.63	Paternal	HC/-
6	*AC069277.2*	3	6532166	6777816	p26.1	+	0.79/0.79/0.80	Paternal	HC/-
7	*NAP1L5*	4	89617066	89619386	q22.1	-	0.97/0.95/0.93	Paternal	HC/LC
8	*PXDC1*	6	3722848	3752260	p25.2	-	0.65/0.64/0.68	Paternal	HC/LC
9	*FAM50B*	6	3849620	3851551	p25.2	+	0.94/0.94/1.00	Paternal	HC/HC
10	*GRB10*	7	50657760	50861159	p12.1	-	0.57/0.57/0.29	Maternal	-/HC
11	*SGCE*	7	94214542	94285521	q21.3	-	0.83/0.83/0.53	Paternal	HC/-
12	*PEG10*	7	94285637	94299007	q21.3	+	0.97/0.97/1.00	Paternal	HC/LC
13	*RP11-134O21.1*	8	2523591	2585991	p23.2	-	0.70/0.68/0.78	Paternal	HC/-
14	*GS1-57L11.1*	8	2584858	2680004	p23.2	+	0.73/0.73/0.91	Paternal	HC/-
15	*H19*	11	2016406	2022700	p15.5	-	0.10/0.26/0.00	Maternal	HC/HC
16	*KCNQ1*	11	2465914	2870339	p15.5	+	0.19/0.35/0.47	Maternal	HC/HC
17	*KCNQ1OT1*	11	2629558	2721224	p15.5	-	0.96/0.94/0.74	Paternal	HC/HC
18	*RB1*	13	48877887	49056122	q14.2	+	0.39/0.39/0.54	Maternal	HC/-
19	*LPAR6*	13	48963707	49018840	q14.2	-	0.87/0.39/0.60	Paternal	HC/LC
20	*MEG3*	14	101245747	101327368	q32.2	+	0.21/0.26/0.01	Maternal	LC/HC
21	*MKRN3*	15	23810454	23873064	q11.2	+	0.90/0.90/1.00	Paternal	HC/LC
22	*SNRPN*	15	25068794	25223870	q11.2	+	0.98/0.98/1.00	Paternal	HC/HC
23	*SNURF*	15	25200181	25245423	q11.2	+	0.98/0.98/1.00	Paternal	HC/HC
24	*SNHG14*	15	25223730	25664609	q11.2	+	0.98/0.91/0.90	Paternal	HC/HC
25	*IGF1R*	15	99192200	99507759	q26.3	+	0.59/0.59/0.50	Paternal	HC/-
26	*PRR25*	16	855443	863861	p13.3	+	0.69/0.67/0.66	Paternal	HC/-
27	*ZNF597*	16	3486104	3493542	p13.3	-	0.04/0.06/0.05	Maternal	HC/HC
28	*NAA60*	16	3493611	3536963	p13.3	+	0.15/0.28/0.40	Maternal	HC/HC
29	*ZNF714*	19	21264965	21308073	p12	+	0.62/0.62/0.63	Paternal	HC/-
30	*ZNF613*	19	52430400	52452012	q13.41	+	0.50/0.50/0.67	Paternal	-/HC
31	*ZNF331*	19	54024235	54083523	q13.42	+	0.81/0.81/0.70	Paternal	HC/HC

32	*PEG3*	19	57321445	57352096	q13.43	-	0.98/0.98/1.00	Paternal	HC/-
33	*HM13*	20	30102231	30157370	q11.21	+	0.57/0.58/0.63	Paternal	HC/HC
34	*L3MBTL1*	20	42136320	42179590	q13.12	+	0.95/0.95/0.97	Paternal	HC/HC
35	*SGK2*	20	42187608	42216877	q13.12	+	0.92/0.91/0.93	Paternal	HC/HC
36	*GNAS-AS1*	20	57393974	57425958	q13.32	-	0.97/0.97/0.98	Paternal	HC/HC
37	*NHP2L1*	22	42069934	42086508	q13.2	-	0.57/0.57/0.62	Paternal	HC/HC

High confidence imprinted genes were classified as those transcripts showing significant evidence of parental expression bias (at 10% FDR) by both statistical tests used in at least one of the two cohorts studied. Rows shaded grey indicate novel imprinted loci not reported in previous studies.

Table 2. Low confidence imprinted genes identified in either LCLs or whole blood.

	Gene name	Chr	Start (hg19)	End (hg19)	Cytogenetic Band	Strand	Pat Ratio (LCL_S/LCL_U/WB)	Preferentially expressed allele	Confidence (LCL/WB)
1	*NEK10*	3	27151576	27410951	p24.1	-	0.48/0.48/0.18	Maternal	-/LC
2	*EHHADH*	3	184908412	184999778	q27.2	-	0.58/0.58/0.88	Paternal	-/LC
3	*IGF2BP3*	7	23349828	23510086	p15.3	-	0.54/0.54/1.00	Paternal	LC/-
4	*RPS2P32*	7	23530092	23530983	p15.3	+	0.88/0.79/0.64	Paternal	LC/-
5	*TRAPPC9*	8	140742586	141468678	q24.3	-	0.50/0.50/0.57	Paternal	-/LC
6	*IGF2*	11	2150342	2170833	p15.5	-	NA/ NA/0.89	Paternal	-/LC
7	*(unannotated transcript)*	13	60794418	60853802	q21.2	+	NA/0.86/NA	Paternal	LC/-
8	*RP11-64J4.2*	17	3182069	3289633	p13.3	-	0.27/0.30/0.49	Maternal	LC/-
9	*CHRNE*	17	4801069	4806369	p13.2	-	0.56/0.59/0.70	Paternal	-/LC

Low confidence imprinted genes were classified as those transcripts showing significant evidence of parental expression bias (at 10% FDR) by just one statistical test in one of the two cohort studied. Rows shaded grey indicate novel imprinted loci not reported in previous studies.

Figures

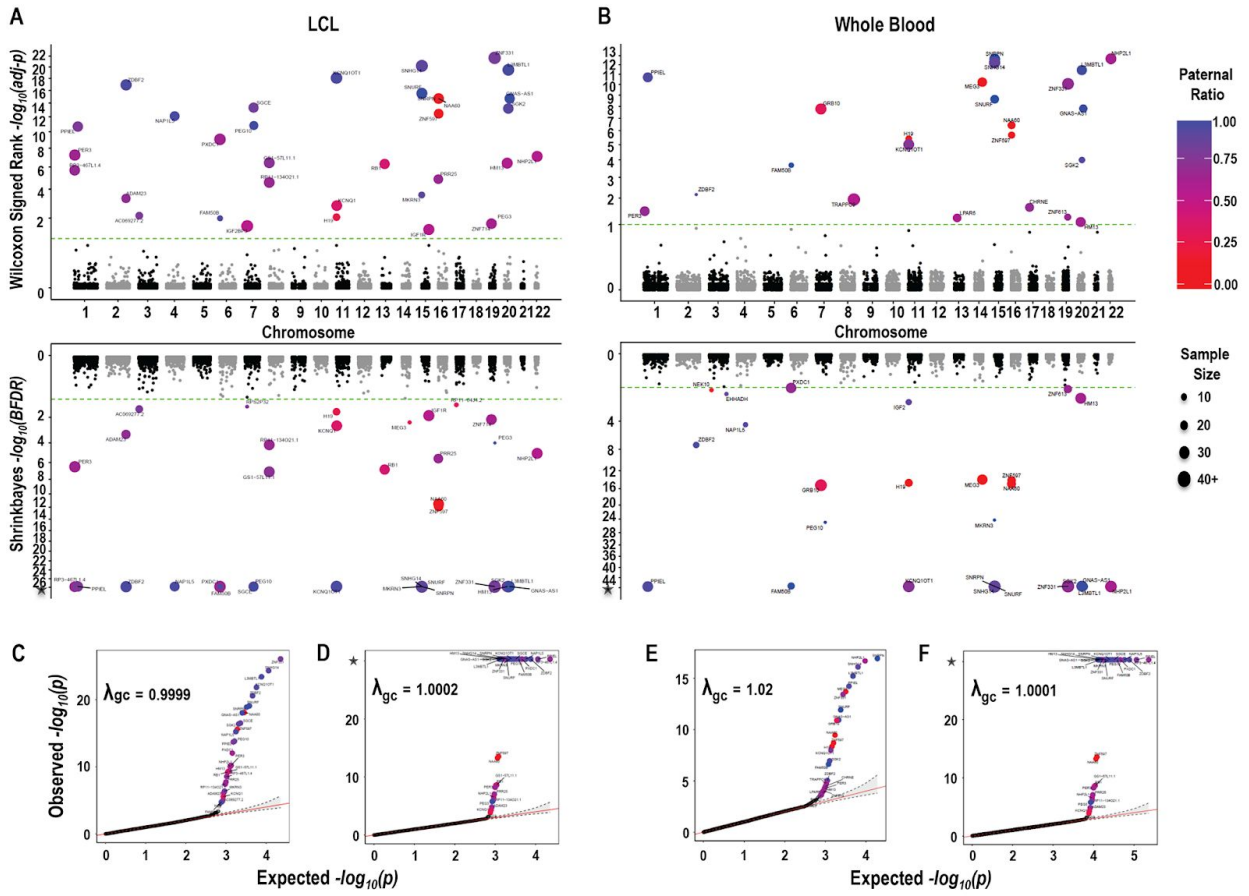

Figure 1. Miami and Quantile-Quantile plots of genome-wide results for parentally biased gene expression in 165 lymphoblastoid cell lines (LCL) and 131 whole blood (WB) samples. All data shown are based on bidirectional RNA-Seq data. In both **(A)** LCLs and **(B)** whole blood two statistical tests for parental bias were used: the upper panel in each cohort shows results from the paired Wilcoxon Signed Rank test, and the lower panel shows results

from the *Shrinkbayes* test. -$\log_{10}$ transformed adjusted p-values are shown on the y-axis, and chromosome and position on the x-axis: the dotted green lines indicate a statistical threshold of 10% FDR, with all genes exceeding this highlighted and labeled according to their paternal expression ratio, and number of informative samples (see legend). These plots show the results of analysis based on known transcript annotations, and thus do not include the novel unannotated transcript at 13q21.2 identified by sliding window analysis. **(C and E)** QQ plots for the paired Wilcoxon Signed Rank test in LCLs and whole blood. **(D and F)** QQ plots for *Shrinkbayes* in LCLs and whole blood. Note for *Shrinkbayes*, some of the observed –$\log_{10}$ p-values are infinite, indicated by an asterisk on the y-axis. In each plot, the top 30 genes are highlighted and colored according to their paternal ratio. For both cell cohorts and statistical tests the genomic inflation factor is approximately equal to 1.

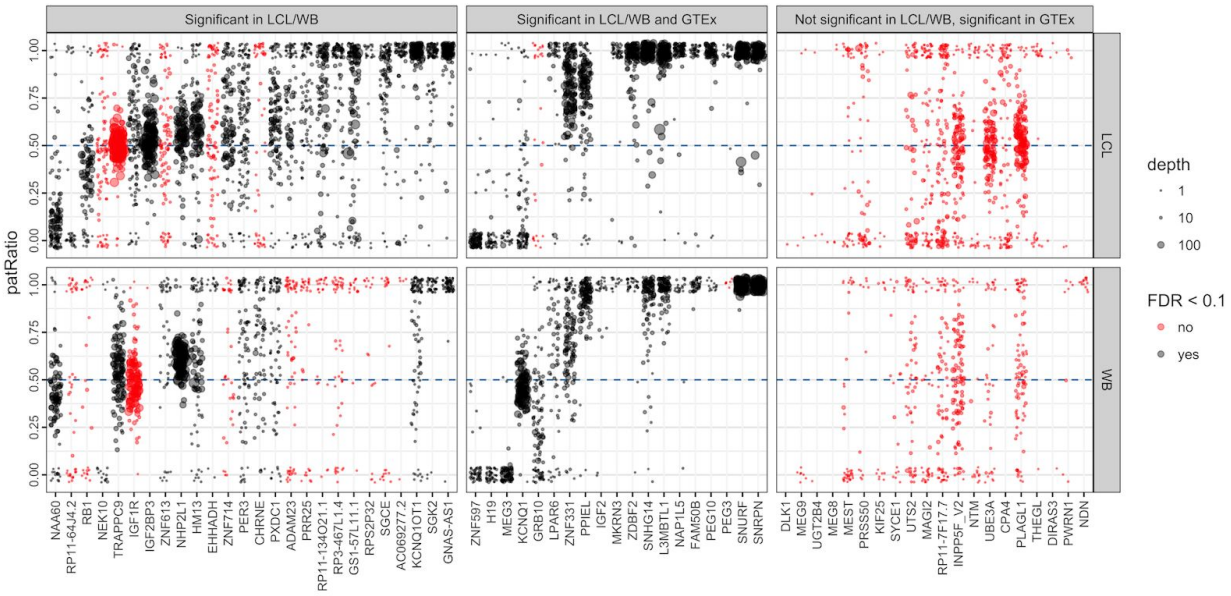

Figure 2. Varying degrees of parental bias among imprinted genes detected in LCLs, WB and GTEx. Each point represents the PatRatio (the mean fraction of reads transcribed from the paternal allele) in each informative individual per gene, with the point size indicating total read

depth over all heterozygous transcribed SNVs in that sample. Genes are ordered left to right by increasing mean PatRatio. The upper panel shows stranded data from LCLs, while the lower panel shows unstranded data from WB samples. Note that due to very low read depth in some genes/individuals, several genes showed highly variable PatRatios within the population. A small X- and Y-axis jitter was added to reduce overplotting effects. Genes shown in black were significant (FDR <0.1), while those in red did not pass this statistical threshold for significance. The figure is divided into three panels left,middle and right panel. Genes in the middle panel were found significant in LCL and/or WB and reported as putatively imprinted in GTEx (Baran et al. 2015); genes shown in the left panel were found significant in LCL and/or WB but not reported in GTEx; and genes shown in the right panel represent those reported as putatively imprinted in GTEx but were not identified as showing significant evidence of imprinting in either LCL or WB. Some genes in the right panel such as *DLK1*, *MEG9*, *THEGL*, *DIRAS3*, *PWRN1* and *NDN* show evidence of parental expression bias, but the limited number of informative samples meant we did not consider these in our formal analysis.

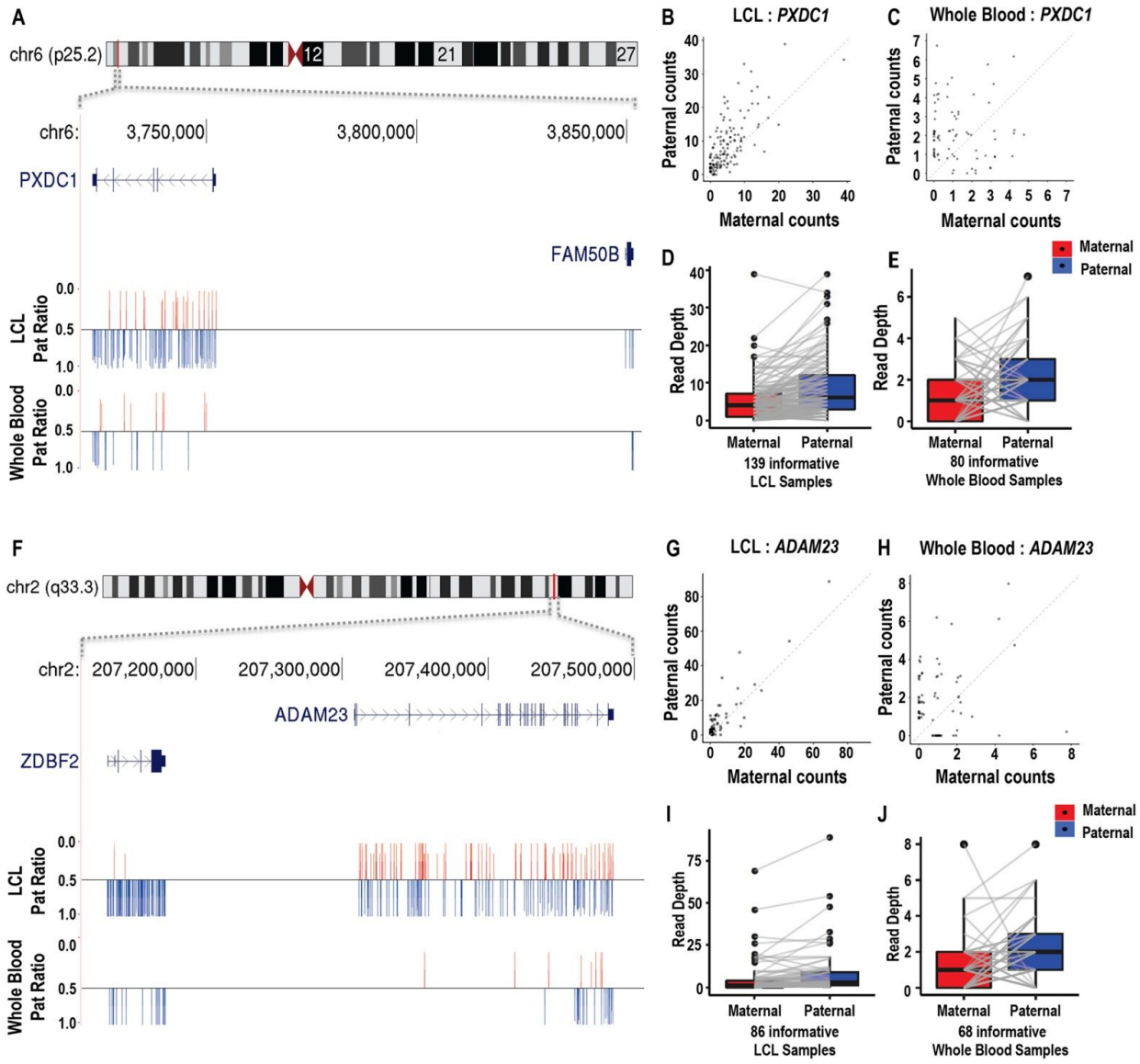

Figure 3. *PXDC1* and *ADAM23* are novel incompletely imprinted genes that lie adjacent to known imprinted genes. (A-E) *PXDC1* lies ~100kb distal to the known paternally expressed gene *FAM50B* at 6p25.2, and, although bi-allelically expressed, shows approximately 2-fold higher expression from the paternal allele in both LCLs (**B** and **D**) and WB (**C** and **E**). **(F-J)** *ADAM23* lies ~130kb distal to the known paternally expressed gene *ZDBF2* at 2q33.3, and also exhibits ~2-fold over-expression from the paternal allele in LCLs (**G** and **I**) and WB (**H** and **J**). In

(A) and **(F)**, the mean fraction of reads transcribed from the paternal allele at every informative SNV position (the Pat ratio) is shown as bar, using a baseline of 0.5 (corresponding to equal expression of the two parental alleles). SNVs with preferential paternal expression (Pat ratio >0.5) are shown in blue, while SNVs with preferential maternal expression (Pat ratio <0.5) are shown in red. In **(D/E)** and **(I/J)**, vectors join the allelic expression values within each informative individual based on the sum of total RNA-seq reads overlapping phased heterozygous SNVs within each gene.

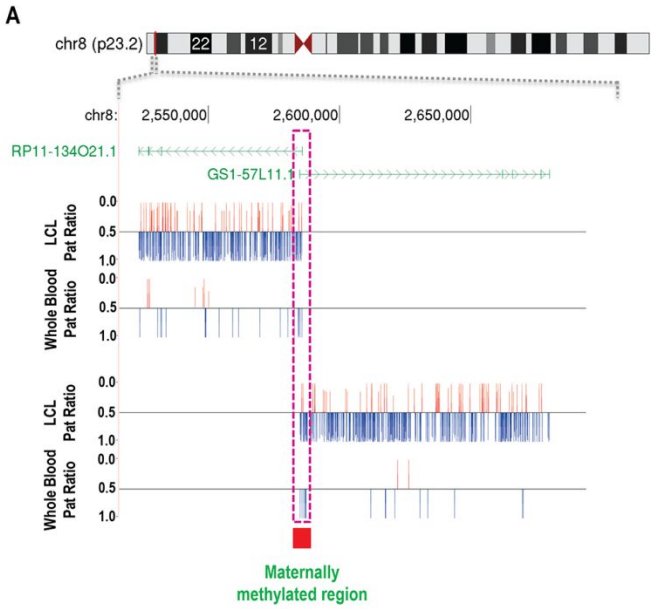

Figure 4. Two novel imprinted transcripts located at 8p23.2 share a bidirectional promoter that coincides with a maternally methylated locus. *RP11-134O21.1* and *GS1-57L11.1* lie in an antisense orientation, and both show ~2-fold expression from the paternal versus maternal allele in LCLs. Prior DNA methylation studies (Joshi et al. 2016) identified a region of increased maternal methylation located at the shared promoter of these two transcripts, confirming parent-of-origin effects at this locus, and indicating this as the likely regulatory element controlling imprinted expression at this locus.

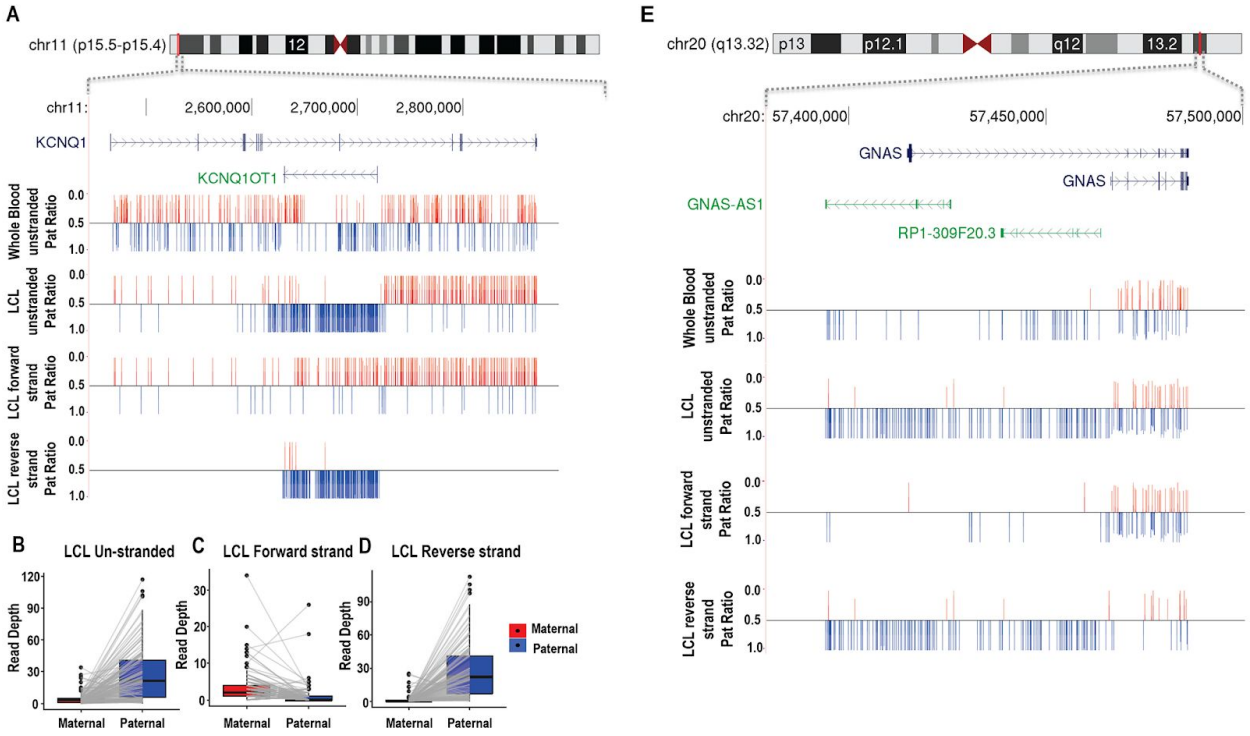

Figure 5. Stranded RNA-Seq data provides improved resolution of imprinting at overlapping antisense genes. Several loci in the genome contain multiple imprinted transcripts, including pairs of overlapping antisense genes with opposite imprinting patterns. Strand-specific RNA-Seq provided considerably improved ability to discern the correct imprinting patterns at these loci when compared to the use of unstranded libraries. **(A-D)** *KCNQ1* and *KCNQ1OT1* lie within the 11p15.5 imprinted region. *KCNQ1* on the plus strand is maternally expressed, while *KCNQ1OT1* on the negative strand is paternally expressed. In whole blood where only unstranded data was available, no significant parental bias was detected from either transcript, likely due to the combined signal from the two overlapping transcripts giving the

appearance of biparental expression. However, the use of stranded RNA-Seq in LCLs clearly shows the that the two transcripts are antisense and have opposite imprinting patterns. **(E)** Similarly, *GNAS* and *GNAS-AS1* are antisense transcripts located in 20q13.32. In LCLs, the stranded RNA-Seq data shows that while *GNAS-AS1* is a paternally expressed imprinted gene, *GNAS* shows biparental expression.

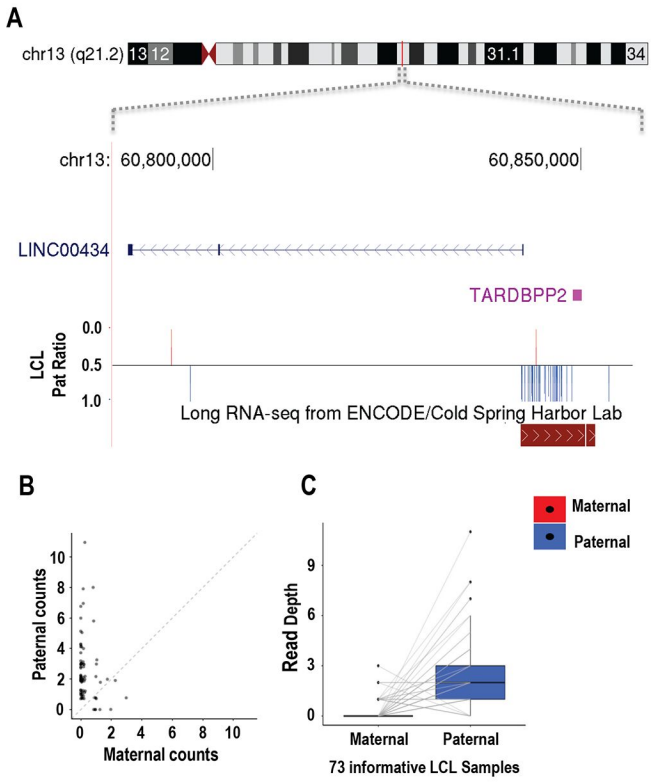

Figure 6. A novel putative imprinted lncRNA at 13q21.2. Using a sliding window analysis to interrogate the genome independent of gene annotations, we identified a cluster of 35 SNVs located in 13q21.2 (chr13:60,841,936-60,848,791, hg19) that showed a strong paternal expression bias. The putative transcript containing these SNVs is located on the forward strand, and apparently shares a bidirectional promoter with the non-coding RNA *LINC00434*. This SNV cluster overlaps a putative anonymous transcript identified in LCLs by the ENCODE project.

List of Supplementary Tables

Supplementary Table 1. All Unique Gene Fragments with FDR q<0.1, prior to manual curation.

Supplementary Table 2. Data for all Unique Gene Fragments in genome.

Supplementary Table 3. Comparison of imprinted genes detected in the current study with those reported by Baran et al. in GTEx samples.

Supplementary Table 4. Data for 25kb sliding windows with FDR q<0.1.

Supplementary Table 5. 165 LCLs used for RNAseq analysis, and their parents.

Supplementary Table 6. 131 whole blood samples used for RNAseq analysis.

List of Supplementary Figures

Supplementary Figure 1. Reference bias can cause false positive signals of imprinting.

Supplementary Figure 2. Significant signals of imprinted expression extend downstream of known imprinted genes.

Supplementary Figure 3. The effect of masking SNV positions and utilizing WASP on reference genome mapping bias.

Supplementary Figure 4. A summary of the analytical pipeline used for identifying parental bias in gene expression in whole blood samples.

Supplementary Figure 5. Definition of unique gene fragments (UGFs).

Supplementary Figure 6. Power estimates for ShrinkBayes and the paired Wilcoxon Signed Rank test on number of individuals and genes.

Supplementary Figure 7. Putative imprinted UGFs identified by ShrinkBayes and/or the paired Wilcoxon Signed Rank test as a function of underlying sample size (L) and mean expression (R).