

# Transcriptome-Wide Survey of Mouse CNS-Derived Cells Reveals Monoallelic Expression within Novel Gene Families

Sierra M. Li<sup>1</sup>, Zuzana Valo<sup>2</sup>, Jinhui Wang<sup>2</sup>, Hanlin Gao<sup>2</sup>, Chauncey W. Bowers<sup>3</sup>, Judith Singer-Sam<sup>2\*</sup>

**1** Division of Biostatistics, Beckman Research Institute, City of Hope National Medical Center, Duarte, California, United States of America, **2** Division of Biology, Beckman Research Institute, City of Hope National Medical Center, Duarte, California, United States of America, **3** Division of Computational Biology, Beckman Research Institute, City of Hope National Medical Center, Duarte, California, United States of America

## Abstract

Monoallelic expression is an integral component of regulation of a number of essential genes and gene families. To probe for allele-specific expression in cells of CNS origin, we used next-generation sequencing (RNA-seq) to analyze four clonal neural stem cell (NSC) lines derived from *Mus musculus* C57BL/6 (B6) × *Mus musculus molossinus* (JF1) adult female mice. We established a JF1 cSNP library, then ascertained transcriptome-wide expression from B6 vs. JF1 alleles in the NSC lines. Validating the assay, we found that 262 of 268 X-linked genes evaluable in at least one cell line showed monoallelic expression (at least 85% expression of the predominant allele,  $p$ -value < 0.05). For autosomal genes 170 of 7,198 genes (2.4% of the total) showed monoallelic expression in at least 2 evaluable cell lines. The group included eight known imprinted genes with the expected pattern of allele-specific expression. Among the other autosomal genes with monoallelic expression were five members of the glutathione transferase gene superfamily, which processes xenobiotic compounds as well as carcinogens and cancer therapeutic agents. Monoallelic expression within this superfamily thus may play a functional role in the response to diverse and potentially lethal exogenous factors, as is the case for the immunoglobulin and olfactory receptor superfamilies. Other genes and gene families showing monoallelic expression include the annexin gene family and the *Thy1* gene, both linked to inflammation and cancer, as well as genes linked to alcohol dependence (*Gabrg1*) and epilepsy (*Kcnma1*). The annotated set of genes will provide a resource for investigation of mechanisms underlying certain cases of these and other major disorders.

**Citation:** Li SM, Valo Z, Wang J, Gao H, Bowers CW, et al. (2012) Transcriptome-Wide Survey of Mouse CNS-Derived Cells Reveals Monoallelic Expression within Novel Gene Families. PLoS ONE 7(2): e31751. doi:10.1371/journal.pone.0031751

**Editor:** Amanda Ewart Toland, Ohio State University Medical Center, United States of America

**Received:** November 15, 2011; **Accepted:** January 12, 2012; **Published:** February 22, 2012

**Copyright:** © 2012 Li et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This project was supported in part by NCI grant P30 CA033572. The work is solely the responsibility of the authors and does not represent the views of the NCI or NIH. No additional external funding was received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: jsam@coh.org

## Introduction

Monoallelic expression (also termed allelic exclusion) refers to the epigenetic silencing of one of the two copies (i.e., alleles) of a gene. Well-studied examples of monoallelic expression in placental mammals include X-inactivation, which affects most genes on an entire chromosome, and imprinting, the parent-of-origin silencing of some autosomal genes. In the case of X-inactivation in females, there is random silencing of one X chromosome in individual cells early in development, and the pattern of silencing is stably inherited through all subsequent somatic cell divisions.

Other genes known to show monoallelic expression can be broadly classified into two categories. The first group includes immunoglobulin and olfactory receptor genes [1,2]. Genes in this group are expressed at high level from a single allele following induction/differentiation. While the mechanisms leading to allele-specific expression differ between immunoglobulin and olfactory receptor gene families, they share intriguing similarities. Monoallelic expression in both cases increases the specificity of a given responder cell as well as the potential diversity within a mixed population of such cells. The two superfamilies also share a similar

chromosomal distribution, with multi-gene clusters distributed on multiple chromosomes. Notably, both superfamilies form part of an individual's defense against highly variable, potentially lethal external agents.

A second category of genes is expressed at low levels prior to induction (e.g. IL-4 or IL-5 in latent T cells) [3]. Following stimulation, the expression level may be greatly increased, accompanied by a shift from monoallelic to biallelic expression. In such cases, allele-specific silencing is said to be stochastic, i.e., variable, in cells of the same lineage [4].

Hybridization-based studies have shown that at least 1% of autosomal genes show a pattern of random monoallelic expression in human lymphoblastoid cells [5] and in mouse neural stem cells [6,7]. In this study we used next-generation RNA sequencing (RNA-seq), which offers major advantages over previous methods for transcriptome-wide analysis of random allele-specific expression. In addition to its accuracy and sensitivity [8], RNA-seq allows expression profiling and cSNP analysis to be accomplished in the same experiment, so that allele-specific expression can be correlated with mRNA abundance (and potential biological significance). Using the Illumina sequencing platform for RNA-

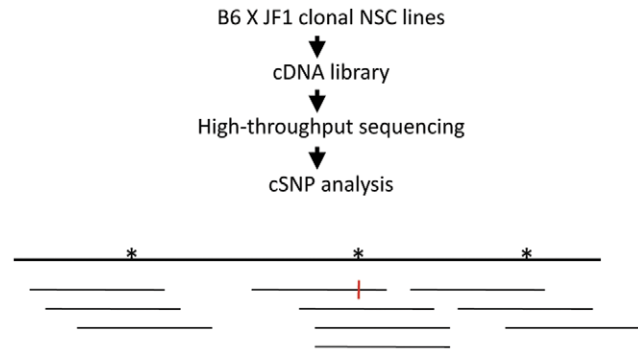
seq analysis of mouse tissues, Mortazavi et al. used “reads per kilobase of exon model per million mapped reads” (RPKM) to quantify transcript levels, and showed that the linear range of transcript detection extended over five orders of magnitude (0.1~10,000 RPKM) [9]. More recent studies using the same platform include an analysis of escape from X-inactivation [10] and surveys of imprinting in mouse brain [11–13].

In this study, we analyzed RNA-seq data using multiple clonal cell lines, and identified monoallelic expression with a low false discovery rate (FDR). Use of these clonal neural stem cell (NSC) lines derived from F<sub>1</sub> mice of both reciprocal crosses allowed us to distinguish imprinting from random allelic exclusion. It also allowed us to detect monoallelic expression that might be missed in a single cell line, since, as discussed above, allelic exclusion is frequently stochastic, i.e., variable in different founder cells (5–7). Our results confirm allele-specific expression of ~2.5% of autosomal genes in CNS-derived stem cells [7]. Furthermore, the annotated list includes genes whose monoallelic expression may be relevant to a number of major inherited disorders of the CNS, as well as disorders that can arise from somatic mutations, including various cancers.

## Results and Discussion

### Transcriptome-wide Profiling of Allele-specific Expression by cSNP-seq

To perform a transcriptome-wide search for random monoallelic expression in cells originating in the CNS, we made use of four clonal NSC lines isolated as previously described [7]. The cells are derived from adult forebrain of female mice resulting from the cross *Mus musculus* C56BL/6 (B6) × *Mus musculus* molossinus JF1 (JF1), and retain the potential to differentiate to populations of neurons and astrocytes [6]. Of the NSC lines analyzed, 2A1, 3A1 and 4A5a were derived from B6<sup>MAT</sup> × JF1<sup>PAT</sup> hybrid mice; the cell line 2A5 was derived from the reciprocal cross, JF1<sup>MAT</sup> × B6<sup>PAT</sup>. We used next-generation sequencing, first to create a cSNP library for JF1 brain, then to analyze allele-specific expression in each of the four NSC lines (“cSNP-seq analysis”). In each case, polyA<sup>+</sup> RNA was converted to double-stranded cDNA. Following addition of linkers specified by Illumina, 250–300 bp fragments were analyzed. For JF1, we analyzed ~132 million single or paired-end reads of length 40–80 bases. Of these, approximately 25–30% high quality reads without gaps could be uniquely aligned to the Refseq transcriptome. We used a threshold of RPKM >= 3 to distinguish expressed genes from background [9]. Of 22,088 Refseq genes expressed in JF1 brain or hybrid cell lines, ~10,000 genes in the JF1 sample had an RPKM >= 3, while ~8,500 genes met the same threshold in each NSC line. Of these, we considered genes to be statistically evaluable if the pooled SNP depth of coverage was at least 10 (Figure 1). Approximately 7100 genes in NSC cells (range: 6925–7373 genes) met these two thresholds, i.e., RPKM >= 3 and pooled SNP depth of coverage >= 10. For these genes, we computed relative B6 and JF1 expression ( $P_{B6}$  and  $P_{JF1}$ , respectively). Genes were considered to show monoallelic expression if  $P_{B6}$  or  $P_{JF1}$  was >= 0.85, with a  $p$ -value <= 0.05 (exact binomial test, null hypothesis:  $P_{B6} = P_{JF1} = 0.5$ ). With the same  $p$ -value threshold, another subset of genes for which  $0.7 <= P_{B6}$  or  $P_{JF1} < 0.85$  were considered to show a trend towards monoallelic expression. Genes for which  $P_{B6}$  or  $P_{JF1} < 0.7$  were considered to show biallelic expression. A small set of genes (~40–90 in each cell line) were excluded from classification because although  $P_{B6}$  or  $P_{JF1} >= 0.7$ , the results were not statistically significant ( $p$ -value > 0.05), mostly due the low depth



**Figure 1. Outline of the cSNP-seq assay.** The diagram shows a hypothetical single-exon gene with 3 SNPs (asterisks), a pooled SNP depth of 10, and a 90% preference for the B6 allele (the red vertical bar shows a JF1 sequence at the cSNP indicated.). doi:10.1371/journal.pone.0031751.g001

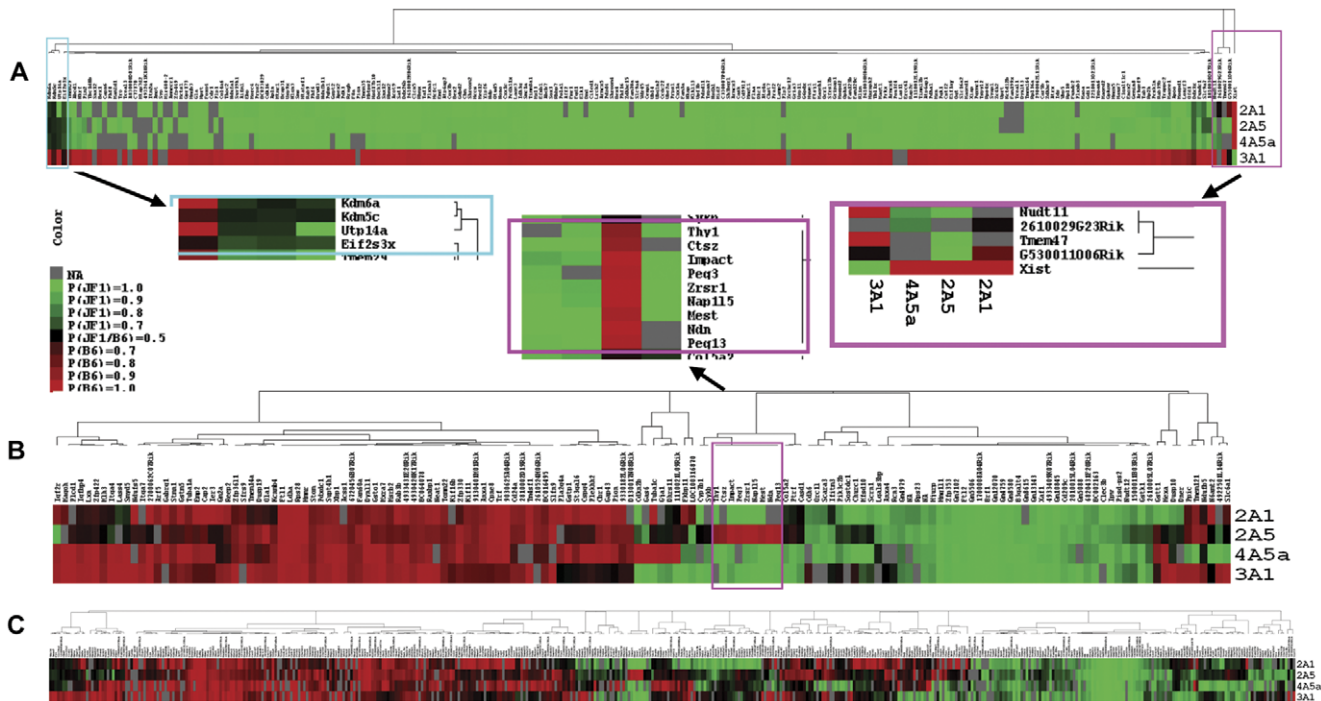
of SNP coverage of the genes. Details of the statistical analysis are in Methods S1. Results for all evaluable genes as well as summaries of Illumina sequencing and mapping are shown in Table S1, Figure S1 and Dataset S1.

### Proof-of-principle: X-linked Genes

To establish the validity of the method, we first analyzed X-linked genes in the NSC lines. A total of 215, 222, 243 and 211 Refseq genes were evaluable in cell lines 2A1, 2A5, 3A1 and 4A5a, using the criteria described above. Taken together, 268 genes were evaluable in at least one cell line, of which 262 genes showed monoallelic expression (97.8%), and 235 genes were evaluable in at least 2 cell lines, of which 231 and 229 showed monoallelic expression in at least one (98.3%) or two (97.4%) lines. The estimated FDR ranged from 0.24% (cell line 2A5) to 2.25% (cell line 4A5a) as shown in Table S1. Figure 2A shows the heatmaps for the expression pattern for the 235 genes evaluable in at least 2 cell lines, with clustering based on strength of monoallelic expression. Each cell line shows predominant expression from only one X chromosome: the JF1-derived X chromosome for cell lines 2A1, 2A5 and 4A5a, and the B6-derived X chromosome for cell line 3A1. *Xist* is in a subcluster by itself since, as expected, it shows the reverse pattern of the other X-linked genes in each cell line. Several genes, clustered at the left corner of the heatmap, show biallelic or less strong monoallelic expression patterns. The cluster contains 3 previously known escape genes: *Kdm6a*, *Kdm5c* and *Eij2s3X*, and also includes *Utp14a*, a small nucleolar p53-binding protein involved in 18 s rRNA maturation [14]. This gene, which shows biallelic expression in two of the four cell lines, maps to XqA4, not in the vicinity of the other known escape genes.

### Allele-specific Expression of Autosomal Genes

On autosomes, ~7000 genes were evaluable in each cell line (range: 6724–7151), with ~200 genes (range: 172–244) showing monoallelic expression. The FDR varied from 1.33% to 3.75% among the 4 hybrid cell lines. Of the 7,198 autosomal genes evaluable in at least 2 cell lines, 476 showed monoallelic expression in at least 1 cell line, of which 170 genes (2.4%) showed monoallelic expression in at least 2 cell lines. Among the 170 genes, 152 genes were evaluable in at least 3 cell lines. The heatmap shown in Figure 2B summarizes results for these 152 genes. Among the genes was a cluster that includes the known paternally expressed genes, *Impact*, *Peg3*, *Zrsr1*, *Nap115*, *Mest*, *Ndn* and *Peg13* (Figure 2B inset). We found that the two novel



**Figure 2. Heatmaps showing results of cSNP analysis.** A. X linked genes ( $n = 235$ ) evaluable for at least 2 out of the 4 cell lines. Insets expand results of hierarchical clustering. *Right inset*: Unique pattern of the *Xist* gene. *Left inset*: Genes that escape X-inactivation (see text). B. Autosomal genes ( $n = 152$ ) that show monoallelic expression in at least 2 of 3 evaluable cell lines. *Inset*: Genes that show imprinted gene expression. C. Autosomal genes ( $n = 476$ ) with monoallelic expression in at least 1 of 2 evaluable cell lines.  
doi:10.1371/journal.pone.0031751.g002

potentially imprinted genes in the cluster, *Ctshz* and *Thy1*, are not imprinted in mouse brain. The known maternally expressed gene, *Igf2r* is in a unique subcluster at the extreme left of the heatmap.

Of the remaining  $\sim 140$  genes, most show a strain- or sequence-specific pattern, i.e., the same allele (B6 or JF1) is favored in each cell line that shows monoallelic expression, similar to a pattern noted previously [15]. An additional group of  $\sim 20$  genes shows a random pattern, i.e., the gene is expressed from the B6 allele, the JF1 allele or both alleles, depending upon the cell line. The variable pattern of monoallelic and biallelic expression across several cell lines is consistent with previous studies of autosomal monoallelic expression [4–6]. The heatmap in Figure 2C shows a similar pattern for 476 genes with monoallelic expression in at least 1 of 2 evaluable cell lines.

Previous studies have shown that for specific genes expressed from either one or both alleles, the expression level is frequently higher for cells showing biallelic expression [5,6]. Our study allowed us to examine transcriptome-wide differences in expression levels associated with monoallelic vs. biallelic expression. We found, by two statistical analyses, that monoallelic expression is correlated with a  $\sim 30$ – $35\%$  reduction in expression level across a broad range of RPKM.

In the first analysis, for the 313 genes that showed a mixed pattern of biallelic expression in some cell lines and monoallelic expression in others, we compared average expression levels in the respective cell lines. Figure 3 shows that there is a trend towards lower RNA abundance in cell lines expressing only one allele. The mean transcript level of a gene in cells with monoallelic expression was 64.17% that of cells with biallelic expression (i.e., the mean  $\log_2(\text{RPKM})$  difference was 0.64 (95% CI 0.51–0.78), which is statistically significant by the paired t-test at a  $p$ -value of  $2.2e^{-16}$ ). If the number of expressed alleles were the only factor, the

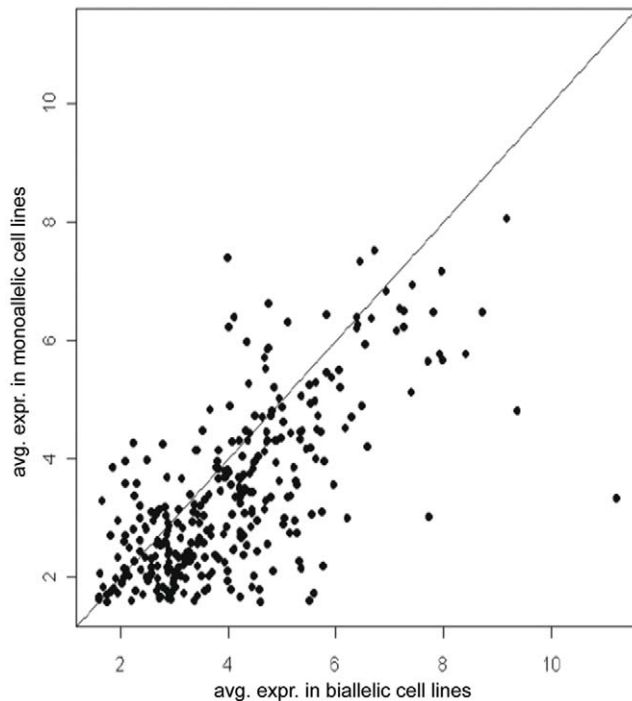
predicted mean transcript ratio would be 0.5. Therefore, our finding of a ratio of 0.64 suggests modulation by additional regulatory elements. These results are comparable to the effect of gene dosage on global gene expression [16,17].

We found a similar statistically significant  $\sim 30\%$  difference when we compared expression levels (RPKM) for genes that showed monoallelic vs. those that showed biallelic expression in each cell line (Figure S2). In both statistical analyses, the broad distribution of RPKM for genes showing monoallelic as well as biallelic expression demonstrates that a sizable number of genes with allelic exclusion are expressed at moderate to high levels.

To determine whether the genes showing monoallelic expression were clustered on particular autosomes, we plotted the chromosomal distribution of these genes in each NSC line (Figure S3). The genes showing single-allele expression are broadly distributed on all autosomes, rather than confined to specific chromosomes or chromosomal regions. These results are consistent with an absence of chromosome-wide coordination for autosomal monoallelic expression [5].

### Enriched Monoallelic Expression of Specific Gene Families: Glutathione Transferases, Annexins and Protocadherins

We used the DAVID Bioinformatics Resource to search for enrichment of particular categories of genes. Considering protein domains and Kegg pathways, we found several categories of genes to be enriched among the genes showing monoallelic expression in at least 2 cell lines (Table S2). Included were five genes of the glutathione superfamily ( $\sim 17$ -fold enrichment, FDR 1.21% to 2.37%). When genes showing monoallelic expression in at least 1 cell line were analyzed ( $n = 577$ ), additional genes appeared, including 4 genes of the annexin gene superfamily ( $\sim 12$ -fold



**Figure 3. Expression levels vs. monoallelic or biallelic expression.** Results are plotted for 313 genes that show monoallelic expression in some NSC lines and biallelic expression in others. For each gene the average  $\log_2(\text{RPKM})$  in NSC lines with monoallelic expression (y-axis) is plotted vs. the  $\log_2(\text{RPKM})$  in cell lines with biallelic expression. doi:10.1371/journal.pone.0031751.g003

enrichment, FDR  $\sim 6\%$ ). Notably, protocadherins also appeared (7–15 fold enrichment, FDR  $< 10^{-6}\%$ ). These members of the cadherin superfamily provide additional positive controls, since they are known to show monoallelic expression [18]. A recent study suggests that monoallelic expression of protocadherins may contribute to specificity of neuronal circuitry [19].

### Monoallelic Expression of Gst superfamily Genes Following Differentiation of NSC Lines

Proteins of the abundant glutathione transferase (Gst) superfamily catalyze the addition of the tri-peptide glutathione to electrophilic compounds, resulting in detoxification of chemical carcinogens and pollutants, and also a decrease in the efficacy of some anticancer therapeutics (reviewed in Hayes et al. [20]). Gst proteins have a major additional role in adaptation to oxidative and inflammatory stress, resulting from the same catalytic activity applied to endogenous intermediates of the stress response.

Gst proteins are found in the cytosol (including the Gsta, Gstm, Gstp, Gstt gene families), mitochondria (the Gstk family) and microsomes. The different families, which are categorized based on sequence similarity, map to different clusters on multiple chromosomes, but the genes within each family are clustered at the same location. Gst proteins are typically dimers, with proteins in the Gsta and Gstm families capable of forming heterodimers with other subunits in the same group.

There have been many reports of altered Gst expression in tumors (see, for example, Hayes and Pulford [21]). Because of the broad, overlapping specificity of Gst genes, knockout experiments of a single subunit have not always resulted in dramatic defects. In an important experiment, homozygous knockout of the 2 subunits of the mouse Gstm complex was shown to result in increased

skin tumorigenesis in response to carcinogens and tumor promoters [22].

Because of the compelling interest of the Gst superfamily, we used RT-PCR to analyze further five Gst genes that showed monoallelic expression by Illumina sequencing: *Gstk1*, *Gstm5*, *Gsto1*, *Gstp1* and *Gstt1*, each from a different subfamily. Table 1 confirms that the genes show monoallelic expression in multiple cell lines. Moreover, when the cell lines are differentiated to astrocytes and neurons, the allele-specific pattern of expression is almost entirely preserved. An example of the results obtained by the two independent methods is shown for the *Gstp1* gene in Figure 4. Figure 4A shows the location of the three cSNPs (in the 3<sup>rd</sup>, 6<sup>th</sup> and 7<sup>th</sup> exons, respectively) for which data was obtained by Illumina sequencing. The inset shows that two of the cell lines express the B6 allele only, while two others show biallelic expression. The RPKM indicate high levels of expression, approaching that of housekeeping genes in the four cell lines, even in the cell lines showing B6 expression only. On average, the RPKM are two-fold higher in the two lines showing biallelic expression, consistent with the results shown in Figure 3. Figure 4B shows the results of automated sequencing following RT-PCR with primers flanking one of the *Gstp1* cSNPs. Results are shown for NSC lines prior to and following differentiation to astrocytes and neurons. Complete RT-PCR results for all of the Gst genes assayed are shown in Figure S4.

### Additional Genes of Interest Showing Monoallelic Expression

Table 2 lists the five Gst genes as well as additional genes of interest validated by allele-specific RT-PCR. Results for these

**Table 1. RT-PCR analysis of allele-specific expression of the Gst gene family in undifferentiated NSCs, and following differentiation to neurons and astrocytes.**

Gene	Cell type	Neural stem cell line				
		2A1	2A5	3A1	4A5	4B3
<i>Gstk1</i>	<i>undiff</i>	100% J††	biallelic†	100% J††	100% J	100% J
	<i>astro**</i>	100% J	89% J	100% J	100% J	100% J
	<i>neurons**</i>	100% J	92% J	100% J	100% J	100% J
<i>Gstm5</i>	<i>undiff</i>	98% B††	biallelic†	86% B††	100% B	99% B
	<i>astro**</i>	100% B	80% B	biallelic	95% B	100% B
	<i>neurons**</i>	100% B	94% B	86% B	100% B	100% B
<i>Gsto1</i>	<i>undiff</i>	100% B††	84% B†	100% B††	100% B	100% B
	<i>astro**</i>	100% B	86% B	100% B	90% B	100% B
	<i>neurons**</i>	100% B	100% B	98% B	100% B	100% B
<i>Gstp1</i>	<i>undiff</i>	100% B††	biallelic†	biallelic†	100% B	biallelic
	<i>astro**</i>	100% B	biallelic	biallelic	100% B	biallelic
	<i>neurons**</i>	100% B	biallelic	biallelic	100% B	biallelic
<i>Gstt1</i>	<i>undiff</i>	100% J††	biallelic†	biallelic†	96% J	92% B
	<i>astro**</i>	97% J	91% B	<i>n.d.</i>	100% J	100% B
	<i>neurons**</i>	94% J	biallelic	biallelic	100% J	100% B

Numerical values are shown for all RT-PCR samples with at least a 0.8 preference for the B6 (B) or JF1 (J) allele. Other values are listed as biallelic (see Figure S4 for representative data and technical replicates.) *n.d.*, not detectable.

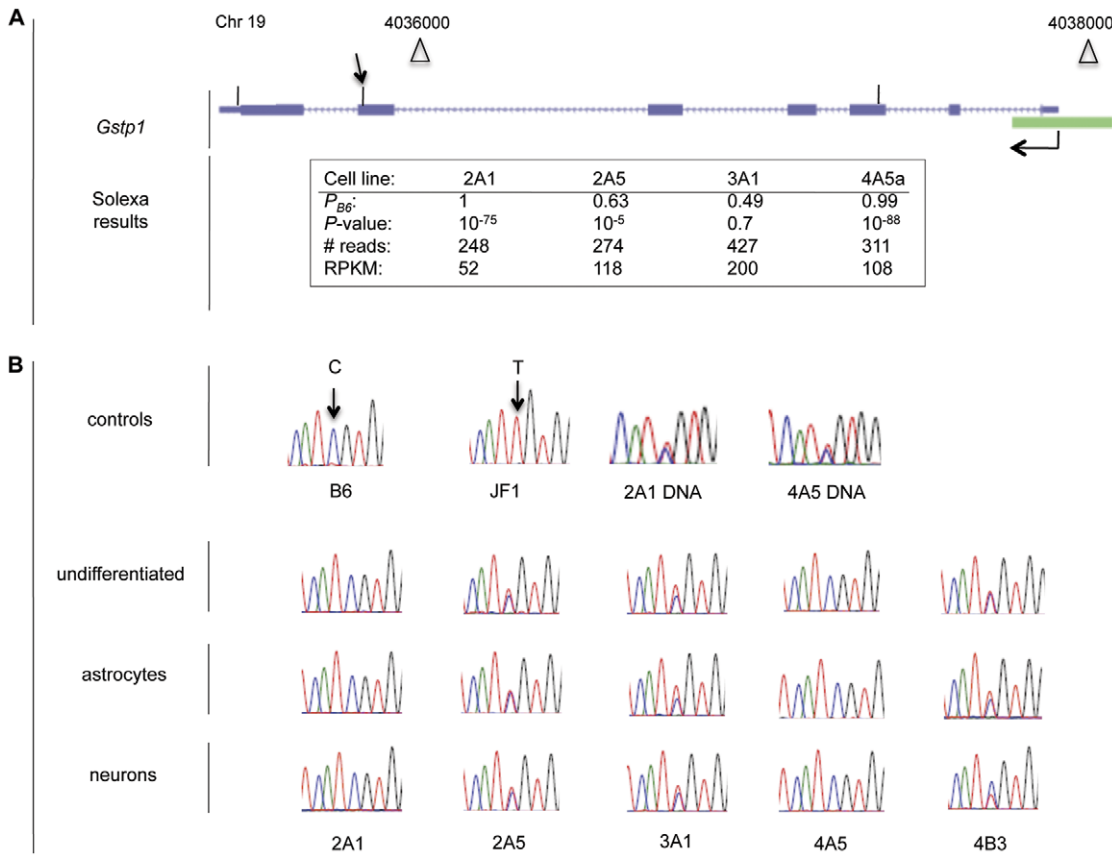
\*\*Asterisks denote concordance with undifferentiated NSCs,  $p$ -value  $< 0.001$ .

Daggers indicate results for samples evaluated by Illumina sequencing:

† Biallelic expression at  $P_{B6}$  or  $P_{JF1} < 0.85$ ,

†† Monoallelic expression at  $P_{B6}$  or  $P_{JF1} > 0.85$ .

doi:10.1371/journal.pone.0031751.t001



**Figure 4. Allele-specific expression of *Gstp1*.** A. Structure of the gene and results of Illumina sequencing in four NSC lines. The chromosomal location, start site and orientation of transcription, the 7 exons, and a CpG island (green rectangle) at the 5' terminus are shown. The cSNPs in exons 3, 6 and 7 are indicated by the vertical lines above the respective exons. The table shows that the B6 allele is expressed in cell lines 2A1 and 4A5a, while both alleles are expressed in cell lines 2A5 and 3A1. B. RT-PCR results. First row: Chromatograms show a C/T cSNP between B6 and JF1 transcripts, and the presence of both alleles in genomic DNA of NSC lines 2A1 and 4A5. The arrow in Figure 4A indicates the location of the cSNP. Second to fourth rows: Analysis of RNA isolated from five NSC lines in the undifferentiated state, and differentiated to astrocytes and neurons, as indicated. The B6 allele is expressed in cell lines 2A1 and 4A5, while both alleles are expressed in cell lines. 2A5, 3A1 and 4B3. doi:10.1371/journal.pone.0031751.g004

**Table 2.** Selected list of autosomal genes with monoallelic expression by Illumina sequencing validated by RT-PCR.

Gene	Mono/Eval NSCs <sup>†</sup>	RPKM		Chr	Description
		Mono NSCs	JF1 brain		
<i>Anxa1</i>	3/4 (B6)	10 to 78	2	19	Annexin I, anti-inflammatory pathway
<i>Anxa2</i>	1/4 (B6)	77	63	9	Annexin II, cell growth, signal transduction
<i>Chl1</i>	1/4 (JF1)	37	48	6	Cell adhesion molecule (L1CAM homolog)
<i>Gabrg1</i>	2/3 (B6)	5 to 6	18	5	GABA A receptor, gamma 1 (alcoholism)
<i>Gm2a</i>	2/4 (B6)	39 to 66	55	11	GM2-ganglioside activator
<i>Gstk1</i>	3/4 (JF1)	11 to 22	15	6	Glutathione S-transferase, kappa 1
<i>Gstm5</i>	3/4 (B6)	49 to 91	70	3	Glutathione S-transferase, mu 5
<i>Gsto1</i>	3/4 (B6)	17 to 20	30	19	Glutathione S-transferase, omega 1
<i>Gstp1</i>	2/4 (B6)	52 to 200	108	19	Glutathione S-transferase, pi 1
<i>Gstt1</i>	2/4 (B6/JF1)	4 to 64	9	10	Glutathione S-transferase, theta 1
<i>Hexa</i>	3/4 (B6/JF1)	80 to 111	36	9	Hexosaminidase A (Tay-Sachs Disease)
<i>Kcna1</i>	1/4 (JF1)	8	54	14	Calcium-activated potassium channel
<i>Thy1</i>	3/3 (B6/JF1)	14 to 173	535	9	Thy-1 cell surface antigen

<sup>†</sup>Number of cell lines with monoallelic expression/number of evaluable cell lines. In column 2, the expressed allele(s) are shown in parentheses. The values shown for columns 2–4 are from Illumina sequencing (see text). doi:10.1371/journal.pone.0031751.t002

additional genes are shown in Figure S4 and summarized in Table S3.

The additional genes analyzed are potentially relevant to diverse genetic and somatic disorders. *Chil1*, “close homologue of L1”, is a cell adhesion gene, and member of the immunoglobulin superfamily. *Chil1* plays a role in neuronal migration and survival [23], and in adult mice, it plays a necessary role in the cycling process of synaptic vesicles [24]. *Chil1* knockout mice show behavioral defects [25], and in human the homologous *CALL* gene has been implicated in schizophrenia [26]. Although only one cell line, 3A1, showed monoallelic expression for this gene (in the undifferentiated state as well as following differentiation to neurons and astrocytes), its expression pattern warrants further study because of its importance in the CNS.

*Gabrg1* codes for a subunit of the GABA-A subgroup of receptors for the inhibitory neurotransmitter  $\gamma$ aminobutyric acid (GABA). It is one of two such genes that has been implicated in alcohol dependence in humans [27,28]. *Gabrg1* shows monoallelic expression in 2 of 3 evaluable NSC lines (2A1 and 4A5a). Furthermore, RT-PCR results demonstrate allele-specific expression (2A1, 4A5) or a trend (4B3), which is maintained during *in vitro* differentiation.

The functions of the *Kcnma1* gene are varied, with generalized epilepsy as one of the major phenotypes associated with gene malfunction [29]. Our cSNP-seq analysis shows that the *Kcnma1* gene is expressed from the JF1 allele in cell line 2A1; RT-PCR demonstrates the same pattern in NSC lines 2A5 and 4A5, with a trend toward JF1 expression in cell lines 2A1 and 3A1. Monoallelic expression or a strong trend is preserved in these lines as the cells differentiate. The *Kcnma1* gene is large, and alternate splicing has been shown to affect function. We therefore examined in detail the cSNP-seq results for all 7 exons containing evaluable SNPs, and found the same pattern of preference for the JF1 allele.

The annexin family, comprising a group of phospholipid-calcium binding proteins, plays an important role in signal transduction (reviewed in Perretti and D’Acquisto [30]). *Anxa1*, a well-characterized member of the family is a component of the anti-inflammatory pathway, including in the CNS [31]. The enzyme also protects against DNA damage in breast cancer cells [32]. RT-PCR analysis confirms monoallelic expression (B6 allele) in cell lines 2A1 and 2A5, and, in addition, shows expression of the same allele in cell line 4B3. Two of the cell lines (2A1 and 4B3) preserve the same pattern following differentiation to primitive neurons. We also confirmed monoallelic expression of a second member of the annexin gene family, *Anxa2*.

*Hexa* and *Gm2a* code for two interacting protein subunits, mutations of which result in “GM2 gangliosidosis”, Tay-Sachs Disease in the case of Hexa, and “Tay-Sachs-like Disease” in the case of Gm2a. The *Hexa* gene codes for a subunit of  $\beta$ -hexosaminidase A, a lysosomal enzyme involved in the breakdown of gangliosides, while the *Gm2a* gene, codes for a small glycolipid transport protein that is a co-factor for  $\beta$ -hexosaminidase A. Together they catalyze the lysosomal breakdown of the ganglioside GM2 (or other N-acetyl hexosamines). The subunit most commonly mutated is Hexa, resulting in an autosomal recessive disease that affects neurons in the CNS [33]. This inheritance pattern and the relative lack of phenotypic variation in heterozygotes for Tay Sachs Disease suggest that random monoallelic expression does not noticeably influence (human) disease development. However understanding the role of monoallelic expression in regulation of these genes may be important in the ultimate design of gene-based therapies.

*Gm2a* shows monoallelic expression (B6 allele) in three of the cell lines assayed by RT-PCR, with 1 and 2 cell lines maintaining the

pattern, in neurons and astrocytes, respectively. The *Hexa* gene shows more robust preservation of the monoallelic pattern. Of the three cell lines showing monoallelic expression, 2A1 and 4A5 express the JF1 allele, while 3A1 expresses the B6 allele. This pattern is maintained in both differentiated cell types except for 4A5, which shows a trend towards JF1 monoallelic expression.

Allele-specific expression of the *Thy1* gene (CD90, thymocyte differentiation antigen 1) in various cells and states of differentiation is of interest because of its potential broad-based functions in apoptosis, axon growth regulation, cell adhesion, inflammation and metastasis [34]. cSNP-seq analysis revealed monoallelic expression of *Thy1* in the three NSC lines in which it was detected (JF1 allele, 2A1 and 4A5a; B6 allele, 2A5). RT-PCR confirmed expression of the JF1 allele in cell line 2A1, and a trend towards expression of the B6 allele in cell line 2A5. RT-PCR also showed expression of the JF1 allele in three additional cell lines, 3A1, 4A5 and 4B3. The pattern is generally maintained following differentiation to astrocytes and neurons (Table S3).

### Random vs. haplotype-dependent monoallelic expression

Figure 2B suggests that haplotype-dependent (i.e., sequence-specific or strain specific) expression may account for a significant proportion of the 152 autosomal genes showing monoallelic expression in at least 2 of 3 evaluable cell lines. Of these genes, 16 genes showed monoallelic expression from the B6 allele only and 59 genes showed expression from the B6 allele or biallelic expression; 28 genes showed expression from the JF1 allele only, while 23 genes showed expression from the JF1 allele or biallelic expression. In addition to the 8 known imprinted genes, 18 genes showed monoallelic expression from both the B6 and JF1 alleles, providing a clear demonstration of haplotype-independence. These are listed in Table S5.

The apparent bias towards expression of the same allele in multiple cell lines is consistent with previous reports of haplotype-dependent monoallelic expression [15,35]. However, we note that our present data cannot easily distinguish between strain-dependent and random monoallelic expression for the following reasons: 1) The 4 NSC lines we analyzed are derived from the same inbred hybrid mouse strain, yet the majority of genes showing a strain-specific bias also show biallelic expression in some cell lines. This variability between genetically identical cell lines suggests a stochastic component. 2) Our previous results suggest that there may be selection of a small subpopulation of cells in the parental NSC lines prior to generation of clones [7]. This may result in an impression of bias towards one allele across 4 cell lines even in cases of random monoallelic expression. 3) Of the genes we analyzed by RT-PCR, only 3 genes show monoallelic expression from either the B6 or JF1 allele, while 10 genes show bias towards one allele; the pattern is preserved in most cell lines during *in vitro* differentiation to neurons and glia (Table 1, Table S3). However, there is essentially equal expression from both parental alleles in F<sub>1</sub> hybrid adult cortex, the tissue from which the NSC lines were derived (Figure S4). These results are not consistent with haplotype-specific expression that extends to mature neurons and/or glia. In conclusion, for most genes, our current results are consistent with strain-specific and/or random monoallelic expression. Future studies should allow us to distinguish between these possibilities.

### Perspective: Relevance of Results to Human Disease

We report here the first transcriptome-wide direct analysis of monoallelic expression in clonal cells of neural origin. We have previously noted the potential connection between inheritance

patterns of a number of major human disorders and monoallelic expression of relevant genes [6,7]. For example, the diseases associated with *Chl1* (schizophrenia), *Gabrg1* (alcohol dependence) and *Kcnma1* (epilepsy) all display twin discordance consistent with random monoallelic expression. In addition, Table 2 illustrates a potential link between genes that show allele-specific expression and signal transduction, mutagenesis, and the response to stress and inflammation. Consistent with these functions, there have been numerous studies linking changes in expression of glutathione transferases, particularly *Gspt1* and annexin genes to various cancers (reviewed in Tew et al. [36] and Mussunoor and Murray [37]). In these cases, heritable monoallelic expression in a given clone of cells may be considered as the equivalent of loss of heterozygosity. In this scenario, mutation of the one active allele of a tumor suppressor or potential oncogene might lead to rapid carcinogenesis, depending upon the pathway.

## Methods

Detailed experimental procedures are in Supporting Information. The JF1 cSNP library is available at the dbSNP database ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). Results at each cSNP for the 4 cell lines analyzed by Illumina sequencing are available online at <http://genome.ucsc.edu>.

## Ethics statement

Procedures on mice involved little or no pain or distress and were approved by the Institutional Animal Care and Use Committee at the City of Hope (IACUC #97013, approved until 12/20/2011.) City of Hope is accredited by AAALAC.

## Mice, cell lines and cSNP-seq analysis

B6×JF1 hybrid mice and the clonal NSC lines 2A1, 2A5, 3A1, 4A5 and 4B3 have been previously described [7]. NSC line 4A5a is a subclone of 4A5. Next-generation sequencing was carried out by the City of Hope DNA sequencing/Solexa Core. Details of sequencing, alignment and data analysis are in Methods S1.

## RT-PCR

Conditions of RT-PCR, and quantitative automated sequencing of amplified products were previously described [7]. Primers used are listed in Table S4. Differentiation of NSC lines to astrocytes and neurons was previously described [6].

## Supporting Information

**Figure S1** cSNP-seq analysis of JF1 and NSC transcripts. (A) Short reads sequenced and mapped with different filtering criteria. (B) Lower-left panels: scatter plots of log<sub>2</sub>-transformed RPKM for JF1 and 4 hybrid cell lines. Lowess smoothing is used to draw the lines. Upper-right panels: Calculated Pearson correlation among the samples. \*\*\* indicates the correlations are significant at  $p$ -value < 0.001 from a t-test. (C) Distribution of JF1 SNP depth of coverage for 118,579 cSNPs identified on 12,416 Refseq genes. SNPs with less than 3 or larger than 2000 coverage have been removed. (D) Chromosomal distribution of Refseq genes with monoallelic expression in the 4 NSC lines. (PDF)

**Figure S2** Quantile plot for Refseq genes with biallelic or monoallelic expression. (A) 2A1, (B) 2A5, (C) 3A1 and (D) 4A5a. In each case, the log<sub>2</sub>(RPKM) for genes with biallelic expression is consistently higher than for genes with monoallelic expression up to

the 80<sup>th</sup> to 90<sup>th</sup> quantile. We observe an approximately 0.5 reduction of log<sub>2</sub>(RPKM) for genes with monoallelic expression compared to genes with biallelic expression ( $\sim 4.0$  log<sub>2</sub>(RPKM) vs.  $\sim 4.5$  log<sub>2</sub>(RPKM)). This equals about a 30% reduction in raw RPKM, i.e. transcript levels, for genes with monoallelic expression. The result is statistically significant for all 4 hybrid cell lines, the largest  $p$ -value being 0.0002 (two sample t-test) for 2A5. (PDF)

**Figure S3** Chromosomal location of autosomal genes with monoallelic expression in NSC lines. The cell lines are indicated in panels A–D. (PDF)

**Figure S4** RT-PCR results for selected genes with monoallelic expression. RT-PCR primers were selected to include a B6/JF1 SNP within the amplified product. For each gene, the first row shows the results of automated sequencing of RNA from brain tissue of B6 and JF1 mice and F1 hybrid progeny, as well as DNA of selected NSC lines. The following 3 rows show results for the NSC lines indicated, both for undifferentiated NSCs and following differentiation to astrocytes or neurons. In each case, the % expression of the predominant allele is shown. Technical replicates are in parentheses. For  $n=2$ , the range is shown; for  $n=3$ , the SEM is given. A1, *Anxa1*; A2, *Anxa2*; B, *Chl1*; C, *Gabrg1*; D, *Gm2a*; E, *Gstk1*; F, *Gstm5*; G, *Gsto1*; H, *Gstp1*; I, *Gstt1*; J, *Hexa*; K, *Kcnma1*; L, *Thy1*. (PDF)

**Table S1** Number of Refseq genes in each category of allele-specific expression by cell line. (DOC)

**Table S2** Enrichment of autosomal gene clusters showing monoallelic expression by use of the DAVID Bioinformatics Site. (DOC)

**Table S3** Allele-specific expression of selected genes in NSC lines before and after differentiation to astrocytes or neurons. (DOC)

**Table S4** List of primers. (DOC)

**Table S5** Genes with random (i.e., haplotype-independent) monoallelic expression detected by RNA-seq. (DOC)

**Methods S1** Details of Illumina sequencing, alignment and data analysis. (DOC)

**Dataset S1** Expression and classification of monoallelic expression for genes evaluable in NSC cell lines 2A1, 2A5, 3A1 and 4A5a. (XLS)

## Acknowledgments

We thank Jackson Champer for analysis of X-linked escape genes and Haiqing Li of the Bioinformatics Core for his help.

## Author Contributions

Conceived and designed the experiments: SML JSS. Performed the experiments: ZV JW. Analyzed the data: SML CWB JSS. Contributed reagents/materials/analysis tools: JW HG. Wrote the paper: SML JSS.

## References

- Cedar H, Bergman Y (2008) Choreography of Ig allelic exclusion. *Current Opinion in Immunology* 20: 308–317.
- Chess A, Simon I, Cedar H, Axel R (1994) Allelic inactivation regulates olfactory receptor gene expression. *Cell* 78: 823–834.
- Kelly BL, Locksley RM (2000) Coordinate regulation of the IL-4, IL-13, and IL-5 cytokine cluster in Th2 clones revealed by allelic expression patterns. *J Immunol* 165: 2982–2986.
- Guo L, Hu-Li J, Paul WE (2005) Probabilistic regulation in TH2 cells accounts for monoallelic expression of IL-4 and IL-13. *Immunity* 23: 89–99.
- Gimelbrant A, Hutchinson JN, Thompson BR, Chess A (2007) Widespread monoallelic expression on human autosomes. *Science* 318: 1136–1140.
- Wang J, Valo Z, Bowers CW, Smith DD, Liu Z, et al. (2010) Dual DNA methylation patterns in the CNS reveal developmentally poised chromatin and monoallelic expression of critical genes. *PLoS ONE* 5: e13843.
- Wang J, Valo Z, Smith D, Singer-Sam J (2007) Monoallelic Expression of Multiple Genes in the CNS. *PLoS ONE* 2: e1293.
- Bjornsson HT, Albert TJ, Ladd-Acosta CM, Green RD, Rongione MA, et al. (2008) SNP-specific array-based allele-specific expression analysis. *Genome Res* 18: 771–779.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5: 621–628.
- Yang F, Babak T, Shendure J, Disteche CM (2010) Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* 20: 614–622.
- Gregg C, Zhang J, Butler JE, Haig D, Dulac C (2010) Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* 329: 682–685.
- Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, et al. (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* 329: 643–648.
- Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, et al. (2008) Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS ONE* 3: e3839.
- Hu L, Wang J, Liu Y, Zhang Y, Zhang L, et al. (2011) A small ribosomal subunit (SSU) processome component, the human U3 protein 14A (hUTP14A) binds p53 and promotes p53 degradation. *J Biol Chem* 286: 3119–3128.
- Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, et al. (2008) Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat Genet* 40: 904–908.
- Kahlem P, Sultan M, Herwig R, Steinfath M, Balzereit D, et al. (2004) Transcript level alterations reflect gene dosage effects across multiple tissues in a mouse model of down syndrome. *Genome Res* 14: 1258–1267.
- Deng X, Disteche CM (2010) Genomic responses to abnormal gene dosage: the X chromosome improved on a common strategy. *PLoS Biol* 8: e1000318.
- Esumi S, Kakazu N, Taguchi Y, Hirayama T, Sasaki A, et al. (2005) Monoallelic yet combinatorial expression of variable exons of the protocadherin-alpha gene cluster in single neurons. *Nat Genet* 37: 171–176.
- Zipursky SL, Sanes JR (2010) Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly. *Cell* 143: 343–353.
- Hayes JD, Flanagan JU, Jowsey IR (2005) Glutathione transferases. *Annu Rev Pharmacol Toxicol* 45: 51–88.
- Hayes JD, Pulford DJ (1995) The glutathione S-transferase supergene family: regulation of GST and the contribution of the isoenzymes to cancer chemoprotection and drug resistance. *Crit Rev Biochem Mol Biol* 30: 445–600.
- Henderson CJ, Smith AG, Ure J, Brown K, Bacon EJ, et al. (1998) Increased skin tumorigenesis in mice lacking pi class glutathione S-transferases. *Proc Natl Acad Sci U S A* 95: 5275–5280.
- Chen S, Mantei N, Dong L, Schachner M (1999) Prevention of neuronal cell death by neural adhesion molecules L1 and CHL1. *J Neurobiol* 38: 428–439.
- Andreyeva A, Leshchynska I, Knepper M, Betzel C, Redecke L, et al. (2010) CHL1 is a selective organizer of the presynaptic machinery chaperoning the SNARE complex. *PLoS ONE* 5: e12018.
- Morellini F, Lepsveridze E, Kahler B, Dityatev A, Schachner M (2007) Reduced reactivity to novelty, impaired social behavior, and enhanced basal synaptic excitatory activity in perforant path projections to the dentate gyrus in young adult mice deficient in the neural cell adhesion molecule CHL1. *Mol Cell Neurosci* 34: 121–136.
- Sakurai K, Migita O, Toru M, Arinami T (2002) An association between a missense polymorphism in the close homologue of L1 (CHL1, CALL) gene and schizophrenia. *Mol Psychiatry* 7: 412–415.
- Ray LA, Hutchison KE (2009) Associations among GABRG1, level of response to alcohol, and drinking behaviors. *Alcohol Clin Exp Res* 33: 1382–1390.
- Covault J, Gelernter J, Jensen K, Anton R, Kranzler HR (2008) Markers in the 5'-region of GABRG1 associate to alcohol dependence and are in linkage disequilibrium with markers in the adjacent GABRA2 gene. *Neuropsychopharmacology* 33: 837–848.
- Du W, Bautista JF, Yang H, Diez-Sampedro A, You SA, et al. (2005) Calcium-sensitive potassium channelopathy in human epilepsy and paroxysmal movement disorder. *Nat Genet* 37: 733–738.
- Perretti M, D'Acquisto F (2009) Annexin A1 and glucocorticoids as effectors of the resolution of inflammation. *Nat Rev Immunol* 9: 62–70.
- McArthur S, Cristante E, Paterno M, Christian H, Roncaroli F, et al. (2010) Annexin A1: a central player in the anti-inflammatory and neuroprotective role of microglia. *J Immunol* 185: 6317–6328.
- Nair S, Hande MP, Lim LH (2010) Annexin-1 protects MCF7 breast cancer cells against heat-induced growth arrest and DNA damage. *Cancer Lett* 294: 111–117.
- Beutler E, Kuhl W, Comings D (1975) Hexosaminidase isozyme in type O Gm2 gangliosidosis (Sandhoff-Jatzkewitz disease). *Am J Hum Genet* 27: 628–638.
- Rege TA, Hagood JS (2006) Thy-1 as a regulator of cell-cell and cell-matrix interactions in axon regeneration, apoptosis, adhesion, migration, cancer, and fibrosis. *FASEB J* 20: 1045–1054.
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464: 768–772.
- Tew KD, Manevich Y, Grek C, Xiong Y, Uys J, et al. (2011) The role of glutathione S-transferase P in signaling pathways and S-glutathionylation in cancer. *Free Radic Biol Med*.
- Mussunoor S, Murray GI (2008) The role of annexins in tumour development and progression. *J Pathol* 216: 131–140.