Copy number variation modifies expression time courses

Evelyne Chaignat,[1] Emilie Aït Yahya-Graison,[1] Charlotte N. Henrichsen,[1] Jacqueline Chrast,[1] Frédéric Schütz,[1,2] Sylvain Pradervand,[1,2] and Alexandre Reymond[1,3]

[1]*Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland;* [2]*Swiss Institute of Bioinformatics (SIB), 1015 Lausanne, Switzerland*

A preliminary understanding into the phenotypic effect of DNA segment copy number variation (CNV) is emerging. These rearrangements were demonstrated to influence, in a somewhat dose-dependent manner, the expression of genes that map within them. They were also shown to modify the expression of genes located on their flanks and sometimes those at a great distance from their boundary. Here we demonstrate, by monitoring these effects at multiple life stages, that these controls over expression are effective throughout mouse development. Similarly, we observe that the more specific spatial expression patterns of CNV genes are maintained through life. However, we find that some brain-expressed genes mapping within CNVs appear to be under compensatory loops only at specific time points, indicating that the effect of CNVs on these genes is modulated during development. Notably, we also observe that CNV genes are significantly enriched within transcripts that show variable time courses of expression between strains. Thus, modifying the copy number of a gene may potentially alter not only its expression level, but also the timing of its expression.

[Supplemental material is available online at http://www.genome.org. The expression array data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE16675.]

Copy number variation (CNV) of genomic segments among phenotypically normal individuals was recently shown to be surprisingly frequent in human (Iafrate et al. 2004; Sebat et al. 2004) and model organisms such as mouse (Cutler et al. 2007; Graubert et al. 2007; She et al. 2008; Cahan et al. 2009; Henrichsen et al. 2009b), rat (Guryev et al. 2008), or *Drosophila* (Dopman and Hartl 2007; Emerson et al. 2008). For example, more than 58,000 human CNVs mapping to 14,500 regions (CNVRs) (http://projects.tcag.ca/variation/) and encompassing hundreds of genes (Redon et al. 2006) have so far been identified. They significantly contribute to genetic variation, covering more nucleotide content per genome than single nucleotide polymorphisms (e.g., ~0.8% of the length of the human genome differs between two human individuals) (Conrad et al. 2010). A certain synteny has been observed among related species, with 20%–25% of chimpanzee and macaque CNVs overlapping those found in human (Perry et al. 2006; Lee et al. 2008). The existence of hotspots for copy number variation was suggested, as CNVs identified in multiple macaques were frequently observed in multiple human samples (Lee et al. 2008). Functional categories that are overrepresented among genes mapping within these regions include immune response, sensory perception, neurotransmission, and metabolism (Redon et al. 2006; Henrichsen et al. 2009b; Orozco et al. 2009).

A growing number of associations between these structural changes and susceptibility to disease have been uncovered (for review, see Ionita-Laza et al. 2009; Zhang et al. 2009; Carvalho et al. 2010; Fanciulli et al. 2010; Lee and Scherer 2010). CNVs impact tissue transcriptomes on a global scale by modifying the expression of genes that localize within the CNV and on its flanks, an effect that can extend over hundreds of kilobases from the breakpoints (Merla et al. 2006; Stranger et al. 2007; Cahan et al. 2009; Henrichsen et al. 2009b; Orozco et al. 2009). Multiple genes show a correlation (sometimes a negative one) between copy number and expression (Stranger et al. 2007; Cahan et al. 2009; Henrichsen et al. 2009a,b; Orozco et al. 2009). Genes within CNVs were shown to have more specific spatial expression than other genes and to be under tissue-specific differential constraints (Dopman and Hartl 2007; Henrichsen et al. 2009b). These first genome-wide analyses provided initial evidence into the effects of CNVs on gene expression (for review, see Reymond et al. 2007; Henrichsen et al. 2009a), but they did not gauge their functional impact during development. Here, we present a comprehensive analysis of their influence throughout the life of an organism.

Results

Expression patterns of CNV genes at embryonic stage E14.5

We have previously cataloged mouse CNVs (Henrichsen et al. 2009b). In brief, using a hidden Markov model–based approach that incorporates cross-sample information, we predicted around 7000 autosomal CNVs in the sampled animals from 13 inbred strains and wild-caught *Mus musculus domesticus*. They ranged from 43 to 3345 kb in size (median 61 kb) and could be grouped in ~3800 CNV regions, three-quarters of which (77%) could be validated by rehybridization on a higher resolution array (Henrichsen et al. 2009b). This validated set of CNVs was used for all subsequent expression analyses.

Recent analyses of mouse microarray data suggested that CNV genes, defined as genes with half or more of their transcription unit overlapping a CNV, were expressed at lower levels and more

[3]**Corresponding author.**
E-mail Alexandre.Reymond@unil.ch; fax 41-21-692-3965.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.112748.110.

ISSN 1088-9051/11; www.genome.org

specifically than genes that do not vary in copy number (Henrichsen et al. 2009b). We exploited the recently released high-resolution transcriptome atlas of expression in the mouse, a collection of in situ hybridizations (ISH) of 18,000 genes at embryonic stage E14.5 (14.5 d post coitum) (http://www.eurexpress.org/ee/), to assess the expression patterns of genes that map to validated CNVs (CNV genes) during fetal life. We divided the genes that displayed a regional expression pattern at this developmental stage between CNV genes and non-CNV genes, and counted the number of anatomical structures in which they were expressed. We found that CNV genes were detected in a smaller number of anatomical structures on average (3.0, median: 2) relative to genes mapping elsewhere (3.7, median: 3), a statistically significant difference (two-tailed Mann-Whitney *U* test, $P = 0.04$; sample sizes: $n = 81$ [CNV genes] and $n = 3913$ [non-CNV genes]). CNV transcripts were never found in more than nine distinct anatomical structures, whereas non-CNV transcripts were detected in more regions of the embryo (Fig. 1). As CNV genes are expressed at significantly lower levels than non-CNV genes (Henrichsen et al. 2009b), we repeated the above comparison using non-CNV genes that were expressed at comparable levels. Again, CNV genes were detected in significantly less anatomical structures than non-CNV genes (two-tailed Mann-Whitney *U* test, $P = 0.036$; sample sizes, $n = 81$ [CNV genes] and $n = 3538$ [non-CNV genes]), showing that the observed differences were not merely due to differences in detection capacities. Furthermore, the results are consistent whether or not we use cellular resolution expression data, i.e., ISH and microarrays (this study; Henrichsen et al. 2009b). Thus, CNV genes exhibit more specific expression patterns than non-CNV genes, not only in adulthood (Henrichsen et al. 2009b), but also at embryonic stage E14.5.

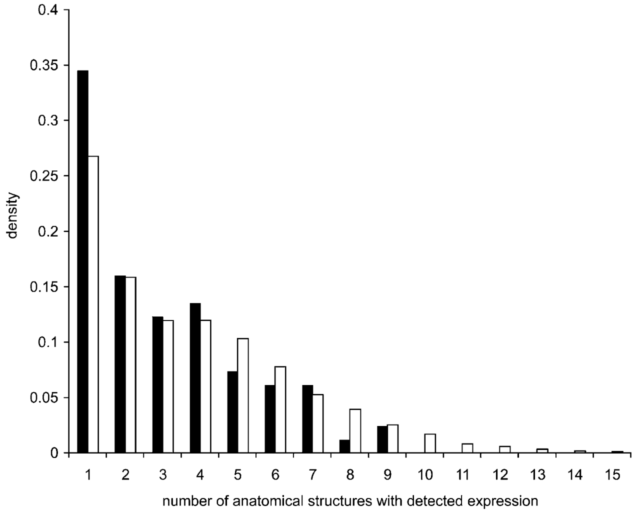

Figure 1. Spatial expression patterns of CNV genes distribution of CNV (black) and non-CNV (white) expressed transcripts in function of the number of anatomical structures in which they are detected at embryonic stage E14.5 by in situ hybridization performed by the EURexpress consortium (http://www.eurexpress.org/ee/). CNV genes are expressed in a significantly smaller number of anatomical structures (two-tailed Mann-Whitney *U*-test, $P = 0.04$).

Influence of CNVs on the expression of genes mapping within them and on their flanks

To further assess the effect of structural variants during development, we analyzed genome-wide expression levels of two major organs, brain and liver, at four different time points (E14.5 plus postnatal days P1, P7, and P90). Liver was selected, as it shows great individual expression variation as well as a dosage-compensation mechanism comparable to that of most tissues that we have tested in the past. On the other hand, brain was chosen, as brain-expressed CNV genes were shown to be more tightly regulated than other CNV genes, i.e., a smaller fraction of their expression variance could be attributed to changes in gene dosage, suggesting that a stronger dosage-compensation mechanism controlling the expression of CNV genes is at play in this tissue (Henrichsen et al. 2009b). At each stage, the transcriptome of three males from three representative mouse strains (C57BL/6J, DBA/2J, and 129S2) were profiled using Affymetrix expression arrays (see Methods). We found that the expression variance of genes mapping within our validated set of CNVs that vary in number between the three assessed strains was significantly larger than that of genes elsewhere in the genome at all developmental stages and in both analyzed tissues (largest two-tailed Mann-Whitney *U*-test, $P = 4 \times 10^{-4}$ [brain] and $P = 1.3 \times 10^{-7}$ [liver]) (Table 1; Fig. 2A). Consistently, transcripts that map within CNVs were significantly overrepresented among the most differentially expressed transcripts at each studied time point and tissue (two-tailed Mann-Whitney *U*-test, $P < 0.001$) (Fig. 2B; Methods). This effect was not merely due to an overrepresentation of genes belonging to large families that are potentially more prone to cross-hybridization, such as olfactory receptors, as the statistically significant increase in expression variation of CNV genes remains even after removing those transcripts that hybridize to the 7% of probesets interrogating these families, (largest two-tailed Mann-Whitney *U*-test, $P = 3 \times 10^{-3}$ [brain] and $P = 1.4 \times 10^{-4}$ [liver]). CNVs, therefore, provide a significant contribution to the gene-expression differences observed between tissues of developing mice.

Previous studies have shown that structural variants not only alter the expression of genes within their boundaries, but also that of genes located on their flanks (Merla et al. 2006; Stranger et al. 2007; Guryev et al. 2008; Molina et al. 2008; Cahan et al. 2009; Henrichsen et al. 2009b; for reviews, see Reymond et al. 2007; Henrichsen et al. 2009a). To determine whether this influence was effective throughout development, we evaluated the expression variation of transcripts in the vicinity of CNVs within the time-course data (see Methods). These analyses showed that transcripts mapping 50–250 kb and 250–450 kb from the CNV breakpoints (but not further) showed a significantly higher expression variance in all tested tissues and time points than more distant transcripts (e.g., 50–250 kb: largest two-tailed Mann-Whitney *U*-test, $P = 0.04$ [brain] and $P = 0.02$ [liver]; Table 1; Fig. 2A for all *P*-values; even when considering multiple testing corrections, the majority of tests remain significant). Transcripts that mapped within 50 kb of the CNV boundary were conservatively excluded to avoid any possible erroneous inclusion of transcripts that do, in fact, vary in copy number. Thus, our time-course data shows that CNVs significantly affect tissue transcriptomes throughout development by altering gene dosage and exerting long-distance effects on neighboring genomic regions.

The effect of copy number variation on brain expression varies during development

To estimate the proportion of expression variation that is explained by copy number changes alone, we dissected the expression variance of CNV and non-CNV genes within (intra) and between (inter) strains in more detail with standard analysis of variance (ANOVA;

Table 1. Gene expression variance in CNV and CNV flanks versus other (non-CNV) genes in the genome

	Brain				Liver			
Distance from nearest breakpoint (kb)	E14.5 *P*	P1 *P*	P7 *P*	P90 *P*	E14.5 *P*	P1 *P*	P7 *P*	P90 *P*
Inside CNV	0.0004	1.7×10^{-9}	7.9×10^{-4}	4.5×10^{-9}	1.3×10^{-7}	6.8×10^{-18}	6.2×10^{-13}	3.2×10^{-13}
Window 50–250	0.0408	1.5×10^{-5}	0.0006	0.0224	0.0079	0.0221	0.0003	2.3×10^{-7}
Window 250–450	0.0536	1.3×10^{-8}	0.0002	0.0003	0.0093	0.0076	0.0005	2.7×10^{-5}

see Methods). The significant increase in gene expression variance of CNV transcripts between strains ($P < 0.001$, Supplemental Fig. S1) may be due to either genetic background (strain) or copy number changes. We assessed the contribution of each of these factors and used these values to calculate the proportion of expression variance of CNV genes that was solely due to changes in copy number, as previously described (Henrichsen et al. 2009b; see Methods and Supplemental Table S1 for sum of sum of squares values and formula used). We found that in liver a similar substantial proportion (67%–77%) of the interstrain expression variance of CNV genes through development could be attributed to copy number changes (Supplemental Table S1). Interestingly, in brain, the percentage of the CNV gene expression variance between strains that could be attributed to changes in gene dosage varied during development (Supplemental Table S1). This value was low (35%) prenatally when neuronal proliferation is at its highest (E14.5), reached about two-thirds in the first postnatal week when dynamic outgrowth, neuronal differentiation, and synaptogenesis take place (P1 and P7; 58% and 59%, respectively), before decreasing again in adulthood (P90) when the neuronal circuit is mature (11%). These observations suggest that the extent of the influence of CNV genes on the brain transcriptome changes during development. It is important to specify here that the varying influence of CNVs on gene expression in the brain throughout development, and the constant influence observed in liver, can singularly and confidently be attributed to changes in copy number, as the "strain effect" produced similar variation in both tissues across developmental stages (Supplemental Table S1). Similarly, these differences cannot be explained by different sets of genes being expressed at each assessed time point, as the vast majority (71%–81%) of genes expressed at one stage are also expressed at the other three. These percentages are further increased if we consider the pairwise comparison E14.5/P1 and P7/P90 (81% and 89%, respectively).

This differential constraint on brain CNV genes throughout development is possibly actuated by variable restrictions on the individual anatomical substructures of that tissue. To challenge this hypothesis, we took advantage of the extensive GSE4734 expression data of five different brain regions (bed nucleus of the stria terminalis, hippocampus, hypothalamus, periaqueductal gray, and pituitary gland) (Hovatta et al. 2007) from 7-wk-old males of six inbred mouse lines and our CNV data for these strains (see Methods). Again, we found an increased expression variance of transcripts mapping within CNVs and up to 250 kb from the nearest boundary (Supplemental Fig. S1), independently confirming the conclusions obtained from our own and the BXD expression data (see Henrichsen et al. 2009b and above). We estimated the proportion of expression variation explained solely by copy number changes for each of these five brain regions (see above and Methods) and found that a variable proportion of the interstrain expression variance of CNV genes (29%–57%) could be attributed to copy number changes (Supplemental Table S2). These proportions are, however, always appreciably lower than those registered in other tissues (Henrichsen et al. 2009b; see above), confirming that CNV transcripts expressed in the brain are more tightly regulated than other genes. About half of the interstrain expression variance of CNV genes could be attributed to copy number changes in the stria terminalis (57%), hippocampus (56%), and periaqueductal gray (53%). Interestingly, we observed a tighter transcriptional regulation for CNV transcripts expressed in the hypothalamus (29%), which links the nervous system to the endocrine system via the pituitary gland, which is itself tightly regulated (46%), suggesting that these latter structures may be under stronger regulatory control. Taken together, these expression studies of brain regions indicate that the apparent modification of the constraints imposed on whole-brain CNV transcripts throughout development cannot be attributed (or can only be partially attributed) to the composite nature of the brain.

CNV transcripts show interstrain differences in temporal expression patterns

The development of multicellular organisms is exquisitely regulated by transcriptional networks, which act to specify cell types and provide positional information. Thus, cell and tissue identity are defined not only by which genes are expressed, but also by their level and timing of expression (temporal expression). We used an unsupervised approach to identify dominant relative expression patterns through developmental time in brain and liver. Mfuzz, a noise-robust soft clustering algorithm (Futschik and Carlisle 2005), clustered expressed genes by their temporal patterns of expression and identified three clusters with strong support in each of the tissues (Supplemental Fig. S2). We then investigated whether the time courses registered for a given transcript in the three different mouse inbred strains (C57BL/6J, DBA/2J, and 129S2) were all included in a single or in multiple clusters, hence assessing which transcripts showed significant changes in their expression time course between strains. We observed that, in liver, CNV transcripts were significantly depleted amongst transcripts that showed the same expression profile over time between strains (Fisher's exact test, $P = 0.02$; odds ratio = 0.69). Thus, liver-expressed CNV genes showed significantly more interstrain differences in their expression time courses than genes that did not vary in copy number (for individual examples, see Fig. 3). Some of the CNV transcripts showed no correlation between expression and gene dosage (for examples, see Fig. 3A,E), while a larger proportion displayed a modified time course of expression between strains. In some cases these perturbations appeared to be directly correlated with copy number (see examples in Fig. 3B–E), whereas others were not (see examples in Fig. 3F–I). We assessed whether the CNV genes that showed different or similar time courses of expression in different strains were enriched or depleted for particular Gene Ontology (GO) categories, but found no significant correlations (see Methods). The depletion of CNV genes within transcripts having the same profile between strains is not found in brain ($P = 0.24$), consistent with our findings above as well as previous studies, showing that brain CNV transcripts are under tighter regulation (Henrichsen et al. 2009b).

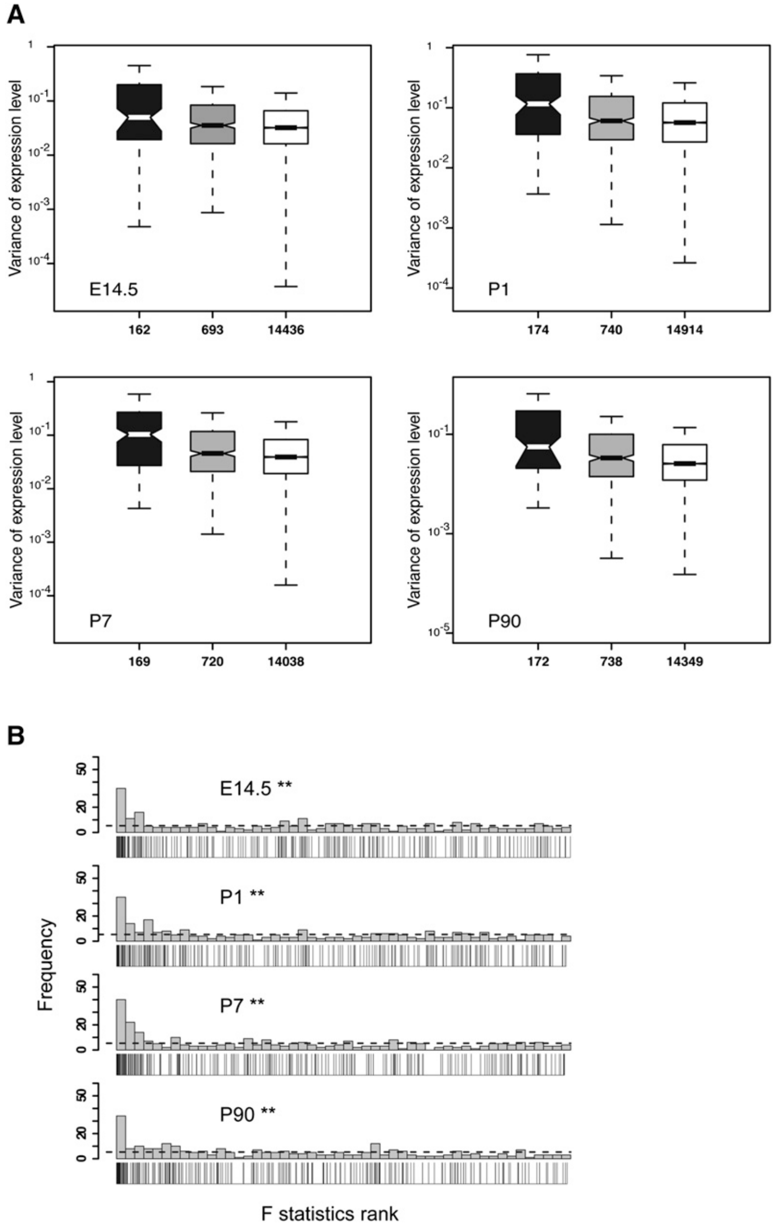

Figure 2. (*A*) Expression throughout development of genes within CNVs, in neighboring regions, and elsewhere in the genome. Boxplot distribution of signal variances (nine individuals, three strains) of transcripts expressed in the liver at E14.5, P1, P7, and P90, and mapping within CNVs (black), 50–250 kb from the nearest CNV breakpoint (gray), or further away (white). The black (largest two-tailed Mann-Whitney *U* test, $P = 1.3 \times 10^{-7}$) and the gray distributions (largest two-tailed Mann-Whitney *U* test, $P = 0.02$) are significantly different from the white in all monitored tissues. The numbers of transcripts for which expression could be detected are indicated. Similar results were obtained for brain transcripts (Supplemental Table S1). (*B*) Overrepresentation of CNV genes among differentially expressed genes. For each gene, we calculated the F statistics representing the differential expression of transcripts in each tissue and developmental time point using the Bioconductor *limma* package. We then ranked genes by their F-statistic and binned the ranked genes into 50 bins. For each tissue and time point, we display the number of CNV genes in each bin, ordering bins from the highest F-statistic on the *left* to the lowest F-statistic on the *right*; the number of CNV genes is given by the height of the bar and CNV genes in each are indicated by tick marks *below* the histogram. Data from the liver at each time point are shown here as examples; similar results were obtained for brain transcripts (see text for details). Under the assumption that genes are equally likely to be CNV genes independent of their expression differences, the number of CNV genes should be uniformly distributed among bins, as indicated by the dashed line. We tested this assumption using a two-tailed Mann-Whitney *U* test. The assumption of uniform distribution was rejected for both brain and liver at all developmental stages assessed (**$P < 0.001$), indicating an overrepresentation of CNV genes among differentially expressed genes throughout development.

To further assess the influence of gene dosage modification on temporal expression patterns, we monitored the expression of 61 CNV and 41 randomly selected non-CNV transcripts at 10 different developmental time points (E12.5, E14.5, E16.5, E17.5, E18.5, P1, P7, P14, P30, and P90) in all three inbred strains (C57BL/6J, DBA/2J, and 129S2) using real-time quantitative PCR. The list of genes and assays used are presented in Supplemental Table S3. We identified transcripts that showed highly similar expression profiles through development and others that were divergent in the three mouse strains (some examples are shown in Fig. 4A–D). To gauge whether CNV transcripts were more prone to dissimilar expression patterns between strains, we computed the sum of squared deviations from the mean between strains for each developmental time point and used the sum of these values to rank the assessed transcripts (Fig. 4E; Supplemental Fig. S3). We observed, in both brain and liver, a statistically significant enrichment of CNV transcripts among the transcripts with the highest score, i.e., the transcripts that vary more between strains (Wilcoxon signed-rank test, $P = 6 \times 10^{-6}$ [liver] and $P = 10^{-3}$ [brain]). Similar results were obtained if, instead, we used the median of the sum of squared deviations from the mean between strains for each developmental time point to rank the transcripts (Wilcoxon signed-rank test, $P = 3 \times 10^{-5}$ [liver] and $P = 5 \times 10^{-4}$ [brain]).

Our results suggest that CNV genes are more likely to show different temporal expression patterns between strains than genes that do not vary in copy number, showing that CNVs shape tissue transcriptomes globally in both space and time.

Discussion

To obtain a global view of the impact of CNVs on gene expression patterns throughout development, we characterized the transcriptome of two major organs at different stages. We found that CNVs shape tissue transcriptomes throughout development by modifying gene dosage, but also by profoundly affecting the expression of genes located in their vicinity, thus

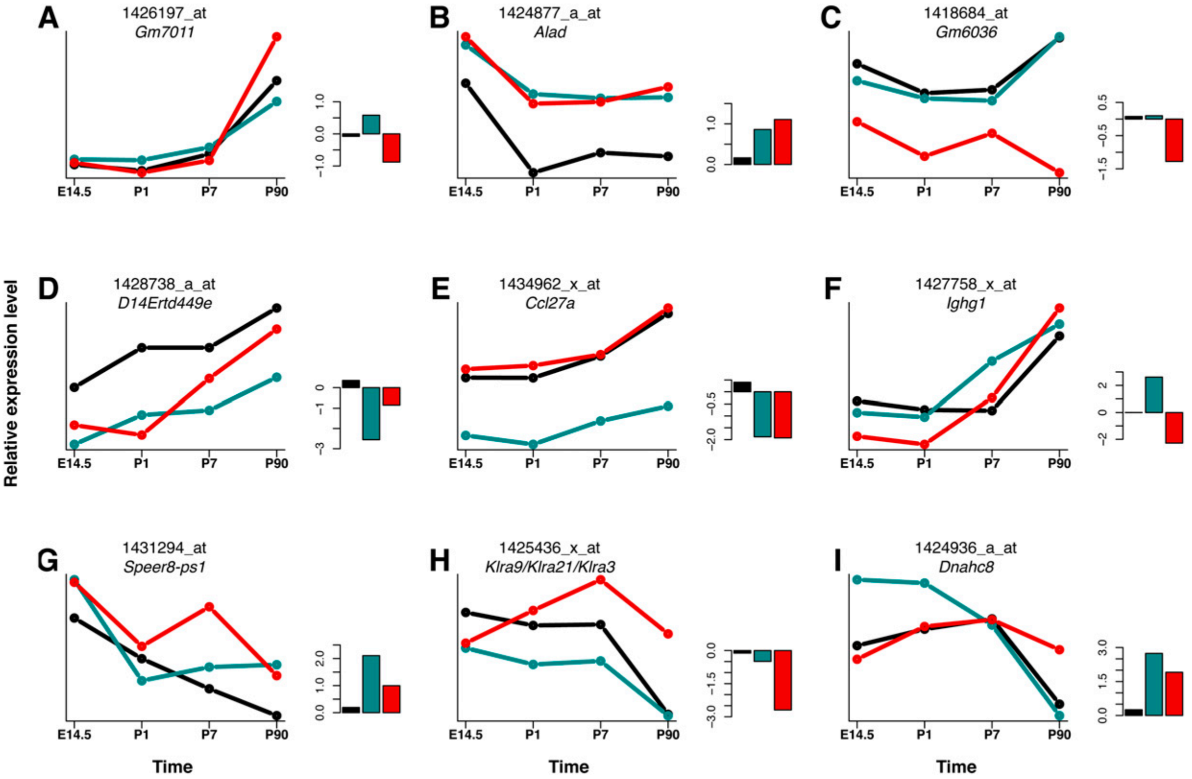

Figure 3. Examples of time course expression profiles in liver. Relative expression levels of CNV transcripts during development in the three inbred strains: C57BL/6J (black line), DBA/2J (red), and 129S2 (teal). Examples of CNV transcripts under regulatory feedback loops that buffer gene-dosage alterations (transcript monitored by probeset 1426197_at; *A*), showing a positive correlation between relative copy number and expression levels (1424877_a_at, 1418684_at, and 1428738_a_at; *B–D*, respectively), showing a positive correlation between relative copy number and expression level in one strain and buffering of gene dosage in another (1434962_x_at; *E*), or with modified time courses of expression (1427758_x_at, 1431294_at, 1425436_x_at, and 1424936_a_at; *F–I*, respectively). Bar graphs on the *right* show the $\log_2$ ratios for the CNV encompassing the transcript considered in C57BL/6J (black), DBA/2J (red), and 129S2 strains (teal) relative to the C57BL/6J reference as determined by array CGH in Henrichsen et al. (2009b). We note that the noise-robust soft clustering algorithm Mfuzz (Futschik and Carlisle 2005) included in the same cluster the time courses registered for the CNV transcript 1426197_at in *A*) in the three different mouse inbred strains, while all other CNV transcripts shown here were incorporated in multiple clusters, thus illustrating the propensity of CNV genes to change their timing of expression between strains (see text for details).

complementing previous data obtained in adult tissues or cell lines (Merla et al. 2006; Guryev et al. 2008; Cahan et al. 2009; Henrichsen et al. 2009b). Interestingly, recent results have shown that structural variants can have an effect on normal copy number genes positioned along the entire length of a chromosome (Harewood et al. 2010; Ricard et al. 2010). Our analyses also corroborate the finding that genes within CNVs have particular properties with respect to their spatial expression patterns and dosage sensitivity (Henrichsen et al. 2009b). For instance, non-CNV genes are expressed in a greater number of anatomical structures than CNV genes during embryonic stages. Similarly, we confirmed that CNV genes expressed in the brain are more tightly regulated than other CNV genes, as previously shown (Henrichsen et al. 2009b). Remarkably, these regulatory mechanisms appear to be alleviated at specific developmental stages, which raises some intriguing questions: Is the tight regulation of these brain-expressed CNV genes deleterious during some phases of life? Are some of the regulatory feedback loop proteins lacking during certain phases of development? Are they down-regulated at these time points? Interestingly, the strict regulation imposed upon the expression levels of CNV genes in the brain is reduced when the central nervous system cells are outgrowing, differentiating, and creating synapses, a "critical" period during which neurons and synapses are competing for growth factors and are subject to pruning. As processes that are inhibited or unused during this early brain development phase will not develop later, we can hypothesize that the release of control over the expression of CNV-genes might facilitate this process either by favoring some outgrowth or alternatively by facilitating the formation of neuronal junctions.

We also found that CNV transcripts are enriched among those transcripts that show varying time courses of expression between strains, suggesting that CNV genes are more likely to show different temporal patterns of expression in different individuals. Hence, segmental copy number variation shapes tissue transcriptomes, not only by altering the dosage of genes that map within the CNV and affecting the expression of neighboring genes, but also by modifying the timing of expression of the former class of transcripts. Further studies are warranted to pinpoint whether the same effects are brought about by smaller CNVs, which represent the majority of structural changes (Zhang et al. 2009).

Methods

Specificity of expression pattern

We counted the number of anatomical structures in which genes were expressed using the EUrexpress collection of ISH of sagittal sections of E14.5 mouse embryos (http://www.eurexpress.org/ee/)

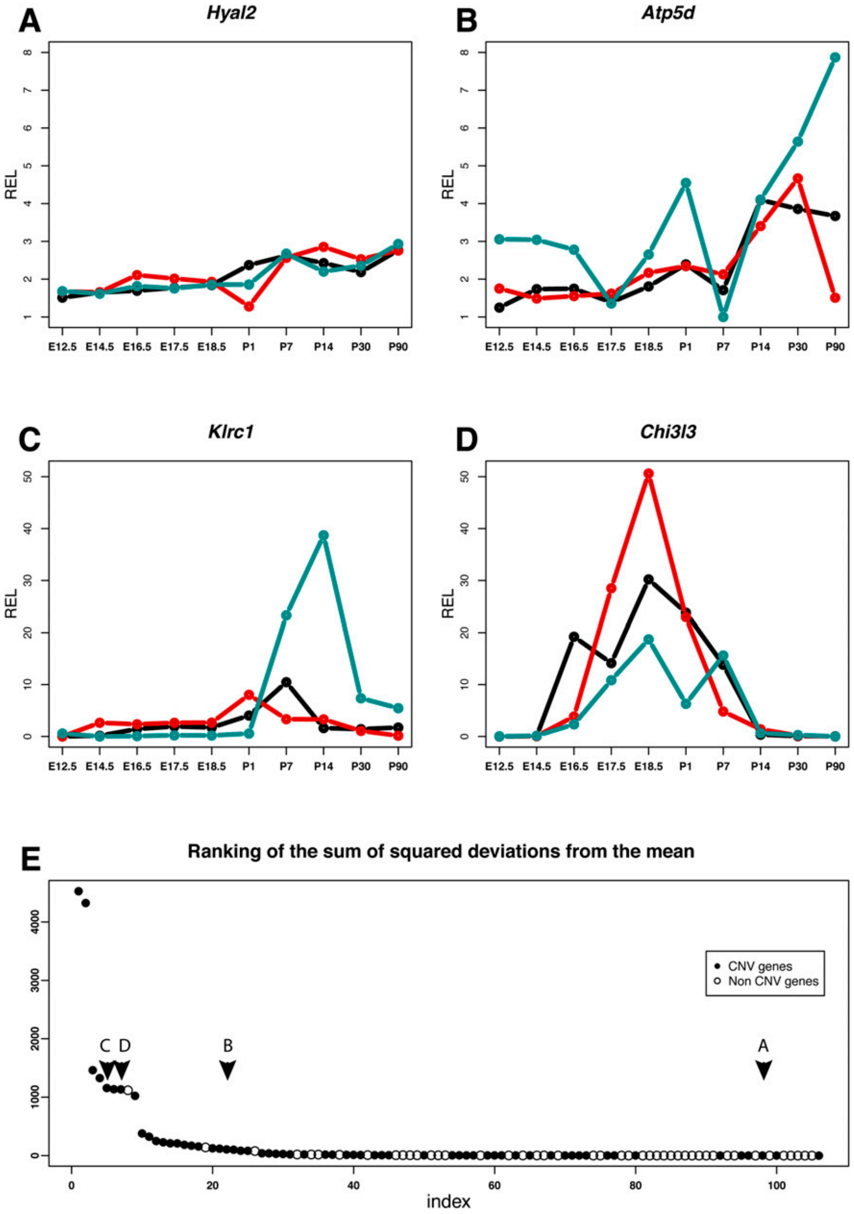

Figure 4. Ten-stage expression profiles of liver CNV transcripts. Real-time quantitative PCR-measured relative expression levels of CNV transcripts during development (E12.5, E14.5, E 16.5, E17.5, E18.5, P1, P7, P14, P30, P90 time points) in the three inbred strains: C57BL/6J (black line), DBA/2J (red), and 129S2 (teal). Examples of liver CNV transcripts showing a similar (*A*) or a divergent expression profile between strains are shown (*B–D*). (REL) Relative expression level. (*E*) Assessed transcripts were ranked decreasingly by the sum of squared deviations from the mean between strains for each developmental time point. (Filled circles) CNV transcripts; (open circles) non-CNV transcripts. Data from liver are shown, while data from brain are presented in Supplemental Figure S4. We observe in both brain and liver a statistically significant enrichment of CNV transcripts among the transcripts with the highest score, i.e., the transcripts that vary more between strains (see text for details; Wilcoxon signed-rank test $P = 6 \times 10^{-6}$ [liver] and $P = 10^{-3}$ [brain]). The position within the ranking of the CNV-transcript profiles presented as examples in *A–D* are indicated.

and compared the results obtained for CNV and non-CNV genes. As CNV genes are expressed at significantly lower levels than non-CNV genes (Henrichsen et al. 2009b), we repeated the above comparison using non-CNV genes expressed at levels comparable to those of CNV genes. We identified these non-CNV genes by computing the mean normalized expression level (NEL) for each gene across the 108 adult samples described in Henrichsen et al. (2009b) (three replicates, six strains, and six adult tissues; accession no. GSE10744) and filtering out non-CNV transcripts that lay outside the range of the mean NEL of CNV transcripts.

Animals and sexing

Inbred mice of the strains C57BL/6J and DBA/2J were obtained directly from The Jackson Laboratory (Bar Harbor, MN), whereas strain 129S2 animals were purchased from Charles River Laboratories (Wilmington, MA). Tissue samples used for the experiments described in this report are from purchased animals (P90 stage) or from the F_1 generation (E12.5, E14.5, E16.5, E17.5, E18.5, P1, P7, P14, P30 stages). E12.5, E14.5, E16.5, E17.5, E18.5, P1, P7, and P14 individuals were sexed molecularly and morphologically upon dissection. Briefly, genomic DNA was extracted from tissues not used for expression profiling (see below) using Maxwell cartridges following the manufacturer's instructions (Promega). Presence of the Y chromosome was assessed by multiplexed PCR with pairs of primers specific for the MMUY *Sry* (5′-TCATGAG ACTGCCAACCACAG-3′ and 5′-CATGAC CACCACCACCACCAA-3′) and the MMU1 *Myog* genes (5′-TTACGTCCATCGTGGACA GC-3′ and 5′-TGGGCTGGGTGTTAGTCT TA-3′) as described (McClive and Sinclair 2001).

Gene expression profiling

Whole brain and liver from E14.5, P1, and P7 males were dissected and immediately frozen on dry ice. Total RNA was extracted using TRIzol reagent (Invitrogen), cleaned on RNeasy columns (Qiagen) according to the manufacturers' protocols, and used as a template for complementary DNA (cDNA) synthesis and biotinylated antisense cRNA preparation. The synthesis of cDNA and cRNA, labeling, hybridization, and scanning of the samples were performed as described by Affymetrix (http://www.affymetrix.com). GeneChip Mouse Genome 430 2.0 arrays, each interrogating 45,101 target sequences with appropriate probesets (Affymetrix), were used to hybridize the labeled cRNA. For each developmental time point, three individuals each of the three inbred strains C57BL/6J, 129S2, and DBA/2J were processed for a total of 54 expression arrays. Expression data analysis was performed in R using the Affymetrix Bioconductor package for low-level analysis and MAS5 and RMA normalizations. RMA normalization was performed separately for each tissue and time point. The 18 expression arrays for the P90 time point were generated previously (Henrichsen et al. 2009b) and renormalized using only the three inbred strains investigated here. The RMA normalized expression data were filtered by the detection *P*-values computed by MAS5, using a 0.01 cut-off. A target sequence (transcript) was considered expressed if the signal of its corresponding probeset passed the detection threshold in at least two mice of at least one strain at one time point. Of the 45,101 interrogated target sequences, we identified 14,927–15,828 and 18,372–21,806 expressed transcripts in liver and brain, respectively. For both tissues, these expression data allowed clustering of the samples by strain. CNVs were derived from previous data (these can be recovered from Supplemental Table S4 of Henrichsen et al. 2009b) by selecting CNVs confirmed in DBA/2J and 129S2. A transcript was defined as mapping to a CNV if its target sequence overlapped by at least 50% with the CNV region coordinates. For flanking genes, we note that we conservatively excluded signals from probe sets that mapped to the first 50 kb that flank validated CNV boundaries to avoid possible erroneous inclusion of transcripts of genes that vary in copy number in our analyses of the influence on expression on neighboring genes.

For each tissue and developmental time point and for each transcript an analysis of variance (standard ANOVA) was performed to calculate the between strain and the within strain variance. The analysis of variance was computed in the statistical software R using the *anova()* function on the linear model $Y_i = \mu + S_k + e_i$ fitted with the *lm()* function, where Y_i is the $\log_2$ expression for probeset *i* in strain S_k. In order to compare the between and within strain variances of CNV transcripts versus non-CNV transcripts, we used both a Student *t*-test and a random sampling approach, in which a number of transcripts equivalent to the number of CNV transcripts were randomly chosen 1000 times in order to calculate a null distribution. A corresponding *P*-value was then calculated for the within-strain variance of CNV transcripts by comparing its value with its null distribution.

The significant increase in gene expression variance of CNV transcripts between strains might be due to the genetic background (strain) or to copy number changes. To assess the contribution of each of these factors, we estimated the strain effect based on differences in the expression variance of non-CNV genes within and between strains for all tissues and used this value to calculate the proportion of expression variance of CNV genes due to changes in copy number (see Supplemental Tables S1 and S2 for sum of sum of squares values and formula used). To further identify transcripts that vary between strains, we used the Bioconductor package *limma* instead of standard ANOVA (Smyth 2004). Briefly, *limma* calculates a moderated *F*-statistic using an empirical Bayesian method that has been specifically designed for microarray data. To test whether CNV transcripts tend to be differentially expressed between strains, we used the *geneSetTest* function of *limma.*

The transcriptome profiles of five different brain regions (bed nucleus of the stria terminalis, hippocampus, hypothalamus, periaqueductal, and pituitary gland) of six mouse strains (129S6/SvEvTac, A/J, C57BL/6J, C3H/HeJ, DBA/2J, and FVB/NJ) were extracted from the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; accession no. GSE4734) (Hovatta et al. 2007). We considered the set of validated CNVs between C57BL6/J and strains 129S2 (as a proxy for 129S6/SvEvTac), DBA/2J, A/J, and/or C3HeB/FeJ (as a proxy for C3H/HeJ) to determine whether a transcript was mapping within, close to, or far away from a structural rearrangement (these can be recovered from Supplemental Table S4 of Henrichsen et al. 2009b). We note that the results were not significantly modified, whether or not we included the expression data from FVB/NJ individuals, and if we only considered validated CNVs between strains C57BL/6J, DBA/2J, and A/J.

Clustering

We selected all of the expressed transcripts and clustered a total of ~26,000 time courses for liver and brain by their temporal expression pattern using the noise-robust soft clustering algorithm Mfuzz (Futschik and Carlisle 2005). This R package includes a standardization step applied to the whole data set before the clustering itself. We optimized the "*m*" fuzzification parameter and "*c*" number of clusters, as suggested by the authors, and found that the optimized values were "*c* = 3" for both studied tissues with *m* values of 4.15 (liver) and 4.05 (brain).

The transcripts were then separated in two classes: first, the ones that showed the same time course (i.e., belonged strictly to the same cluster with membership values of ≥0.5) for each of the three analyzed strains; and second, the remaining ones. We tested whether any of these classes were enriched or depleted for CNV transcripts. Similarly, we assessed enrichment within the CNV transcripts of each of the classes of functional GO categories (biological process, molecular function, cellular component) and KEGG pathways using Babelomics (http://babelomics.bioinfo.cipf.es/; Al-Shahrour et al. 2006, 2008). We found no significant correlations, possibly due to a lack of statistical power, i.e., relatively low number of GO annotated genes expressed in the considered tissues that vary in number between the strains assessed.

Real-time quantitative PCR

Quantitative polymerase chain reaction was performed using SYBR_GREEN PCR Master Mix (Roche) following the manufacturer's specifications. Briefly, whole brain and liver total RNA was converted to cDNA using Superscript III (Invitrogen) primed with a mix of oligo(dT) and random hexamers. Primers were designed using the Getprime program (http://updepla1srv1.epfl.ch/getprime/) with default parameters. The list of assessed genes and the assays used are presented in Supplemental Table S3. The efficiency of each assay was tested in a cDNA dilution series as described (Livak and Schmittgen 2001). All RT-PCR reactions were performed in a 10-µL final volume, and three replicates per sample set up in a 384-well plate format using a Freedom EVO robot (TECAN) and run in an ABI 7900 Sequence Detection System (Applied Biosystems) with the following amplification conditions: 50°C for 2 min, 95°C for 10 min, and 40 cycles of 95°C 3 sec/60°C 40 sec.

Each plate included assays for the appropriate normalization genes to control for any variability between the different plate runs. Raw threshold cycles (Ct) values were obtained using SDS2.4 (Applied Biosystems). To calculate the normalized relative expression ratio, we followed the method described in Merla et al. (2006) and Molina et al. (2008), and exploited the geNorm method (Vandesompele et al. 2002) and the qBase pipeline (Hellemans et al. 2007) to select *Actb*, *Eef1a1*, and *Rpl13* and *Actb, Eef1a1, Hprt*, and *Tbp* as brain and liver normalization genes, respectively. At each developmental time point, the relative expression was measured in three males of each of the three studied strains.

Selected transcripts met the following criteria: (1) they were expressed both in brain and liver according to the microarray analysis; and (2) non-CNV transcripts mapped at least 500 kb away from the nearest CNV breakpoint.

Acknowledgments

We thank Sophie Wicker, Otto Hagenbüchle, and Keith Harshman from the Lausanne DNA Array Facility for technical help; Graciana Diez-Roux, Sandro Banfi, Andrea Ballabio, and the members of the EURexpress consortium for sharing unpublished results; and Christelle Borel, Louise Harewood, Henrik Kaessmann, and Nicolas Vinckenbosch for comments. This work was supported by grants from the Telethon Action Suisse, the Jérôme Lejeune Foundation, the Swiss National Science Foundation, and the European Commission anEUploidy Integrated Project (grant 037627) to A.R.

Author contributions: A.R. designed the study and obtained the necessary financial support; E.C., C.N.H., and J.C. prepared the necessary materials and produced the data; E.C., E.A.Y.G., C.N.H., F.S., and S.P. conducted the statistical analyses guided by A.R. A.R. wrote the manuscript with contributions from E.C., E.A.Y.G., and C.N.H.

References

Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J, Dopazo J. 2006. BABELOMICS: A systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res* **34:** W472–W476.

Al-Shahrour F, Carbonell J, Minguez P, Goetz S, Conesa A, Tarraga J, Medina I, Alloza E, Montaner D, Dopazo J. 2008. Babelomics: Advanced functional profiling of transcriptomics, proteomics and genomics experiments. *Nucleic Acids Res* **36:** W341–W346.

Cahan P, Li Y, Izumi M, Graubert TA. 2009. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet* **41:** 430–437.

Carvalho CM, Zhang F, Lupski JR. 2010. Genomic disorders: A window into human gene and genome evolution. *Proc Natl Acad Sci* **107** (**Suppl 1**): 1765–1771.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464:** 704–712.

Cutler G, Marshall LA, Chin N, Baribault H, Kassner PD. 2007. Significant gene content variation characterizes the genomes of inbred mouse strains. *Genome Res* **17:** 1743–1754.

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster. Proc Natl Acad Sci* **104:** 19920–19925.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster. Science* **320:** 1629–1631.

Fanciulli M, Petretto E, Aitman TJ. 2010. Gene copy number variation and common human disease. *Clin Genet* **77:** 201–213.

Futschik ME, Carlisle B. 2005. Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol* **3:** 965–988.

Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, et al. 2007. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* **3:** e3. doi: 10.1371/journal.pgen.0030003.

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, et al. 2008. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* **40:** 538–545.

Harewood L, Schutz F, Boyle S, Perry P, Delorenzi M, Bickmore WA, Reymond A. 2010. The effect of translocation-induced nuclear reorganization on gene expression. *Genome Res* **20:** 554–564.

Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J. 2007. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol* **8:** R19. doi: 10.1186/gb-2007-8-2-r19.

Henrichsen CN, Chaignat E, Reymond A. 2009a. Copy number variants, diseases and gene expression. *Hum Mol Genet* **18:** R1–R8.

Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A. 2009b. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* **41:** 424–429.

Hovatta I, Zapala MA, Broide RS, Schadt EE, Libiger O, Schork NJ, Lockhart DJ, Barlow C. 2007. DNA variation and brain region-specific expression profiles exhibit different relationships between inbred mouse strains: Implications for eQTL mapping studies. *Genome Biol* **8:** R25. doi: 10.1186/gb-2007-8-2-r25.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949–951.

Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C. 2009. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics* **93:** 22–26.

Lee C, Scherer SW. 2010. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* **12:** e8. doi: 10.1017/S1462399410001390.

Lee AS, Gutierrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, Miller GM, Korbel JO, Lee C. 2008. Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum Mol Genet* **17:** 1127–1136.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* **25:** 402–408.

McClive PJ, Sinclair AH. 2001. Rapid DNA extraction and PCR-sexing of mouse embryos. *Mol Reprod Dev* **60:** 225–226.

Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zabot MT, Antonarakis SE, Reymond A. 2006. Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am J Hum Genet* **79:** 332–341.

Molina J, Carmona-Mora P, Chrast J, Krall PM, Canales CP, Lupski JR, Reymond A, Walz K. 2008. Abnormal social behaviors and altered gene expression rates in a mouse model for Potocki-Lupski syndrome. *Hum Mol Genet* **17:** 2486–2495.

Orozco LD, Cokus SJ, Ghazalpour A, Ingram-Drake L, Wang S, van Nas A, Che N, Araujo JA, Pellegrini M, Lusis AJ. 2009. Copy number variation influences gene expression and metabolic traits in mice. *Hum Mol Genet* **18:** 4118–4129.

Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Caceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, et al. 2006. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci* **103:** 8006–8011.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.

Reymond A, Henrichsen CN, Harewood L, Merla G. 2007. Side effects of genome structural changes. *Curr Opin Genet Dev* **17:** 381–386.

Ricard G, Molina J, Chrast J, Gu W, Gheldof N, Pradervand S, Schütz F, Young JI, Lupski JR, Reymond A, et al. 2010. Phenotypic consequences of copy number variation: Insights from Smith-Magenis and Potocki-Lupski syndrome mouse models. *PLoS Biol* **8:** e1000543. doi: 10.1371/journal.pbio.1000543.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

She X, Cheng Z, Zollner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and copy number variation. *Nat Genet* **40:** 909–914.

Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat App Genet Mol Biol* **3:** doi: 10.2201/1544-6115.1027.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315:** 848–853.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3:** RESEARCH0034. doi: 10.1186/gb-2002-3-7-research0034.

Zhang F, Gu W, Hurles ME, Lupski JR. 2009. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10:** 451–481.

Received July 13, 2010; accepted in revised form November 9, 2010.