

Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles

Carl E.G. Bruder,^{1,*} Arkadiusz Piotrowski,¹ Antoinet A.C.J. Gijsbers,^{2,3} Robin Andersson,⁴ Stephen Erickson,⁵ Teresita Diaz de Ståhl,⁶ Uwe Menzel,⁶ Johanna Sandgren,⁷ Desiree von Tell,¹ Andrzej Poplawski,¹ Michael Crowley,¹ Chiquito Crasto,¹ E. Christopher Partridge,¹ Hemant Tiwari,⁵ David B. Allison,^{1,5} Jan Komorowski,⁴ Gert-Jan B. van Ommen,^{2,3} Dorret I. Boomsma,⁸ Nancy L. Pedersen,⁹ Johan T. den Dunnen,^{2,3} Karin Wirdefeldt,⁹ and Jan P. Dumanski^{1,6}

The exploration of copy-number variation (CNV), notably of somatic cells, is an understudied aspect of genome biology. Any differences in the genetic makeup between twins derived from the same zygote represent an irrefutable example of somatic mosaicism. We studied 19 pairs of monozygotic twins with either concordant or discordant phenotype by using two platforms for genome-wide CNV analyses and showed that CNVs exist within pairs in both groups. These findings have an impact on our views of genotypic and phenotypic diversity in monozygotic twins and suggest that CNV analysis in phenotypically discordant monozygotic twins may provide a powerful tool for identifying disease-predisposition loci. Our results also imply that caution should be exercised when interpreting disease causality of de novo CNVs found in patients based on analysis of a single tissue in routine disease-related DNA diagnostics.

Monozygotic (MZ) twins represent an important resource in genetic studies related to normal development and disease. Numerous twin registries exist,¹ often specializing in collection of phenotypically discordant MZ twins. Consequently, twin research has become a powerful tool for studying various diseases and endophenotypes, evaluating quantitative-trait loci, estimating heritability, studying differences in gene expression, and testing hypotheses regarding gene-environment interactions.¹ It is generally presumed that MZ twins are genetically identical and that phenotypic differences between twins are mainly due to environmental factors. Examples of genetic and, more recently, epigenetic differences between MZ twins have, however, been described.^{2–5} The former are mainly related to aneuploidies. Somatic mosaicism is usually defined by the presence of genetically distinct populations of somatic cells in a single organism. Any genetic difference between MZ twins represents an extreme example of somatic mosaicism. Earlier reports have shown somatic mosaicism for mutations in specific disease-related genes or chromosomal aberrations that are connected with a disease and can, for instance, result in a milder disease phenotype.^{6–9} This steadily growing body of data indicates that somatic mosaicism for pathogenic mutations affecting known disease genes might be seen as a rule rather than as an exception. In addition, it was recently demonstrated that the frequency of inversions is altered between different populations of normal somatic cells in a healthy sub-

ject.¹⁰ However, the frequency of in vivo somatic mosaicism for copy-number variations (CNVs) in populations of apparently normal cells is so far unexplored.

A recent and important development in human genetics is the discovery of substantial large-scale structural variation (SV) changing the chromosomal architecture (such as deletions, duplications, insertions, inversions, and more complex rearrangements) and occurring both in phenotypically normal and in diseased subjects. The most explored subtype of SV involves changes affecting copy number of DNA segments (denoted here as Copy-Number Variation, [CNV]), often involving fragments of chromosomes that are considerable in size. Although approximately three years have passed since the initial reports,^{11,12} current publications still suggest that CNV is an underestimated aspect of the human genome in health and disease.¹³ In one comprehensive recent study,¹⁴ it has been suggested that the total amount of sequence variation involving CNVs between two normal subjects is actually higher than that for single-nucleotide polymorphisms (SNPs). This conclusion has been reinforced by the recent increase in resolution of CNV discovery.¹⁵ The advent of massively parallel sequencing will soon bridge the remaining gap between efficient global analysis of SNPs and assessments of SVs, so that in the next few years we will be able to fully explore the “SV plasticity” of our genome and its relation to normal and pathogenic variation.

The rationale of this study was to test whether phenotypically unselected and concordant MZ twins, as well as

¹Department of Genetics, University of Alabama at Birmingham, Birmingham, AL 35294-0024, USA; ²Center for Human and Clinical Genetics, Leiden University Medical Center (LUMC), 2300 RC Leiden, The Netherlands; ³Center for Medical Systems Biology, 2300 RA Leiden, The Netherlands; ⁴Linnaeus Centre for Bioinformatics, Uppsala University, SE-75124 Uppsala, Sweden; ⁵Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Ryals Public Health Building, Suite 327, Birmingham, Alabama 35294-0022, USA; ⁶Department of Genetics and Pathology, Rudbeck Laboratory, ⁷Department of Surgical Sciences, Uppsala Academic Hospital, Uppsala University, SE-751 85 Uppsala, Sweden; ⁸Department of Biological Psychology, VU University, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands; ⁹Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden

*Correspondence: cbruder@genetics.uab.edu

DOI 10.1016/j.ajhg.2007.12.011. ©2008 by The American Society of Human Genetics. All rights reserved.

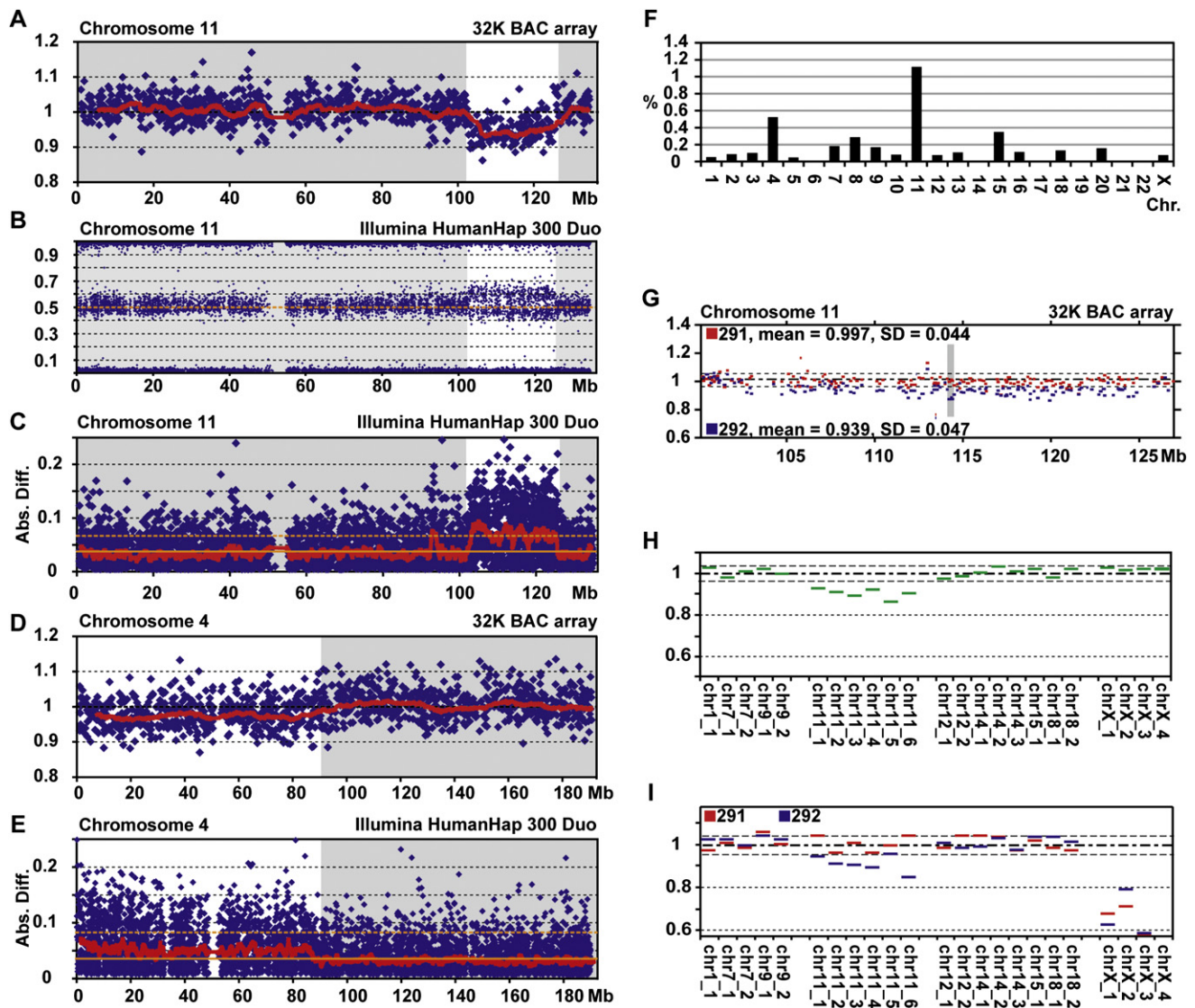


Figure 1. Copy-Number Variation (CNV) Analysis of Twin Pair 291/292

We applied 32K BAC array (A, D, F, and G), Illumina HumanHap 300 Duo beadchip (B, C, and E), and high-resolution array based on repeat-free nonredundant PCR fragments (H and I).

(A) and (D) illustrate combined array-CGH profiles of chromosomes 11 and 4, respectively, which were derived from two dye-swap experiments on 32K BAC array for twin 292 versus 291. Relative fluorescent intensity was plotted as a straight ratio. The data points shown were scored in both experiments and do not deviate more than one global standard deviation (SD) from each other. The global mean and one standard deviation for these two combined experiments are 1.003 and 0.039, respectively. Two large aberrations on chromosomes 11 (~22 Mb) and 4 (~85 Mb), present in a minority of blood cells from twin 292, are highlighted by white areas. The red line shows the moving average with a period of 30 data points.

(B) shows the values of SNP allele frequencies, obtained by the Illumina beadchip, in twin 292 on chromosome 11. The region corresponding to the 22 Mb deletion is highlighted in the white box. Heterozygous SNPs are distributed around a value of 0.5.

(C) and (E) plot the values of absolute difference (Abs. Diff.) between the heterozygous SNP allele frequencies in twin 292 versus 291, as measured by the Illumina beadchip for chromosomes 11 and 4, respectively. The solid and the dotted orange lines represent the average and the standard deviation of the absolute difference, respectively. The larger the absolute difference between the allele frequencies in two tested individuals, the larger the variance in copy number. The red line shows the moving average with a period of 30 data points. The data from (A), (C), (D), and (E) were used to compute the statistical significance of the observed aberrations (Figure 4).

(F) displays the relative fraction of BACs per chromosome that deviate more than two SDs from the mean, point in the same direction (i.e., show deletion or gain), and are located less than 1 Mb from each other. The two combined 32K array experiments used for these calculations are displayed in (A) and (D). A higher relative score was indicative of a larger region of deviation for the particular chromosome, as seen for chromosomes 11 and 4. Although the 11q deletion was approximately four times smaller than the deletion on chromosome 4, it was present in a higher number of cells and thus contains higher number of BACs deviating by more than 2 SDs.

(G) displays the summary of five 32K array experiments between each of twins 291 and 292 against healthy female control (F1), displaying 155 clones within the 100–127 Mb interval of 11q. The blue dots represent the average intensity ratios of data points scored in at least

selected MZ twins discordant for a neurodegenerative phenotype, display CNVs. We used two different genome-wide platforms—the 32K BAC array and the Illumina Human-Hap 300 Duo beadchip system—for cross-validation of the most salient findings. We present evidence for large-scale CNVs among MZ twins and suggest that these variations may be common, notably in somatic development. Our results question the long-standing notion that MZ twins are essentially genetically identical and open up new possibilities in the use of discordant MZ twins for identifying regions harboring disease- or trait-influencing loci.

Arrays constructed with genomic clones are widely used as a tool for CNV detection.^{14,16–18} The 32K BAC-array platform is a sensitive system to assess subtle imbalances and has been validated by use of a large number of samples with known genotypes. The samples used in this study were hybridized to the 32K BAC array and washed and scanned according to previously published protocols.¹⁹ All hybridizations were done in duplicate, with the dyes being swapped between the twins in the second hybridization to eliminate a possible dye-specific bias. The results were deposited in and analyzed with the Linnaeus Centre for Bioinformatics (LCB) environment for microarray-data management.²⁰ Here, nonoptimal array features were filtered out, and spatial artifacts were addressed with the use of print-tip loess normalization.¹⁹ Comparative analysis was done in Microsoft Excel. Here, we excluded all clones that were not reliably scored in both hybridizations for each twin. We then calculated the global standard deviation (SD) of the remaining clones. Clones with an internal SD larger than the global SD were excluded for further analysis. Finally, we performed a *t* test to test whether the clones from the two duplicated experiments deviated significantly from the value of one, which is the theoretical value for a normal copy-number ratio. Computational analysis of statistical significance was carried out with the R statistical computing environment.²¹ We further applied the SNP-based Illumina HumanHap 300 Duo beadchip, which can also be used for copy-number analysis.²² The Illumina hybridizations were performed at the Leiden Genome Technology Center according to the manufacturer's instructions (Illumina, San Diego, CA). Image analysis was performed in Illumina's BeadStudio software, and the raw data were exported into Microsoft Access. Here, we filtered out all data that had allele frequency values representing homozygous

SNPs. The absolute difference of the remaining heterozygous SNPs between the twins in each twin pair was then calculated and plotted. The data were finally deposited at the GEO main page at NCBI under the accession number GSE9609.

We studied a total of 19 pairs of MZ twins by using peripheral-blood-derived DNA (Table S1, available online). Nine MZ twin pairs were previously collected and assessed for Parkinson disease (PD [MIM 168600]), parkinsonism [MIM 168600], or Lewy body dementia [MIM 127750] by the Swedish Twin Registry (STR).^{23,24} Six of these nine MZ twin pairs were discordant for probable or possible PD, two pairs were discordant for parkinsonism of unknown cause, and one pair was discordant for Lewy body dementia. In two of the six MZ twin pairs discordant for PD, the co-twin had essential tremor, and in one pair, the co-twin had a mixed form of parkinsonism (Table S1). Ten phenotypically unselected and concordant normal Dutch MZ twins were collected as part of the Netherlands Twin Register (NTR).²⁵ The monozygosity of the STR twins was determined on the basis of questionnaires, which alone have a 98% probability of correct zygosity establishment,²³ and was further confirmed by genotyping with 13–18 SNPs distributed throughout the genome. On the basis of allele frequencies from 249 control subjects,²⁶ the probability of being concordant at that number of unlinked loci, assuming dizygosity, was less than 0.03 (Table S1 and data not shown). The ten NTR MZ twins were genotyped on the SNP beadchip containing more than 300,000 SNPs, and genotypes were concordant for all SNPs.

The primary platform for analysis of STR twin pairs was the 32K BAC array. We compared one twin versus its co-twin and each twin against a genetically well-characterized normal female control (F1).¹⁹ The latter comparison serves to assess the true genotype of the twins (i.e., presence of putative shared CNV), whereas the former experiments show putative imbalance within each twin pair. BAC-array profiles from experiments between any twin versus F1 are consistently noisier, with higher standard deviations, compared to experiments using DNA from one twin versus another from the same pair (not shown). Comparisons within the STR twin pairs revealed a considerable number of loci suggestive of putative CNV. For instance, the profile from twin pair 291/292 points to a deletion encompassing 22 Mb of 11q in subject 292 on the basis of profiles from seven experiments (Figure 1). This finding was confirmed

two 292-versus-F1 hybridizations. Red dots show the average value of two 291-versus-F1 experiments. The mean and SD shown were calculated with data points from the 102–124 Mb interval for the respective hybridizations. These experiments show that the CNV detected on 11q represents a deletion in twin 292.

(H) Array-CGH analysis of DNA from twin 292 versus 291 on the confirmatory PCR-based array covering the four shaded clones. The mean and SD for the six chromosome 11 data points are 0.899 and 0.022, respectively. Corresponding numbers for the nonchromosome 11 data points are 1.006 and 0.018.

(I) Summary of two array-CGH hybridizations of twin 291 against F1 (red) and twin 292 against F1 (blue) on the same array shown in (E). The average intensity ratio for autosomal PCR fragments for twin 291 is 1.000, SD = 0.031. The average intensity ratio for nonchromosome 11 autosomal fragments for twin 292 is 1.011, SD = 0.021. The average ratio for the chromosome 11 clones for 292 is 0.907, SD = 0.039.

and refined on the Illumina beadchip (Figure 1). The centromeric breakpoint of the 22 Mb deletion, located at ~102 Mb, overlaps with a known intrachromosomal segmental duplication, and this deletion encompasses the *ATM* gene [MIM 607585] (see below). Comparing the BAC-array ratios with experiments in which all cells are affected by an imbalance (e.g., ratios for chromosome X in hybridization of male versus female), we estimated that the deletion in twin 292 is present in ~20% of nucleated blood cells.¹⁶ We also confirmed this CNV in quintuple in a high-resolution, repeat-free, nonredundant, PCR-based array (Figure 1), a test we previously developed for the detection of pathogenic gains and deletions and the discovery of CNV.^{27–29} Details regarding this confirmatory array are available from the authors upon request.

In addition to the 11q rearrangement, another large deletion, affecting 4p and a considerable part of 4q, was found in twin 292 by both 32K BAC analysis and Illumina beadchip (Figure 1). This ~85 Mb deletion was present in ~10–15% of cells. On the basis of the striking conjunction of two very large deletions, we searched the literature and found that chromosome 4 deletion and an 11q deletion, which targets the *ATM* gene, are common for chronic lymphocytic leukemia (CLL [MIM 151400]).^{30,31} We were not previously aware of any additional disease affecting twin 292, but consultation of the medical records confirmed that subject 292 had been diagnosed with CLL prior to sampling of his blood in the course of our study (Table S1). We have thus “rediaagnosed” a clone of CLL cells containing two somatic and pathogenic CNVs in this twin. In this case, because we are studying the correct target tissue, we are confident that these two changes are truly pathogenic. The fact that we were able to detect this, in spite of a low percentage of cells containing aberrations, is a good illustration of the power of both genome-wide array platforms. As far as we are aware, the chromosome 4 aberration in subject 292 represents the lowest reported number of cells containing a change that has been detected via an array-based genome-wide platform for CNV screening. There are two main challenges in scoring “somatic CNVs” in monozygotic twins or other types of samples used for analysis of somatic mosaicism for copy-number changes. One concern is usually the small size of these CNVs. The second is that these aberrations will typically occur in only a proportion of cells. Both of these aspects make the analysis more challenging than scoring “germline CNVs” that are expected to be present in 100% of studied cells. The other eight STR MZ pairs also showed additional imbalances on the 32K BAC array, and this was replicated on the Illumina beadchip for two pairs. For instance, Figure 2 displays five small deviating loci, concordant between 32K array and Illumina, in twin pairs 491/492 and 701/702.

Several other putative CNV loci were found with the 32K arrays in six additional STR MZ pairs (Figures S1A and S1B). These loci were defined by two dye-swap experiments with at least two overlapping BAC clones deviating by >2 global

standard deviations, the same criteria used when scoring the 11 Mb and 85 Mb deletions in twin 292. Tables S2–S4 further summarize the deviating loci, defined as single BACs or neighboring/overlapping BACs in all STR-twins with different cutoffs of statistical significance (difference of ≥ 2 or ≥ 3 global SDs). There were 31 loci that deviated in three phenotypically discordant MZ pairs and four loci that deviated in four such pairs (Tables S3 and S4). These can be viewed as candidates for containment of genes involved in the development of PD.

For the analysis of the ten phenotypically concordant twin pairs from NTR, the Illumina system was the primary platform used (Table S1). In addition to the concordant SNP genotypes (see above), including several CNVs that were shared by both twins of a pair, a few discordances in A and B allele frequencies were also found, suggesting putative de novo somatic CNV events. Figure 3 shows a clear example of a large CNV, which covers ~1.6 Mb on chromosome 2 in twin pair D and extends from SNPs rs2304429 to rs1662987, implying a deletion in twin D8. On the basis of the quantification, it too is found to be present in less than 100% of the cells. Two additional methods were applied to validate this finding: high-resolution Melting Curve Analysis and pyrosequencing (Figures 3C and 3D). Both methods confirmed the presence of the deletion and indicated that it was present in approximately 70–80% of blood cells from D8. Circumstantial evidence was obtained for other, posttwinning CNVs, but the statistics of identifying somatically mosaic (i.e., incomplete) CNVs by SNP concordance analysis requires further methodological development beyond the scope of this report.

Analysis of CNVs is a generally understudied aspect of human genetic variation, particularly in somatic cells. MZ twins represent an excellent focus for such studies because any genotypic difference between twins derived from the same zygote highlights an irrefutable case of somatic variation. It is likely that the confirmed CNVs shown here represent only the “tip of an iceberg” of all CNVs that are actually present in the studied twins. The notion of somatic variation being far more common than previously assumed agrees well with our other, recent results showing CNVs between normal, fully differentiated tissues within an individual human subject (A.P., C.E.G.B., R.A., T.D.d.S., U.M., J.S., D.v.T., A.P., C.C., E.C.P., J.K., and J.P.D., unpublished data).

Our findings influence the understanding of phenotypic and genotypic diversity in MZ twins. First, we wish to stress that although our data seem to suggest a difference, it is not justified to conclude that phenotypically discordant MZ twins are more frequently affected by CNV than are concordant MZ twins. To assess this question, larger cohorts of both twin categories need to be analyzed, preferably by use of DNA extracted from multiple tissues. On the other hand, some of the CNVs in twins discordant for Parkinson disease, parkinsonism, or Lewy body dementia might be pathogenic for these phenotypes, especially when they would also be present in the affected cells of

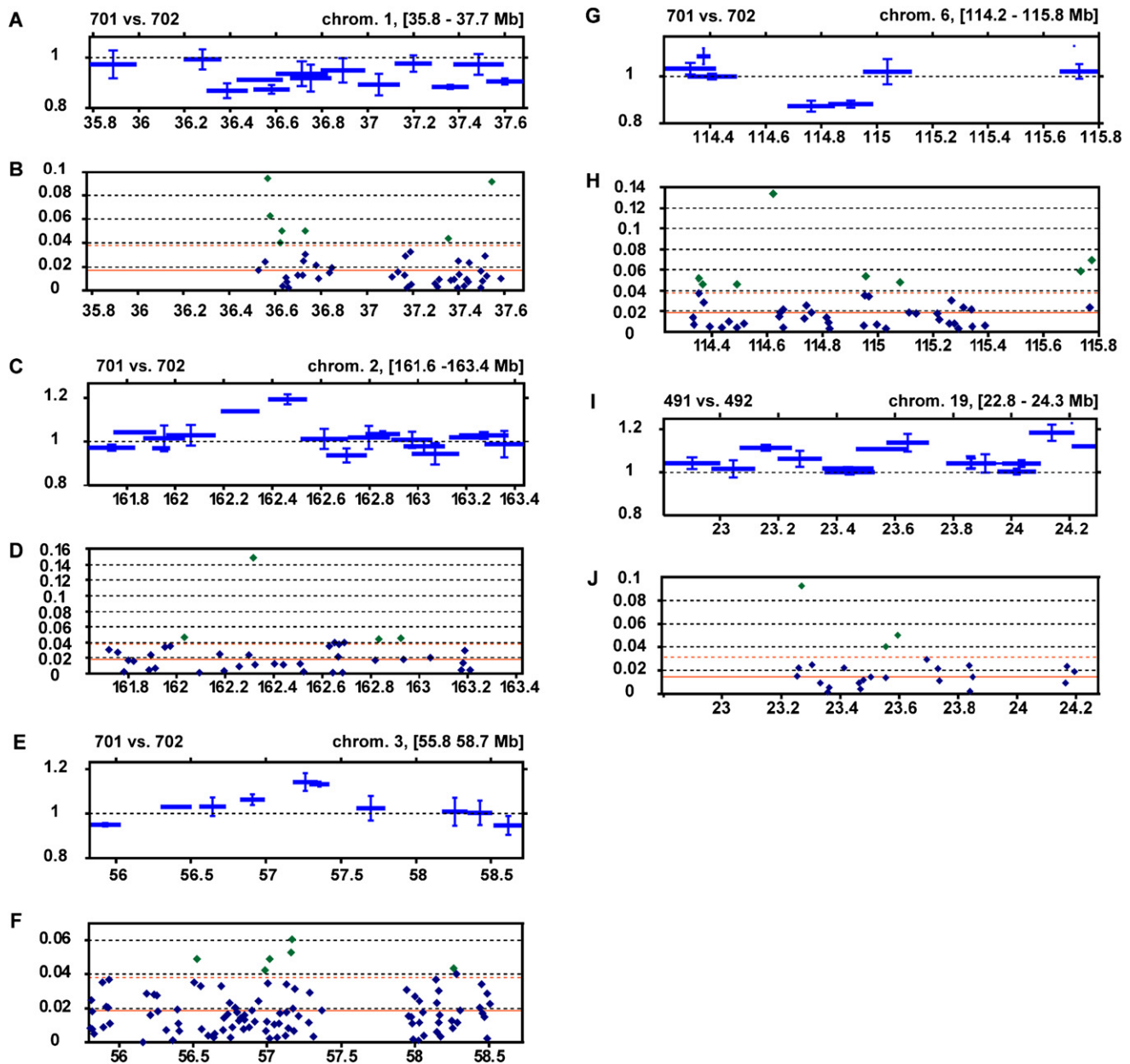


Figure 2. Parallel CNV Analysis with Two CNV Platforms

We show five loci deviating in twin pairs 701/702 and 491/492 on the 32K BAC array (A, C, E, G, and I) and the Illumina HumanHap 300 Duo beadchip (B, D, F, H, and J). Regions altered between the two twins in each pair were defined as having at least two overlapping BAC clones that differ by ≥ 2 global standard deviations and are supported by at least one corresponding and one deviating SNP on the Illumina beadchip.

32K BAC array-CGH profiles were derived from two dye-swap experiments. Relative fluorescent intensity was plotted as a straight ratio. The data points shown were scored in both experiments and were not deviating more than one global SD from each other. The global mean and one SD for these two combined experiments were 1.003 and 0.039, respectively. The horizontal blue lines represent the average straight ratio, and error bars for each BAC show SD between two measurements in dye-swap experiments.

(B), (D), (F), (H), and (J) show the absolute difference in allele frequencies for the heterozygous SNPs on the beadchip for the corresponding chromosomal regions. The solid orange lines indicate the average absolute difference for all heterozygous SNPs for each of the twin pairs, and the dotted lines indicate the standard deviation (SD) for these. The green diamonds indicate SNPs deviating from the average with more than one SD. All genomic positions on the x axis are in million base pairs (Mb) and according to the hg18 build.

the central nervous system. Indeed, a frequent limitation of CNV analyses as generally performed in the context of somatic disorders is that the genotype of one tissue analyzed (usually blood) might not always be relevant to the

genotype of the tissue responsible for the disease, a phenomenon well known for other disorders.³² Second, our findings issue a note of caution. Because genome-wide tools show de novo CNVs in healthy individuals, we

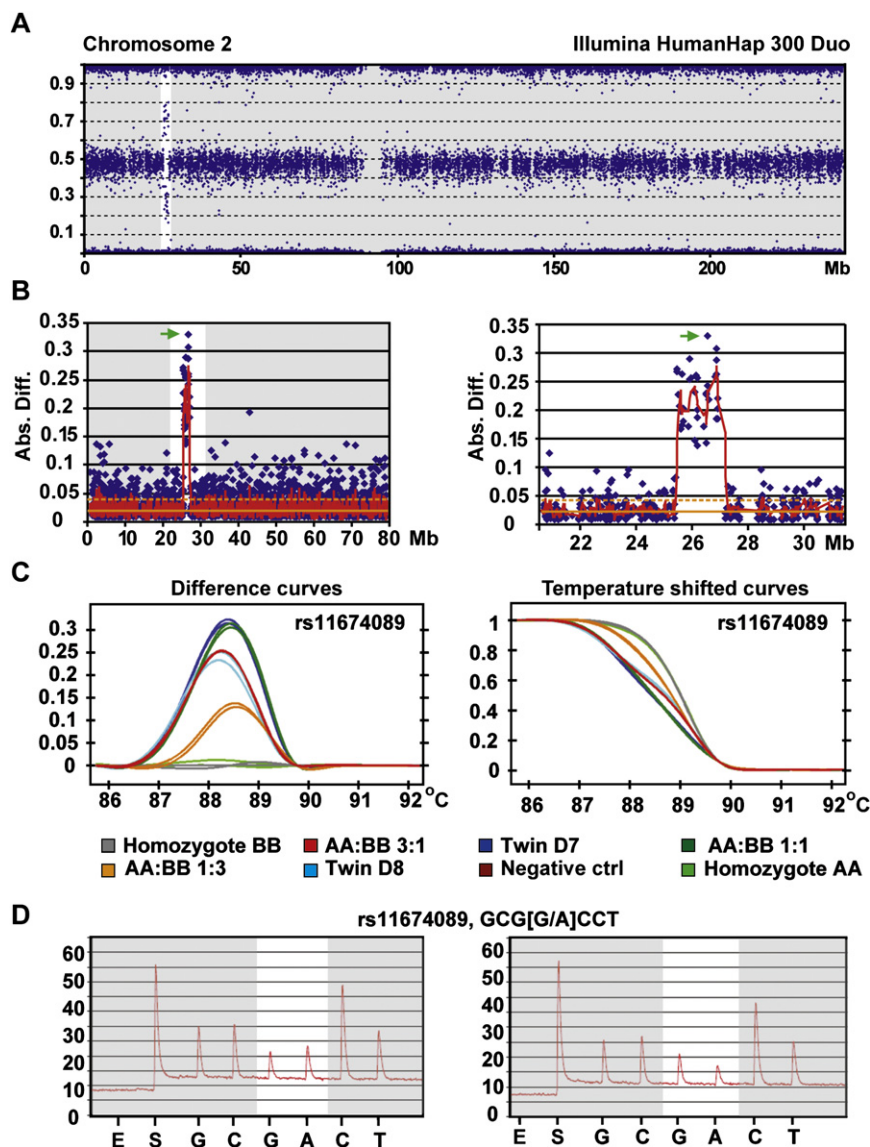


Figure 3. CNV Analysis of Twin D8 Showing the 1.6 Mb Deletion on Chromosome 2

(A) Profile of the entire chromosome 2 from Illumina HumanHap 300 Duo beadchip showing the values of SNP allele ratios. True heterozygous SNPs are expected to be distributed around a value of 0.5. In the highlighted region (white box), the allele ratios differ significantly from 0.5, indicating an imbalance in the allele signals caused by a 1.6 Mb deletion.

(B) Two enlarged views of the deleted region, plotted as values of absolute difference between the heterozygous SNP allele frequencies in twin D8 versus twin D7, calculated in a similar way as shown in Figures 1C and 1E. The red line in both graphs displays the moving average, with a period of 30 and five data points in the left- and right-hand graphs, respectively. The solid and the dotted orange lines represent the average and the standard deviation of the absolute difference, respectively. The green arrows indicate the SNP rs11674089, used for confirmatory experiments using Melting Curve Analysis (C) and pyrosequencing (D).

To confirm the presence of the deletion and to estimate the number of cells carrying it, we performed high-resolution Melting Curve Analysis (hrMCA, [C]) and pyrosequencing (D). For hrMCA, reference samples were made from homozygous AA and BB individuals and mixes of AA:BB of 1:1, 3:1 and 1:3. MCA shows the signals for the homozygous AA and BB samples as nearly identical, whereas the signal of twin D7 (heterozygous for this SNP) matches that of the reference AB sample. The signal for twin D8 does not follow

that of the reference AB sample but follows that of the AA:BB 3:1 sample instead, confirming the deletion and estimating it as being present in 70%–80% of the cells.

(D) displays pyrosequencing analysis of the SNP rs11674089 for twins D7 (left panel) and D8 (right panel). Taking the G to A ratio of twin D7 as normal (50:50), the G to A ratio in twin D8 shifts to 3:1, suggesting that ~70%–80% of the cells analyzed carry the 1.6 Mb deletion.

should expect such changes when analyzing patients. The detection of de novo aberrations has so far been considered evidence for a causal link between mutation and disease status of a patient. We show that one should be careful when drawing such conclusions.

Finally, estimating the frequency with which de novo CNVs occur is currently difficult. On the basis of the detection of one megabase-range, confirmed CNV among ten unselected, phenotypically concordant MZ twin pairs next to several potential ones with a lower degree of cellular mosaicism and/or smaller size, the de novo posttwinning CNV frequency could be as high as 5% on a per-individual basis or 10% per twinning event. The detection rate among the discordant Parkinson twins suggests an even higher figure, but this is within a selected, discordant

group and thus might represent a mixture of phenotypically relevant, as well as random, and phenotypically neutral CNV events. This figure is clearly only a first approximation, and a larger cohort of MZ twins needs to be studied in order to more accurately define the frequency. It is not straightforward to derive supporting evidence from other studies, because parents were typically not analyzed in these studies. One such estimate, based on the de novo deletion frequency in the *DMD* gene [MIM 300377], which lacks conspicuous low-copy repeats (segmental duplications) enhancing propensity for rearrangements, arrives at one de novo deletion per eight newborns and one duplication per 50 newborns.³³ Sebat et al.³⁴ mention the detection of one de novo deletion when 28 CEPH individuals were studied, an essentially similar frequency. We

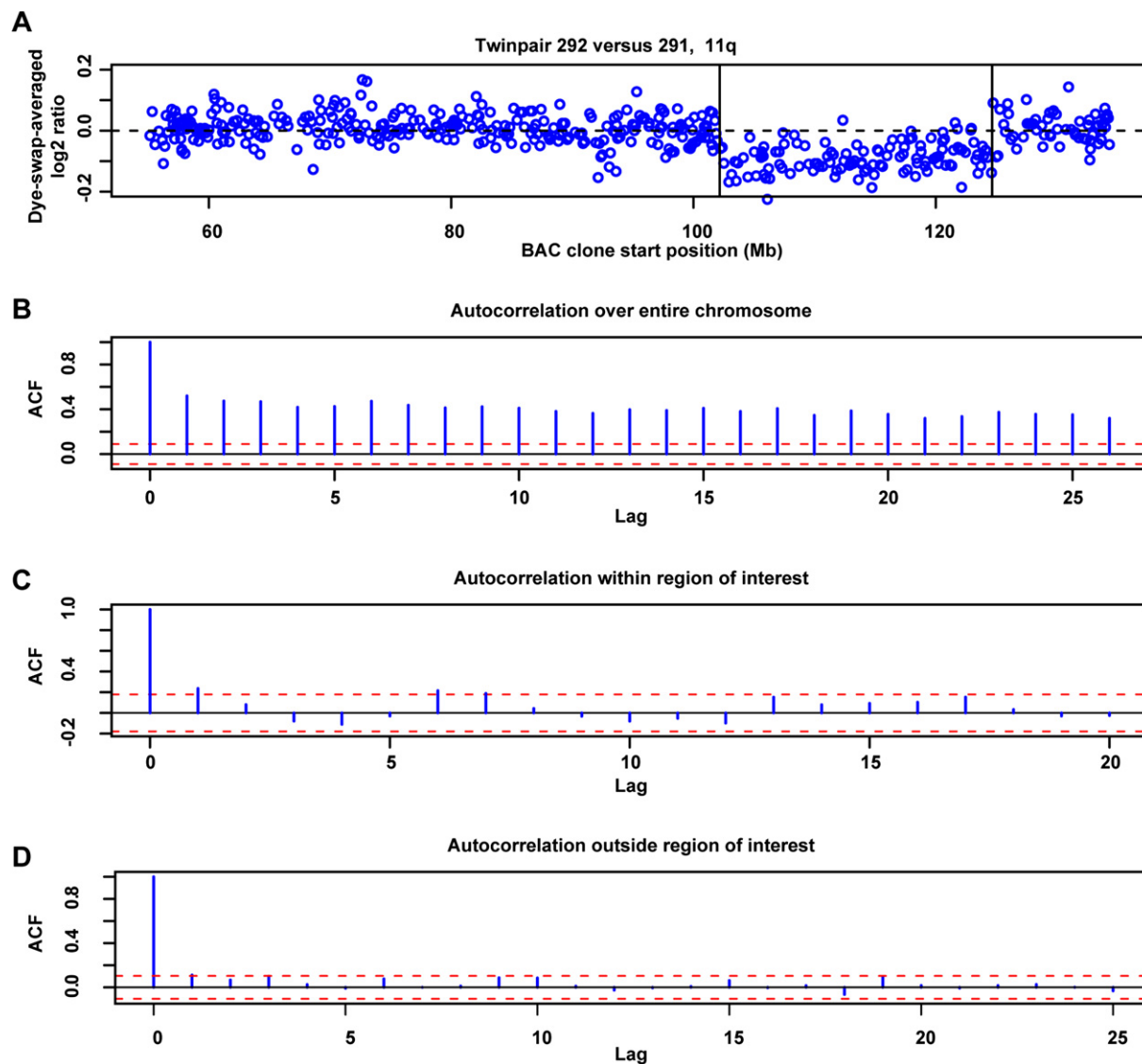


Figure 4. Statistical Analysis of Two Dye-Swap Experiments from Twin 291 versus 292

(A) Dye-swap-averaged log₂ ratios of the data points from the long arm of chromosome 11 (11q) between twins 291 and 292. Boundaries of the inferred region of deletion are marked. A nonparametric Wilcoxon rank sum test comparing the values of the 120 data points within the region to the 359 data points outside the region yields a *p* value less than 2.2×10^{-16} . In a series of *n* points, the number of contiguous regions of any length was $n(n-1)/2$, because a contiguous region was defined by its boundaries. The number of possible regions in the data shown in (A) was therefore 114,481. Thus, even the Bonferroni-corrected *p* value was on the order of 10^{-11} . This shows that the region of deletion was not merely the result of capitalizing on chance.

(B), (C), and (D) display the autocorrelation function (ACF) computed on the data shown in (A). (B) was computed with all data points from 11q, and (C) and (D) were computed on data points within and outside the deleted region, respectively. The ACFs show that the autocorrelation evident in (B) was due almost entirely to the deletion and was negligible after controlling for it.

note that the frequency of de novo CNVs strongly depends on the cutoff criteria set to detect these, especially when the size of an aberration and the number of cells affected by it diminish. In the above circumstances, the confirmation of CNVs via alternative techniques becomes particularly challenging. Our work clearly suggests a need for new, better methods for confirmatory analyses, of which massively parallel sequencing seems promising.¹⁵ On the other hand, our results also point to the feasibility of studies targeting larger cohorts of MZ twins discordant for var-

ious phenotypes as a way to characterize genetic factors predisposing one to disease. Future studies should ideally be performed with the use of DNA extracted from more than one tissue, or from cells derived from the same developmental lineage as the target tissue that is responsible for generation of the discordant phenotype.

Supplemental Data

Two supplemental figures and four supplemental tables are available at <http://www.ajhg.org/>.

Acknowledgments

This study was supported by funds from the Health Services Foundation and the General Endowment Fund (HSF, GEF) from the University of Alabama at Birmingham Medical School, the Swedish Cancer Society, the Swedish Children's Cancer Foundation, and the U.S. Army Medical Research and Materiel Command (award no. W81XWH-04-1-0269 to J.P.D.), as well as by Netherlands Genomics Initiative funds to the Center for Medical Systems Biology supporting A.A.C.H.G., J.T.d.D., D.I.B., and G.J.B.v.O. The twin specimens and phenotypic data were collected with the support of National Institutes of Health (NIH) grant ES10758 to N.L.P. S.E. acknowledges the support of NIH grant 5T32HL072757-04, and D.B.A. acknowledges the support of NIH grants 3P30DK056336-05S2 and 5R01ES009912-08. We thank Drs. Eline Slagboom and Claudia Ruivenkamp for sample contribution, guidance, and discussions, as well as Drs. Bruce A. Korf, Lisa Guay-Woodford, Jay McDonald, and Marco Marra for critical comments on the manuscript. We also thank Dr. Doug Horsman for valuable advice about the genetics of CLL.

Received: October 11, 2007

Revised: December 6, 2007

Accepted: December 19, 2007

Published online: February 14, 2008

Web Resources

The URLs for data presented herein are as follows:

CHORI BACAPAC Resources, <http://bacpac.chori.org/genomicRearrangs.php>

Database of Genomic Variants, <http://projects.tcag.ca/variation/>
GEO omnibus main page, <http://www.ncbi.nlm.nih.gov/projects/geo/>

Human Genome Browser, <http://genome.ucsc.edu/cgi-bin/hgGateway>

Linnaeus Centre for Bioinformatics (LCB) environment for microarray-data management, <http://www.lcb.uu.se/lcbdw.php>

Online Mendelian Inheritance in Man, <http://www.ncbi.nlm.nih.gov/Omim>

Wilcox.Test Function for R, <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html>

Accession Numbers

The array data were deposited at the GEO main page at NCBI under the accession number GSE9609.

References

1. Busjahn, A., and Hur, Y.M. (2006). Twin registries: An ongoing success story. *Twin Res. Hum. Genet.* 9, 705.
2. Machin, G.A. (1996). Some causes of genotypic and phenotypic discordance in monozygotic twin pairs. *Am. J. Med. Genet.* 61, 216–228.
3. Gringras, P., and Chen, W. (2001). Mechanisms for differences in monozygous twins. *Early Hum. Dev.* 64, 105–117.
4. Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballesstar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. USA* 102, 10604–10609.
5. Petronis, A. (2006). Epigenetics and twins: Three variations on the theme. *Trends Genet.* 22, 347–350.
6. Gratacos, M., Nadal, M., Martin-Santos, R., Pujana, M.A., Gago, J., Peral, B., Armengol, L., Ponsa, I., Miro, R., Bulbena, A., et al. (2001). A polymorphic genomic duplication on human chromosome 15 is a susceptibility factor for panic and phobic disorders. *Cell* 106, 367–379.
7. Youssoufian, H., and Pyeritz, R.E. (2002). Mechanisms and consequences of somatic mosaicism in humans. *Nat. Rev. Genet.* 3, 748–758.
8. Erickson, R.P. (2003). Somatic gene mutation and human disease other than cancer. *Mutat. Res.* 543, 125–136.
9. Gollob, M.H., Jones, D.L., Krahn, A.D., Danis, L., Gong, X.Q., Shao, Q., Liu, X., Veinot, J.P., Tang, A.S., Stewart, A.F., et al. (2006). Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. *N. Engl. J. Med.* 354, 2677–2688.
10. Flores, M., Morales, L., Gonzaga-Jauregui, C., Dominguez-Vidana, R., Zepeda, C., Yanez, O., Gutierrez, M., Lemus, T., Valle, D., Avila, M.C., et al. (2007). Recurrent DNA inversion rearrangements in the human genome. *Proc. Natl. Acad. Sci. USA* 104, 6099–6106.
11. 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.
12. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat. Genet.* 36, 949–951.
13. Sebat, J. (2007). Major changes in our DNA lead to major changes in our thinking. *Nat. Genet.* 39, s3–s5.
14. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444–454.
15. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. (2007). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science* 318, 420–426.
16. Buckley, P.G., Mantripragada, K.K., Benetkiewicz, M., Tapia-Páez, I., Diaz de Ståhl, T., Rosenquist, M., Ali, H., Jarbo, C., de Bustos, C., Hirvelä, C., et al. (2002). A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Hum. Mol. Genet.* 11, 3221–3229.
17. Ishkanian, A.S., Malloff, C.A., Watson, S.K., deLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., et al. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* 36, 299–303.
18. Wong, K.K., deLeeuw, R.J., Dosanjh, N.S., Kimm, L.R., Cheng, Z., Horsman, D.E., MacAulay, C., Ng, R.T., Brown, C.J., Eichler, E.E., et al. (2007). A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* 80, 91–104.
19. Diaz de Ståhl, T., Sandgren, J., Piotrowski, A., Nord, H., Andersson, R., Menzel, U., Bogdan, A., Thuresson, A.-C., Poplawski, A., von Tell, D., et al. (2007). Profiling of copy number variations (CNVs) in healthy individuals in three ethnic groups using a human genome 32K BAC-clone-based array. *Hum. Mutat.*, in press.

20. Ameer, A., Yankovski, V., Enroth, S., Spjuth, O., and Komorowski, J. (2006). The LCB Data Warehouse. *Bioinformatics* 22, 1024–1026.
21. R Development Core Team (2006). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).
22. Peiffer, D.A., Le, J.M., Steemers, F.J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C.A., Belmont, J., et al. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* 16, 1136–1148.
23. Lichtenstein, P., De Faire, U., Floderus, B., Svartengren, M., Svedberg, P., and Pedersen, N.L. (2002). The Swedish Twin Registry: A unique resource for clinical, epidemiological and genetic studies. *J. Intern. Med.* 252, 184–205.
24. Wirdefeldt, K., Gatz, M., Bakaysa, S., Fiske, A., Flensburg, M., Petzinger, G., Widner, H., Lew, M., Welsh, M., and Pedersen, N. (2007). Complete ascertainment of Parkinson disease in the Swedish Twin Registry. *Neurobiol. Aging*, in press.
25. Boomsma, D.I., de Geus, E.J., Vink, J.M., Stubbe, J.H., Distel, M.A., Hottenga, J.J., Posthuma, D., van Beijsterveldt, T.C., Hudziak, J.J., Bartels, M., et al. (2006). Netherlands Twin Register: From twins to twin families. *Twin Res. Hum. Genet.* 9, 849–857.
26. Elbaz, A., Nelson, L.M., Payami, H., Ioannidis, J.P., Fiske, B.K., Annesi, G., Carmine Belin, A., Factor, S.A., Ferrarese, C., Hadjigeorgiou, G.M., et al. (2006). Lack of replication of thirteen single-nucleotide polymorphisms implicated in Parkinson's disease: A large-scale international study. *Lancet Neurol.* 5, 917–923.
27. Mantripragada, K.K., Buckley, P.G., Jarbo, C., Menzel, U., and Dumanski, J.P. (2003). Development of NF2 gene specific, strictly sequence defined diagnostic microarray for deletion detection. *J. Mol. Med.* 81, 443–451.
28. Mantripragada, K.K., Thuresson, A.C., Piotrowski, A., Diaz de Stahl, T., Menzel, U., Grigelionis, G., Ferner, R.E., Griffiths, S., Bolund, L., Mautner, V., et al. (2006). Identification of novel deletion breakpoints bordered by segmental duplications in the NF1 locus using high resolution array-CGH. *J. Med. Genet.* 43, 28–38.
29. de Bustos, C., Diaz de Stahl, T., Piotrowski, A., Mantripragada, K.K., Buckley, P.G., Darai, E., Hansson, C.M., Grigelionis, G., Menzel, U., and Dumanski, J.P. (2006). Analysis of copy number variation in the normal human population within a region containing complex segmental duplications on 22q11 using high-resolution array-CGH. *Genomics* 88, 152–162.
30. Summersgill, B., Thornton, P., Atkinson, S., Matutes, E., Shipley, J., Catovsky, D., Houlston, R.S., and Yuille, M.R. (2002). Chromosomal imbalances in familial chronic lymphocytic leukaemia: A comparative genomic hybridisation analysis. *Leukemia* 16, 1229–1232.
31. Ripolles, L., Ortega, M., Ortuno, F., Gonzalez, A., Losada, J., Ojanguren, J., Soler, J.A., Bergua, J., Coll, M.D., and Caballin, M.R. (2006). Genetic abnormalities and clinical outcome in chronic lymphocytic leukemia. *Cancer Genet. Cytogenet.* 171, 57–64.
32. Helderma-van den Enden, A.T., Ginjaar, H.B., Kneppers, A.L., Bakker, E., Breuning, M.H., and de Visser, M. (2003). Somatic mosaicism of a point mutation in the dystrophin gene in a patient presenting with an asymmetrical muscle weakness and contractures. *Neuromuscul. Disord.* 13, 317–321.
33. van Ommen, G.J. (2005). Frequency of new copy number variation in humans. *Nat. Genet.* 37, 333–334.
34. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445–449.