Twin Research and Human Genetics
Volume 18 ▌ Number 1 ▌ pp. 13–18 ▌ © The Author(s) 2015 ▌ doi:10.1017/thg.2014.85

Large Autosomal Copy-Number Differences within Unselected Monozygotic Twin Pairs are Rare

Allan F. McRae,[1,2,3] Peter M. Visscher,[1,2] Grant W. Montgomery,[3] and Nicholas G. Martin[3]
[1]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia
[2]University of Queensland Diamantina Institute, Brisbane, Queensland, Australia
[3]QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia

Monozygotic (MZ) twins form an important system for the study of biological plasticity in humans. While MZ twins are generally considered to be genetically identical, a number of studies have emerged that have demonstrated copy-number differences within a twin pair, particularly in those discordant for disease. The rate of autosomal copy-number variation (CNV) discordance within MZ twin pairs was investigated using a population sample of 376 twin pairs genotyped on Illumina Human610-Quad arrays. After CNV calling using both QuantiSNP and PennCNV followed by manual annotation, only a single CNV difference was observed within the MZ twin pairs, being a 130 KB duplication of chromosome 5. Five other potential discordant CNV were called by the software, but excluded based on manual annotation of the regions. It is concluded that large CNV discordance is rare within MZ twin pairs, indicating that any CNV difference found within phenotypically discordant MZ twin pairs has a high probability of containing the causal gene(s) involved.

Keywords: CNV, de novo, monozygotic twins, post-zygotic

MZ twin pairs result from the fertilization and subsequent splitting of a single embryo and, thus, are widely considered to be genetically identical. Being genetically identical makes MZ twins an important resource for the study of developmental plasticity in humans. A growing number of studies are investigating epigenetic differences in phenotypically discordant MZ twins, particularly those discordant for disease (Bell & Spector, 2011; Kato et al., 2005; Van Dongen et al., 2012; Zwijnenburg et al., 2010).

While MZ twins are largely genetically identical, somatic mutations have been observed in one twin of a pair. These tend to be observed in twin pairs that are discordant for disease, although it is likely that such differences also exist unnoticed within MZ pairs with 'normal' phenotypes. For example, mutations have been found in *ATP2A2* in the affected twin of a pair discordant for Darier disease (Sakuntabhai et al., 1999), in *IRF6* in a pair discordant for Van der Woude syndrome (Kondo et al., 2002) and in *SCN1A* in Dravet's syndrome (Vadlamudi et al., 2010).

Copy-number variants (CNV) are becoming recognized as affecting a wide range of human traits (Girirajan et al., 2011; Zhang et al., 2009) and thus there is potential for MZ discordance to be caused by de novo CNVs. An initial investigation of 19 MZ twins pairs — 9 pairs discordant for disease and 10 phenotypically unselected or concordant normal pairs — found three copy-number differences within pairs, two large deletions in one individual of a discordant pair, and one large deletion in a member of a phenotypically concordant pair (Bruder et al., 2008). All these CNVs were present in only a portion of cells in the carrier, with frequencies of 20% and 15% for the CNV in the discordant pair and in 80% of cells in the individual from the concordant pair. A study of CNV in MZ twin pairs selected for concordance and discordance for attention problems identified two de novo CNVs, one occurring pre-twinning (and thus present in both members of the twin pairs) and one post-twinning (Ehli et al., 2012). A number of other studies of MZ twins with discordant disease phenotypes have failed to identify any CNV discordance within the pair (Lasa et al., 2010; Ono et al., 2010; Veenma et al., 2012).

A thorough quantification of the rate of genetic differences in a phenotypically unselected set of MZ twins is required in order to determine the population rate of copy-number differences within MZ pairs. If such copy-number

RECEIVED 4 November 2014; ACCEPTED 11 November 2014. First published online 12 January 2015.

ADDRESS FOR CORRESPONDENCE: Allan F. McRae, Centre for Neurogenetics and Statistical Genomics, Queensland Brain Institute, The University of Queensland, QBI Building (#79), St Lucia, QLD 4072, Australia. E-mail: a.mcrae@uq.edu.au

differences are common within unselected MZ pairs, the probability of any observed copy-number differences in an MZ pair discordant for a disease being causal is greatly reduced. The frequency of copy-number differences is also an important consideration when investigation developmental plasticity in epigenetic studies of MZ twins as such studies assume genetic identity within the twin pair. We address this issue through the investigation of CNV in a population-based sample of 376 pairs of MZ twins.

Materials and Methods

Cohort

After quality control filtering (described below), 376 pairs of MZ twins — 169 male and 207 female — were available for CNV calling. The genotyping and initial cleaning of these samples are described at length elsewhere (Medland et al., 2009). Briefly, all samples were genotyped on Illumina Human610-Quad arrays. Samples with greater than 5% missing data were excluded from further analysis. MZ twin status and sex was confirmed using the SNP genotypes.

CNV Calling

The primary CNV calls were made using the software QuantiSNP v2 (Colella et al., 2007) using the default settings ($L =$ 2 M, 10 EM iterations) and using the GC correction option. A second set of CNV calls was made using PennCNV (Wang et al., 2007), also using the GC correction model (Diskin et al., 2008).

Poor quality samples were excluded from the study if they failed any of the stringent quality controls metric from either CNV calling program. For QuantiSNP, samples with more than 200 CNV calls (without filtering on log Bayes factor scores), an outlier rate > 0.0075, spread of log R ratio values > 0.2 or measured spread of distribution of B allele frequencies for heterozygote genotypes > 0.07 were excluded. Further samples were excluded based on PennCNV output with greater than 70 CNV calls, a log R ratio standard deviation > 0.28, the B allele frequency standard deviation > 0.045, and an absolute waviness factor > 0.02. Thresholds for each measurement were chosen by examining their distribution for the obvious outliers. Overall, $\sim$6% of samples failed one of these stringent quality metrics, resulting in 46 out of the original 422 twin pairs being excluded from the final dataset (due to at least one of the pair failing quality control). A strong overlap was observed between the samples excluded using each metric for both CNV calling programs.

CNV Differences within MZ Pairs

Differences within MZ pairs were first detected using QuantiSNP CNV calls, followed by confirmation of the CNV calls using PennCNV. QuantiSNP CNV calls were filtered to have at least five probes, a length of 1 kb and to be autosomal. As CNVs in highly repetitive regions — in particular centromeric regions — are known to have high false positive rates, any CNV that was annotated as having greater than 70% of the region as repetitive DNA in the UCSC version hg18 RepeatMasker annotation was removed.

For each individual, all CNVs with a QuantiSNP log Bayes factor greater than 30 were investigated for presence in the QuantiSNP CNV calls in the co-twin, regardless of the log Bayes factor. CNV calls for the twin pair are considered concordant if there is any overlap of CNV calls of the same category (deletion or duplication), irrespective of the actual called copy number and the CNV end points. Discordant CNV calls were further investigated using PennCNV, confirming both the CNV call in the original individual and the lack of CNV in the co-twin. This was repeated using a log Bayes factor threshold of 10 for the initial QuantiSNP CNV calls.

Results

CNVs in the MZ twins were called using QuantiSNP v2 on data from Illumina Human610-Quad arrays. After stringent quality control filters were applied, 376 pairs of MZ twins remained. All CNVs called with a log Bayes factor of greater than 30 were investigated for presence in the co-twin, regardless of their significance, providing 126 candidate regions for copy-number differences within the MZ pairs. A large number of these differences were repeated across other pairs and also seen at high frequency in a larger cohort (data not shown), indicating they were false positive CNVs. This was confirmed by noting that many of these CNVs fell in centromeric and other highly repetitive DNA regions that are known to generate false positive CNV calls (Wang et al., 2007). Based on CNVs in centromeric regions, any CNV across a region with greater than 70% of the sequence in repeat regions was removed, reducing the number of candidate regions to 16.

The remaining 16 regions were further investigated using CNV calls from PennCNV (Wang et al., 2007) and all called CNV regions were confirmed. For 9 of the 16 CNVs, PennCNV also identified a corresponding CNV in the co-twin. All the CNVs identified in the co-twin had a low confidence score compared to that in the primary twin, and many were called to have a shorter length or as part of a number of CNVs across a region, indicating likely false positive differences due to limitations of the calling algorithms. Additionally, one CNV identified by QuantiSNP was called as part of a larger CNV by PennCNV. In the co-twin, a CNV was found that partially overlapped the larger CNV but did not overlap the original CNV call, and so this was excluded from further analysis.

The details of the remaining six identified CNV differences are given in Table 1. There was strong evidence for one duplication on chromosome 5 and the measured log R ratios and B allele frequencies are given in Figure 1. This CNV covered 37 probes on the genotyping array and was highly significant, with a QuantiSNP log Bayes factor of 54.1 and

TABLE 1

Details of the Six CNV Differences Identified Across Six Different MZ Pairs by Both QuantiSNP and PennCNV

Chr	Copy number	QuantiSNP details*	PennCNV details*
2	1	41,092,148–41,101,972, 10, 35.9	41,092,148–41,101,972, 10, 44.01
2	1	89,731,562–89,885,025, 15, 42.5	89,731,562–89,885,025, 15, 40.8
5	**3**	**78,333,027–78,460,944, 37, 54.1**	**78,333,027–78,460,944, 37, 77.6**
8	1	40,304,029–40,308,706, 8, 35.8	40,304,029–40,308,706, 8, 39.5
15	1	19,157,192–19,407,285, 32, 33.3	19,158,166–19,228,416, 10, 22.1
17	1	41,792,236–41,914,286, 36, 30.8	41,814,370–41,914,286, 34, 31.2

Note: *start – stop, # probes, log Bayes factor/confidence score. The copy-number variant showing the strongest evidence for differences within an MZ pair is highlighted in bold type.

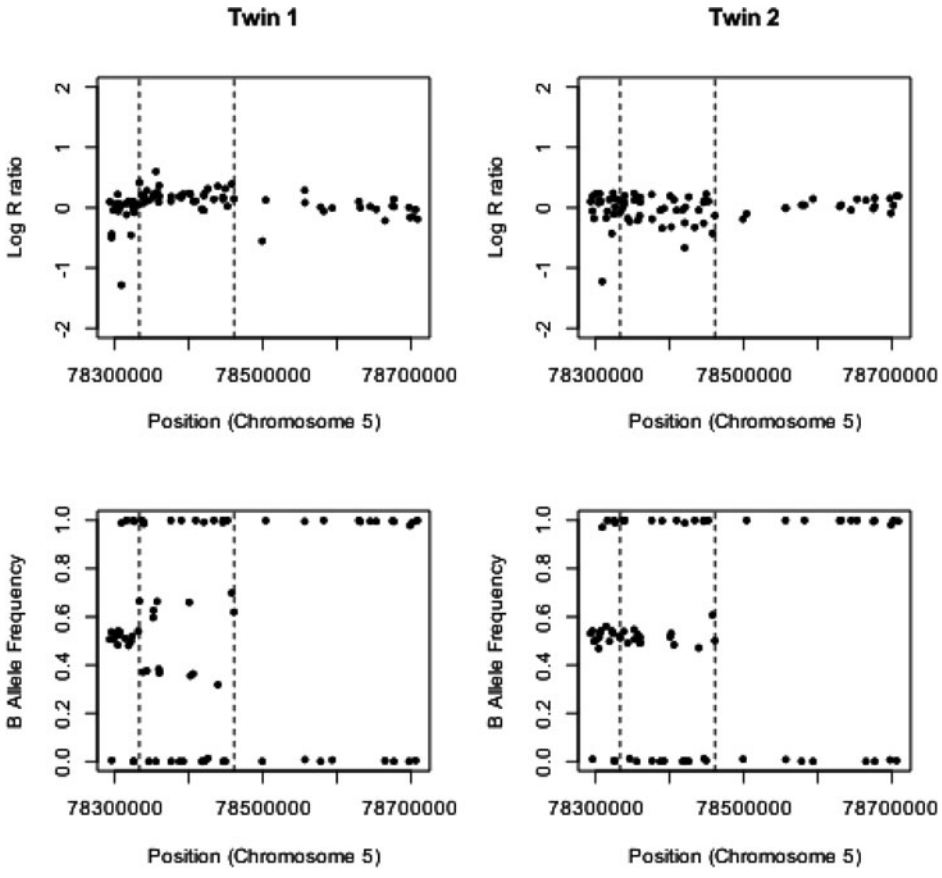

FIGURE 1

Copy-number discordance within a MZ twin pair. A ~130 KB duplication is observed between 78,333,027–8,460,944 bp (indicated by dashed lines) on chromosome 5 in twin 1. This duplication shows the characteristic increase in log R ratio above the value of zero expected when two copies of the chromosome are present. There is also clear evidence of the characteristic four genotypes (AAA, AAB, ABB, and BBB) that are seen in a duplication event in the plot of the B allele frequencies (bottom). No evidence of a duplication is observed in the co-twin (right).

PennCNV confidence score of 77.6. The log R ratio in the region of the insertion is slightly raised above zero, although not to the extent of the expected value of $\log_2(3/2)$ ($= 0.58$) if three copies were present in all cells. The B allele frequencies for the heterozygous SNP clearly show two bands that indicate a duplication, but deviate towards 0.5 from their expected values of 0.33 and 0.67, indicating that this duplication is potentially a mosaic. Similar plots of log R ratio and B allele frequencies for the five identified deletions are given in supplementary Figures S1–S5. All of these deletions show the requisite reduction in log R ratio values, but none to the extent of the expected value of -1. The longer deletion on chromosome 2 (supplementary Figure S1) covers an immunoglobulin region, which is known to generate false positive CNV calls (Wang et al., 2007). The region of chromosome 17 identified as having a putative deletion (supplementary Figure S2) was in a region of (monomorphic) CNV probes only, so B allele frequency cannot be used as a

confirmatory measure and we rely on intensity data alone. Also, the log R ratios in the surrounding regions are relatively noisy, indicating a difficult region to call CNVs using these SNP arrays. The called deletion on chromosome 15 (supplementary Figure S3) is not well supported by the B allele frequencies in the region. While these do not cluster around 0.5, the B allele frequency values are consistent across the co-twins indicating the copy-number state does not differ within the pair. The two remaining CNV (supplementary Figures S4 and S5) are both less than 10 KB and do not have any heterozygous SNPs in the region to confirm that the deletion is present only in one twin using differences in B allele frequencies. The validity of these CNV differences would require further investigation using alternative technology. Overall, after stringent quality control and manual inspection of CNV differences within MZ twin pairs, we only find strong evidence for an ~130 KB duplication on chromosome 5.

As the initial threshold of a QuantiSNP log Bayes factor of greater than 30 may result in false negative CNV calls, the CNV filtering procedure was repeated with a threshold of 10. This identified 1,300 potential CNV differences within the MZ pairs, with 586 remaining after filtering out regions in repetitive sequence. Of these 586 CNVs, 499 were also called in the same individual in PennCNV. PennCNV also called an overlapping CNV in the co-twin in 107 of these CNV calls, leaving a total of 392 CNV differences called by both methods, consisting of 307 deletions and 85 duplications. Manual inspection of the 64 CNV having a log Bayes factor >20 showed similar patterns to the CNVs in supplementary Figures S1–S5, where the log R ratio in CNV region was not strongly deviated from zero and the same region in the co-twin showed similar patterns for both log R ratio and B allele frequency measurements indicating that there was no genuine difference in copy-numbers identified within the twin pairs.

Discussion

We have performed a survey of copy-number differences within 376 pairs of MZ twins using Illumina SNP arrays. Overall, there was evidence for only one copy-number difference within a pair of MZ twins, consisting of a ~130 KB duplication. Other regions had CNVs called by both QuantiSNP and PennCNV in a single member of the twin pair, but close inspection of the log R ratio and B allele frequency measurements in the co-twin indicated likely false negative CNV calls generating spurious discordance.

The use of SNP arrays for CNV calling is a relatively noisy process, with even the best calling algorithms suffering from moderate false positive and negative rates (Dellinger et al., 2010; Pinto et al., 2011). In this study, the concordance between two CNV calling software packages, QuantiSNP v2 and PennCNV, was checked for all CNV calls. After filtering out CNVs called in highly repetitive regions, the two CNV calling algorithms demonstrated a strong concordance, particularly for CNVs called with high confidence, indicating a low false positive rate. However, the false negative rate appears high, with a number of apparent CNV differences between MZ co-twins found by QuantiSNP v2 being rejected on the basis of the CNV also being called in the co-twin with PennCNV. Furthermore, manual inspection of intensity levels in the regions of putative CNV discordance indicated that many CNVs were not called by either piece of software. Inspection of the concordant CNV calls within MZ twin pairs shows, as would be expected, that longer CNVs are called more accurately. Thus, despite the limitations of CNV calling from SNP arrays, it is possible to conclude that large copy-number differences within unselected MZ twin pairs are rare. Different technology will be needed to accurately assess small CNVs differing within an MZ pair.

The one CNV with strong evidence for a discordance in an MZ pair was a ~130 kb duplication between 78,333,027–78,460,944 bp on chromosome 5. This duplication demonstrated the expected four bands in the B allele frequency (corresponding to genotypes AAA, AAB, ABB, and BBB), although the two heterozygous classes deviated towards 0.5 from their expected values of 0.33 and 0.67. The log R ratio also is closer to zero than would be expected for a duplication. This indicates that the duplication is probably a mosaic, in that not all cells in the individual contain it. Given the CNV is only seen in one of the twin pair, it must have occurred post fertilization. While the copy-number data are most parsimoniously consistent with mutation post embryo splitting, we are unable to rule out mutation pre-splitting with the 'non-carrier' co-twin either being a genuine non-carrier or being a mosaic below detection levels. Regardless of whether the CNV occurred pre- or post-splitting, it would be highly likely that cells with and without the duplication are present in the carrier, as is consistent with the observed log R ratio and B allele frequencies. As low rates of mosaicism have been observed in the general population (Jacobs et al., 2012; Laurie et al., 2012), it is not possible to attribute the splitting of the embryo to this duplication.

The lack of evidence for frequent CNV differences within unselected MZ pairs supports the continued use of discordant MZ pairs to investigate the developmental etiology of disease. Several studies have provided evidence for copy-number differences within pairs of MZ twins discordant for disease (Bruder et al., 2008; Ehli et al., 2012), the implication being that these are causal variants. This study demonstrates that such differences are rare in a population-based sample of MZ twins, raising the chance that the identified differences are causal. However, given that such CNV differences do occur in a population sample and genuine replication of a CNV difference in an independent pair is likely to be rare, it is important to remain cautious in conclusions about causality and to pursue other biological support for the role

of a CNV region in disease. Epigenome association studies that look for discordant DNA methylation in pairs of MZ twins that are discordant for disease also implicitly make an assumption of no genetic differences within the twin pair. While this study shows that such an assumption will hold for a general population sample, the frequency of copy-number differences in pairs selected for disease discordance remains to be determined and whether frequency may vary between diseases.

Acknowledgments

We thank our twin sample for their participation; Marlene Grace and Ann Eldridge for sample collection; Anjali Henders, Megan Campbell, Lisa Bowdler, Steven Crooks, and staff of the QIMR Molecular Epidemiology Laboratory for DNA sample processing and preparation; Kerrie McAloney for study co-ordination; and Harry Beeby, Daniel Park, and David Smyth for IT support. This work was supported by grants from the Australian Research Council (ARC: A7960034, A79906588, A79801419, DP0212016, DP0343921, DP0664638, DP1093900, FT0991360) and the Australian National Health and Medical Research Council (NHMRC: Medical Bioinformatics Genomics Proteomics Program, 389891). P.M.V. and G.W.M. are supported by the NHMRC Fellowship Scheme.

Supplementary Material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/thg.2014.85.

References

Bell, J. T., & Spector, T. D. (2011). A twin approach to unraveling epigenetics. *Trends in Genetics*, *27*, 116–125.

Bruder, C. E. G., Piotrowski, A., Gijsbers, A. A. C. J., Andersson, R., Menzel, U., Sandgren, J., ... Dumanski,, J. P. (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *American Journal of Human Genetics*, *82*, 763–771.

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., ... Ragoussis, J. (2007). QuantiSNP: An objective bayes hidden-markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, *35*, 2013–2025.

Dellinger, A. E., Saw, S.-M., Goh, L. K., Seielstad, M., Young, T. L., & Li, Y.-J. (2010). Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Research*, *38*, e105.

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., ... Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, *36*, e126.

Ehli, E. A., Abdellaoui, A., Hu, Y., Hottenga, J. J., Kattenberg, M., van Beijsterveldt, T., ... Davies,, G. E. (2012). De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on attention problems. *European Journal of Human Genetics*, *20*, 1037–1043.

Girirajan, S., Campbell, C. D., & Eichler, E. E. (2011). Human copy number variation and complex genetic disease. *Annual Review of Genetics*, *45*, 203–226.

Jacobs, K. B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z., Rodriguez-Santiago, B., ... Chanock, S. J. (2012). Detectable clonal mosaicism and its relationship to aging and cancer. *Nature Genetics*, *44*, 651–658.

Kato, T., Iwamoto, K., Kakiuchi, C., Kuratomi, G., & Okazaki, Y. (2005) Genetic or epigenetic difference causing discordance between monozygotic twins as a clue to molecular basis of mental disorders. *Molecular Psychiatry*, *10*, 622–630.

Kondo, S., Schutte, B. C., Richardson, R. J., Bjork, B. C., Knight, A. S., Watanabe, Y., ... Murray,, J. C. (2002). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nature Genetics*, *32*, 285–289.

Lasa, A., Ramón y Cajal, T., Llort, G., Suela, J., Cigudosa, J. C., Cornet, M., ... Baiget, M. (2010). Copy number variations are not modifiers of phenotypic expression in a pair of identical twins carrying a BRCA1 mutation. *Breast Cancer Research and Treatment*, *123*, 901–905.

Laurie, C. C., Laurie, C. A., Rice, K., Doheny, K. F., Zelnick, L. R., McHugh, C. P., ... Weir, B. S. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics*, *44*, 642–650.

Medland, S. E., Nyholt, D. R., Painter, J. N., McEvoy, B. P., McRae, A. F., Zhu, G., ... Martin, N. G. (2009). Common variants in the trichohyalin gene are associated with straight hair in Europeans. *American Journal of Human Genetics*, *85*, 750–755.

Ono, S., Imamura, A., Tasaki, S., Kurotaki, N., Ozawa, H., Yoshiura, K., ... Okazaki, Y. (2010). Failure to confirm CNVs as of etiological significance in twin pairs discordant for schizophrenia. *Twin Research and Human Genetics*, *13*, 455–460.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., ... Feuk, L. (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology*, *29*, 512–520.

Sakuntabhai, A., Ruiz-Perez, V., Carter, S., Jacobsen, N., Burge, S., Monk, S., ... Hovnanian, A. (1999). Mutations in ATP2A2, encoding a Ca2+ pump, cause Darier disease. *Nature Genetics*, *21*, 271–277.

Vadlamudi, L., Dibbens, L. M., Lawrence, K. M., Iona, X., McMahon, J. M., Murrell, W., ... Berkovic, S. F. (2010). Timing of de novo mutagenesis — A twin study of sodium-channel mutations. *New England Journal of Medicine*, *363*, 1335–1340.

Van Dongen, J., Slagboom, P. E., Draisma, H. H. M., Martin, N. G., & Boomsma, D. I. (2012). The continuing value of

twin studies in the omics era. *Nature Review Genetics*, *13*, 640–653.

Veenma, D., Brosens, E., de Jong, E., van de Ven, C., Meeussen, C., Cohen-Overbeek, T., ... de Klein, A. (2012). Copy number detection in discordant monozygotic twins of congenital diaphragmatic hernia (CDH) and esophageal atresia (EA) cohorts. *European Journal of Human Genetics*, *20*, 298–304.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., ... Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, *17*, 1665–1674.

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009) Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, *10*, 451–481.

Zwijnenburg, P. J. G., Meijers-Heijboer, H., & Boomsma, D. I. (2010). Identical but not the same: the value of discordant monozygotic twins in genetic research. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *153B*, 1134–1149.