Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Brooke Carin Weckselblatt Date

Sequence Analysis of Chromosome Translocations in Neurodevelopmental Disorders

By

Brooke Carin Weckselblatt
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

_____
M. Katharine Rudd, PhD, FACMG
Advisor

_____
Victor Corces, PhD
Committee Member

_____
Maureen Powers, PhD
Committee Member

_____
Zhaohui Qin, PhD
Committee Member

_____
Michael Zwick, PhD
Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____
Date

Sequence Analysis of Chromosome Translocations in Neurodevelopmental Disorders

By

Brooke Carin Weckselblatt
B.A., Bryn Mawr College, 2010

Advisor: M. Katharine Rudd, PhD, FACMG

An abstract of
a dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Sciences,
Genetics and Molecular Biology
2015

Abstract

Sequence Analysis of Chromosome Translocations in Neurodevelopmental Disorders

By Brooke Carin Weckselblatt

Translocation is one of the most common structural chromosome abnormalities observed in humans. Constitutional unbalanced translocations result in partial monosomy and partial trisomy of many genes, which may lead to neurodevelopmental disorders. We analyzed the breakpoints of 57 unique unbalanced translocations to investigate the mechanisms of how they form. 51 are simple unbalanced translocations between two different chromosome ends, and six rearrangements have more than three breakpoints involving two to five chromosomes. Sequencing 37 breakpoint junctions revealed that simple translocations have between zero and four basepairs (bp) of microhomology (n=26), short inserted sequences (n=8), or paralogous repeats (n=3) at the junctions, indicating that translocations do not arise primarily from non-allelic homologous recombination, but instead form most often via non-homologous end joining or microhomology-mediated break-induced replication. Three complex translocations have inversions, insertions, and multiple breakpoint junctions between only two chromosomes. Whole-genome sequencing and fluorescence in situ hybridization analysis of two *de novo* translocations revealed at least 18 and 33 breakpoints involving five different chromosomes. Breakpoint sequencing of one maternally inherited translocation involving four chromosomes uncovered multiple breakpoints with inversions and insertions. All of these breakpoint junctions had zero to four bp of microhomology consistent with chromothripsis, and both *de novo* events occurred on paternal alleles. Together with other studies, these data suggest that constitutional chromothripsis arises in the paternal genome, and may be transmitted maternally. In addition, we analyzed genes at the breakpoints of these interchromosomal translocations and at the breakpoints of intrachromosomal duplication CNVs. Three simple translocations fuse genes that are predicted to produce in-frame transcripts, and we predicted six in-frame fusion genes at sequenced duplication breakpoints; four gene fusions were formed by tandem duplications, one by two interconnected duplications, and one by duplication inserted at another locus. These unique fusion genes could be related to clinical phenotypes and warrant further study. Breakpoint sequencing of our large collection of chromosome rearrangements provides a comprehensive analysis of the molecular mechanisms behind translocation formation.

Sequence Analysis of Chromosome Translocations in Neurodevelopmental Disorders

By

Brooke Carin Weckselblatt
B.A., Bryn Mawr College, 2010

Advisor: M. Katharine Rudd, PhD, FACMG

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Sciences,
Genetics and Molecular Biology
2015

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Portions of this chapter have been published in *Trends in Genetics* (doi: 10.1016/j.tig.2015.05.010) as a review article and reformatted for this document.

Genomic structural variation (SV) refers to abnormalities in chromosome structure. The first human chromosome rearrangements were observed down the microscope in cells from tumors (neoplastic) or blood (constitutional). SVs are important for human health because they contribute to genetic diversity and evolution (Redon et al. 2006; Conrad et al. 2010; Kidd et al. 2010; Mills et al. 2011) but also can drive disease (Stankiewicz and Lupski 2010; Cooper et al. 2011; Coe et al. 2014). Approximately 15-20% of those with intellectual disability and autism spectrum disorders have a clinically relevant SV (Miller et al. 2010; Cooper et al. 2011; Kaminsky et al. 2011). Exome sequencing studies estimate that single nucleotide variation (SNV) is responsible for another ~25% of neurodevelopmental disorders (Lee et al. 2014; Yang et al. 2014). Constitutional SV arises in premeiotic, meiotic, or post-zygotic cells, and in most cases the timing of SV formation is not known.

Analysis of SV identifies genes related to disease, breakage hotspots, parent-of-origin biases, and common mutational mechanisms. Together, these data point to risk factors for SV formation and critical genes responsible for genetic syndromes. DNA sequence at SV breakpoint junctions reveals signatures of diverse DNA repair mechanisms that shape human chromosome rearrangements. Sequencing breakpoints also has the potential to uncover more complex genomic structures that are missed by low-resolution methods. Recently, whole genome sequencing (WGS) studies of pathogenic SV have revealed many genomic breakpoints in complex rearrangements that arise from one catastrophic event. This introductory chapter discusses the mechanisms and consequences of simple and complex constitutional SV.

Methods for SV detection

Standard SV detection methods include chromosome banding, fluorescence *in situ* hybridization (FISH), and array comparative genome hybridization (CGH). For SV detection by chromosome banding, chromosomes are prepared from dividing cells, stained, and viewed with a microscope. Large deletions, duplications, and translocations are detected if the banding pattern or chromosome structure is altered. However, smaller microdeletions and microduplications are not observed at this low-resolution.

A microarray-based method such as array CGH will detect copy number differences between abnormal and reference genomes. Though an array won't determine copy number variation (CNV) location and SV organization, FISH may resolve the location of chromosomal segments originally identified by microarray and/or any next-generation sequencing (NGS). To do so, fluorescently labeled DNA probes hybridize to metaphase or interphase cells to visualize a locus on a chromosome and determine copy number. FISH can also detect copy-neutral SV like inversions and balanced translocations that do not result in changes in copy number.

Recently, targeted NGS and WGS technology has been applied to detect CNV and copy-neutral SV using sequence read-depth and paired reads that span breakpoints. The most comprehensive SV analysis comes from sequencing the whole genome, but complex genomic structures identified by WGS may require FISH or chromosome banding to place rearranged segments. Filtering discordant sequence reads to identify paired ends that span SV breakpoints is one successful strategy to pinpoint breakpoints; however, it is difficult to uniquely map short reads to repetitive DNA. One way to capture breakpoints in repeats is by creating large-insert "jumping" libraries that

sequence the ends of DNA fragments several kilobases (kb) long (Korbel et al. 2007; Hanscom and Talkowski 2014). These large-fragment mate-pair libraries increase the likelihood of detecting SV that has breakpoints within interspersed repeats. This is especially useful for inversions and balanced translocations that are not detected by copy number methods.

DNA repair mechanisms

At a sequenced SV junction, long stretches of homologous sequence shared between breakpoints indicate that the rearrangement may be a product of non-allelic homologous recombination (NAHR) (Figure 1.1). NAHR is the recombination between regions with high sequence similarity but different genomic positions. Genomic regions susceptible to NAHR include segmental duplications (SDs), long interspersed nuclear elements (LINEs), and human endogenous retroviral elements (HERVs). SDs, also known as low-copy repeats, are genomic segments that are at least 1 kb in length and share greater than 90% sequence identity. LINEs are retrotransposons interspersed throughout the genome that are ~6 kb when full-length. Derived from ancient retroviruses, HERV sequences are flanked by long terminal repeats.

On the other hand, the absence of sequence homology points to repair by nonhomologous end joining (NHEJ), where broken DNA ends ligate together without a homologous template following a double-strand break (DSB) (Figure 1.1) (Lupski 1998; Lieber 2010). The presence of inserted or inverted sequences at breakpoints suggests that the error occurred during DNA replication, like in microhomology-mediated break-induced replication (MMBIR) or in fork stalling and template switching (FoSTeS)

(Figure 1.1). During MMBIR, **a** broken DNA strand at a collapsed replication fork uses microhomology to invade a nearby replication fork (Hastings et al. 2009). In the FoSTeS model, at a stalled replication fork, the lagging strand disengages and invades a nearby replication fork, then reinitiates DNA synthesis (Zhang et al. 2009b).

Simple intrachromosomal SV

Simple intrachromosomal deletions, duplications, and inversions involve only one chromosome and are the product of one or two DSBs (Figure 1.2). Deletions and duplications are easily detected by array-based methods that measure differences in copy number between subject and reference genomes. These CNVs may also be detected by measuring NGS read depth, since relative to the rest of the genome, a region with half of the coverage is inferred to be a deletion, and a region with ~50% more read depth is inferred to be a duplication (Chiang et al. 2009; Abyzov et al. 2011; Haraksingh et al. 2011).

Because inversions are copy-neutral, they escape detection by microarray and read depth methods. Recent use of mate-pair and fosmid/BAC end sequencing enabled the identification of hundreds of inversion polymorphisms in the human genome (Tuzun et al. 2005; Kidd et al. 2008; Williams et al. 2012; Rasekh et al. 2015). Though most inversions are not associated with an abnormal phenotype, some alter the orientation of repetitive DNA in a way that predisposes the chromosome to rearrangement in the future. Recurrent deletions and duplications of Chromosomes 5q35, 8p23.1, 16p12.1, and 17q21.31 occur via NAHR and only arise in parents with an inversion of these chromosomes. Thus, inversion carriers have an increased risk for offspring with genomic

disorders (Giglio et al. 2001; Antonacci et al. 2009; Antonacci et al. 2010; Watson et al. 2014).

Deletions can lie either within a chromosome arm (interstitial) or truncate the end of a chromosome (terminal) (Figure 1.2B) (Luo et al. 2011). Terminal deletions have been described on almost every human chromosome end, and in some cases these CNVs result in a recognizable genomic disorder. For example, Wolf-Hirshhorn (Battaglia et al. 1999), Cri-du-chat (Zhang et al. 2005), Kleefstra (Kleefstra et al. 2009), Jacobsen (Grossfeld et al. 2004), and Phelan-McDermid (Durand et al. 2007) syndromes are caused by terminal deletions of Chromosomes 4p, 5p, 9q, 11q, and 22q, respectively. Sequence analysis of terminal deletions revealed guanine-rich motifs overrepresented at breakpoints. This suggests that either G-rich sequences are risk factors for chromosome breakage, or that once a DSB occurs, G-rich DNA is an ideal substrate for *de novo* telomere synthesis and terminal deletion formation (Luo et al. 2011; Bose et al. 2014).

Interstitial deletions and duplications may be caused by NAHR, NHEJ, or MMBIR. The genomic organization of interstitial deletions is relatively simple, and haploinsufficiency for genes within the deleted segment can lead to abnormal outcomes. The phenotypic significance of interstitial duplications is more difficult to interpret since genes at breakpoints may or may not be disrupted depending on the orientation of the duplicated segment. Sequence analysis of a diverse collection of interstitial duplications revealed that they are almost always tandem, in direct orientation relative to the original locus (Figure 1.2B) (Newman et al. 2015).

Most deletion and duplication CNVs have non-recurrent breakpoints, with blunt ends or microhomology at breakpoint junctions (Vissers et al. 2009; Luo et al. 2011;

Verdin et al. 2013). Although this microhomology may seem coincidental, many CNV sequencing studies have revealed greater microhomology than expected by chance (Vissers et al. 2009; Conrad et al. 2010; Newman et al. 2015). Recurrent deletions and duplications make up ~20% of pathogenic intrachromosomal rearrangements (Luo et al. 2011; Itsara et al. 2012). Genomic disorders caused by these recurrent CNVs are ideal for genotype-phenotype correlations because the same contiguous genes are deleted or duplicated in unrelated individuals (Watson et al. 2014). The earliest recurrent deletions and duplications discovered turned out to be mediated by NAHR between SDs hundreds of kb in length on the same chromosome (Lupski 1998). More recent studies have used genomic approaches to predict intrachromosomal CNVs mediated by long (>10 kb) SDs with high sequence identity (>95%) (Sharp et al. 2006; Liu et al. 2012; Dittwald et al. 2013). NAHR frequency is positively correlated with SD length, proximity, and sequence identity, so the most common CNVs are flanked by long stretches of near-perfect homology (Liu et al. 2011b).

Shorter paralogous repeats can also mediate NAHR, albeit less frequently than long SDs. Sequencing across interspersed repeats is challenging, and until recently, many of these breakpoints were missed by CNV sequencing studies. This year, recombination between LINE pairs was discovered at the breakpoints of 44 pathogenic CNVs. High sequence identity appears to be a requirement for LINE-LINE rearrangements because the minimum identity between recombining LINEs was 96%, and most pairs were greater than 97% identical (Startek et al. 2015). Some LINEs had less than 1 kb of homology, suggesting that even fragmented LINEs can participate in NAHR. Recombination between human endogenous retrovirus (HERV) elements can also give rise to recurrent

CNVs. Deletions and duplications mediated by HERV-HERV recombination at three intrachromosomal loci were sequenced in a recent study (Campbell et al. 2014). Like other HERV-mediated chromosome rearrangements (Hermetz et al. 2012; Robberecht et al. 2013; Weckselblatt et al. 2015), all of the CNVs are flanked by HERV-H elements that are at least 3 kb long and 93-96% identical. The longer length and significant sequence identity of intact HERV-H elements may make them particularly recombinogenic.

On the other hand, *Alu*s, the most abundant class of repeats in the human genome, are only ~300 bp long. *Alu* pairs that flank deletions and duplications are 75-91% identical (Vissers et al. 2009; Shlien et al. 2010; Luo et al. 2011; Verdin et al. 2013; Boone et al. 2014; Carvalho et al. 2014; Newman et al. 2015). Sequencing breakpoints of 54 CNVs at the *Alu*-rich *SPAST* locus revealed 38 that spanned hybrid *Alu*s (Boone et al. 2014). Lower sequence identity between *Alu* pairs suggests that these CNVs may not be the product of NAHR, but rather are the result of homeologous, or near-homologous, recombination that occurs between more divergent sequences (Rossetti et al. 2004). Compared to deletions and duplications mediated by LINE-LINE and HERV-HERV events, (30 kb-5.5 megabases (Mb); median 523 kb) those flanked by *Alu*s tend to be smaller (1.9 kb-4.2 Mb; median 65.4 kb) (Vissers et al. 2009; Shlien et al. 2010; Luo et al. 2011; Verdin et al. 2013; Boone et al. 2014; Carvalho et al. 2014; Newman et al. 2015).

Simple interchromosomal SV

Translocation is the exchange of genomic material between two different chromosomes (Figure 1.2). The initial event that gives rise to translocations is usually reciprocal, producing two derivative chromosomes that are balanced. However, derivative translocation chromosomes may segregate in a balanced or an unbalanced manner. Balanced translocations are copy-neutral and do not cause a phenotype unless they disrupt developmentally important gene(s) at breakpoints. On the other hand, unbalanced translocations result in trisomy and monosomy of chromosome ends and are usually found in individuals with developmental delay, intellectual disability, and/or birth defects, depending on the genes affected by the CNVs. Unbalanced translocations are easily detected by a number of methods, whereas detecting balanced translocations requires techniques that capture breakpoints, such as WGS or targeted NGS.

Like intrachromosomal rearrangements, most constitutional translocations are non-recurrent. In studies of sequenced balanced translocations, microhomology is the most common feature at breakpoint junctions (Chen et al. 2008; Higgins et al. 2008; Chiang et al. 2012). A recent study of nine unbalanced translocations revealed that six were mediated by NAHR between 6-kb LINE, 3-kb HERV, or 1.7-kb SD pairs that are each >90 identical (Robberecht et al. 2013). Although in this group NAHR between paralogous repeats appeared to be the "driver" of unbalanced translocations, our larger-scale study determined that NAHR is unlikely to be the major mechanism of translocation formation (see Chapter 2 of this dissertation). In both unbalanced translocation studies, LINE and HERV elements were capable of NAHR, whereas no *Alu-Alu* events were detected (Robberecht et al. 2013; Weckselblatt et al. 2015). Indeed, *Alu-Alu* recombination has been reported in only three translocations (Rouyer et al. 1987;

Higgins et al. 2008; Fruhmesser et al. 2013). This trend suggests that, for NAHR-mediated rearrangements, those that are interchromosomal may require longer stretches of homology and greater sequence identity than those that are intrachromosomal.

Recurrent translocations are caused by NAHR between homologous sequences on different chromosomes, or by breakage hotspots in palindromic AT-rich repeats (PATRRs). The same SDs responsible for reciprocal deletions and duplications of the short arm of Chromosome 8 also underlie recurrent translocations between Chromosomes 4, 8, and 12 (Giglio et al. 2002; Ou et al. 2011; Goldlust et al. 2013). A recurrent translocation between Chromosomes 4 and 18 is also caused by NAHR between 92% identical HERV-H repeats (Hermetz et al. 2012). PATRRs on Chromosomes 3, 8, 11, 17, and 22 give rise to recurrent translocations, the most well known of which is the der(22)t(11:22), which causes Emanuel syndrome (Kato et al. 2012).

Complex chromosome rearrangement

Complex chromosome rearrangements have three or more breakpoints and may lead to a balanced or an unbalanced copy number state (Zhang et al. 2009a; Quinlan and Hall 2012). Recent NGS breakpoint studies have paved the way to understanding the mutational mechanisms and defining the genomic structure of these rearrangements. Here we describe insights into the major classes of complex chromosome rearrangement.

Inverted duplication adjacent to terminal deletion

Inverted duplication next to terminal deletion is a common type of rearrangement that has been recognized in cancer and constitutional genomes (Tanaka et al. 2007;

Hermetz et al. 2014). Several models have been put forth to explain these CNVs, and all include a dicentric chromosome that goes through a breakage-fusion-bridge cycle. Analysis of 34 sequenced breakpoints revealed spacers with normal copy number (median size 3 kb) between the inverted duplications (Figure 1.3A) and short inverted homology at the edges of the inverted segments. These molecular features support a model whereby the initial DSB leads to a terminal deletion, followed by fold-back of the truncated chromosome, formation of a dicentric chromosome, and a second DSB between the two centromeres that is repaired by addition of a new telomere (Hermetz et al. 2014). The disomic spacers between inverted duplications correspond to the fold-back portion of the chromosome and their discovery provided important insight in the formation of these complex chromosome rearrangements. Spacers are too small to detect by array-based methods, so sequencing breakpoint junctions was a major advance in understanding this rearrangement mechanism.

Inverted duplications adjacent to deletions have also been described in ring chromosomes (Murmann et al. 2009), an interstitial chromosome rearrangement (Milosevic et al. 2014), and unbalanced translocations (Hermetz et al. 2014). These rearrangements are also are formed through a dicentric chromosome step, but instead of resolving as a terminal deletion, the second DSB is repaired by an internal site on the same chromosome or capture of a nonhomologous chromosome (Figure 1.3E).

Duplication-normal-duplication (DUP-NML-DUP)

Adjacent duplications with a normal copy number region between them have a characteristic “DUP-NML-DUP” pattern by array CGH (Figure 1.3B). Sequencing DUP-

NML-DUP junctions revealed that most are interconnected with duplications in direct or inverted orientation (Liu et al. 2011a; Carvalho et al. 2013; Brand et al. 2014; Newman et al. 2015). These interstitial duplications are derived from regions of the same chromosome arm that are hundreds of kb to Mb apart. DUP-NML-DUPs are not associated with a particular syndrome since they are derived from diverse genomic loci and involve different genes. Depending on the spacing of probes, some DUP-NML-DUPs may appear as a single duplication by array CGH, so their prevalence is likely underestimated. DUP-NML-DUPs have the potential to duplicate, fuse, and/or disrupt genes at breakpoints; therefore, determining their genomic structure is essential to identify genes involved in disease. For example, in Chapter 4 we describe a DUP-NML-DUP that fuses the *KCNH5* and *FUT8* genes at an inverted junction that is predicted to produce an in-frame fusion transcript (Newman et al. 2015).

Triplication

Triplications are often recognized by array or NGS as segments with increased copy number within a duplicated segment. Type I triplications are oriented head-to-tail, without flanking duplications, and are formed via NAHR between SDs (Liu et al. 2012) (Figure 1.3C). Type II triplications lie within larger duplications and may or may not involve SDs at breakpoints (Figure 1.3C). In most Type II CNVs, the triplicated segment is inverted relative to the duplications, a structure known as DUP-TRP/INV-DUP (Figure 1.4) (Carvalho et al. 2011; Shimojima et al. 2012; Ishmukhametova et al. 2013; Beck et al. 2015; Carvalho et al. 2015).

DUP-TRP/INV-DUP of the *PLP1* gene on Chromosome X causes Pelizaeus-Merzbacher disease, and these complex triplications lead to a more severe clinical phenotype than *PLP1* duplications that also cause the disease (Beck et al. 2015). Triplication breakpoints cluster at inverted SDs distal of *PLP1* and sequence analysis of 17 *PLP1* DUP-TRP/INV-DUPs revealed that a recurrent breakpoint junction lies within these inverted repeats (Beck et al. 2015). Such DUP-TRP/INV-DUPs are proposed to form via a two-step process involving replication fork collapse and strand invasion between inverted repeats, followed by MMBIR or NHEJ (Figure 1.4) (Carvalho et al. 2013; Beck et al. 2015). DUP-TRP/INV-DUPs of *MECP2* also have recurrent breakpoints within inverted repeats and cause a more severe form of *MECP2* duplication syndrome (Carvalho et al. 2011).

Triplications in the same orientation as flanking duplications have been described at other loci (Newman et al. 2015). Whereas inverted triplications tend to have inverted repeats at junctions, direct triplications lack inverted repeats (Carvalho et al. 2013; Beck et al. 2015; Newman et al. 2015). Recently, terminal regions of absence of heterozygosity were detected distal of some triplications. Extended absence of heterozygosity adjacent to triplications is likely due to MMBIR template switching between homologous chromosomes, which leads to regional uniparental disomy at end of the chromosome (Carvalho et al. 2015).

Insertional translocation

Like other complex chromosome rearrangements, insertional translocations have more than two breakpoints. As opposed to more common translocations of chromosome

ends, these translocated segments are inserted interstitially into a nonhomologous chromosome (Figure 1.3D). Insertional translocations often appear to be simple interstitial duplications by copy number studies; however, FISH and breakpoint analyses revealed that ~2% of genomic gains detected by array CGH are inserted in another chromosome (Kang et al. 2010; Neill et al. 2011; Nowakowska et al. 2012; Newman et al. 2015). Unbalanced insertional translocations may be inherited from parents with the balanced form of the rearrangement (Kang et al. 2010; Nowakowska et al. 2012), and in some cases, the insertional translocation includes multiple segments in direct or inverted orientation (Chiang et al. 2012; Newman et al. 2015; Weckselblatt et al. 2015).

Chromoanagenesis

The most severe forms of genomic reorganization are described as "chromoanagenesis," or chromosome rebirth, since chromosomes are rearranged beyond recognition (Holland and Cleveland 2012). Chromosome shattering, "chromothripsis (Stephens et al. 2011)," and chromosome reconstitution, "chromoanasynthesis (Liu et al. 2011a)," are two types of chromoanagenesis, and their underlying mechanisms are just beginning to be understood.

Chromothripsis

Chromothripsis was originally detected in chronic lymphocytic leukemia where dozens of breakpoints were clustered on a single chromosome arm (Stephens et al. 2011). Chromothripsis is present in ~2% of cancer genomes (Kim et al. 2013) and has been reported at similar frequencies in constitutional chromosome rearrangements.

Chromothripsis involving up to five different chromosomes has been described in children with neurodevelopmental disorders (Figure 1.5A). Long stretches of homology are absent from the breakpoint junctions, so DNA repair likely occurs via NHEJ (Kloosterman et al. 2011; Chiang et al. 2012; Kloosterman et al. 2012; Genesio et al. 2013; Kloosterman and Cuppen 2013; Macera et al. 2014; Nazaryan et al. 2014; van Heesch et al. 2014; Weckselblatt et al. 2015). Despite tens of breakpoints per genome, constitutional chromothripsis is largely copy neutral. Retention of essentially normal copy number in chromothripsis genomes could be mechanistically important, or could simply reflect selective pressure in liveborn individuals (Kloosterman and Cuppen 2013). Some breakpoints have adjacent deletions, and many are inverted, but duplications are rare (Kloosterman et al. 2011; Kloosterman et al. 2012; Weckselblatt et al. 2015). WGS is ideal to capture tens of breakpoints in one experiment, including balanced translocations and inversions in chromothripsis genomes that go unnoticed by other methods (Kloosterman et al. 2011; Chiang et al. 2012; Kloosterman et al. 2012; van Heesch et al. 2014; Weckselblatt et al. 2015). However, visualization of chromosomes is still necessary to localize rearranged segments and determine the contiguous structure of chromosomes scrambled by chromothripsis (Macera et al. 2014; Nazaryan et al. 2014; Weckselblatt et al. 2015). Breakpoint analysis of a growing number of complex rearrangements has revealed that translocations involving three or more different chromosomes are likely formed via chromothripsis (Kloosterman et al. 2011; Kloosterman et al. 2012; Weckselblatt et al. 2015). In Chapter 3, we describe chromothripsis translocations involving five different chromosomes and investigate their origin.

In addition to cancer and constitutional situations, chromothripsis has been observed upon integration of a transgene (Chiang et al. 2012) and in a hematopoietic stem cell lineage (McDermott et al. 2015). Somatic chromothripsis was recently described in a woman with WHIM syndrome, a rare immunodeficiency disorder resulting from a mutated copy of the *CXCR4* gene. In this case, chromothriptic deletion of her dominant *CXCR4* mutation led to reversion of the disease (McDermott et al. 2015).

Chromoanasynthesis

Chromosome reconstitution confined to a single chromosome or locus has been termed chromoanasynthesis (Liu et al. 2011a). Whereas chromothripsis is limited to two copy number states, has features of NHEJ at breakpoints, and may involve multiple chromosomes, chromoanasynthesis leads to deletions, duplications, and triplications along a single chromosome (Figure 1.5B) (Liu et al. 2011a; Plaisancie et al. 2014; Zanardo et al. 2014). Constitutional chromoanasynthesis has been recognized in rearrangements that involve eight to 33 breakpoints (Liu et al. 2011a; Plaisancie et al. 2014), and sequenced junctions bear signatures of FoSTeS and MMBIR (Liu et al. 2011a). Going forward, as WGS is more widely applied to SV, we expect to better define the features and origins of these highly complex chromosome rearrangements.

Consequences of SV

Fusion genes

In many cases, SV breakpoints intersect open reading frames of genes. Though the transcriptional consequences of most SVs have not been investigated, breakpoints that disrupt or fuse genes have the potential to wreak havoc on normal development. Fusion genes are predicted at the breakpoints of constitutional deletions (Rippey et al. 2013; Boone et al. 2014), duplications (Rippey et al. 2013; Newman et al. 2015), balanced translocations (Backx et al. 2011; Utami et al. 2014), unbalanced translocations (Weckselblatt et al. 2015), an insertional translocation (Newman et al. 2015), an inverted DUP-NML-DUP (Newman et al. 2015), and chromothriptic rearrangements (van Heesch et al. 2014). Genes disrupted or fused at the breakpoints of balanced rearrangements are excellent candidates for neurodevelopmental disorders because the rest of the genome is intact. However, fusion genes in unbalanced rearrangements also have the potential to acquire new functions related to phenotypic outcomes.

Mutations adjacent to breakpoint junctions

Although DNA at breakpoints is known to be altered by resection, insertion, and inversion, recent studies suggest that regions further from junctions are also mutated. Complex duplications of the *MECP2* locus have SNV within 50 bp of breakpoint junctions that arose at the same time as the *de novo* duplications (Carvalho et al. 2013). Similar "micro-mutations" have been detected adjacent to pathogenic deletions of five different chromosomes (Wang et al. 2015). It remains to be determined whether these mutations occur at other SV, but this phenomenon may be similar to the mutations induced by APOBEC cytosine deaminase associated with somatic mutations in cancer (Roberts et al. 2013).

Position effect

SVs may also exert position effects that alter the expression of intact genes near breakpoints. Position effects have been noted at the *FOXL2* (Fantes et al. 1995; Beysen et al. 2005; Bhatia et al. 2013), *PLP1* (Lee et al. 2006), *SHOX* (Fukami et al. 2006), and *SOX9* (Hill-Harfe et al. 2005; Velagaleti et al. 2005) genes, among others. In the recurrent translocation between Chromosomes 11 and 22, aberrant nuclear positioning of translocated regions results in differential expression of many genes on different chromosomes (Harewood et al. 2010). Future studies of *cis* and *trans* position effects related to SV may inform phenotypes when even thorough breakpoint analysis by WGS fails to pinpoint genes related to disease (Gilissen et al. 2014).

Research Objectives

Translocations are among the earliest recognized chromosome rearrangements, and both the balanced and unbalanced forms of translocations are routinely detected in individuals with neurodevelopmental or other congenital disorders. Balanced translocation breakpoints have been studied to determine rearrangement mechanisms and for candidate gene discovery, but unbalanced translocations have not been well characterized. Although unbalanced translocations carry monosomic and trisomic regions that complicate genotype-phenotype correlation, they offer the same opportunity to study mutational mechanism because both balanced and unbalanced translocations arise from the same initial events. To this end, we have analyzed constitutional unbalanced translocations from 57 subjects with neurodevelopmental disorders using a combination

of array CGH, targeted NGS, and WGS. We aim to establish the mechanisms, structures, and consequences of this particular class of SV. In Chapter 2, we describe sequenced simple unbalanced translocation junctions to infer how they formed. For several rearrangements, WGS revealed extreme structural complexity indicative of chromothripsis, as explained in Chapter 3. In addition, we performed breakpoint junction sequencing of interchromosomal translocations and intrachromosomal duplications to predict the formation of fusion genes at their SV junctions (Chapter 4). Finally, Chapter 5 discusses the impact of this work in the context of the chromosome rearrangement field and provides future directions to further explore the causes and effects of translocations.

References

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**(6): 974-984.

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M et al. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* **42**(9): 745-750.

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE. 2009. Characterization of six human disease-associated inversion polymorphisms. *Hum Mol Genet* **18**(14): 2555-2566.

Backx L, Seuntjens E, Devriendt K, Vermeesch J, Van Esch H. 2011. A balanced translocation t(6;14)(q25.3;q13.2) leading to reciprocal fusion transcripts in a patient with intellectual disability and agenesis of corpus callosum. *Cytogenet Genome Res* **132**(3): 135-143.

Battaglia A, Carey JC, Cederholm P, Viskochil DH, Brothman AR, Galasso C. 1999. Natural history of Wolf-Hirschhorn syndrome: experience with 15 cases. *Pediatrics* **103**(4 Pt 1): 830-836.

Beck CR, Carvalho CMB, Banser L, Gambin T, Stubbolo D, Yuan B, Sperle K, McCahan SM, Henneke M, Seeman P et al. 2015. Complex Genomic Rearrangements at the PLP1 Locus Include Triplication and Quadruplication. *PLoS Genet* **3**: e1005050.

Beysen D, Raes J, Leroy BP, Lucassen A, Yates JR, Clayton-Smith J, Ilyina H, Brooks SS, Christin-Maitre S, Fellous M et al. 2005. Deletions involving long-range conserved nongenic sequences upstream and downstream of FOXL2 as a novel disease-causing mechanism in blepharophimosis syndrome. *Am J Hum Genet* **77**(2): 205-218.

Bhatia S, Bengani H, Fish M, Brown A, Divizia MT, de Marco R, Damante G, Grainger R, van Heyningen V, Kleinjan DA. 2013. Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia. *Am J Hum Genet* **93**(6): 1126-1134.

Boone PM, Yuan B, Campbell IM, Scull JC, Withers MA, Baggett BC, Beck CR, Shaw CJ, Stankiewicz P, Moretti P et al. 2014. The Alu-Rich Genomic Architecture of SPAST Predisposes to Diverse and Functionally Distinct Disease-Associated CNV Alleles. *Am J Hum Genet*.

Bose P, Hermetz KE, Conneely KN, Rudd MK. 2014. Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. *PLoS One* **9**(7): e101607.

Brand H, Pillalamarri V, Collins RL, Eggert S, O'Dushlaine C, Braaten EB, Stone MR, Chambert K, Doty ND, Hanscom C et al. 2014. Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. *Am J Hum Genet* **95**(4): 454-461.

Campbell IM, Gambin T, Dittwald P, Beck CR, Shuvarikov A, Hixson P, Patel A, Gambin A, Shaw CA, Rosenfeld JA et al. 2014. Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination. *BMC Biol* **12**: 74.

Carvalho CM, Pehlivan D, Ramocki MB, Fang P, Alleva B, Franco LM, Belmont JW, Hastings PJ, Lupski JR. 2013. Replicative mechanisms for CNV formation are error prone. *Nat Genet* **45**(11): 1319-1326.

Carvalho CM, Pfundt R, King DA, Lindsay SJ, Zuccherato LW, Macville MVE, Liu P, Johnson D, Stankiewicz P, Brown CW et al. 2015. Absence of Heterozygosity due to Template Switching during Replicative Rearrangements. *Am J Hum Genet*: 1-10.

Carvalho CM, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, McCall A, Pivnick EK, Hines-Dowell S, Seaver LH et al. 2011. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet* **43**(11): 1074-1081.

Carvalho CM, Vasanth S, Shinawi M, Russell C, Ramocki MB, Brown CW, Graakjaer J, Skytte AB, Vianna-Morgante AM, Krepischi AC et al. 2014. Dosage changes of a segment at 17p13.1 lead to intellectual disability and microcephaly as a result of complex genetic interaction of multiple genes. *Am J Hum Genet* **95**(5): 565-578.

Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G et al. 2008. Mapping translocation breakpoints by next-generation sequencing. *Genome Res* **18**(7): 1143-1149.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR et al. 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet* **44**(4): 390-397, S391.

Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**(1): 99-103.

Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LE et al. 2014. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46**(10): 1063-1071.

Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, Turner DJ, Hurles ME. 2010. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet* **42**(5): 385-391.

Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V et al. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**(9): 838-846.

Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, Rodriguez Rojas LX, Elton LE, Scott DA, Schaaf CP et al. 2013. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res* **23**(9): 1395-1409.

Durand CM, Betancur C, Boeckers TM, Bockmann J, Chaste P, Fauchereau F, Nygren G, Rastam M, Gillberg IC, Anckarsater H et al. 2007. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nat Genet* **39**(1): 25-27.

Fantes J, Redeker B, Breen M, Boyle S, Brown J, Fletcher J, Jones S, Bickmore W, Fukushima Y, Mannens M et al. 1995. Aniridia-associated cytogenetic

rearrangements suggest that a position effect may cause the mutant phenotype. *Hum Mol Genet* **4**(3): 415-422.

Fruhmesser A, Blake J, Haberlandt E, Baying B, Raeder B, Runz H, Spreiz A, Fauth C, Benes V, Utermann G et al. 2013. Disruption of EXOC6B in a patient with developmental delay, epilepsy, and a de novo balanced t(2;8) translocation. *Eur J Hum Genet* **21**(10): 1177-1180.

Fukami M, Kato F, Tajima T, Yokoya S, Ogata T. 2006. Transactivation function of an approximately 800-bp evolutionarily conserved sequence at the SHOX 3' region: implication for the downstream enhancer. *Am J Hum Genet* **78**(1): 167-170.

Genesio R, Ronga V, Castelluccio P, Fioretti G, Mormile A, Leone G, Conti A, Cavaliere ML, Nitsch L. 2013. Pure 16q21q22.1 deletion in a complex rearrangement possibly caused by a chromothripsis event. *Mol Cytogenet* **6**(1): 29.

Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R et al. 2001. Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am J Hum Genet* **68**(4): 874-883.

Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, Giorda R, Ragusa A, Guerneri S, Selicorni A, Stumm M et al. 2002. Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am J Hum Genet* **71**(2): 276-285.

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A et al. 2014. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**(7509): 344-347.

Goldlust IS, Hermetz KE, Catalano LM, Barfield RT, Cozad R, Wynn G, Ozdemir AC, Conneely KN, Mulle JG, Dharamrup S et al. 2013. Mouse model implicates GNB3 duplication in a childhood obesity syndrome. *Proc Natl Acad Sci U S A* **110**(37): 14990-14994.

Grossfeld PD, Mattina T, Lai Z, Favier R, Jones KL, Cotter F, Jones C. 2004. The 11q terminal deletion disorder: a prospective study of 110 cases. *Am J Med Genet A* **129A**(1): 51-61.

Hanscom C, Talkowski M. 2014. Design of large-insert jumping libraries for structural variant detection using illumina sequencing. *Curr Protoc Hum Genet* **80**: 7 22 21-29.

Haraksingh RR, Abyzov A, Gerstein M, Urban AE, Snyder M. 2011. Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. *PLoS One* **6**(11): e27859.

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**(5): 554-564.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**(8): 551-564.

Hermetz KE, Newman S, Conneely KN, Martin CL, Ballif BC, Shaffer LG, Cody JD, Rudd MK. 2014. Large inverted duplications in the human genome form via a fold-back mechanism. *PLoS Genet* **10**(1): e1004139.

Hermetz KE, Surti U, Cody JD, Rudd MK. 2012. A recurrent translocation is mediated by homologous recombination between HERV-H elements. *Mol Cytogenet* **5**(1): 6.

Higgins AW, Alkuraya FS, Bosco AF, Brown KK, Bruns GA, Donovan DJ, Eisenman R, Fan Y, Farra CG, Ferguson HL et al. 2008. Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. *Am J Hum Genet* **82**(3): 712-722.

Hill-Harfe KL, Kaplan L, Stalker HJ, Zori RT, Pop R, Scherer G, Wallace MR. 2005. Fine mapping of chromosome 17 translocation breakpoints > or = 900 Kb upstream of SOX9 in acampomelic campomelic dysplasia and a mild, familial skeletal dysplasia. *Am J Hum Genet* **76**(4): 663-671.

Holland AJ, Cleveland DW. 2012. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat Med* **18**(11): 1630-1638.

Ishmukhametova A, Chen JM, Bernard R, de Massy B, Baudat F, Boyer A, Mechin D, Thorel D, Chabrol B, Vincent MC et al. 2013. Dissecting the structure and mechanism of a complex duplication-triplication rearrangement in the DMD gene. *Hum Mutat* **34**(8): 1080-1084.

Itsara A, Vissers LE, Steinberg KM, Meyer KJ, Zody MC, Koolen DA, de Ligt J, Cuppen E, Baker C, Lee C et al. 2012. Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. *Am J Hum Genet* **90**(4): 599-613.

Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, Kunig D, Moreno-De-Luca D, Moreno-De-Luca A, Mulle JG, Warren ST et al. 2011. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet Med* **13**(9): 777-784.

Kang SH, Shaw C, Ou Z, Eng PA, Cooper ML, Pursley AN, Sahoo T, Bacino CA, Chinault AC, Stankiewicz P et al. 2010. Insertional translocation detected using FISH confirmation of array-comparative genomic hybridization (aCGH) results. *Am J Med Genet A* **152A**(5): 1111-1126.

Kato T, Kurahashi H, Emanuel BS. 2012. Chromosomal translocations and palindromic AT-rich repeats. *Curr Opin Genet Dev* **22**(3): 221-228.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**(7191): 56-64.

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**(5): 837-847.

Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, Park PJ. 2013. Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res* **23**(2): 217-227.

Kleefstra T, van Zelst-Stams WA, Nillesen WM, Cormier-Daire V, Houge G, Foulds N, van Dooren M, Willemsen MH, Pfundt R, Turner A et al. 2009. Further clinical and molecular delineation of the 9q subtelomeric deletion syndrome supports a major contribution of EHMT1 haploinsufficiency to the core phenotype. *J Med Genet* **46**(9): 598-606.

Kloosterman WP, Cuppen E. 2013. Chromothripsis in congenital disorders and cancer: similarities and differences. *Curr Opin Cell Biol* **25**(3): 341-348.

Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M et al. 2011. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* **20**(10): 1916-1924.

Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, van Binsbergen E, Renkens I, Duran K, Ballarati L, Vergult S, Giardino D, Hansson K et al. 2012. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep* **1**(6): 648-655.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**(5849): 420-426.

Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, Das K, Toy T, Harry B, Yourshaw M et al. 2014. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312**(18): 1880-1887.

Lee JA, Madrid RE, Sperle K, Ritterson CM, Hobson GM, Garbern J, Lupski JR, Inoue K. 2006. Spastic paraplegia type 2 associated with axonal neuropathy and apparent PLP1 position effect. *Ann Neurol* **59**(2): 398-403.

Lieber MR. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem* **79**: 181-211.

Liu P, Carvalho CM, Hastings PJ, Lupski JR. 2012. Mechanisms for recurrent and complex human genomic rearrangements. *Curr Opin Genet Dev* **22**(3): 211-220.

Liu P, Erez A, Nagamani SC, Dhar SU, Kolodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA et al. 2011a. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146**(6): 889-903.

Liu P, Lacaria M, Zhang F, Withers M, Hastings PJ, Lupski JR. 2011b. Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. *Am J Hum Genet* **89**(4): 580-588.

Luo Y, Hermetz KE, Jackson JM, Mulle JG, Dodd A, Tsuchiya KD, Ballif BC, Shaffer LG, Cody JD, Ledbetter DH et al. 2011. Diverse mutational mechanisms cause pathogenic subtelomeric rearrangements. *Hum Mol Genet* **20**(19): 3769-3778.

Lupski JR. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**(10): 417-422.

Macera MJ, Sobrino A, Levy B, Jobanputra V, Aggarwal V, Mills A, Esteves C, Hanscom C, Pereira S, Pillalamarri V et al. 2014. Prenatal diagnosis of chromothripsis, with nine breaks characterized by karyotyping, FISH, microarray and whole-genome sequencing. *Prenat Diagn* **35**(3): 299-301.

McDermott DH, Gao JL, Liu Q, Siwicki M, Martens C, Jacobs P, Velez D, Yim E, Bryke CR, Hsu N et al. 2015. Chromothriptic cure of WHIM syndrome. *Cell* **160**(4): 686-699.

Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ et al. 2010. Consensus statement: chromosomal

microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* **86**(5): 749-764.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**(7332): 59-65.

Milosevic J, El Khattabi L, Roubergue A, Coussement A, Doummar D, Cuisset L, Le Tessier D, Flageul B, Viot G, Lebbar A et al. 2014. Inverted duplication with deletion: first interstitial case suggesting a novel undescribed mechanism of formation. *Am J Med Genet A* **164A**(12): 3180-3186.

Murmann AE, Conrad DF, Mashek H, Curtis CA, Nicolae RI, Ober C, Schwartz S. 2009. Inverted duplications on acentric markers: mechanism of formation. *Hum Mol Genet* **18**(12): 2241-2256.

Nazaryan L, Stefanou EG, Hansen C, Kosyakova N, Bak M, Sharkey FH, Mantziou T, Papanastasiou AD, Velissariou V, Liehr T et al. 2014. The strength of combined cytogenetic and mate-pair sequencing techniques illustrated by a germline chromothripsis rearrangement involving FOXP2. *Eur J Hum Genet* **22**(3): 338-343.

Neill NJ, Ballif BC, Lamb AN, Parikh S, Ravnan JB, Schultz RA, Torchia BS, Rosenfeld JA, Shaffer LG. 2011. Recurrence, submicroscopic complexity, and potential clinical relevance of copy gains detected by array CGH that are shown to be unbalanced insertions by FISH. *Genome Res* **21**(4): 535-544.

Newman S, Hermetz KE, Weckselblatt B, Rudd MK. 2015. Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints. *Am J Hum Genet* **96**(2): 208-220.

Nowakowska BA, de Leeuw N, Ruivenkamp CA, Sikkema-Raddatz B, Crolla JA, Thoelen R, Koopmans M, den Hollander N, van Haeringen A, van der Kevie-Kersemaekers AM et al. 2012. Parental insertional balanced translocations are an important cause of apparently de novo CNVs in patients with developmental anomalies. *Eur J Hum Genet* **20**(2): 166-170.

Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, Szafranski P, Cooper ML, Rao M, Shao L et al. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res* **21**(1): 33-46.

Plaisancie J, Kleinfinger P, Cances C, Bazin A, Julia S, Trost D, Lohmann L, Vigouroux A. 2014. Constitutional chromoanasynthesis: description of a rare chromosomal event in a patient. *Eur J Med Genet* **57**(10): 567-570.

Quinlan AR, Hall IM. 2012. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet* **28**(1): 43-53.

Rasekh ME, Chiatante G, Miroballo M, Tang J, Ventura M, Amemiya CT, Eichler EE, Antonacci F, Alkan C. 2015. Discovery of large genomic inversions using pooled clone sequencing. In *bioRxiv*.

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**(7118): 444-454.

Rippey C, Walsh T, Gulsuner S, Brodsky M, Nord AS, Gasperini M, Pierce S, Spurrell C, Coe BP, Krumm N et al. 2013. Formation of chimeric genes by copy-number

variation as a mutational mechanism in schizophrenia. *Am J Hum Genet* **93**(4): 697-710.

Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, Vermeesch JR. 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res* **23**(3): 411-418.

Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G et al. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**(9): 970-976.

Rossetti LC, Goodeve A, Larripa IB, De Brasi CD. 2004. Homeologous recombination between AluSx-sequences as a cause of hemophilia. *Hum Mutat* **24**(5): 440.

Rouyer F, Simmler MC, Page DC, Weissenbach J. 1987. A sex chromosome rearrangement in a human XX male caused by Alu-Alu recombination. *Cell* **51**(3): 417-425.

Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC et al. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* **38**(9): 1038-1042.

Shimojima K, Mano T, Kashiwagi M, Tanabe T, Sugawara M, Okamoto N, Arai H, Yamamoto T. 2012. Pelizaeus-Merzbacher disease caused by a duplication-inverted triplication-duplication in chromosomal segments including the PLP1 region. *Eur J Med Genet* **55**(6-7): 400-403.

Shlien A, Baskin B, Achatz MI, Stavropoulos DJ, Nichols KE, Hudgins L, Morel CF, Adam MP, Zhukova N, Rotin L et al. 2010. A common molecular mechanism underlies two phenotypically distinct 17p13.1 microdeletion syndromes. *Am J Hum Genet* **87**(5): 631-642.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**: 437-455.

Startek M, Szafranski P, Gambin T, Campbell IM, Hixson P, Shaw CA, Stankiewicz P, Gambin A. 2015. Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. *Nucleic Acids Res* **43**(4): 2188-2198.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**(1): 27-40.

Tanaka H, Cao Y, Bergstrom DA, Kooperberg C, Tapscott SJ, Yao MC. 2007. Intrastrand annealing leads to the formation of a large DNA palindrome and determines the boundaries of genomic amplification in human cancer. *Mol Cell Biol* **27**(6): 1993-2002.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D et al. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**(7): 727-732.

Utami KH, Hillmer AM, Aksoy I, Chew EG, Teo AS, Zhang Z, Lee CW, Chen PJ, Seng CC, Ariyaratne PN et al. 2014. Detection of chromosomal breakpoints in patients with developmental delay and speech disorders. *PLoS One* **9**(6): e90852.

van Heesch S, Simonis M, van Roosmalen MJ, Pillalamarri V, Brand H, Kuijk EW, de Luca KL, Lansu N, Braat AK, Menelaou A et al. 2014. Genomic and functional

overlap between somatic and germline chromosomal rearrangements. *Cell Rep* **9**(6): 2001-2010.

Velagaleti GV, Bien-Willner GA, Northup JK, Lockhart LH, Hawkins JC, Jalal SM, Withers M, Lupski JR, Stankiewicz P. 2005. Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. *Am J Hum Genet* **76**(4): 652-662.

Verdin H, D'Haene B, Beysen D, Novikova Y, Menten B, Sante T, Lapunzina P, Nevado J, Carvalho CM, Lupski JR et al. 2013. Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. *PLoS Genet* **9**(3): e1003358.

Vissers LE, Bhatt SS, Janssen IM, Xia Z, Lalani SR, Pfundt R, Derwinska K, de Vries BB, Gilissen C, Hoischen A et al. 2009. Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. *Hum Mol Genet* **18**(19): 3579-3593.

Wang Y, Su P, Hu B, Zhu W, Li Q, Yuan P, Li J, Guan X, Li F, Jing X et al. 2015. Characterization of 26 deletion CNVs reveals the frequent occurrence of micro-mutations within the breakpoint-flanking regions and frequent repair of double-strand breaks by templated insertions derived from remote genomic regions. *Hum Genet*.

Watson CT, Marques-Bonet T, Sharp AJ, Mefford HC. 2014. The genetics of microdeletion and microduplication syndromes: an update. *Annu Rev Genomics Hum Genet* **15**: 215-244.

Weckselblatt B, Hermetz KE, Rudd MK. 2015. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res* **25**(7): 937-947.

Williams LJ, Tabbaa DG, Li N, Berlin AM, Shea TP, Maccallum I, Lawrence MS, Drier Y, Getz G, Young SK et al. 2012. Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res* **22**(11): 2241-2249.

Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C et al. 2014. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**(18): 1870-1879.

Zanardo EA, Piazzon FB, Dutra RL, Dias AT, Montenegro MM, Novo-Filho GM, Costa TV, Nascimento AM, Kim CA, Kulikowski LD. 2014. Complex structural rearrangement features suggesting chromoanagenesis mechanism in a case of 1p36 deletion syndrome. *Mol Genet Genomics* **289**(6): 1037-1043.

Zhang F, Carvalho CM, Lupski JR. 2009a. Complex human chromosomal and genomic rearrangements. *Trends Genet* **25**(7): 298-307.

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. 2009b. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* **41**(7): 849-853.

Zhang X, Snijders A, Segraves R, Zhang X, Niebuhr A, Albertson D, Yang H, Gray J, Niebuhr E, Bolund L et al. 2005. High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *Am J Hum Genet* **76**(2): 312-326.

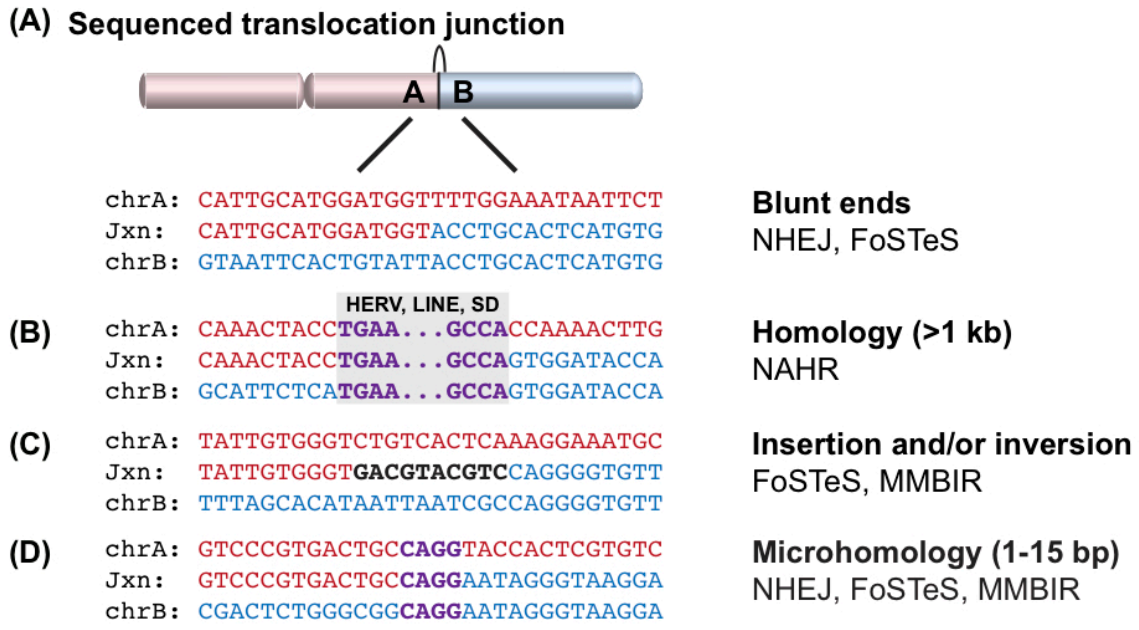


Figure 1.1: Signatures of mutational mechanisms

DNA sequence that spans the translocation breakpoint junction is aligned to the pink reference chromosome A (chrA) and blue reference chromosome B (chrB). The breakpoints are located where the junction (Jxn) sequence transitions from chrA to chrB.

(A) Jxn with blunt ends at the breakpoints points to repair by NHEJ or FoSTeS.

(B) Homology, shown in purple, >1 kb long and shared between chrA and chrB breakpoints suggests NAHR between paralogous HERVs, LINEs, or SDs.

(C) The presence of inverted and/or inserted sequence (shown in black), at the breakpoints are signatures of replicative mechanisms like FoSTeS and MMBIR.

(D) 1-15 bp of microhomology between chromosome breakpoints is common and may be due to NHEJ, FoSTeS, or MMBIR.

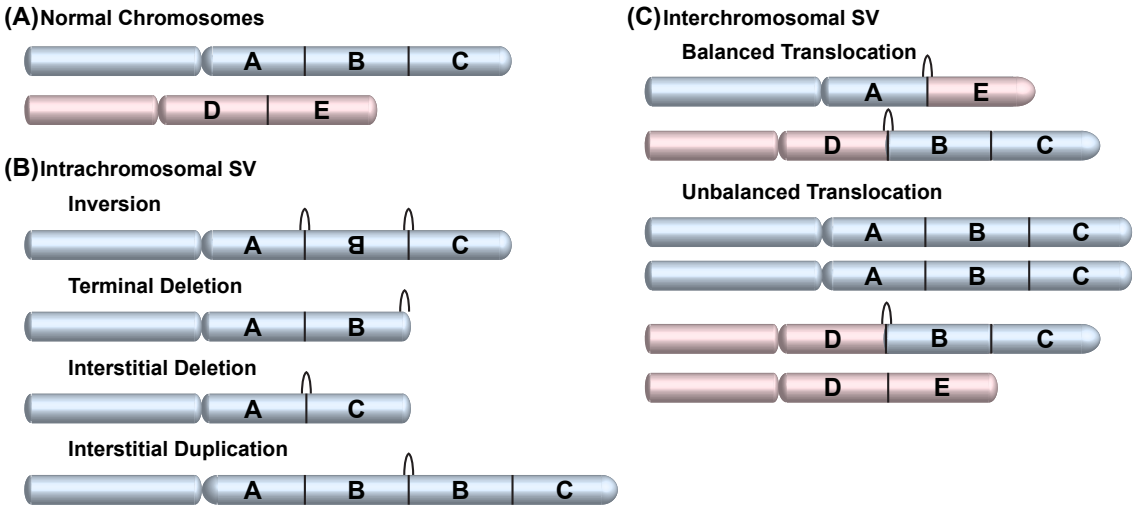


Figure 1.2: Simple chromosome rearrangements

(A) Two nonhomologous chromosomes shown in blue and pink. Segments are labeled with letters A-E. Black arches indicate SV breakpoint junctions.

(B) Intrachromosomal rearrangements include inversions, interstitial and terminal deletions, and interstitial duplications.

(C) Simple translocations between two different chromosome ends. Balanced translocations do not result in CNV, but unbalanced translocations have partial monosomy (segment E) and partial trisomy (segments B-C).

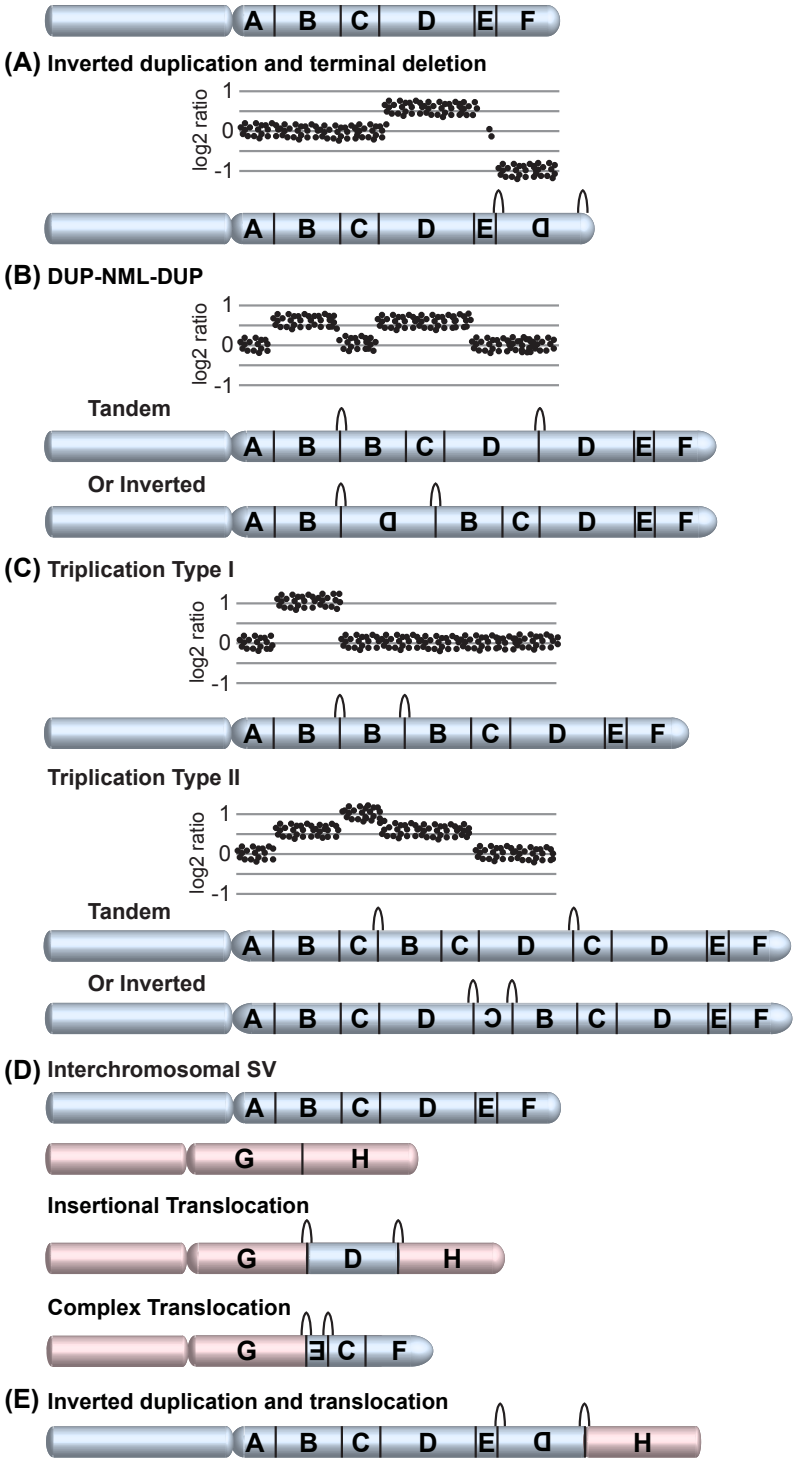

A B C D E F
(A) Inverted duplication and terminal deletion
log2 ratio
1
0
-1
A B C D E D
(B) DUP-NML-DUP
Tandem
A B B C D D E F
Or Inverted
A B D B C D E F
(C) Triplication Type I
A B B B C D E F
Triplication Type II
Tandem
A B C B C D C D E F
Or Inverted
A B C D C B C D E F
(D) Interchromosomal SV
G H
Insertional Translocation
G D H
Complex Translocation
G E C F
(E) Inverted duplication and translocation
A B C D E D H

Figure 1.3: Complex chromosome rearrangements

Complex rearrangements and their array CGH signatures are shown relative to the blue reference chromosome (top) divided into segments A-F.

(A) Inverted duplications adjacent to terminal deletions have a short disomic spacer region (segment E) between inverted duplications.

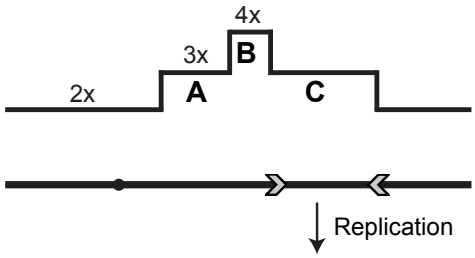
(B) A DUP-NML-DUP appears by array CGH as two copy number gains (segments B and D). The duplications may be in direct orientation, or one duplicated segment (D) may be inverted between two copies of the other (B).

(C) Triplication Type I has three direct copies of B. In Triplication Type II, the triplication (C) is embedded within a duplicated region (B-D). The triplicated segment may be in direct or inverted orientation.

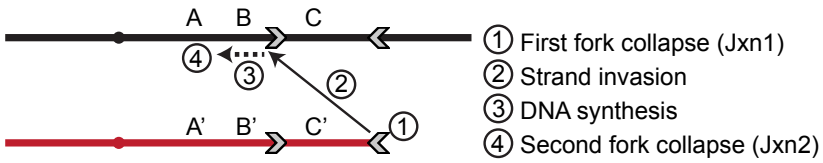
(D) Complex interchromosomal rearrangements occur between the blue and pink chromosomes. An insertional translocation involves the interstitial insertion of one chromosome segment (D) into another chromosome. Some complex translocations have multiple chromosome segments and/or inversion at the breakpoint junction.

(E) An inverted duplication with terminal deletion may end with the translocated end of a nonhomologous chromosome.

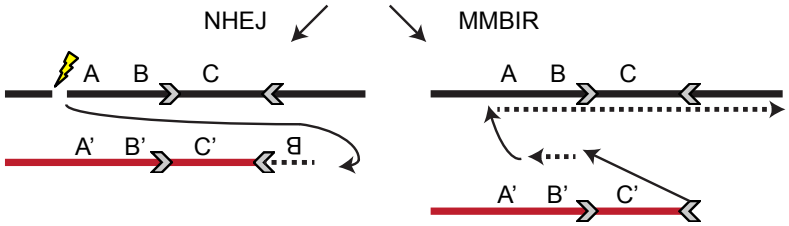
(A) Mechanism of DUP-TRP/INV-DUP



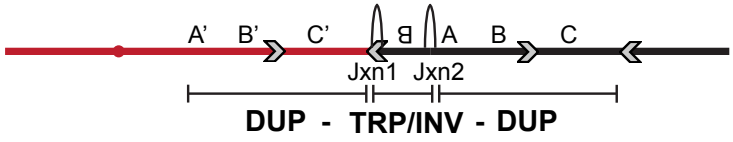
(B)



(C)



(D)



DUP - TRP/INV - DUP

Figure 1.4: DUP-TRP/INV-DUP formation

(A) Copy number changes are detected relative to the black reference chromosome. 2x indicates normal disomic copy number, while 3x genomic copies of A and C are duplications and 4x total copies of segment B is a triplication. Inverted repeats (grey arrows) are present at the edges of segment C.

(B) At a collapsed replication fork, sequence homology drives strand invasion from one inverted repeat into one from the opposite strand. DNA synthesis is re-initiated until the occurrence of a second collapsed replication fork.

(C) This second junction may arise from an NHEJ or MMBIR mechanism. In NHEJ, a DSB occurs on the original DNA strand and is repaired by joining the to end of the replicated strand. In MMBIR, the lagging strand disengages, invades upstream sequence, and synthesizes DNA along the rest of the chromosome.

(D) The resulting structure is a duplication, inverted triplication, and duplication. Orientation of the triplicated "B" is confirmed by sequencing across Jxn1 and Jxn2.

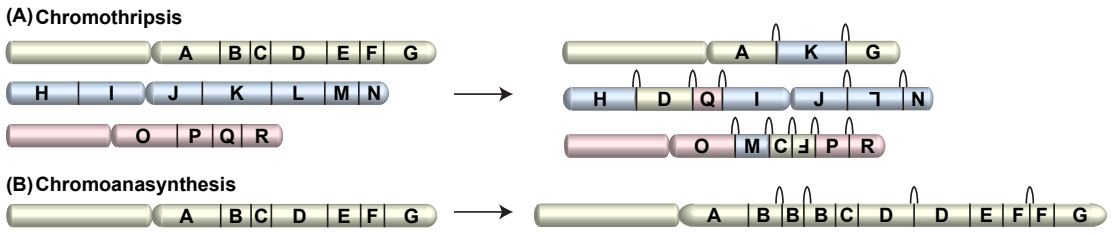


Figure 1.5: Massive genomic reorganization

(A) Chromothripsis shatters three nonhomologous chromosomes. The only CNVs are deletions of B and D, but translocating segments and inversions have shuffled the contents of the three chromosomes. The 12 breakpoint junctions have blunt ends or short microhomology.

(B) Chromoanagenesis leads to triplication (B) and duplications (D and F) across one chromosome. These breakpoint junctions contain microhomology and insertions that suggest a DNA replication-based mechanism of repair.

Chapter 2

Unbalanced translocations arise from diverse mutational mechanisms

Portions of this chapter have in published in *Genome Research* (2015, 25:937-947) as a research article and have been reformatted for this document.

Introduction

Translocation is one of the most common structural chromosome abnormalities found in humans, with a *de novo* frequency of 1 in 2,000 (Warburton 1991). Unbalanced translocations lead to monosomy and trisomy for segments of different chromosomes and account for ~1% of cases of developmental delay and intellectual disability (Ravnan et al. 2006; Ballif et al. 2007; Shao et al. 2008). The initial exchange of genetic material between two non-homologous chromosomes can occur during premeiotic mitoses, meiotic recombination in the parental germline, or postzygotic mitoses in the early embryo (Vanneste et al. 2009; Robberecht et al. 2013). Unbalanced translocations detected in affected children may be inherited from a parent who carries the balanced form of the rearrangement or may arise *de novo*.

Recurrent translocations may be mediated by NAHR between segmental duplications (Giglio et al. 2002; Ou et al. 2011) or paralogous interspersed repeats (Luo et al. 2011; Hermetz et al. 2012; Robberecht et al. 2013). Palindromic AT-rich repeats on Chromosomes 3, 8, 11, 17, and 22 also generate recurrent translocations, the most common of which is the recurrent t(11;22) that causes Emanuel syndrome (Edelmann et al. 2001; Kurahashi et al. 2003; Gotter et al. 2007; Kato et al. 2012; Kato et al. 2014). Most constitutional translocations, however, are not recurrent, and sequencing of translocation breakpoints has revealed features of NHEJ and MMBIR at more than 60 unique translocation junctions (Chen et al. 2008; Higgins et al. 2008; Sobreira et al. 2011; Chiang et al. 2012; Robberecht et al. 2013). Recently, a study of 12 *de novo* unbalanced translocations, nine of which were sequenced, concluded that NAHR between paralogous

repeats is the predominant mechanism of *de novo* unbalanced translocation formation (Robberecht et al. 2013).

Here we investigate rearrangement mechanisms of 57 constitutional unbalanced translocations isolated from subjects with neurodevelopmental phenotypes. In this group. 51 translocations are a simple rearrangement between two different chromosome ends, and this chapter is focused on their formation. Using a combination of array CGH, SureSelect sequence capture, and WGS, we provide a comprehensive sequence analysis of unbalanced translocations.

Results

Sequencing unbalanced translocation junctions

We recruited subjects with developmental delays, autism, intellectual disability (ID), and/or congenital anomalies after routine cytogenetics testing at Emory Genetics Laboratory (EGL). For 57 unrelated individuals with a previous diagnosis of an unbalanced translocation we extracted DNA from peripheral blood for further study. In this cohort, translocation breakpoints are spread across all of the autosomes and the X Chromosome. From the 57 subjects, 51 carry rearrangements that are simple unbalanced translocations with one derivative chromosome that fuses two chromosome breakpoints; six rearrangements have more than one breakpoint junction that joins multiple segments from two or more chromosomes (see Chapter 3).

To fine-map breakpoints, we designed custom oligonucleotide microarrays with dense probe coverage in 1-megabase (Mb) windows centered around the breakpoints determined by diagnostic chromosomal microarray analysis (CMA) (Figure 2.1). High-

density arrays resolve unbalanced translocation breakpoints to 200-1000 bp, but do not detect copy-neutral structural variation. Next, we attempted SureSelect Target Enrichment to capture 40-kilobase (kb) regions surrounding 44 fine-mapped translocations (40 simple and four complex). Since none of the breakpoints were shared between individuals, we pooled genomic DNA from five to seven subjects per SureSelect library and separated subject-specific junctions after NGS using Illumina HiSeq. We sequenced 100-bp paired-end reads and analyzed discordant reads where paired-ends map to different chromosomes, map too close together, or map too far apart relative to the GRCh37/hg19 reference genome (Figure 2.2).

Discordant reads spanned 19 of 40 simple translocations targeted by SureSelect and Illumina HiSeq. To confirm NGS results, we PCR-amplified translocation junctions predicted by discordant reads and Sanger sequenced amplicons. We confirmed 18/19 of simple translocations supported by discordant reads. One translocation junction that failed PCR confirmation (EGL313) was supported by discordant reads between unique sequence and a segmental duplication. For the 21/40 simple translocations where SureSelect plus Illumina HiSeq did not yield discordant reads, we attempted long-range PCR using breakpoint estimates from high-resolution array CGH and successfully sequenced 12. We PCR-amplified and sequenced an additional seven simple translocations without attempting SureSelect, leading to a total of 37 simple translocation junctions confirmed by Sanger sequencing.

Simple unbalanced translocations

We confirmed the junctions of 37 simple unbalanced translocations by Sanger sequencing (Table 2.1) (Weckselblatt et al. 2015). Six junctions had blunt ends and 20 junctions had one to four basepairs (bp) of microhomology shared between the two sides of the translocation. Eight translocations had short insertions or inversions at the breakpoint junction, ranging in length from 2-209 bp. In four translocations the inserted sequence is a copy of adjacent sequence, indicating DNA slippage (Viguera et al. 2001). Like other DNA replication-based rearrangements (Lee et al. 2007; Zhang et al. 2009; Conrad et al. 2010; Luo et al. 2011; Newman et al. 2015), two of these local duplications are in an inverted orientation relative to the reference genome, and two are in direct orientation. Insertions in LM219, EGL366, and EGL087 map to regions 210 bp, 1.5 kb, and 56 kb from the breakpoint, respectively. The origin of EGL089's 7-bp insertion is unknown.

Three translocations have at least 335 bp of perfect homology shared between the two sides of the junction, consistent with NAHR. EGL051's translocation occurs between segmental duplications on Chromosomes 5 and 14 that are 95% identical over 1.5 kb. In EGL080, the translocation breakpoint spans a L1PA2 on Chromosome 8 and a L1PA3 on Chromosome 1 that are 93% identical across the 6.0-kb repeats. EGL083's junction lies in HERV-H elements on Chromosomes 8 and 12 that are 92% identical across the 3.2-kb and 3.0-kb repeats. In each of these translocations, recombination occurred at paralogous sites within repeats and created a hybrid repeat element at the breakpoint junction. Breakpoints in LM219's unbalanced translocation fall in *Alu*Sx and *Alu*Sx1 repeats; however, the junction does not lie in homologous parts of the *Alu*s.

Discussion

Unbalanced translocation mechanisms

We analyzed translocations from 57 individuals with unique chromosome rearrangements and found that most junctions have little or no sequence homology. For the 37 simple unbalanced translocations we sequenced, 70% have 0-4 bp of microhomology, 22% have insertions or inversions, and only 8% have long stretches of homology shared between translocating segments, suggesting that NHEJ and MMBIR are the predominant mechanisms of translocation formation (Hastings et al. 2009; Zhang et al. 2009). Recently, Robberecht et al. sequenced the junctions of nine *de novo* unbalanced translocations and found that six were mediated by NAHR between LINEs, HERVs, or segmental duplications (Robberecht et al. 2013). They concluded that NAHR between these longer repeats drives *de novo* unbalanced translocation formation. We determined translocation inheritance in 20 trios and found that eight were *de novo*, seven were maternally inherited, and five were paternally inherited (Table 2.1). Similar to the 30% observed by Robberecht et al., 40% of our unbalanced translocations were *de novo*; however, only two out of eight *de novo* unbalanced translocations in our study were mediated by NAHR. As in Robberecht et al., these two junctions lie in homologous LINE or HERV repeats. Nonetheless, most *de novo* translocations in our study lack extensive sequence homology at junctions. Like other structural variation in the human genome (Conrad et al. 2010; Luo et al. 2011; Chiang et al. 2012; Newman et al. 2015), most *de novo* unbalanced translocations are the product of NHEJ or MMBIR.

It is possible at least some of the 14 simple translocations that failed junction sequencing have repetitive DNA or cryptic complexity at the breakpoints that prevented

SureSelect, NGS, or junction PCR. Even if all 14 translocations were the product of NAHR, junctions without significant sequence homology still outnumber those formed by NAHR. Translocations in EGL045 and EGL315 may be NAHR-mediated since breakpoints determined by high-resolution array CGH map to homologous repeats (HERV-H and L1PA2/L1PA3, respectively) (Table 2.1). However, breakpoints of the remaining 12 translocations map to regions that lack homology between both sides of the junction. Furthermore, breakpoints that fine-map to homologous interspersed repeats are not guaranteed to be the product of NAHR. For example, array CGH mapped both breakpoints in EGL103's translocation to *Alu*Sx1 repeats, but sequencing revealed that breakpoints were outside of the repeats and the junction lacked significant sequence homology.

Forty-eight percent (49/102) of sequenced breakpoints from simple translocations lie within repeats (Table 2.1). This is not surprising since approximately half of the human genome is repetitive (Lander et al. 2001), and similar repeat content has been reported at other CNV breakpoints (Vissers et al. 2009; Bose et al. 2014). Translocation junctions of EGL051, EGL080, and EGL083 are located in paralogous segmental duplications, L1s, and HERV-H elements, respectively. Robberecht et al. found the same classes of repeats at breakpoint junctions of unbalanced translocations. These repeats are more than 1-kb long, are found only in primates, and are greater than 92% identical. While recombination between *Alu*s have been described for numerous interstitial deletions and duplications (Luo et al. 2011; Boone et al. 2014; Newman et al. 2015), *Alu-Alu* events rarely mediate germline translocations (Rouyer et al. 1987; Chen et al. 2008; Luo et al. 2011; Chiang et al. 2012; Fruhmesser et al. 2013; Robberecht et al. 2013).

These data suggest that specific types of repeats may be favored in aberrant homologous recombination that gives rise to translocations.

We identified two breakpoints shared between our translocations and those described in Robberecht et al. Translocations in EGL083 and Robberecht Case 3 are mediated by NAHR and have a breakpoint on Chromosome 12 in the same HERV-H (hg19; Chr 12:4,128,160-4,131,129). However, the translocation partners are different chromosomes. Recombination between HERV-H repeats has been implicated in other translocations and deletions (Hermetz et al. 2012; Shuvarikov et al. 2013; Campbell et al. 2014). Robberecht Case 7 has an unbalanced translocation likely mediated by NAHR between L1PA4 elements on Chromosomes 9 and 10. EGL319's translocation has a breakpoint in the same Chromosome 9 L1PA4 (hg19; Chr 9:15,595,148-15,601,275), although the translocation partner is different and the junction has microhomology rather than features of NAHR. It is possible that this L1PA4 is a breakage hotspot that may be resolved by diverse DNA repair mechanisms.

Translocation annotation and technical limitations

Mapping translocation breakpoints at the nucleotide level required a tiered approach consisting of high-resolution array CGH, targeted sequence capture with NGS, WGS, and confirmation by junction PCR followed by Sanger sequencing. We successfully confirmed the breakpoints of 37/51 simple unbalanced translocations. Fourteen translocation junctions could not be verified by the above methods, and this is due to a combination of technical limitations, lack of genomic DNA, and the nature of the rearrangements.

FISH analysis revealed that copy number gains in EGL354, EGL357, and EGL358 were unbalanced translocations to the short arms of Chromosomes X, 21, and 22, respectively (Table 2.1). However, we did not detect genomic losses of those chromosome arms by array CGH. This is consistent with small deletions of ends of the derivative chromosomes that may lie in segmental duplications or other repetitive DNA not included in microarray analysis (Rudd 2012). Though we targeted the breakpoints corresponding to the terminal gains of these unbalanced translocations, SureSelect plus Illumina HiSeq did not identify reads that cross the translocation breakpoints.

We fine-mapped 14 breakpoint regions from 12 translocations to LINEs and attempted to capture these loci by SureSelect. Discordant reads spanned the junction from LINE to unique sequence in only five of these breakpoints (EGL002, EGL064, EGL306, EGL317, and EGL319), which is consistent with the previously recognized limitation in LINE breakpoint sequencing (Talkowski et al. 2011). Surprisingly, our SureSelect approach was successful in mapping informative reads to three segmental duplications. Discordant reads and Sanger sequencing supported EGL051's junction between two 95%-identical segmental duplications. EGL313's junction was supported by discordant reads that anchor the segmental duplication at the breakpoint to unique sequence; however, we were not able to confirm this junction by Sanger sequencing. EGL062's breakpoint failed SureSelect, but we sequenced this junction from segmental duplication to unique sequence by long-range PCR. In this large-scale analysis of unbalanced translocations, we report a paucity of sequence homology at breakpoint junctions and conclude that NAHR is unlikely to be the primary driver of this type of rearrangement.

This comprehensive analysis revealed that most unbalanced translocations are simple, and likely formed by NHEJ and MMBIR repair processes.

Methods

Custom array CGH

This study was approved by the Institutional Review Board (IRB) at Emory University. Subjects had CMA testing with a version of the EmArray oligonucleotide array (Baldwin et al. 2008), followed by confirmation by chromosome banding or FISH. G-banding of chromosomes from peripheral blood has a resolution of 550-700 bands and FISH was performed as described (Baldwin et al. 2008). For most subjects, DNA extracted from whole blood was used for all microarray and breakpoint sequencing experiments. We used DNA from lymphoblastoid cell lines for EGL316, EGL382, and LM219. To fine-map unbalanced translocation breakpoints, we performed high-resolution array CGH. We designed custom 4 x 180K oligonucleotide arrays with ~200-bp probe spacing using eArray from Agilent Technologies (Santa Clara, CA; https://earray.chem.agilent.com/earray/). The array design ID (AMADID) identifiers are 018181, 021634, 021635, 021636, 021637, 034386, 037387, 035709, 035730, 037646, 040718, and 063584. Subject DNA was co-hybridized with reference DNA from either GM10851 or GM15510. Arrays were scanned using the Agilent high-resolution C scanner (Agilent Technologies, Santa Clara, CA), and signal intensities were evaluated using Feature Extraction Version 9.5.1.1 software (Agilent Technologies, Santa Clara, CA). We used Agilent Genomic Workbench 6.0 software (Agilent Technologies, Santa Clara, CA) to analyze the array data and call breakpoints.

Sequencing unbalanced translocations

We used Agilent SureSelect Target Enrichment to pull down 40-kb regions around breakpoints fine-mapped by custom array CGH. SureSelect followed by Illumina HiSeq sequencing was performed at Hudson Alpha Genomic Services Lab. After NGS, we aligned 100-bp paired-end reads from fastq files to the GRC37/hg19 reference genome using Burrows-Wheeler Alignment (BWA) tool 0.5.9 and identified misaligned pairs using the SAMTools 0.1.18 filter function. Paired-end reads that aligned to the reference genome too far apart, too close together, in the wrong orientation/genome order, or to different chromosomes were clustered to predict structural variation,

We performed long-range PCR and Sanger sequencing to confirm breakpoints. We used the Qiagen LongRange PCR Kit (Catalog # 206403), following the manufacturer's protocol. Sanger sequencing was performed by Beckman Coulter Genomics (Danvers, MA), and the reads were aligned to the human genome reference assembly (GRC37/hg19) using the BLAT tool on the UCSC Genome Browser (http://genome.ucsc.edu/).

Whole-genome sequencing

WGS of genomic DNA from EGL382 was performed by Complete Genomics (Mountain View, CA) as described (Drmanac et al. 2010). Complete Genomics provided the individual reads, quality scores, and initial mappings to the GRCh37 reference genome in .tsv format. To identify discordant read pairs, we converted Reads and Mappings flagged as structural variant candidates to SAM format with the map2sam

command in CGATools 1.7.1 (http://cgatools.sourceforge.net/). We used SAMTools (Li et al. 2009) to sort, index, and convert files to BAM. To account for intra-read gaps, we used a custom Perl script that extracts discordant read pairs that map aberrantly relative to the reference genome. We viewed discordant reads with Integrative Genomics Viewer (Robinson et al. 2011) to identify and interpret structural variation.

Data access

Agilent array CGH data have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE68019. Breakpoint junction sequences have been submitted to GenBank (http://www.ncbi.nlm.nih.gov/genbank/) under accession numbers KR072894 - KR072971. Illumina sequencing data have been submitted to the NCBI Sequence Read Archive (SRA;http://www.ncbi.nlm.nih.gov/sra) under SRP057518 and Complete Genomics whole genome sequencing data have been submitted to the database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap) under accession number phs000845.v1.p1.

References

Baldwin EL, Lee JY, Blake DM, Bunke BP, Alexander CR, Kogan AL, Ledbetter DH, Martin CL. 2008. Enhanced detection of clinically relevant genomic imbalances using a targeted plus whole genome oligonucleotide microarray. *Genet Med* **10**(6): 415-429.

Ballif BC, Sulpizio SG, Lloyd RM, Minier SL, Theisen A, Bejjani BA, Shaffer LG. 2007. The clinical utility of enhanced subtelomeric coverage in array CGH. *Am J Med Genet A* **143A**(16): 1850-1857.

Boone PM, Yuan B, Campbell IM, Scull JC, Withers MA, Baggett BC, Beck CR, Shaw CJ, Stankiewicz P, Moretti P et al. 2014. The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *Am J Hum Genet* **95**(2): 143-161.

Bose P, Hermetz KE, Conneely KN, Rudd MK. 2014. Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. *PLoS One* **9**(7): e101607.

Campbell IM, Gambin T, Dittwald P, Beck CR, Shuvarikov A, Hixson P, Patel A, Gambin A, Shaw CA, Rosenfeld JA et al. 2014. Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination. *BMC Biol* **12**: 74.

Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G et al. 2008. Mapping translocation breakpoints by next-generation sequencing. *Genome Res* **18**(7): 1143-1149.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR et al. 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet* **44**(4): 390-397, S391.

Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, Turner DJ, Hurles ME. 2010. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat Genet* **42**(5): 385-391.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**(5961): 78-81.

Edelmann L, Spiteri E, Koren K, Pulijaal V, Bialer MG, Shanske A, Goldberg R, Morrow BE. 2001. AT-rich palindromes mediate the constitutional t(11;22) translocation. *Am J Hum Genet* **68**(1): 1-13.

Fruhmesser A, Erdel M, Duba HC, Fauth C, Amberger A, Utermann G, Zschocke J, Kotzot D. 2013. Combined Dup(7)(q22.1q32.2), Inv(7)(q31.31q31.33), and Ins(7;19)(q22.1;p13.2p13.2) in a 12-year-old boy with developmental delay and various dysmorphism. *Eur J Med Genet* **56**(7): 383-388.

Giglio S, Calvari V, Gregato G, Gimelli G, Camanini S, Giorda R, Ragusa A, Guerneri S, Selicorni A, Stumm M et al. 2002. Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am J Hum Genet* **71**(2): 276-285.

Gotter AL, Nimmakayalu MA, Jalali GR, Hacker AM, Vorstman J, Conforto Duffy D, Medne L, Emanuel BS. 2007. A palindrome-driven complex rearrangement of 22q11.2 and 8q24.1 elucidated using novel technologies. *Genome Res* **17**(4): 470-481.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**(8): 551-564.

Hermetz KE, Surti U, Cody JD, Rudd MK. 2012. A recurrent translocation is mediated by homologous recombination between HERV-H elements. *Mol Cytogenet* **5**(1): 6.

Higgins AW, Alkuraya FS, Bosco AF, Brown KK, Bruns GA, Donovan DJ, Eisenman R, Fan Y, Farra CG, Ferguson HL et al. 2008. Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. *Am J Hum Genet* **82**(3): 712-722.

Kato T, Franconi CP, Sheridan MB, Hacker AM, Inagakai H, Glover TW, Arlt MF, Drabkin HA, Gemmill RM, Kurahashi H et al. 2014. Analysis of the t(3;8) of hereditary renal cell carcinoma: a palindrome-mediated translocation. *Cancer Genet* **207**(4): 133-140.

Kato T, Kurahashi H, Emanuel BS. 2012. Chromosomal translocations and palindromic AT-rich repeats. *Curr Opin Genet Dev* **22**(3): 221-228.

Kurahashi H, Shaikh T, Takata M, Toda T, Emanuel BS. 2003. The constitutional t(17;22): another translocation mediated by palindromic AT-rich repeats. *Am J Hum Genet* **72**(3): 733-738.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**(7): 1235-1247.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Luo Y, Hermetz KE, Jackson JM, Mulle JG, Dodd A, Tsuchiya KD, Ballif BC, Shaffer LG, Cody JD, Ledbetter DH et al. 2011. Diverse mutational mechanisms cause pathogenic subtelomeric rearrangements. *Hum Mol Genet* **20**(19): 3769-3778.

Newman S, Hermetz KE, Weckselblatt B, Rudd MK. 2015. Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints. *Am J Hum Genet* **96**(2): 208-220.

Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, Szafranski P, Cooper ML, Rao M, Shao L et al. 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome Res* **21**(1): 33-46.

Ravnan JB, Tepperberg JH, Papenhausen P, Lamb AN, Hedrick J, Eash D, Ledbetter DH, Martin CL. 2006. Subtelomere FISH analysis of 11 688 cases: an evaluation of the frequency and pattern of subtelomere rearrangements in individuals with developmental disabilities. *J Med Genet* **43**(6): 478-489.

Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, Vermeesch JR. 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res* **23**(3): 411-418.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**(1): 24-26.

Rouyer F, Simmler MC, Page DC, Weissenbach J. 1987. A sex chromosome rearrangement in a human XX male caused by Alu-Alu recombination. *Cell* **51**(3): 417-425.

Rudd MK. 2012. Structural variation in subtelomeres. *Methods Mol Biol* **838**: 137-149.

Shao L, Shaw CA, Lu XY, Sahoo T, Bacino CA, Lalani SR, Stankiewicz P, Yatsenko SA, Li Y, Neill S et al. 2008. Identification of chromosome abnormalities in subtelomeric regions by microarray analysis: a study of 5,380 cases. *Am J Med Genet A* **146A**(17): 2242-2251.

Shuvarikov A, Campbell IM, Dittwald P, Neill NJ, Bialer MG, Moore C, Wheeler PG, Wallace SE, Hannibal MC, Murray MF et al. 2013. Recurrent HERV-H-mediated 3q13.2-q13.31 deletions cause a syndrome of hypotonia and motor, language, and cognitive delays. *Hum Mutat* **34**(10): 1415-1423.

Sobreira NL, Gnanakkan V, Walsh M, Marosy B, Wohler E, Thomas G, Hoover-Fong JE, Hamosh A, Wheelan SJ, Valle D. 2011. Characterization of complex chromosomal rearrangements by targeted capture and next-generation sequencing. *Genome Res* **21**(10): 1720-1727.

Talkowski ME, Ernst C, Heilbut A, Chiang C, Hanscom C, Lindgren A, Kirby A, Liu S, Muddukrishna B, Ohsumi TK et al. 2011. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am J Hum Genet* **88**(4): 469-481.

Vanneste E, Voet T, Le Caignec C, Ampe M, Konings P, Melotte C, Debrock S, Amyere M, Vikkula M, Schuit F et al. 2009. Chromosome instability is common in human cleavage-stage embryos. *Nat Med* **15**(5): 577-583.

Viguera E, Canceill D, Ehrlich SD. 2001. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J* **20**(10): 2587-2595.

Vissers LE, Bhatt SS, Janssen IM, Xia Z, Lalani SR, Pfundt R, Derwinska K, de Vries BB, Gilissen C, Hoischen A et al. 2009. Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. *Hum Mol Genet* **18**(19): 3579-3593.

Warburton D. 1991. De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am J Hum Genet* **49**(5): 995-1013.

Weckselblatt B, Hermetz KE, Rudd MK. 2015. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res* **25**(7): 937-947.

Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. 2009. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* **41**(7): 849-853.

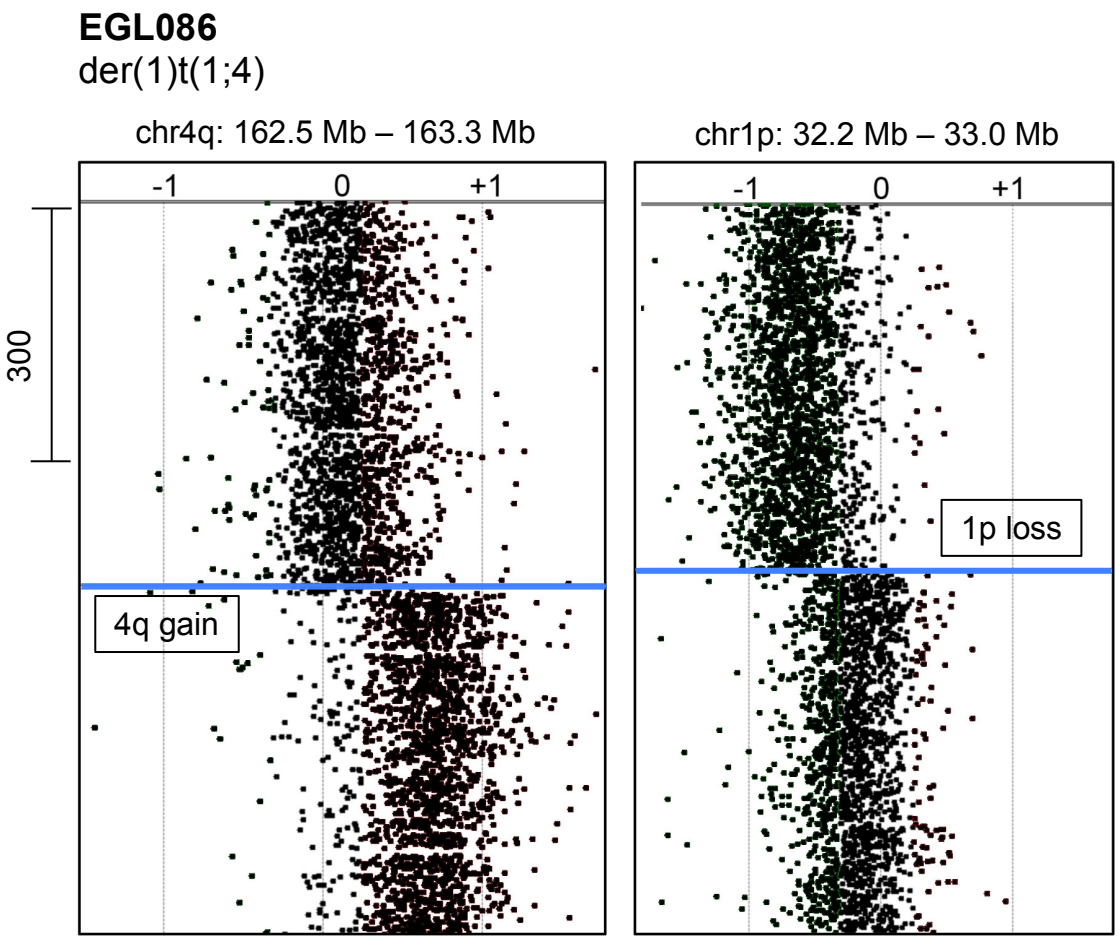


Figure 2.1: Array CGH analysis.

Averaged $\log_2$ ratios of signal intensities calculated using Genomic Workbench software are shown. Grey vertical lines indicate $\log_2$ ratios of -1, 0, and +1. For EGL086's translocation, this confirms that there is a gain of genomic material from chromosome 4q, and estimates that the breakpoint is located where probes have shifted closer to +1 on the log scale (blue horizontal line). On chromosome 1p, there is a loss of genomic material, and the breakpoint is located where probes shift to -1.

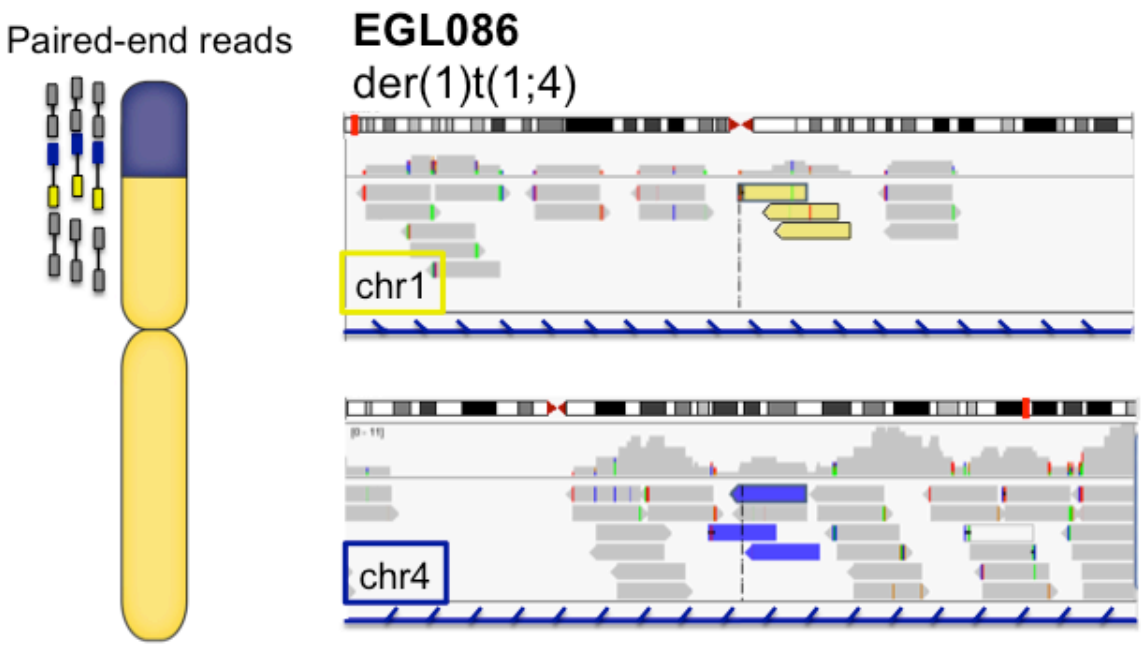


Figure 2.2: Targeted NGS viewed in IGV.

Grey boxes are sequence reads that map to the conventional genomic location, and colored boxes are sequence reads whose mate pairs map aberrantly. Here, the yellow-labeled reads on chromosome 1 have mate pairs that map to chromosome 4 (blue-labeled reads).

Subject	Inheritance	Translocation Karyotype	Capture method	Junction Features	Sequenced Junction chrA	Repeating Elements	Sequenced Junction chrB	Repeating Elements
EGL002	unknown	der(20)t(16;20)(q23;p13)	SureSelect and PCR	local duplication GG	chr20:1622433	L1PA8	chr16:78834284	
EGL007	unknown	der(20)t(7;20)(q32.1;p13)	PCR	Microhomology C	chr20:1848430		chr7:128944868	L1MA2
EGL019	unknown	der(6)t(1;6)(q41;q27)	PCR	Microhomology GG	chr6:168874213		chr1:214162602	
EGL020	unknown	der(18)t(18;20)(q22.1;p13)	SureSelect and PCR	Blunt Ends	chr18:62609097		chr20:1053849	MLT1C
EGL022	unknown	der(8)t(8;18)(p23.2;q21.1)	PCR	Microhomology CAA	chr8:1322883		chr18:47532854	
EGL024	unknown	der(11)t(9;11)(q33.2;q25)	PCR	Blunt Ends	chr11:134572101		chr9:123987159	AluJo
EGL036	unknown	der(5)t(5;7)(p15.33;q32.3)	PCR	Microhomology AA	chr5:4032149	LTR75_1	chr7:132515447	
EGL040	unknown	der(14)t(6;14)(q27;q32.2)	PCR	Microhomology CAGG	chr14:101243847		chr6:170842114	
EGL045	unknown	der(2)t(2;10)(q37.3;p14)	N/A	N/A	N/A	HERVH	N/A	HERVH
EGL047	unknown	der(1)t(1;19)(p36.33;p13.3)	SureSelect and PCR	Microhomology CAG	chr1:6609687		chr19:866173	
EGL051	maternal unbalanced translocation	der(14)t(5;14)(q35.5;q32.33)	SureSelect and Long-Range PCR	Homology 335 bp	chr14:105579242-105578907	Low Complexity DNA: CCCGGGCGGCGCTCGCCGGCGCGGCGGC; SegDup	chr5:177946014-177945678	Low Complexity DNA: CCCGGGCGGCGCTCGCCGGCGCGGCGGC; SegDup

Subject	Inheritance	Translocation Karyotype	Capture method	Junction Features	Sequenced Junction chrA	Repeating Elements	Sequenced Junction chrB	Repeating Elements
EGL052	unknown	der(6)t(6;10)(q27;q25.5)	N/A	N/A	N/A	THE1B	N/A	AT-dinucleotide simple tandem repeat
EGL062	unknown	der(1)t(1;13)(p36.32;q34)	Long-Range PCR	Microhomology AGGC	chr1:2583413-2613042	Segmental duplication; Breakpoint falls within STR GCACCCA CACCCCC AGGTGAG CATCTGA CAGCCTG GAGCA	chr13:114963581	AluSc
EGL064	unknown	der(7)t(7;17)(p22.3;q24.3)	SureSelect and PCR	local duplication TTTTATTGTGGG	chr7:1932215	Tigger4a	chr17:67168372	L1MEe
EGL071	unknown	der(22)t(4;22)(q35.1;q13.32)	Long-Range PCR	Microhomology A	chr22:49672573		chr4:185195043	MIRb
EGL080	*de novo*	der(8)t(1;8)(q41;p23.3)	Long-Range PCR	Homology 1044 bp	chr8:726077	L1PA2	chr1:223343289	L1PA3
EGL083	*de novo*	der(12)t(8;12)(q23.3;p13.32)	SureSelect and Long-Range PCR	Homology 446 bp	chr12:4132412	HERVH	chr8:114527165	HERVH

Subject	Inheritance	Translocation Karyotype	Capture method	Junction Features	Sequenced Junction chrA	Repeating Elements	Sequenced Junction chrB	Repeating Elements
EGL086	unknown	der(1)t(1;4)(p36.32;q32.3)	SureSelect and PCR	Microhomology G	chr1:3265373		chr4:162925111	MIRc
EGL087	unknown	der(5)t(X;5)(q28;p15.33)	Long-Range PCR	Insertion CTACATTCGTGGGTTCAAGCAACTGTGGATTAAAGTATTTGGGAAATAAAGT	chr5:1832733		chrX:149136469	L2
EGL089	unknown	der(6)t(6;7)(p25.1;q11.21)	Long-Range PCR	Insertion TGTCATT	chr6:6072341	L1P2	chr17:76020970	AluSx
EGL091	unknown	der(18)t(4;18)(p14;q23)	N/A	N/A	N/A		N/A	
EGL101	unknown	der(20)t(2;20)(q37.3;q13.33)	N/A	N/A	N/A		N/A	
EGL103	maternal balanced translocation	der(22)t(1;22)(q44;q13.31)	Long-Range PCR	Microhomology GT	chr22:44945164		chr1:245400824	
EGL300	paternal balanced translocation	der(21)t(12;21)(p12.3;q22.2)	PCR	Blunt Ends	chr21:40010315		chr12:13574692	L2b
EGL301	unknown	der(21)t(17;21)(q25.1;q22.3)	N/A	N/A	N/A		N/A	
EGL303	unknown	der(7)t(7;10)(q35;p15.3)	Long-Range PCR	Microhomology CAG	chr7:146907003	L1PA16	chr10:8698496	
EGL306	unknown	der(1)t(1;9)(q43;p21.3)	SureSelect and Long-Range PCR	Microhomology C	chr1:239690347	AluJr	chr9:22096073	L2c

Subject	Inheritance	Translocation Karyotype	Capture method	Junction Features	Sequenced Junction chrA	Repeating Elements	Sequenced Junction chrB	Repeating Elements
EGL308	*de novo*	der(18)t(14;18)(q32.33;p11.22)	SureSelect and PCR	Microhomology GAGG	chr18:11008031		chr14:105922362	
EGL310	paternal balanced translocation	der(17)t(5;17)(p13.3;p13.3)	SureSelect and PCR	Blunt Ends	chr17:1876319		chr5:33501282	MLT1H1
EGL311	unknown	der(11)t(11;14)(q25;q24.3)	SureSelect and PCR	Microhomology GAA	chr11:132776896		chr14:74088057	
EGL313	unknown	der(13)t(13;15)(q12.11;q11.2)	N/A	N/A	N/A		N/A	segmental duplication
EGL314	unknown	der(7)t(X;7)(p22.31;q36.1)	N/A	N/A	N/A		N/A	segmental duplication
EGL315	paternal balanced translocation	der(9)t(9;17)(p24.2;p13.3)	N/A	N/A	N/A	L1PA3	N/A	L1PA2
EGL316	maternal balanced translocation	der(22)t(9;22)(q34.13;q13.31)	SureSelect and PCR	Microhomology AGAT	chr22:44359499		chr9:134988941	AluSg
EGL317	*de novo*	der(5)t(4;5)(q34.1;p15.31)	SureSelect and PCR	Blunt Ends	chr5:6234937		chr4:173400357	L1PB4
EGL318	unknown	der(5)t(5;8)(p14.3;p21.2)	PCR	Microhomology TA	chr5:20468966		chr8:26036621	MLT1A0
EGL319	unknown	der(9)t(5;9)(p15.33;p22.3)	SureSelect and PCR	local inverted duplications TGGATAATATCCTGGGAA	chr9:15599733	L1PA4	chr5:3692567	MIRc

Subject	Inheritance	Translocation Karyotype	Capture method	Junction Features	Sequenced Junction chrA	Repeating Elements	Sequenced Junction chrB	Repeating Elements
EGL320	unknown	der(2)t(2;9)(q37.3;q34.2)	SureSelect and PCR	Blunt Ends	chr2:238766942		chr9:136829180	AT-dinucleotide simple tandem repeat
EGL351	unknown	der(X)t(X;7)(q27.1;q36.1)	Long-Range PCR	Microhomology AAAA	chrX:138592181	MLT1F2	chr7:152021915	L1MEe
EGL352	maternal balanced translocation	der(21)t(13;21)(q32.1;q22.2)	SureSelect and PCR	Local inverted duplication TGTGCAGGGAGGTGG	chr21:41617968	LTR54	chr13:152021915	
EGL353	paternal balanced translocation	der(14)t(3;14)(p26.1;q32.33)	N/A	N/A	N/A		N/A	MIRc
EGL354	maternal unbalanced translocation	der(X)t(X;2)(p22.33;p25.3)	N/A	N/A	N/A		N/A	379 bp Simple tandem repeat
EGL355	paternal balanced translocation	der(3)t(3;16)(p26.1;p13.12)	Long-Range PCR	Microhomology A	chr3:8434357	MIRc	chr16:13436583	L2
EGL357	unknown	der(21)t(16;21)(q23.1;p13)	N/A	N/A	N/A		N/A	
EGL358	*de novo*	der(22)t(7;22)(q32.1;p13)	N/A	N/A	N/A		N/A	

Subject	Inheritance	Translocation Karyotype	Capture method	Junction Features	Sequenced Junction chrA	Repeating Elements	Sequenced Junction chrB	Repeating Elements
EGL366	*de novo*	der(1)t(1;9)(p36.32;q34.3)	PCR	Insertion TGCTCCCACC ACAGTGGCC TCAGCTGCTG AGTGTGTGC ATCAGCTGTC TTCAGTGTAG GCAGATGCTT CTCTTTCTCA AAACATCAA GCCACAGAA TACCTCATCA ATGAATCAG TCCAGTGAA CCAATACCTC CAATGAACC AGTCCCTGG GGCCACGGA GACTCTGTGT GAAGATTAG CCTCTGGCTG GGTTTTTGTC TGTCTCTGA	chr1:4896879		chr9:139907470	
EGL382	unknown	der(6)t(6;11)(q27;p15.5)	WGS and PCR	Microhomology T	chr6:167806971		chr11:971021	X7A_LINE
EGL812	*de novo*	der(12)t(7;12)(p22.1,p13.33)	Long-Range PCR	Microhomology C	chr12:2503370		chr7:7221896	

Subject	Inheritance	Translocation Karyotype	Capture method	Junction Features	Sequenced Junction chrA	Repeating Elements	Sequenced Junction chrB	Repeating Elements
EGL816	maternal balanced translocation	der(20)t(12;20)(p13.31;p13)	N/A	N/A	N/A		N/A	
EGL819	maternal balanced translocation	der(X)t(X;6)(q21.31;q16.3)	N/A	N/A	N/A		N/A	
LM219	*de novo*	der(X)t(X;16)(p22.32;p13.3)	PCR	Insertion ATAGTTAACTG	chrX:4598472	AluSx	chr16:3007620	AluSx1

Table 2.1: Breakpoints of 51 simple translocations. We list the karyotype from initial clinical studies and the methods used to capture each junction. Methods that were not performed or successful are annotated as not applicable (N/A). Breakpoints sequenced by NGS and/or Sanger sequencing are listed for the chromosome loss (chrA) and gain (chrB). Junction features, such as microhomology, blunt ends, and insertions, are included for 37 translocations with Sanger-confirmed breakpoints. Breakpoints in repetitive regions are indicated.

Chapter 3

Next generation sequencing of unbalanced translocations reveals complex chromosome rearrangements including chromothripsis

Portions of this chapter have in published in *Genome Research* (2015, 25:937-947) as a research article and have been reformatted for this document.

Introduction

Complex chromosome rearrangements (CCRs) describe SV with at least three breakpoints. Over 250 CCRs have been reported in the literature, most of which are interchromosomal rearrangements that feature several translocated regions (Zhang et al. 2009; Pellestor et al. 2011). Common structures of intrachromosomal CCRs include triplications, adjacent duplications, and inverted duplications next to terminal deletions (Weckselblatt and Rudd 2015). Sequence analysis of breakpoint junctions can reveal a more complex rearrangement structure than predicted from copy number studies alone (Luo et al. 2011; Chiang et al. 2012; Carvalho et al. 2013; Brand et al. 2014; Newman et al. 2015). Even among rearrangements originally ascertained as apparently balanced CCRs, cryptic CNVs are detected with high-resolution array CGH and NGS (De Gregori et al. 2007; Chiang et al. 2012). Identification of these additional breakpoints is critical for defining these rearrangement structures and for genotype-phenotype correlation.

Though most CCRs involve only two chromosomes, some are the product of many breakpoints on three to five different chromosomes. This severe genomic reorganization is defined as chromothripsis, or chromosome shattering. Chromothripsis was originally seen in cancer (Stephens et al. 2011), and has since been observed in some constitutional translocations (Kloosterman et al. 2011; Chiang et al. 2012; Kloosterman et al. 2012; Nazaryan et al. 2014; Pellestor et al. 2014; de Pagter et al. 2015), upon integration of a transgene (Chiang et al. 2012) and in a hematopoietic stem cell lineage (McDermott et al. 2015). Constitutional chromothripsis is characterized by tens of breakpoints localized to a few regions, a moderately balanced copy number state that may

have short deletions, changes in strand orientation between translocated segments, and blunt ends or microhomology at sequenced breakpoint junctions.

Chromothripsis occurs when one or a few chromosomes are lagging during anaphase, and a micronucleus forms around them to separate these chromosomes from the rest of the chromatin mass. Upon pulverization of the micronucleus during the next cell cycle, these compartmentalized chromosomes experience extensive DNA damage and undergo rapid rearrangement (Zhang et al. 2015). In this chapter, we analyze six CCRs, three of which bear signatures of chromothripsis.

Results

Sequencing complex translocation junctions

From our cohort of 57 individuals with unbalanced translocations (Chapter 2), there are six individuals with CCRs. Their rearrangements have more than one breakpoint junction that joins multiple segments from two or more chromosomes. Subjects EGL312 and EGL356 have complex translocations involving two chromosomes, whereas EGL302, EGL305, and EGL321 have complex translocations between four or five chromosomes. EGL826 has one simple balanced translocation between Chromosomes 1 and 3 and a complex unbalanced translocation between Chromosomes 10 and 17.

Complex translocation breakpoints were fine-mapped by array CGH as described in Chapter 2. We used a targeted NGS approach, SureSelect Target Enrichment with Illumina HiSeq, to capture breakpoint junctions for EGL302, EGL305, EGL312, and EGL321's rearrangements, which successfully captured some breakpoint junctions for complex translocations in EGL305 and EGL321. However, for most complex

translocations, we performed WGS (Complete Genomics) or Nextera mate-pair sequencing to capture multiple junctions in one experiment. We confirmed breakpoints with PCR across the junction followed by Sanger sequencing.

Complex translocations between two chromosomes

EGL312, EGL356, and EGL826 have complex translocations between two chromosomes. Though EGL826 has translocations involving four chromosomes, only two chromosomes form a complex rearrangement. According to array CGH, complex translocation breakpoints in EGL312 and EGL356 border repetitive regions, so we performed Nextera mate-pair sequencing (Illumina; San Diego, CA) of 5-7-kb inserts. This approach is ideal for junctions in repetitive DNA because mate pairs span repeats and map to unique sequence (Kloosterman et al. 2011; Talkowski et al. 2011; Talkowski et al. 2012; Hanscom and Talkowski 2014). We identified discordant reads for one of two junctions expected in EGL312 and for three of four junctions expected in EGL356. In EGL312's rearrangement, CMA and FISH analysis revealed an unbalanced translocation of two regions of Chromosome 9 to the short arm of Chromosome 13 (Figure 3.1A). Mate-pair sequencing captured one inverted junction between the two translocated segments of Chromosome 9. This junction connects an L1PA3 repeat to a segmental duplication, so it is not surprising that we failed to capture this breakpoint by SureSelect. However, we did not sequence junction(s) that connect Chromosomes 9 and 13. EGL356 has an insertional translocation with three segments from Chromosome 13 translocated into the long arm of Chromosome 14 (Figure 3.1B). We confirmed insertions by FISH, and CMA revealed a 1.4-Mb deletion at the insertion site on Chromosome 14. Mate-pair

reads cross two translocation junctions between Chromosomes 13 and 14 and an inverted junction between two segments from Chromosome 13.

We also used Complete Genomics WGS to sequence EGL826's two independent chromosome rearrangements (Figure 3.1C). Her balanced translocation between Chromosomes 1 and 3 was maternally inherited, and her unbalanced translocation between Chromosomes 10 and 17 arose *de novo*. Whereas the balanced translocation has two simple translocation junctions, the unbalanced translocation has a 250-kb inverted triplication of Chromosome 17. Between the two rearrangements we sequenced a total of four translocation junctions. There are blunt ends or up to four bp of microhomology at all breakpoint junctions analyzed in these translocations (Table 3.1).

Chromothripsis translocations

Chromosome banding and FISH analyses of EGL302, EGL305, and EGL321 revealed translocations involving four or five different chromosomes. Translocations between more than two chromosomes may be caused by germline chromothripsis (Kloosterman et al. 2011).

EGL305 has a 4-way translocation that he inherited from his mother, who carries a more balanced form of the rearrangement (Figure 3.2). We sequenced two junctions involving four different chromosomes by SureSelect followed by Illumina HiSeq. The derivative Chromosome 1 has a 530-kb deletion at the 1q21 junction that is connected to an inverted breakpoint on Chromosome 15q22. Since the segment of Chromosome 15 is inverted at the junction, there must be additional breakpoint(s) to account for the correct orientation of the end of the long arm of Chromosome 15. Junction sequencing of the

derivative Chromosome 15 revealed an inverted segment of Chromosome 7 that lies between parts of Chromosomes 15 and 4. FISH analysis confirmed that EGL305's mother is balanced for the Chromosome 7 segment; she has a deletion of Chromosome 7 plus the derivative Chromosome 15 with the insertional translocation of Chromosome 7. EGL305 did not inherit the deleted Chromosome 7, so he has three copies of this 4.2-Mb region. DNA was depleted following targeted sequencing so we could not follow up with WGS to sequence additional breakpoints.

We sequenced complex rearrangements in EGL302 and EGL321 via Complete Genomics WGS. In the original cytogenetic characterization of EGL302, we detected translocations involving Chromosomes 8, 9, 11, and 13 by chromosome banding. CMA revealed a 2.8-Mb deletion of Chromosome 8 and a 6.6-Mb deletion of Chromosome 9 that correspond to translocation breakpoints. SureSelect targeted to the Chromosome 8 and 9 deletion regions did not capture any translocation junctions, but WGS revealed 11 breakpoint junctions between Chromosomes 3, 8, 9, 11, and 13 (Figure 3.3). We infer at least two additional breakpoint junctions by FISH mapping translocated segments. Though all the translocations are *de novo*, they appear to have arisen as two separate events. The reciprocal translocation between Chromosomes 11 and 13 has simple breakpoints on each derivative chromosome. However, derivative Chromosomes 3, 8, and 9 are part of complex translocations with multiple breakpoints and inserted fragments. Aside from the megabase-sized deletions on Chromosome 8 and 9, other breakpoints have only deleted 0-70 bp, for a total of 99 bp deleted.

EGL321 has a complex rearrangement involving Chromosomes 2, 3, 7, 10, and 11 (Figure 3.4). We sequenced 23 breakpoint junctions in five derivative chromosomes

using a combination of SureSelect and Complete Genomics WGS. According to FISH analysis, there are at least another six breakpoints. The translocation between Chromosomes 3 and 11 is restricted to those two chromosomes, and a portion of Chromosome 11 is inverted at both of the translocation junctions. Derivative Chromosomes 2 and 7 have swapped multiple segments of these two chromosomes, and the derivative Chromosome 10 has intermingled insertions of Chromosomes 2 and 7. Four breakpoints are completely balanced to the basepair, and the remaining breakpoints have 1-11-bp deletions. In addition to the 800-kb deletion of Chromosome 7 and the 2.2-Mb deletion of Chromosome 11, there are 55 total bp deleted at breakpoint junctions. The majority of breakpoint junctions in EGL302, EGL305, and EGL321 had no homology, and a few have short insertions (Table 3.1). No breakpoint junctions had more than four bp of microhomology.

To determine the parental origin of *de novo* translocations in EGL302 and EGL321, we genotyped family trios for heterozygous SNPs adjacent to chromosome breakpoints. We isolated SNPs from derivative chromosomes by sequencing junctions in the probands, and then determined the parental origin of the SNP at the breakpoint. Of the seven informative SNPs in EGL302 and six informative SNPs in EGL321, all were derived from paternal alleles (Figures 3.3D and 3.4D, Table 3.2) (Weckselblatt et al. 2015).

Discussion

Complex translocations and chromothripsis

We characterized six chromosome rearrangements with multiple breakpoints. Translocations in EGL312, EGL356, and EGL826 have more than one breakpoint and have inversions at the translocation junctions, but only two chromosomes are involved in the complex rearrangements. EGL302, EGL305, and EGL321 have translocations between at least four different chromosomes and many balanced insertions with altering orientations, all of which had blunt ends or microhomology at the junction. These features are hallmarks of chromothripsis (Kloosterman et al. 2011; Chiang et al. 2012; Kloosterman et al. 2012; Pellestor et al. 2014).

Rearrangements in EGL305 were transmitted from his mother, who carried a more balanced form of the translocations. In addition to EGL305, maternal chromothripsis transmission has recently been observed in three other families (de Pagter et al. 2015). In both EGL302's and EGL321's *de novo* chromothripsis events, rearrangements occurred on paternal alleles. Though our sample size is too small to determine a parent-of-origin bias, these data are consistent with other studies that find an enrichment of paternally derived chromosome rearrangements (De Gregori et al. 2007; Grossmann et al. 2010; Thomas et al. 2010; Hehir-Kwa et al. 2011; Kloosterman et al. 2011; Liu et al. 2011; Kloosterman et al. 2012).

As more germline chromothripsis genomes are being sequenced, common features have begun to emerge. Though there are many breakpoints in chromothripsis, few are accompanied by large copy number changes. CGH, WGS, and FISH revealed that EGL302 has at least 18 breakpoints, but only two large deletions of Chromosomes 8 (2.8 Mb) and 9 (6.6 Mb). EGL321 has at least 33 breakpoints, including two with large deletions of Chromosomes 7 (800 kb) and 11 (2.2 Mb). Other breakpoints have small

deletions (up to 70 bp), insertions (1-7 bp), or inversions, but do not have duplications. Similar breakpoint junction characteristics and "mostly balanced" copy number have been described at other chromothripsis rearrangements (Kloosterman et al. 2011; Chiang et al. 2012; Kloosterman et al. 2012; Macera et al. 2014; Nazaryan et al. 2014; Pellestor et al. 2014; de Pagter et al. 2015). In EGL302 and two other chromothripsis events in the literature, breakpoints disrupt the *PTPRD* gene on Chromosome 9 (Macera et al. 2014; de Pagter et al. 2015), suggesting that this locus may be a chromothripsis hotspot. Clinical features in individuals with germline chromothripsis may be due to loss of genes within deletions, or due to genes disrupted by copy-neutral rearrangements. Thus, copy number studies alone may not pinpoint the genes responsible for phenotypes.

Translocation annotation and technical limitations

Though array CGH, FISH, and chromosome banding do not provide nucleotide resolution of breakpoints, they are essential to interpret CNV breakpoints from NGS data. Following WGS of complex translocations and chromothripsis genomes, we performed iterative rounds of FISH to place insertional translocations on the correct derivative chromosome. Furthermore, initial FISH and/or chromosome banding studies are necessary to distinguish unbalanced translocations from terminal deletions and duplications detected by copy number assays (Rudd 2012). Thus, as NGS and WGS approaches become routine for CNV detection (Xi et al. 2011; Michaelson and Sebat 2012; English et al. 2015), techniques that visualize chromosomes will continue to be important for interpreting structural variation.

WGS identified many copy-neutral rearrangements that were missed by microarray analyses of EGL302 and EGL321. Though the copy number changes were relatively minor in these individuals, chromosome banding revealed multiple translocations, so we were not surprised to find additional breakpoints besides those detected by array CGH. On the other hand, WGS does not always reveal additional complexity at translocation junctions. WGS of EGL382's simple translocation (Chapter 2) and EGL826's complex translocation only identified the breakpoints we had already predicted by array CGH. Thus, it is unlikely that most translocations have cryptic complexity. Chromothripsis is estimated to occur in 2-4% of cancers (Forment et al. 2012; Pellestor et al. 2014), which is similar to the incidence of chromothripsis in germline chromosome rearrangements (Kloosterman et al. 2011; Chiang et al. 2012; Forment et al. 2012; Macera et al. 2014).

Our approach to combine SureSelect, Illumina HiSeq, mate-pair sequencing, and WGS uncovered a tens of breakpoints in this group of rare complex translocations. Combined with other complex chromosome rearrangement studies (Borg et al. 2005; Kloosterman et al. 2011; Chiang et al. 2012; Kloosterman et al. 2012; Macera et al. 2014; Pellestor et al. 2014), these data suggest that translocations involving more than two chromosomes are likely to be the product of chromothripsis.

Methods

Custom array CGH

This study was approved by the Institutional Review Board (IRB) at Emory University. Subjects had CMA testing with a version of the EmArray oligonucleotide array (Baldwin et al. 2008), followed by confirmation by chromosome banding or FISH.

G-banding of chromosomes from peripheral blood has a resolution of 550-700 bands and FISH was performed as described (Baldwin et al. 2008). For most subjects, DNA extracted from whole blood was used for all microarray and breakpoint sequencing experiments. We used DNA from lymphoblastoid cell lines for EGL302, EGL321, and EGL826. To fine-map unbalanced translocation breakpoints, we performed high-resolution array CGH. We designed custom 4 x 180K oligonucleotide arrays with ~200-bp probe spacing using eArray from Agilent Technologies (Santa Clara, CA; https://earray.chem.agilent.com/earray/). The array design ID (AMADID) identifiers are 035709, 035730, 037646, and 063584. Subject DNA was co-hybridized with reference DNA from either GM10851 or GM15510. Arrays were scanned using the Agilent high-resolution C scanner (Agilent Technologies, Santa Clara, CA), and signal intensities were evaluated using Feature Extraction Version 9.5.1.1 software (Agilent Technologies, Santa Clara, CA). We used Agilent Genomic Workbench 6.0 software (Agilent Technologies, Santa Clara, CA) to analyze the array data and call breakpoints.

Sequencing unbalanced translocations

We used Agilent SureSelect Target Enrichment to pull down 40-kb regions around breakpoints fine-mapped by custom array CGH for EGL302, EGL305, EGL312, and EGL321's rearrangements. SureSelect followed by Illumina HiSeq sequencing was performed at Hudson Alpha Genomic Services Lab. After NGS, we aligned 100-bp paired-end reads from fastq files to the GRC37/hg19 reference genome using Burrows-Wheeler Alignment (BWA) tool 0.5.9 and identified misaligned pairs using the SAMTools 0.1.18 filter function. Paired-end reads that aligned to the reference genome

too far apart, too close together, in the wrong orientation/genome order, or to different chromosomes were clustered to predict structural variation,

We performed long-range PCR and Sanger sequencing to confirm breakpoints. We used the Qiagen LongRange PCR Kit (Catalog # 206403), following the manufacturer's protocol. Sanger sequencing was performed by Beckman Coulter Genomics (Danvers, MA), and the reads were aligned to the human genome reference assembly (GRC37/hg19) using the BLAT tool on the UCSC Genome Browser (http://genome.ucsc.edu/).

Whole-genome sequencing

WGS libraries for EGL312 and EGL356 were prepared using the Nextera Mate Pair Sample Prep Kit (Catalog # FC-132-1001) according to the manufacturer's instructions. We used the Gel-Plus protocol to size-select 5-7-kb genomic fragments for sequencing. The two libraries were barcoded and sequenced on one lane of Illumina HiSeq, and the reads were analyzed as described above.

WGS of genomic DNA from EGL302, EGL321, and EGL826 was performed by Complete Genomics (Mountain View, CA) as described (Drmanac et al. 2010). Complete Genomics provided the individual reads, quality scores, and initial mappings to the GRCh37 reference genome in .tsv format. To identify discordant read pairs, we converted Reads and Mappings flagged as structural variant candidates to SAM format with the map2sam command in CGATools 1.7.1 (http://cgatools.sourceforge.net/). We used SAMTools (Li et al. 2009) to sort, index, and convert files to BAM. To account for intra-read gaps, we used a custom Perl script that extracts discordant read pairs that map

aberrantly relative to the reference genome. We viewed discordant reads with Integrative Genomics Viewer (Robinson et al. 2011) to identify and interpret structural variation.

Data access

Agilent array CGH data have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE68019. Breakpoint junction sequences have been submitted to GenBank (http://www.ncbi.nlm.nih.gov/genbank/) under accession numbers KR072894 - KR072971. Illumina sequencing data have been submitted to the NCBI Sequence Read Archive (SRA;http://www.ncbi.nlm.nih.gov/sra) under SRP057518 and Complete Genomics whole genome sequencing data have been submitted to the database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap) under accession number phs000845.v1.p1.

References

Borg K, Stankiewicz P, Bocian E, Kruczek A, Obersztyn E, Lupski JR, Mazurczak T. 2005. Molecular analysis of a constitutional complex genome rearrangement with 11 breakpoints involving chromosomes 3, 11, 12, and 21 and a approximately 0.5-Mb submicroscopic deletion in a patient with mild mental retardation. *Hum Genet* **118**(2): 267-275.

Brand H, Pillalamarri V, Collins RL, Eggert S, O'Dushlaine C, Braaten EB, Stone MR, Chambert K, Doty ND, Hanscom C et al. 2014. Cryptic and complex chromosomal aberrations in early-onset neuropsychiatric disorders. *Am J Hum Genet* **95**(4): 454-461.

Carvalho CM, Pehlivan D, Ramocki MB, Fang P, Alleva B, Franco LM, Belmont JW, Hastings PJ, Lupski JR. 2013. Replicative mechanisms for CNV formation are error prone. *Nat Genet* **45**(11): 1319-1326.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR et al. 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet* **44**(4): 390-397, S391.

De Gregori M, Ciccone R, Magini P, Pramparo T, Gimelli S, Messa J, Novara F, Vetro A, Rossi E, Maraschio P et al. 2007. Cryptic deletions are a common finding in "balanced" reciprocal and complex chromosome rearrangements: a study of 59 patients. *J Med Genet* **44**(12): 750-762.

English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, Davis CF, Dahdouli M, Ma S et al. 2015. Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics* **16**(1): 286.

Forment JV, Kaidi A, Jackson SP. 2012. Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat Rev Cancer* **12**(10): 663-670.

Hanscom C, Talkowski M. 2014. Design of large-insert jumping libraries for structural variant detection using illumina sequencing. *Curr Protoc Hum Genet* **80**: 7 22 21-29.

Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M et al. 2011. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* **20**(10): 1916-1924.

Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, van Binsbergen E, Renkens I, Duran K, Ballarati L, Vergult S, Giardino D, Hansson K et al. 2012. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep* **1**(6): 648-655.

Luo Y, Hermetz KE, Jackson JM, Mulle JG, Dodd A, Tsuchiya KD, Ballif BC, Shaffer LG, Cody JD, Ledbetter DH et al. 2011. Diverse mutational mechanisms cause pathogenic subtelomeric rearrangements. *Hum Mol Genet* **20**(19): 3769-3778.

Macera MJ, Sobrino A, Levy B, Jobanputra V, Aggarwal V, Mills A, Esteves C, Hanscom C, Pereira S, Pillalamarri V et al. 2014. Prenatal diagnosis of

chromothripsis, with nine breaks characterized by karyotyping, FISH, microarray and whole-genome sequencing. *Prenat Diagn*.

McDermott DH, Gao JL, Liu Q, Siwicki M, Martens C, Jacobs P, Velez D, Yim E, Bryke CR, Hsu N et al. 2015. Chromothriptic cure of WHIM syndrome. *Cell* **160**(4): 686-699.

Michaelson JJ, Sebat J. 2012. forestSV: structural variant discovery through statistical learning. *Nat Methods* **9**(8): 819-821.

Newman S, Hermetz KE, Weckselblatt B, Rudd MK. 2015. Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints. *Am J Hum Genet* **96**(2): 208-220.

Pellestor F, Anahory T, Lefort G, Puechberty J, Liehr T, Hedon B, Sarda P. 2011. Complex chromosomal rearrangements: origin and meiotic behavior. *Hum Reprod Update* **17**(4): 476-494.

Pellestor F, Gatinois V, Puechberty J, Genevieve D, Lefort G. 2014. Chromothripsis: potential origin in gametogenesis and preimplantation cell divisions. A review. *Fertil Steril* **102**(6): 1785-1796.

Rudd MK. 2012. Structural variation in subtelomeres. *Methods Mol Biol* **838**: 137-149.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**(1): 27-40.

Talkowski ME, Ernst C, Heilbut A, Chiang C, Hanscom C, Lindgren A, Kirby A, Liu S, Muddukrishna B, Ohsumi TK et al. 2011. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am J Hum Genet* **88**(4): 469-481.

Talkowski ME, Ordulu Z, Pillalamarri V, Benson CB, Blumenthal I, Connolly S, Hanscom C, Hussain N, Pereira S, Picker J et al. 2012. Clinical diagnosis by whole-genome sequencing of a prenatal sample. *N Engl J Med* **367**(23): 2226-2232.

Weckselblatt B, Hermetz KE, Rudd MK. 2015. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res* **25**(7): 937-947.

Weckselblatt B, Rudd MK. 2015. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet*.

Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA et al. 2011. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A* **108**(46): E1128-1136.

Zhang CZ, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, Meyerson M, Pellman D. 2015. Chromothripsis from DNA damage in micronuclei. *Nature* **522**(7555): 179-184.

Zhang F, Carvalho CM, Lupski JR. 2009. Complex human chromosomal and genomic rearrangements. *Trends Genet* **25**(7): 298-307.

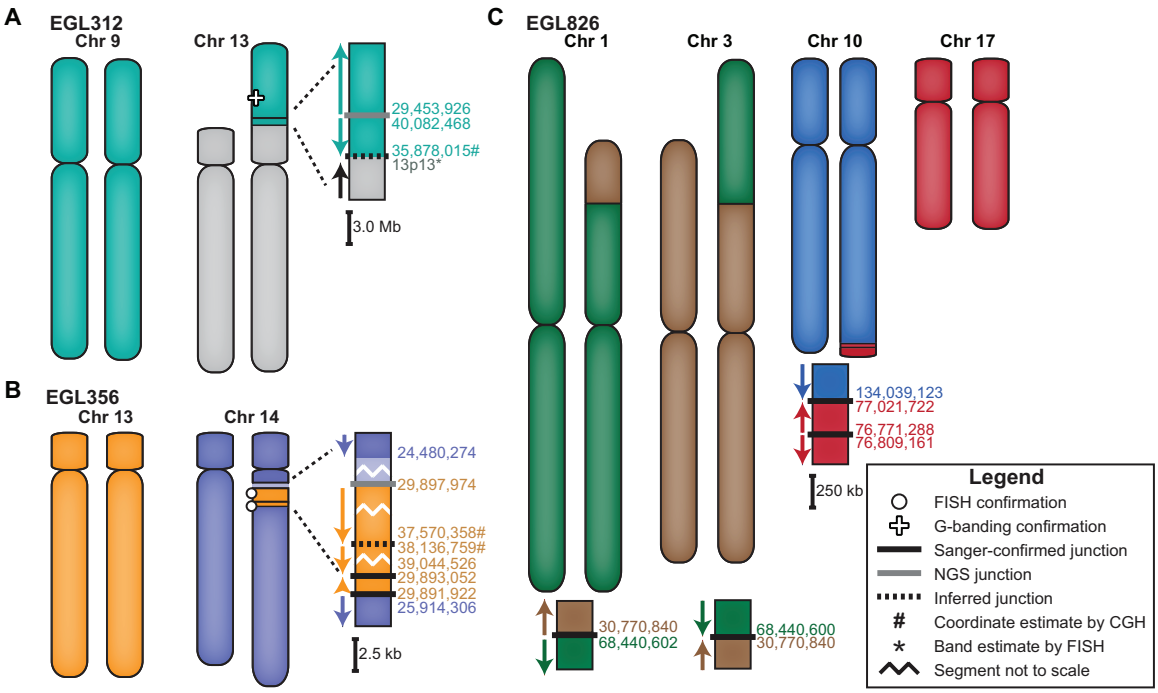

A
EGL312
Chr 9
Chr 13
29,453,926
40,082,468
35,878,015#
13p13*
3.0 Mb
B
EGL356
Chr 13
Chr 14
24,480,274
29,897,974
37,570,358#
38,136,759#
39,044,526
29,893,052
29,891,922
25,914,306
2.5 kb
C
EGL826
Chr 1
Chr 3
Chr 10
Chr 17
134,039,123
77,021,722
76,771,288
76,809,161
250 kb
30,770,840
68,440,602
68,440,600
30,770,840
Legend
FISH confirmation
G-banding confirmation
Sanger-confirmed junction
NGS junction
Inferred junction
Coordinate estimate by CGH
* Band estimate by FISH
Segment not to scale

Figure 3.1: Models of the complex translocations from EGL312, EGL356, and EGL826. See legend for symbol definitions. Zoomed-in junctions point out those confirmed with PCR and Sanger sequencing, supported only by NGS reads, or inferred by FISH. Lighter-colored chromosome segments are deletions at breakpoints. Arrows indicate chromosomal orientation relative to the normal chromosome and are shown proximal to distal.

(A) EGL312 has two regions of Chromosome 9 translocated onto the short arm of Chromosome 13. One NGS breakpoint junction (Nextera mate-pair sequencing) joins the two regions of Chromosome 9, and we infer a second breakpoint junction between Chromosome 9 to Chromosome 13.

(B) EGL356's rearrangement is an insertional translocation of three regions of Chromosome 13 into the long arm of Chromosome 14. There is a 1.5-Mb deletion of Chromosome 14 at the insertion site. Nextera mate-pair sequencing revealed translocation junctions between Chromosomes 13 and 14, and we inferred one connection between two Chromosome 13 regions.

(C) EGL826 has a maternally inherited balanced translocation between Chromosomes 1 and 3, in addition to a complex unbalanced translocation involving Chromosomes 10 and 17. At this translocation junction there is an inverted triplication of a region of Chromosome 17. Breakpoint junctions were detected by WGS (Complete Genomics) and confirmed by PCR and Sanger sequencing.

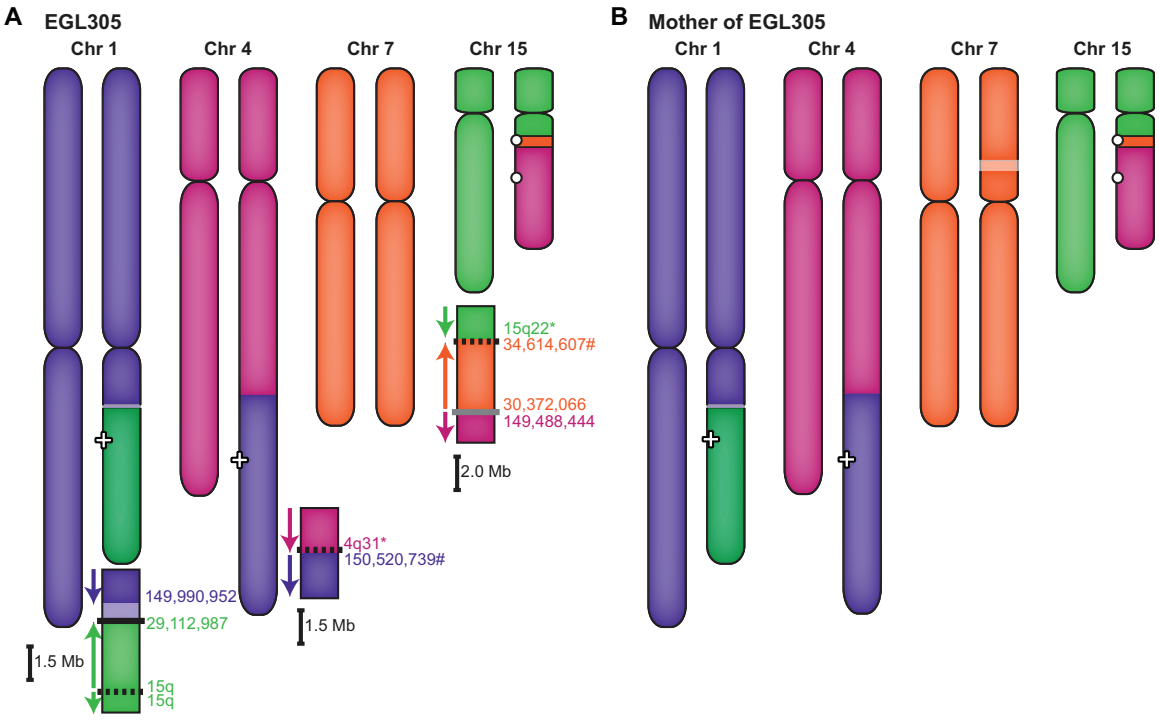


Figure 3.2: Maternal transmission of EGL305's chromothripsis.

(A) A combination of G-banding and FISH revealed EGL305's four-way translocation between Chromosomes 1, 4, 7, and 15. SureSelect and Illumina HiSeq targeted to the Chromosome 1 deletion and Chromosome 7 duplication captured two junctions, and we inferred additional breakpoints.

(B) EGL305's mother carries a more balanced form of the same four-way translocation.

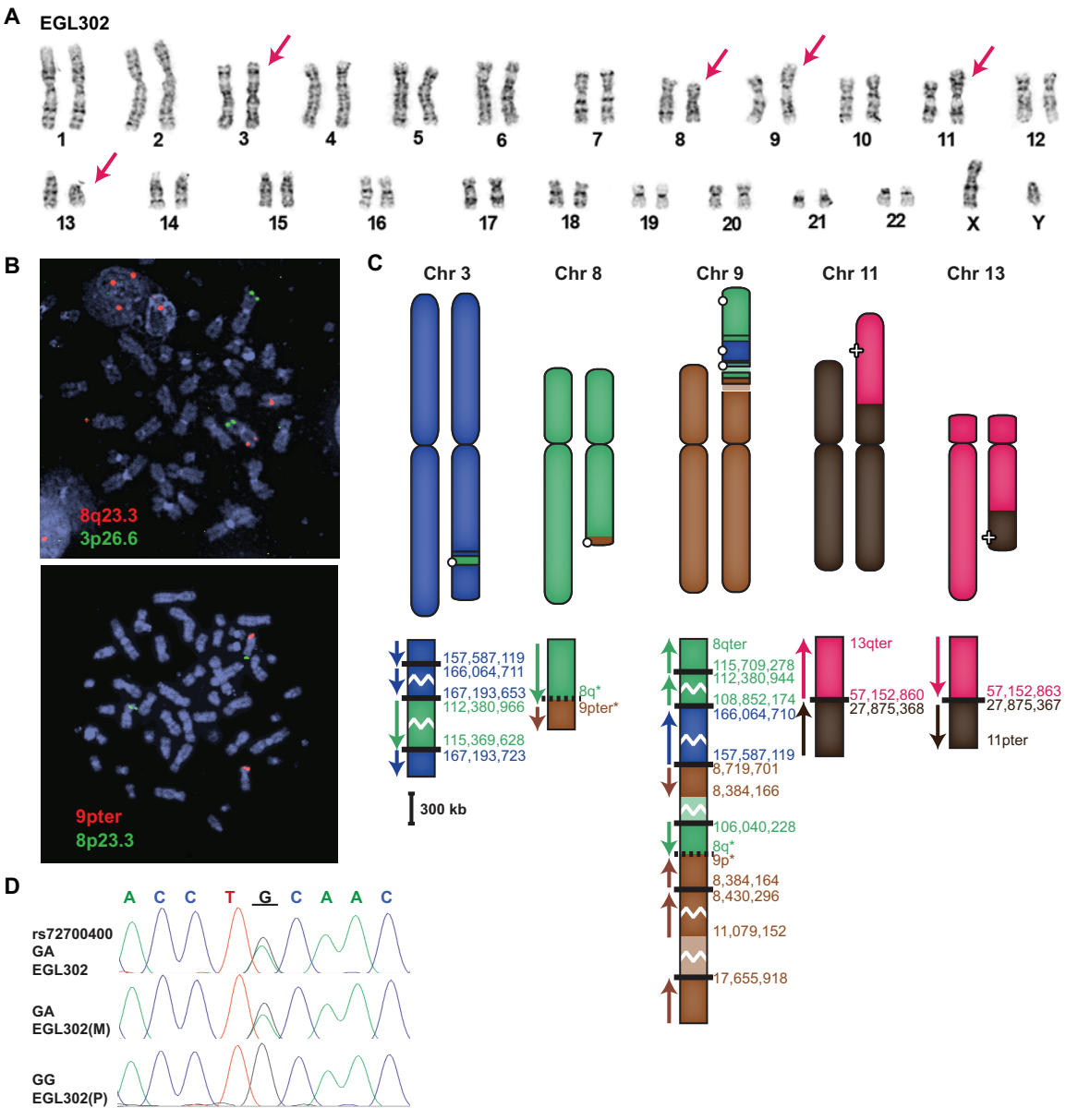

A
EGL302
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
B
8q23.3
3p26.6
9pter
8p23.3
C
Chr 3
Chr 8
Chr 9
Chr 11
Chr 13
157,587,119
166,064,711
167,193,653
112,380,966
115,369,628
167,193,723
8q*
9pter*
300 kb
8qter
115,709,278
112,380,944
108,852,174
166,064,710
157,587,119
8,719,701
8,384,166
106,040,228
8q*
9p*
8,384,164
8,430,296
11,079,152
17,655,918
13qter
57,152,860
27,875,368
57,152,863
27,875,367
11pter
D
A C C T G C A A C
rs72700400
GA
EGL302
GA
EGL302(M)
GG
EGL302(P)

Figure 3.3: EGL302's chromothripsis translocations.

(A) EGL302's karyotype; red arrows indicate translocation chromosomes.

(B) FISH confirms the insertion of 8q23.3 (probe RP11-3A12) to the long arm of Chromosome 3 (3p26.6 control probe CTC-228K22) and the translocation of 9pter (probe CTB-41L13) to the long arm of Chromosome 8 (8p23.3 control probe RP11-410N18).

(C) Model of the rearrangements in EGL302. The balanced translocation between Chromosomes 11 and 13 was confirmed by Sanger sequencing. Chromothripsis between Chromosomes 3, 8, and 9 results in many exchanges between the three chromosomes.

(D) Example of parent-of-origin analysis for EGL302. The underlined guanine (G) at the breakpoint is derived from the paternal (P), not the maternal (M), allele.

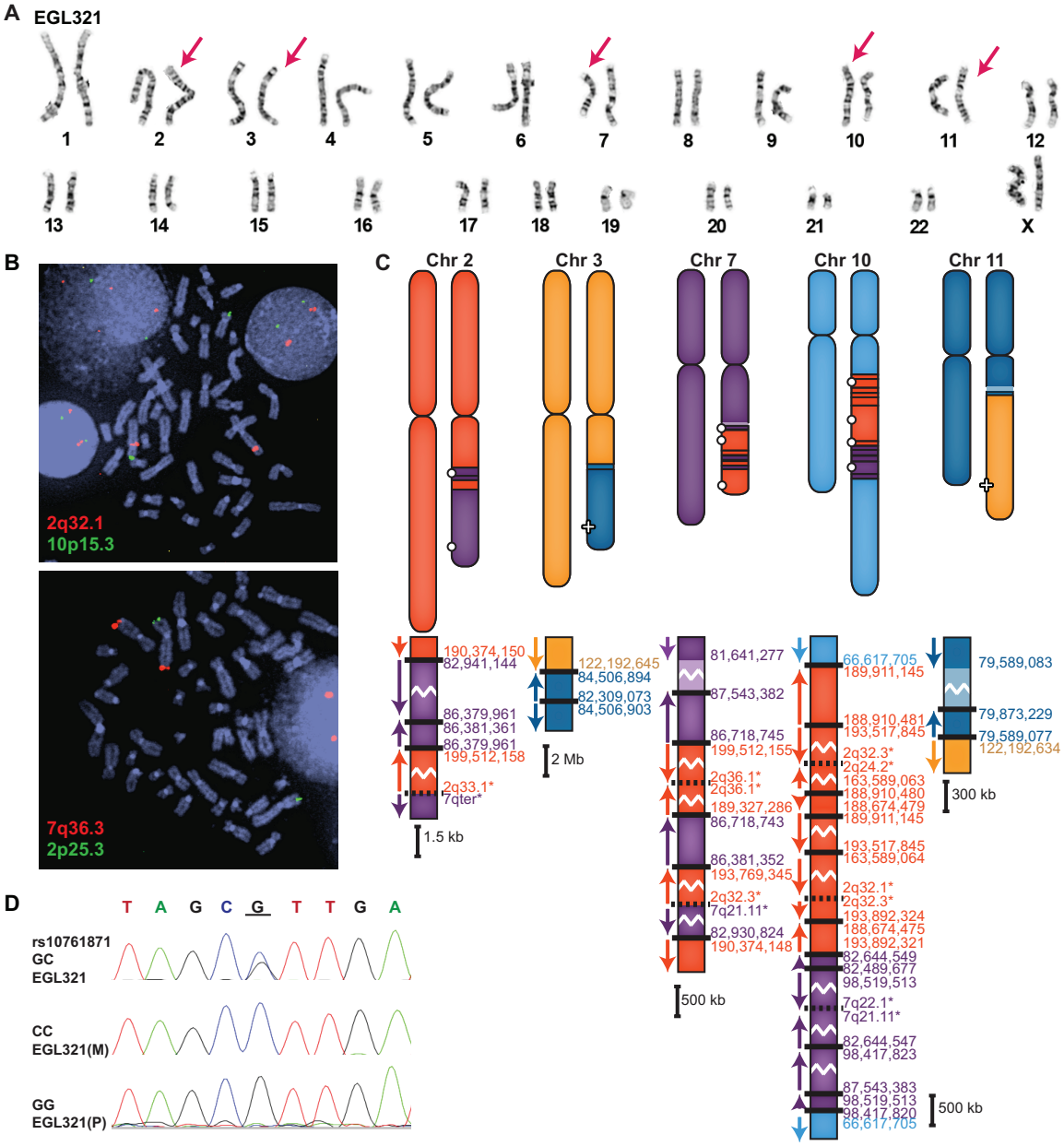

A
EGL321
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
B
2q32.1
10p15.3
7q36.3
2p25.3
C
Chr 2
Chr 3
Chr 7
Chr 10
Chr 11
190,374,150
82,941,144
86,379,961
86,381,361
86,379,961
199,512,158
2q33.1*
7qter*
1.5 kb
122,192,645
84,506,894
82,309,073
84,506,903
2 Mb
81,641,277
87,543,382
86,718,745
199,512,155
2q36.1*
2q36.1*
189,327,286
86,718,743
86,381,352
193,769,345
2q32.3*
7q21.11*
82,930,824
190,374,148
500 kb
66,617,705
189,911,145
188,910,481
193,517,845
2q32.3*
2q24.2*
163,589,063
188,910,480
188,674,479
189,911,145
193,517,845
163,589,064
2q32.1*
2q32.3*
193,892,324
188,674,475
193,892,321
82,644,549
82,489,677
98,519,513
7q22.1*
7q21.11*
82,644,547
98,417,823
87,543,383
98,519,513
98,417,820
66,617,705
500 kb
79,589,083
79,873,229
79,589,077
122,192,634
300 kb
D
T A G C G T T G A
rs10761871
GC
EGL321
CC
EGL321(M)
GG
EGL321(P)

Figure 3.4: EGL321's chromothripsis translocations.

(A) Karyotype of EGL321; red arrows indicate translocation chromosomes.

(B) FISH confirms the translocation of the long arm of Chromosome 2 (probe RP11-89P7) to the long arm of Chromosome 10 (10p15.3 control probe CTB-23B11) and the translocation of the long arm of Chromosome 7 (probe RP11-3K23) to the long arm of Chromosome 2 (2p25.3 control probe RP11-71M21).

(C) Model of EGL321's rearrangements. Zoomed-in translocation junctions show breakpoints on the derivative chromosomes.

(D) Example of parent-of-origin analysis for EGL321. Underlined G is adjacent to a Chromosome 10 breakpoint and is derived from the paternal (P) allele.

	Complex	Chromothripsis	All
Subjects	3	3	6
Total Breakpoints	17	60	77
Total Junctions	10	47	57
Junctions Sequenced	6	35	41
Blunt Ends	1	15	16
Microhomology 1-4 bp	5	17	22
Insertions 1-4 bp	0	3	3
Homology >300 bp	0	0	0

Table 3.1: Features of sequenced breakpoint junctions in complex and chromothripsis translocations.

Subject	SNP	Coordinate	Chromothripsis allele	Maternal genotype	Paternal genotype	Origin
EGL302	rs4604474	chr8:112381404	A	AT	AA	paternal allele
EGL302	rs7866922	chr9:8719260	A	AG	AA	paternal allele
EGL302	rs906636	chr9:8719321	A	AG	AA	paternal allele
EGL302	rs72700399	chr9:8719385	C	CG	CC	paternal allele
EGL302	rs66523670	chr9:8719493	A	AG	AA	paternal allele
EGL302	rs72700400	chr9:8719524	G	GA	GG	paternal allele
EGL302	rs10815870	chr9:8430111	C	AA	CA	paternal allele
EGL321	rs59684283	chr7:81641187	C	TT	CT	paternal allele
EGL321	rs4115404	chr10:66617809	T	GG	TT	paternal allele
EGL321	rs2357461	chr2:193769528	G	AG	GG	paternal allele
EGL321	rs10761871	chr10:66616535	G	CC	GG	paternal allele
EGL321	rs2709905	chr7:82941571	G	AG	GG	paternal allele
EGL321	rs2709904	chr7:82941789	A	GG	AA	paternal allele

Table 3.2: Translocation parent of origin of EGL302 and EGL321.

In these rearrangements, SNPs within one kb of chromosome breakpoints are all derived from paternal alleles. SNPs lie in copy-neutral regions adjacent to breakpoints, so there are two alleles in the parents and the proband.

Chapter 4

Fusion genes are a product of unbalanced translocations and duplication CNVs

Portions of this chapter have in published in *Genome Research* (2015, 25:937-947) and *American Journal of Human Genetics* (2015, 96:208-220) as research articles and have been reformatted for this document.

Introduction

Chromosome rearrangements have the potential to cause human disease through a variety of mechanisms. Balanced rearrangements can physically disrupt genes leading to loss of function, and rearrangements with copy number variation (CNV) alter gene dosage through genomic deletion or duplication. Furthermore, some rearrangements have breakpoint junctions that juxtapose two different genes. A fusion gene forms when both genes are orientated in the same direction and the open reading frame remains preserved. Gene fusions may lead to gain of function of the original genes through altering their activity or regulation.

The most well known fusion gene arises in cancer and is a product of the balanced translocation between Chromosomes 9 and 22 that is present in almost all individuals with chronic myeloid leukemia (CML). The derivative Chromosome 22 joins the *BCR* gene to the *ABL1* gene from Chromosome 9 (de Klein et al. 1982; Heisterkamp et al. 1985). The *BCR-ABL1* product results in a constitutively active ABL1 kinase that acts as a driver of CML. An inhibitor of BCR-ABL1, imatinib, was the first fusion-targeted drug approved for treatment of CML (Druker et al. 1996). Over the last 30 years, more than 8,000 fusion genes have been identified across 16 tumor types (Yoshihara et al. 2014; Mitelman et al. 2015). Many tumor-specific chromosomal rearrangements produce fusion proteins in a way that leads to uncontrolled cell proliferation. Identification of recurrent rearrangements in a given tumor type is important for diagnosis, and pinpointing fusion genes and their protein products is the first step in the development of drug inhibitors (Mitelman et al. 2007; Parker and Zhang 2013).

Much less is known about the role of fusion genes formed via constitutional chromosome rearrangements. The first constitutional gene fusion was reported in 2001, when a balanced translocation led to *PAFAH1B3-CLK2* fusion in a subject with a more complex phenotype than expected from misregulation of *PAFAH1B3* or *CLK2* alone (Nothwang et al. 2001). More recent studies of constitutional breakpoint junctions have revealed fusions at a balanced translocation (Backx et al. 2011), duplications (Rippey et al. 2013), deletions (Boone et al. 2014), and at a chromothriptic junction (van Heesch et al. 2014). In these constitutional rearrangements, a mechanism of disease pathogenesis through gene fusion remains unclear, but could be important for potential therapy.

Genes disrupted or fused at the breakpoints of balanced rearrangements are excellent candidates for neurodevelopmental disorders because the rest of the genome is intact. However, fusion genes in unbalanced rearrangements also have the potential to acquire new functions related to phenotypic outcomes. In this chapter, we analyze the gene content of breakpoints in 57 unbalanced translocations and 184 duplication CNVs. As fusion genes are only beginning to be recognized in constitutional chromosome rearrangements, identifying unique gene fusions is important for understanding their consequences.

Results

Gene content at translocation junctions

As described in Chapters 2 and 3, we fine-mapped breakpoints in a group of 57 unique unbalanced translocations from individuals with neurodevelopmental syndromes. In the 51 simple translocations with 102 sequenced or fine-mapped breakpoints, 44

(43%) of the breakpoints lie in a gene. Thirteen translocations do not disrupt a gene at either chromosome breakpoint, and 32 translocations disrupt a gene at one but not both breakpoints. In six simple translocations, both breakpoints lie in the open reading frame of genes (Table 4.1). Genes juxtaposed by EGL064's and EGL352's translocations are not transcribed in the same direction, and EGL086's fusion gene is predicted to be out-of-frame (Table 4.1). Translocations in EGL002, EGL019, and EGL308, however, are poised to create in-frame fusion transcripts (Figure 4.1) (Weckselblatt et al. 2015).

EGL002's translocation between Chromosomes 16 and 20 joins *SIRPG* exons 1-2 to *WWOX* exon 5. The resulting *SIRPG-WWOX* fusion protein is predicted to retain a SIRPG immunoglobulin domain but lack WWOX WW domains. In EGL019, *SMOC2* exon 1 is joined to *PROX1* exons 2-5, but the fusion protein is not predicted to retain *SMOC2*'s functional domains. EGL308's translocation results in a truncated version of *MTA1*, with exons 8-21 fused to noncoding exons 1-2 of *PIEZO1* upstream. Based on exon phase, all three of these fusion genes are predicted to be in-frame. However, RNA was not available, so we could not confirm the presence of fusion transcripts.

Complex translocations also have the potential to create fusion genes. Sequenced breakpoints in EGL305 and EGL312 do not disrupt genes. In EGL356's rearrangement, a deletion in Chromosome 14 interrupts *DHRS4L1*, and translocations interrupt *MTUS2*, *ALG5*, and *POSTN* on segments of Chromosome 13. EGL826's translocation between Chromosomes 10 and 17 joins *C1QTNF1* and *STK32C* genes in the same orientation, but fusion transcripts are predicted to be out-of-frame. Breakpoints in EGL302's rearrangements disrupt two genes, both on the derivative Chromosome 9. Three different breakpoints interrupt *PTPRD*, and one breakpoint disrupts *SH3GL2*. EGL321's

breakpoints interrupt *GRM3, KPNA1, DLG2, CACNA2D1, GULP1, COL5A2, KCNH7, PCLO,* and *TRRAP*. In both EGL302 and EGL321, functional fusion genes are not predicted due to the fragmentation and orientation of the genes.

Duplication CNV junctions

In addition to interchromosomal translocations, we analyzed genes at the breakpoints of intrachromosomal duplication CNVs. We fine-mapped 184 constitutional duplications ascertained from individuals referred for diagnostic cytogenetics testing (Newman et al. 2015). We included duplications that were reported as pathogenic or of uncertain clinical significance and excluded common CNVs present in the general population (Itsara et al. 2009; Shaikh et al. 2009; MacDonald et al. 2014). Following targeted NGS or WGS to sequence duplications in 112 subjects with 119 CNVs we found that 99 (83%) were tandem duplications in direct orientation.

Intragenic duplications in EGL456 and EGL527 are predicted to result in out-of-frame transcripts of *CNTN4* and *TCOF1*, respectively. EGL456 was referred for testing due to infantile cerebral palsy. *CNTN4* lies within the region deleted in 3p– syndrome, and rearrangements involving *CNTN4* have been described in children with developmental delay, speech delay, or ASD (Fernandez et al. 2004; Roohi et al. 2009; Cottrell et al. 2011). EGL527's referring diagnosis of cleft palate is likely due to loss of function of *TCOF1*, which causes autosomal dominant Treacher Collins syndrome.

In-frame gene fusions are predicted in four tandem duplications (Table 4.2). The phenotypic consequences of the putative *TRPV3-TAXIBP3* (EGL413) and *LTBP1-BIRC6* (EGL415, EGL478) fusions are difficult to predict since these genes have not been

implicated in neurodevelopmental disorders. EGL480's tandem duplication juxtaposes exons 1-6 of *SOS1* to exons 2-33 of *MAP4K3* in frame (Figure 4.2). Gain-of-function missense mutations in *SOS1* cause Noonan syndrome (Tartaglia et al. 2007; Zenker et al. 2007). Although EGL480 does not have a formal diagnosis of Noonan syndrome, he does exhibit hypertelorism, seizures, and developmental delay that could be related to gain of function in the *SOS1-MAP4K3* fusion product.

For several rearrangements, duplication breakpoint junction sequencing revealed greater complexity than recognized by clinical microarray testing. We identified 10 triplications, five adjacent duplications, six duplication-normal-duplication (DUP-NML-DUP), three insertional translocations, an inverted duplication adjacent to a cryptic terminal deletion (LM223), and a duplication with unknown structure (EGL414) (Newman et al. 2015). Though some of triplication breakpoints lie in genes, none are predicted to form fusion transcripts. EGL605's DUP-NML-DUP fuses the *KCNH5* and *FUT8* genes and is predicted to be in frame (Figure 4.2). A *de novo* missense variant in *KCNH5* has been reported in a child with epilepsy (Veeramah et al. 2013). EGL605 was tested due to failure to thrive as an infant and we do not know if she developed seizures later. For EGL701's 522-kb insertion of Xq22.3, junction sequencing revealed an inverted insertion of this Xq22.3 segment into 9q34.11. Breakpoints on Chromosomes 9 and X lie in the *USP20* and *COL4A6* genes, respectively (Figure 4.2). EGL701 has one intact copy of *COL4A6* on his X Chromosome, one intact copy of *USP20* on one Chromosome 9, and disruption of *USP20* on the derivative Chromosome 9 that carries the insertion. Based on the orientation of the genes and the inverted insertion of Xq22.3, this is predicted to result in an in-frame fusion of exons 1-2 of *COL4A6* and exons 4-26

of *USP20*. It remains to be determined whether or not the *COL4A6-USP20* fusion is related to EGL701's referring diagnosis of developmental delay, short stature, and multiple congenital anomalies.

Discussion

Chromosome breakpoints that fuse genes with the same exon phase may create unique in-frame fusion genes. Fusion genes are a hallmark of tumors, and recent studies are beginning to suggest that constitutional fusions play role in the pathogenesis of diseases other than cancer (Nothwang et al. 2001; Backx et al. 2011; Rippey et al. 2013; Boone et al. 2014; van Heesch et al. 2014). A fusion gene that exists in an individual with a balanced rearrangement will be directly linked to their phenotype, but fusion genes in unbalanced rearrangements may still be contributing to phenotypic outcome. To further understand the contribution of fusion genes to these types of rearrangements we analyzed the breakpoints of unbalanced translocation and duplications that bear simple and complex structures.

We predict three novel fusion genes at the junctions of simple unbalanced translocations. Without detailed phenotypic information on the subjects carrying the translocations, it's difficult to predict a role for these fusions. The parent genes of *SIRPG-WWOX, SMOC2-PROX1, PIEZO2-MTA1* are not associated with neurodevelopmental syndromes, and their protein products are not predicted to retain active domains. Thus, clinical phenotypes are most likely due to genes deleted and duplicated as part of the unbalanced translocation rather than a fusion gene at the breakpoint junction. However,

it's important to note that unbalanced translocations are just as likely to generate gene fusions as the more commonly studied balanced translocations.

At duplication CNV junctions, four gene fusions were formed by tandem duplications, one by two interconnected duplications, and one by duplication inserted at another locus (Table 4.2). EGL480's duplication breakpoints join *SOS1* and *MAP4K3*. *SOS1* stimulates the activation of Ras, a small GTPase that regulates cell proliferation, cell differentiation, and apoptosis through the Ras/MAPK pathway (Shapiro 2002). Mutations in genes encoding components of the Ras/MAPK pathway result in developmental syndromes called RASopathies. Missense mutations in *SOS1* lead to an increase in the active form of Ras and cause Noonan syndrome (Tartaglia et al. 2007; Zenker et al. 2007). EGL480 exhibits some phenotypic features associated with Noonan syndrome that that could be related to gain of function in the *SOS1-MAP4K3* fusion product.

Without resolving breakpoints to determine the orientation and location of the duplicated segment, it is impossible to infer the effects of duplications on gene structure. Sequencing EGL605's DUP-NML-DUP revealed an inverted genomic segment that fuses the *KCNH5* and *FUT8* genes in the same orientation. In two out of three insertional translocations, we performed FISH to identify the location of the duplicated material, and in one case the insertion was only detected by breakpoint sequencing. EGL701 inherited this duplication of Xq22.22 from his mother, and based on array CGH we assumed that it was tandem. Instead the duplication is inserted into Chromosome 9 and produces a putative *COL4A6-USP20* fusion at the insertion site. This fusion is predicted to be in frame and could create a unique fusion protein.

Here, we report novel fusion genes at the junctions of tandem duplications, unbalanced translocations, a DUP-NML-DUP, and an insertional translocation. Future mRNA and protein studies are necessary to determine the functional consequences of genes fused at these translocation and duplication breakpoints. In addition, transcripts that we predict to be out of frame may actually produce proteins by alternative splicing using cryptic splice donor and/or acceptor sites. Since many chromosome aberrations that rearrange genes are invisible at the cytogenetic level, determining the orientation and location of CNVs via sequencing is essential to interpret their effects on genes and correlate with phenotypes. These breakpoint analyses, as well as future RNA and protein studies, are essential to determine the functional consequences of constitutional chromosome rearrangements.

Methods

Human Subjects

See Chapter 2 methods for ascertainment of individuals carrying translocations. For duplications, individuals were referred for clinical microarray testing with indications including but not limited to intellectual disability, developmental delay, autism spectrum disorders, congenital anomalies, and dysmorphic features. Duplications were initially identified via diagnostic chromosomal microarray analysis performed at Emory Genetics Laboratory. Clinical microarrays have genome-wide coverage with one oligonucleotide probe per ~75 kilobases and greater probe density in targeted regions (Baldwin et al. 2008). Duplication breakpoints were fine-mapped and sequenced as described for translocations in Chapter 2.

Fusion gene prediction

For breakpoints that interrupt genes oriented in the same direction, we predicted the reading frame of fusion genes. We used all gene isoforms included in the Ensembl release 75 gene transcript database (Flicek et al. 2014) to predict whether the reading frame was preserved following the rearrangement. Juxtaposed exons with the same phase were predicted to be in-frame. We predicted fusion protein motifs by analyzing cDNA sequence from Ensembl 75 with ScanProsite (http://prosite.expasy.org/scanprosite/).

References

Backx L, Seuntjens E, Devriendt K, Vermeesch J, Van Esch H. 2011. A balanced translocation t(6;14)(q25.3;q13.2) leading to reciprocal fusion transcripts in a patient with intellectual disability and agenesis of corpus callosum. *Cytogenet Genome Res* **132**(3): 135-143.

Baldwin EL, Lee JY, Blake DM, Bunke BP, Alexander CR, Kogan AL, Ledbetter DH, Martin CL. 2008. Enhanced detection of clinically relevant genomic imbalances using a targeted plus whole genome oligonucleotide microarray. *Genet Med* **10**(6): 415-429.

Boone PM, Yuan B, Campbell IM, Scull JC, Withers MA, Baggett BC, Beck CR, Shaw CJ, Stankiewicz P, Moretti P et al. 2014. The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *Am J Hum Genet* **95**(2): 143-161.

Cottrell CE, Bir N, Varga E, Alvarez CE, Bouyain S, Zernzach R, Thrush DL, Evans J, Trimarchi M, Butter EM et al. 2011. Contactin 4 as an autism susceptibility locus. *Autism Res* **4**(3): 189-199.

de Klein A, van Kessel AG, Grosveld G, Bartram CR, Hagemeijer A, Bootsma D, Spurr NK, Heisterkamp N, Groffen J, Stephenson JR. 1982. A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* **300**(5894): 765-767.

Druker BJ, Tamura S, Buchdunger E, Ohno S, Segal GM, Fanning S, Zimmermann J, Lydon NB. 1996. Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med* **2**(5): 561-566.

Fernandez T, Morgan T, Davis N, Klin A, Morris A, Farhi A, Lifton RP, State MW. 2004. Disruption of contactin 4 (CNTN4) results in developmental delay and other features of 3p deletion syndrome. *Am J Hum Genet* **74**(6): 1286-1293.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42**(Database issue): D749-755.

Heisterkamp N, Stam K, Groffen J, de Klein A, Grosveld G. 1985. Structural organization of the bcr gene and its role in the Ph' translocation. *Nature* **315**(6022): 758-761.

Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, Krauss RM, Myers RM, Ridker PM, Chasman DI et al. 2009. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* **84**(2): 148-161.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**(Database issue): D986-992.

Mitelman F, Johansson B, Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**(4): 233-245.

Mitelman F, Johansson B, Mertens F. 2015. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.

Newman S, Hermetz KE, Weckselblatt B, Rudd MK. 2015. Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints. *Am J Hum Genet* **96**(2): 208-220.

Nothwang HG, Kim HG, Aoki J, Geisterfer M, Kubart S, Wegner RD, van Moers A, Ashworth LK, Haaf T, Bell J et al. 2001. Functional hemizygosity of PAFAH1B3 due to a PAFAH1B3-CLK2 fusion gene in a female with mental retardation, ataxia and atrophy of the brain. *Hum Mol Genet* **10**(8): 797-806.

Parker BC, Zhang W. 2013. Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment. *Chin J Cancer* **32**(11): 594-603.

Rippey C, Walsh T, Gulsuner S, Brodsky M, Nord AS, Gasperini M, Pierce S, Spurrell C, Coe BP, Krumm N et al. 2013. Formation of chimeric genes by copy-number variation as a mutational mechanism in schizophrenia. *Am J Hum Genet* **93**(4): 697-710.

Roohi J, Montagna C, Tegay DH, Palmer LE, DeVincent C, Pomeroy JC, Christian SL, Nowak N, Hatchwell E. 2009. Disruption of contactin 4 in three subjects with autism spectrum disorder. *J Med Genet* **46**(3): 176-182.

Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'Arcy M et al. 2009. High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res* **19**(9): 1682-1690.

Shapiro P. 2002. Ras-MAP kinase signaling pathways and control of cell proliferation: relevance to cancer therapy. *Crit Rev Clin Lab Sci* **39**(4-5): 285-330.

Tartaglia M, Pennacchio LA, Zhao C, Yadav KK, Fodale V, Sarkozy A, Pandit B, Oishi K, Martinelli S, Schackwitz W et al. 2007. Gain-of-function SOS1 mutations cause a distinctive form of Noonan syndrome. *Nat Genet* **39**(1): 75-79.

van Heesch S, Simonis M, van Roosmalen MJ, Pillalamarri V, Brand H, Kuijk EW, de Luca KL, Lansu N, Braat AK, Menelaou A et al. 2014. Genomic and Functional Overlap between Somatic and Germline Chromosomal Rearrangements. *Cell Rep* **9**(6): 2001-2010.

Veeramah KR, Johnstone L, Karafet TM, Wolf D, Sprissler R, Salogiannis J, Barth-Maron A, Greenberg ME, Stuhlmann T, Weinert S et al. 2013. Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia* **54**(7): 1270-1281.

Weckselblatt B, Hermetz KE, Rudd MK. 2015. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res* **25**(7): 937-947.

Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, Verhaak RG. 2014. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*.

Zenker M, Horn D, Wieczorek D, Allanson J, Pauli S, van der Burgt I, Doerr HG, Gaspar H, Hofbeck M, Gillessen-Kaesbach G et al. 2007. SOS1 is the second most common Noonan gene but plays no major role in cardio-facio-cutaneous syndrome. *J Med Genet* **44**(10): 651-656.

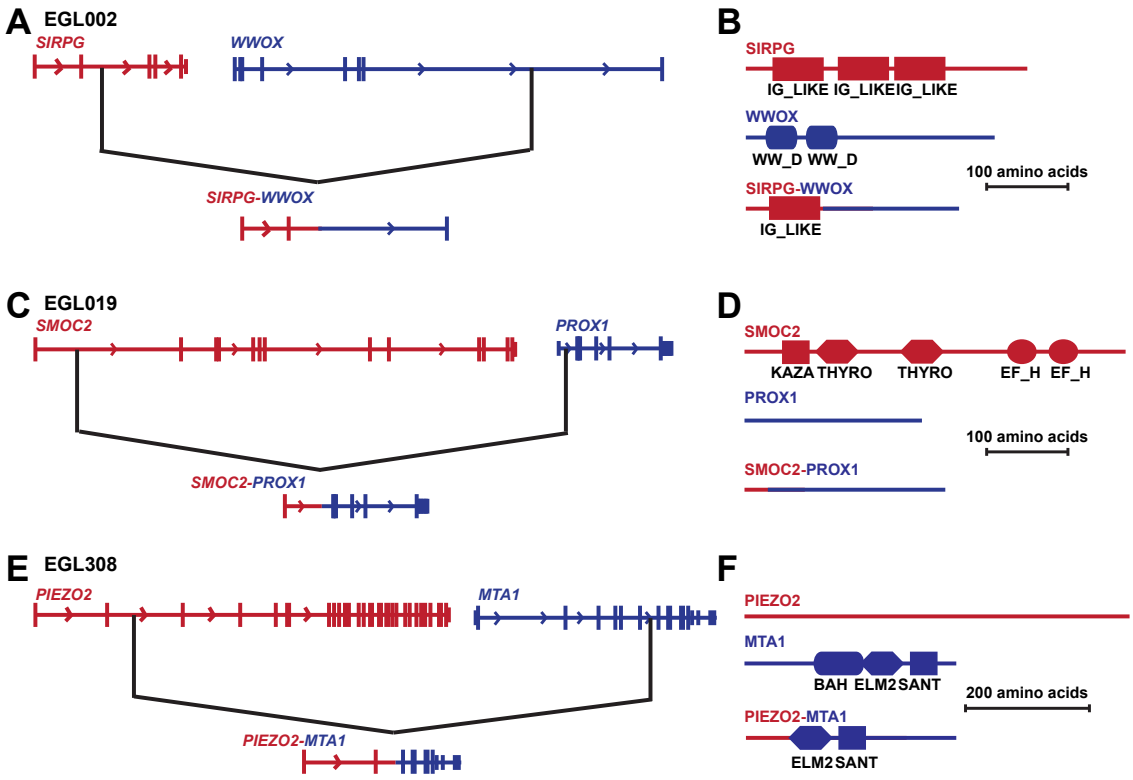


Figure 4.1: Predicted in-frame fusion genes at sequenced translocation junctions. Black lines indicate translocation breakpoints in genes (not drawn to scale). (A,B) Fusion of *SIRPG* and *WWOX* in EGL002. (C,D) EGL019's *SMOC2-PROX1* fusion. (E,F) Fusion of *PIEZO2* and *MTA1* in EGL308.

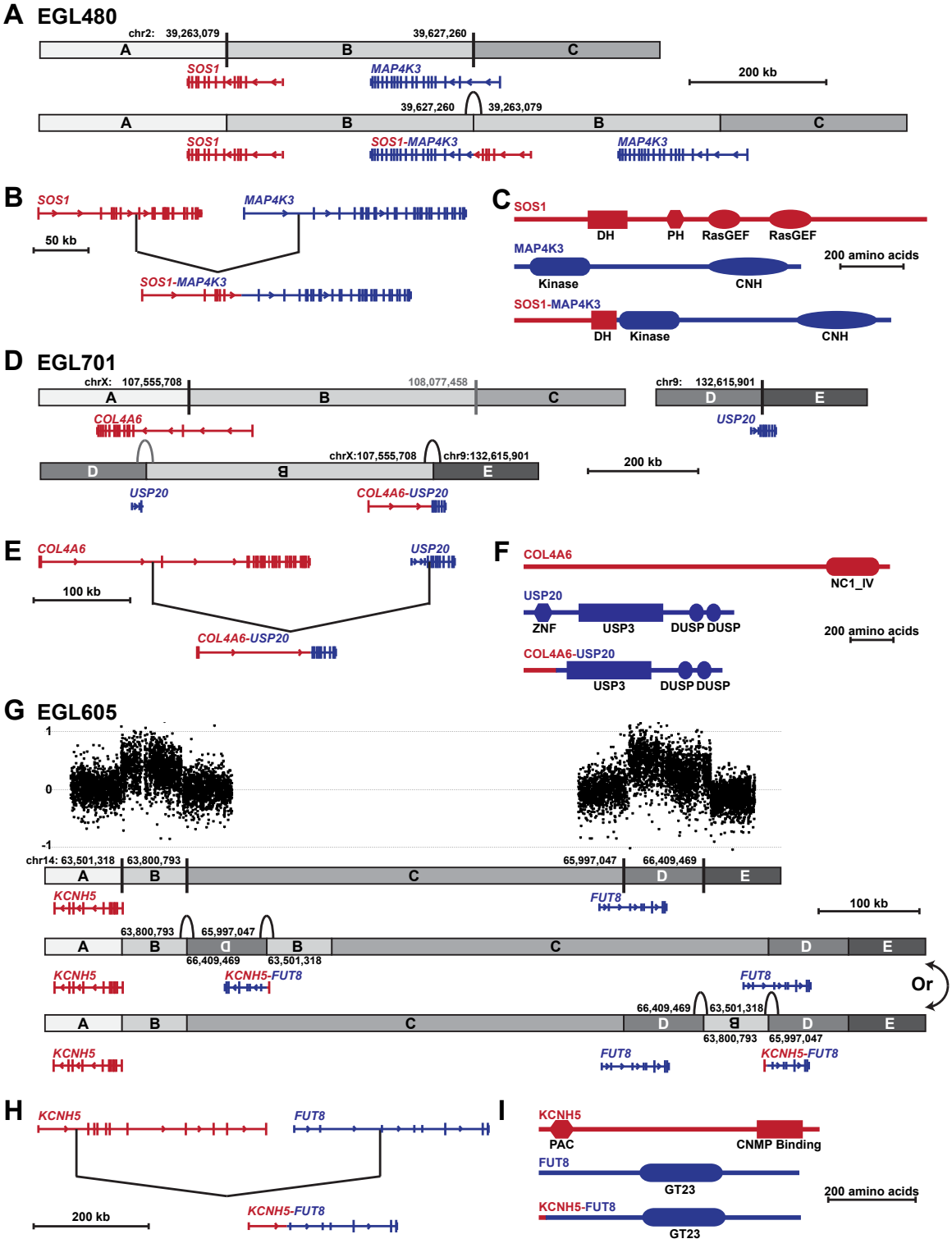

A EGL480
chr2: 39,263,079
39,627,260
A
B
C
SOS1
MAP4K3
200 kb
39,627,260
39,263,079
SOS1-MAP4K3
B
50 kb
SOS1-MAP4K3
C
DH
PH
RasGEF
RasGEF
Kinase
CNH
200 amino acids
DH
Kinase
CNH
D EGL701
chrX: 107,555,708
108,077,458
chr9: 132,615,901
D
E
COL4A6
USP20
chrX:107,555,708
chr9:132,615,901
200 kb
USP20
COL4A6-USP20
E
100 kb
F
NC1_IV
ZNF
USP3
DUSP DUSP
200 amino acids
USP3
DUSP DUSP
G EGL605
1
0
-1
chr14: 63,501,318
63,800,793
65,997,047
66,409,469
KCNH5
FUT8
100 kb
63,800,793
65,997,047
66,409,469
63,501,318
KCNH5-FUT8
Or
63,501,318
63,800,793
65,997,047
H
200 kb
I
PAC
CNMP Binding
GT23
200 amino acids
GT23

Figure 4.2: In-frame fusion genes predicted at duplication junctions. Genes that cross breakpoints are shown relative to the reference genome (above) and the duplication (below). The genomic coordinates of breakpoints have been confirmed by sequencing (black) or high-resolution array CGH (grey). (A) EGL480's direct duplication fuses *SOS1* to *MAP4K3* (B).

(D) EGL701's duplication of the X chromosome is inverted and inserted into Chromosome 9. *COL4A6* is fused to *USP20* at the insertion site (E).

(G) Array CGH (above) and breakpoint sequencing revealed that EGL605's DUP-NML-DUP fuses *KCNH5* to *FUT8* (H) at the inverted junction of the two duplications. There are two possible structures for this rearrangement, and both predict a *KCNH5-FUT8* fusion.

C, F, I show domains of the fusion proteins. We predicted fusion protein motifs by entering fusion cDNA sequence from Ensembl 75 into ScanProsite.

Subject ID	**Predicted frame**	**Fusion gene**
EGL002	In frame	*SIRPG-WWOX*
EGL019	In frame	*SMOC2-PROX1*
EGL308	In frame	*PIEZO2-MTA1*
EGL086	Out of frame	*FSTL5-PRDM16*

Table 4.1. Predicted fusion genes at simple translocation junctions

Subject ID	Duplication structure	Predicted frame	Fusion gene
EGL456	Intragenic, direct	Out of frame	*CNTN4*
EGL527	Intragenic, direct	Out of frame	*TCOF1*
EGL413	Intergenic, direct	In frame	*TRPV3-TAX1BP3*
EGL415	Intergenic, direct	In frame	*LTBP1-BIRC6*
EGL478	Intergenic, direct	In frame	*LTBP1-BIRC6*
EGL480	Intergenic, direct	In frame	*SOS1-MAP4K3*
EGL605	DUP-NML-DUP	In frame	*KCNH5-FUT8*
EGL701	Insertional translocation	In frame	*COL4A6-USP20*
EGL403	Intergenic, direct	Out of frame	*ADD2-EXOC6B*
EGL408	Intergenic, direct	Out of frame	*H6ST2-GPC4*
EGL465	Intergenic, direct	Out of frame	*LPHN2-IFI44*
EGL473	Intergenic, direct	Out of frame	*SHDC-LY9*
EGL492	Intergenic, direct	Out of frame	*BARD1-FN1*
EGL500	Intergenic, direct	Out of frame	*RAF1-TMEM40*
EGL509	Intergenic, direct	Out of frame	*WHSC1-FGFR3*
EGL542	Intergenic, direct	Out of frame	*CACNA2D1-PCLO*
EGL572	Intergenic, direct	Out of frame	*LMX1B-MVB12B*
EGL582	Intergenic, direct	Out of frame	*TEAD1-MICAL2*
EGL598	Intergenic, direct	Out of frame	*PDZRN4-CNTN1*
EGL617	Intergenic, direct	Out of frame	*TRAP1-UBLAD1*
EGL668	Intergenic, direct	Out of frame	*PNPLA4-KAL1*
EGL683	Intergenic, direct	Out of frame	*TAB3-DMD*
EGL692	Intergenic, direct	Out of frame	*XIST-FTX*

Table 4.2. Predicted fusion genes at duplication breakpoints

Chapter 5

Conclusions and Future Studies

Conclusions

This dissertation is focused on understanding how unbalanced translocations, a type of chromosome rearrangement that causes neurodevelopmental disorders, are formed. Analyzing the breakpoint junction sequences from 57 unique translocations revealed molecular mechanisms, and in some cases, structural complexity.

Unbalanced translocation mechanism

Before the start of this thesis, studies of chromosomal SV used chromosome banding, FISH, and array CGH to fine-map breakpoints. Breakpoint junction sequencing is increasingly recognized as a fundamental part of analyzing chromosome SV. At nucleotide resolution, we can make discoveries about DNA breakage and repair mechanisms, as well as identify gene and regulatory elements at the junction. Modern use of NGS such as WGS has enabled more rapid and high-throughput knowledge about SV structure and mechanism.

SV junctions that have extensive sequence homology between breakpoints point to repair by NAHR, while short microhomology or absence of homologous sequence suggests NHEJ ligated the broken DNA ends. More complex features observed at sequenced breakpoint junctions such as template insertions and inversions led to a DNA replication-based model of repair, FoSTeS (Lee et al. 2007). FoSTeS is similar to the MMBIR mechanism that is well understood in yeast (Hastings et al. 2009).

Intrachromosomal CNV studies have described the junctions of a few hundred deletions and duplications (Vissers et al. 2009; Luo et al. 2011; Verdin et al. 2013; Newman et al. 2015). Most intrachromosomal rearrangements are a product of NHEJ and

FoSTeS/MMBIR, but recurrent ones are driven by NAHR between the same loci in unrelated individuals. For deletions, haploinsufficiency for genes within the deleted segment can lead to congenital abnormalities, and genetic triplosensitivity may do the same for duplications. Recurrent CNVs that cause genomic disorders are ideal to correlate genotype with phenotype because the same contiguous genes are deleted or duplicated in unrelated individuals (Watson et al. 2014).

Meanwhile, more than 100 interchromosomal SVs such as translocations have been sequenced (Chen et al. 2008; Higgins et al. 2008; Sobreira et al. 2011; Chiang et al. 2012; Robberecht et al. 2013). Translocations arise from an exchange of genetic material between two non-homologous chromosomes and go on to be inherited in a balanced or unbalanced form. Because balanced translocations do not have genomic copy number changes, it is expected that genes disrupted at the chromosome breakpoint are responsible for a subject's phenotype. Thus, sequencing the breakpoints of *de novo* balanced translocations has led to the discovery of candidate genes related to pediatric disorders (Chen et al. 2008; Higgins et al. 2008).

Though subjects with unbalanced translocations are regularly ascertained, breakpoint junction sequencing for gene discovery is not ideal because genotype-phenotype correlations are complicated by the combined deletions and duplications of hundreds of genes. However, since balanced and unbalanced translocations arise by the same initial events, sequencing the unbalanced form of translocations provides the same opportunity to study rearrangement mechanism. With this dissertation research, we used breakpoint junction sequencing to characterize a diverse group of unbalanced translocations.

As described in Chapter 2, we analyzed 57 unique unbalanced translocations. Using a combination of array CGH and targeted NGS, we found that 34 of 37 sequenced junctions lacked extensive sequence homology. Three unbalanced translocations had breakpoints consistent with NAHR between pairs of LINEs, HERVs, or short SDs; however, most breakpoint junctions had blunt ends, microhomology, inserted sequence, or inversions, indicating that most unbalanced translocations arise by NHEJ or MMBIR (Weckselblatt et al. 2015). These breakpoint signatures are similar to those from over 60 sequenced balanced translocations (Chen et al. 2008; Higgins et al. 2008; Chiang et al. 2012). On the other hand, sequencing junctions of nine unbalanced translocations revealed that six were mediated by NAHR between 6-kb LINE, 3-kb HERV, or 1.7-kb SD pairs that are each >90 identical (Robberecht et al. 2013). Although in this group NAHR between paralogous repeats appeared to be the "driver" of unbalanced translocations, our larger-scale study demonstrated that NAHR is not the major mechanism of translocation formation. We conclude that most unbalanced translocations have simple junctions and form by NHEJ or MMBIR.

Complex translocations and chromothripsis

CCRs are SV that involve at least three breakpoints. Over 250 CCRs have been reported in literature, most of which lack breakpoint sequencing and were initially identified by chromosome banding and/or FISH. Because most CCRs are *de novo* and reported as balanced/copy-neutral, it is expected that the CCR carrier's phenotype is due to genomic alterations in the vicinity of breakpoints (Pellestor et al. 2011). More recent studies that paired FISH with array CGH show that many of these apparently balanced

CCRs harbor microdeletions and/or microduplications (Zhang et al. 2009; Pellestor et al. 2011). Advancing the resolution of CCR structural analysis shows that the abnormal phenotype could be associated with a cryptic genomic imbalance and not necessarily related to the breakpoints. More detailed molecular analysis is required to elucidate the complexity of CCRs and their formation.

Six unbalanced translocations described in this dissertation were originally ascertained as CCRs (Chapter 3). By chromosome banding and FISH, EGL312, EGL356, EGL826, EGL302, EGL321, and EGL305, were recognized as having rearrangements involving multiple segments that translocated between at least two different chromosomes. To characterize their rearrangement structures, we applied NGS to capture breakpoints.

For EGL312 and EGL356's rearrangements, mate-pair sequencing identified an additional translocated segment and an inverted junction, respectively. WGS (Complete Genomics) of EGL826 isolated breakpoints for a maternally inherited simple translocation and mapped an inverted triplication at a *de novo* translocation junction.

Translocations from EGL302, EGL305, and EGL321 involved four or five different chromosomes with as many as 33 breakpoints per rearrangement. These rearrangements had only two large CNVs each, many breakpoints localized to a few genomic regions, translocated segments with oscillating strand orientation, and breakpoint junctions with blunt ends or up to four base pairs of microhomology. These features are hallmarks of constitutional chromothripsis, or chromosome shattering (Kloosterman et al. 2011; Stephens et al. 2011; Kloosterman et al. 2012; Kloosterman and Cuppen 2013; Pellestor et al. 2014).

At the time of this dissertation, one mechanistic explanation of chromothripsis is supported by direct experimental evidence. During anaphase, lagging chromosomes that are far enough away from the main chromatin mass may become compartmentalized into their own micronucleus. Subsequent rupture of the micronucleus causes extensive DNA damage to the missegregated chromosomes (Crasta et al. 2012). By treating cells with a drug that promotes chromosome lag, Zhang et al. performed live imaging to identify cells where a micronucleus formed and then ruptured in the following cell cycle during DNA replication. Single-cell sequencing reveals that only one daughter cell inherits the micronucleated chromosomes, and that their rearrangements can exhibit the defining characteristics of chromothripsis, including clustering of breakpoints, microhomology at many of the breakpoints, and deletion CNVs (Zhang et al. 2015). In this model, one or a few chromosomes will shatter and be reassembled through NHEJ, but the rest of the genome remains intact.

We find that chromothripsis is usually *de novo*, arises on paternal alleles (Kloosterman et al. 2011; Kloosterman et al. 2012; Weckselblatt et al. 2015), and in some cases is transmitted maternally (de Pagter et al. 2015; Weckselblatt et al. 2015). Though more subjects are needed to confirm a paternal bias in constitutional chromothripsis, most *de novo* CNVs are also of paternal origin (Thomas et al. 2006; Thomas et al. 2010), while most familial CCRs are transmitted maternally (Giardino et al. 2006).

Unbalanced translocation consequences

The genomic structural changes from unbalanced translocations may lead to a variety of genetic consequences. In addition to the genomic regions of trisomy and

monosomy, chromosome breakpoints may physically disrupt a gene and cause loss of function. In mostly-balanced chromothripsis genomes, broken genes are particularly likely to be involved in phenotypes because deletions and duplications are relatively minor. The *PTPRD* gene that is broken in three separate constitutional chromothripsis genomes (Macera et al. 2014; de Pagter et al. 2015; Weckselblatt et al. 2015) and is recurrently altered in neuroblastoma chromothripsis (Molenaar et al. 2012; Boeva et al. 2013) warrants further investigation as a chromothripsis hotspot.

It is well-established that translocations are frequently observed in cancers and can result in fusion genes with oncogenic potential. Fusion genes created by germline rearrangements could likewise contribute to intellectual disability and other pediatric phenotypes, but these are rarely reported. At three unbalanced translocation junctions, we observe that the 5' end of one gene is joined to the 3' end of another gene, creating a fusion transcript that is predicted to be in-frame (Weckselblatt et al. 2015). Though the parent genes of *SIRPG-WWOX, SMOC2-PROX1,* and *PIEZO2-MTA1* are not currently implicated in neurodevelopmental syndromes, a fusion may take on novel function.

Duplication CNVs also have the potential to fuse genes at breakpoint junctions. In-frame transcripts are predicted for fusions of *TRPV3-TAX1BP3*, *LTBP1-BIRC6,* and *SOS1-MAP4K3* in simple tandem duplications, *COL4A6-USP20* at the junction of an insertional translocation, and *KCNH5-FUT8* at the inverted junction of two interconnected duplications.

The direct duplication in EGL480 may produce a fusion of *SOS1* and *MAP4K3*. Structural rearrangements that fuse kinase genes are an important class of oncogenes in leukemia and solid tumors (Medves and Demoulin 2012). Here, an activator of Ras,

SOS1, is fused to the kinase *MAP4K3*. EGL480 displays symptoms of Noonan syndrome, which is caused by gain of function mutations in *SOS1*. We hypothesize that the germline *SOS1-MAP4K3* fusion gene also plays a role in EGL480's clinical presentation.

More recent studies of constitutional breakpoint junctions have revealed fusions at duplications (Rippey et al. 2013), deletions (Boone et al. 2014), and at a chromothriptic junction (van Heesch et al. 2014). In the absence of RNA and protein data, a mechanism of disease pathogenesis through these constitutional gene fusions remains unclear, but could be important for potential therapy.

Future directions for translocation studies

Breakpoint junction sequencing has provided insight into the formation of many chromosome rearrangements, including translocations. We've pinpointed genomic locations of breakpoints, identified classes of repetitive DNA that mediate rearrangements, and interpreted genes physically interrupted by breakpoints. However, the current models of translocation formation and other chromosome rearrangement mechanisms are incomplete without considering the role of higher-order genomic organization. Future studies will investigate the association between specific chromatin modifications and nuclear organization of chromosomal regions that rearrange. Expanding this genomic analysis is essential to understand the processes and risk factors involved in translocation formation.

DNA sequence and chromatin at breakpoints

Most constitutional breakpoints do not occur at the same chromosomal location, but common DNA and/or chromatin features underlie some breakpoints. Sequence that has a propensity to form alternative conformations of DNA can predispose breakage (Vissers et al. 2009), such as sites in GC-rich subtelomeres that are predicted to form G-quadruplexes (Bose et al. 2014). Similarly, some common genomic fragile sites are composed of di- or tri-nucleotide repeats that form hairpin and quadruplex secondary structures, leading to replication fork stalling or collapse (Mirkin 2006; Zhang and Freudenreich 2007).

Fragile sites are also an important part of the development of cancer-specific SV. Half of recurrent breakpoints in cancer-associated translocations correspond to fragile sites (Burrow et al. 2009). Two of the most common fragile sites, FRA3B and FRA16D, are located within the tumor suppressor genes *FHIT* and *WWOX,* respectively (Huebner and Croce 2001; Dillon et al. 2010). Furthermore, analysis of six common fragile sites in tumor suppressor genes revealed that they are enriched in histone hypoacetylation and heterochromatic marks relative to the flanking genomic regions (Wang 2006; Jiang et al. 2009).

Other recurrent breakpoints are located in open chromatin, some of which are sites that fuse two different genes together. At the breakpoint junction of the t(9;22) Philadelphia chromosome, the *BCR-ABL1* fusion gene produces a constitutively active *ABL1* tyrosine kinase that leads to acute myeloid leukemia. At the Breakpoint Cluster Region (BCR) locus on Chromosome 22, translocation breakpoints cluster in defined loci. Recurrent translocations form between the BCR and other chromosomes despite an absence of sequence homology between breakpoints; however, these sites do have

chromatin structural elements in common (Zhang and Rowley 2006). BCRs are enriched in DNase I hypersensitivity sites, which are associated with open chromatin and transcription factor binding (Crawford et al. 2006; Zhang and Rowley 2006; Thurman et al. 2012). Studies of chromosome rearrangements in lymphoma, prostate adenocarcinoma, and several breast, ovarian, head and neck, and colorectal cancers concluded that CpG sites at breakpoint regions are hypomethylated relative to adjacent DNA (De and Michor 2011; Grzeda et al. 2014; Lu et al. 2015). It remains to be determined if chromosome breakpoints that form constitutional SV are associated with marks of either active or silenced chromatin. It is tempting to speculate that interchromosomal breakpoints share common chromatin features that are colocalized and/or coregulated in the nucleus.

Spatial organization of chromosomes

For a translocation to occur, breakpoints must happen simultaneously on two different chromosomes, followed by physical contact between double-strand breaks and aberrant joining of nonhomologous chromosomes. Because nuclear proximity between translocating regions is a requirement for interchromosomal SV, we expect that the 3-dimensional organization of the genome can play an important role in how translocations form. The frequencies of experimentally-induced translocations and regions that recurrently translocate in cancers are elevated for loci located near one another in the nucleus (Roix et al. 2003; Chiarle et al. 2011; Klein et al. 2011; Engreitz et al. 2012; Zhang et al. 2012; Roukos et al. 2013), but there is limited information on how this influences constitutional translocations (Bickmore and Teague 2002).

Chromosome location in the nucleus is linked to transcriptional regulation. Silenced genes are localized to the nuclear periphery while active genes are located in the interior of the nucleus (Peric-Hupkes et al. 2010). Regions of the genome composed of active genes have a high density of DNase I hypersensitivity sites and tend to associate with one another in the interphase nucleus (Bickmore 2013; Fanucchi et al. 2013). We hypothesize that translocations where both breakpoints are located within genes may arise because those genes are co-localized in the nucleus. Indeed, our work revealed that 44 out of 102 of simple translocation breakpoints lie in genes (Weckselblatt et al. 2015).

High-resolution Hi-C mapping has identified genome-wide chromatin contacts in fly, mouse, and human chromosomes that are organized in topologically associating domains (TADs). TADs fold into discrete compartments where there is high interaction within a TAD but little to no interaction between different TADs (Dixon et al. 2012; Sexton et al. 2012; Dekker et al. 2013). TAD borders are also enriched for active genes and CTCF, which is implicated in maintaining TAD structure (Dixon et al. 2012; Sexton et al. 2012; Giorgetti et al. 2014).

We predict that linearly distant DNA that is cinched together by the CTCF at TAD borders has the potential to be breakpoint sites of intrachromosomal rearrangements such as interstitial deletions and duplications. Due to the high frequency of interactions within a TAD and at TAD borders, double-strand breaks may favor intra-TAD repair due to spatial constraints. Future studies will compare TAD boundaries and SV breakpoints in search of a correlation, providing insight into the mechanisms of chromosome rearrangements.

Position effect

Unbalanced translocations lead to changes in gene dosage, loss of function of genes physically disrupted by breakpoints, and gain of function of gene hybrids formed at breakpoint junctions. Position effect is another consequence of altering chromosome structure, where intact genes adjacent to breakpoints are subject to new regulatory machinery present at the translocation site. In unbalanced translocations between an autosome and an X Chromosome with an intact X inactivation center (XIC), X inactivation spreads to silence adjacent autosomal DNA (Mattei et al. 1982). *Xist*, a product of the XIC, converts the translocation chromosome to a heterochromatic state, spreading up to 45 megabases from the translocation breakpoint (Sharp et al. 2002). Heterochromatic silencing of the autosomal trisomic segment leads to a milder phenotype. In another case, a balanced translocation between a heterochromatic band of Chromosome 15 and a euchromatic band of Chromosome 16 resulted in a neurological phenotype. Here, the spread of heterochromatin across the translocation junction silenced expression of genes derived from chromosome 16 (Finelli et al. 2012).

When genes are placed into a new nuclear position with anomalous chromatin environments, position effects may also lead to expression changes on a larger scale. In the recurrent unbalanced translocation between Chromosomes 11 and 22, *trans*-effects of aberrant nuclear positioning leads to differential expression of many normal copy number genes on different chromosomes (Harewood et al. 2010). Future studies of *cis* and *trans* position effects related to translocations may inform phenotypes when even thorough breakpoint analysis fails to pinpoint genes related to disease.

Concluding remarks

In this large-scale analysis of unbalanced translocations, breakpoint junction sequencing reveals mutational mechanisms, structural complexity, and novel in-frame fusion genes. With a scarcity of sequence homology at breakpoint junctions, most unbalanced translocations likely formed by NHEJ and MMBIR repair processes. Our approach to combine targeted NGS, mate-pair sequencing, and WGS uncovered a wide range of breakpoints in this diverse cohort. Rarer translocations between four or five chromosomes proved to have tens of breakpoints, most of which were not recognized by standard cytogenetic methods. These chromothripsis rearrangements are usually *de novo*, arise on paternal alleles, and transmit maternally. Future SV studies will combine two-dimensional structural analysis with higher-order genomic organization to bring us closer to elucidating the molecular processes underlying how these rearrangements form.

References

Bickmore WA. 2013. The spatial organization of the human genome. *Annu Rev Genomics Hum Genet* **14**: 67-84.

Bickmore WA, Teague P. 2002. Influences of chromosome size, gene density and nuclear position on the frequency of constitutional translocations in the human population. *Chromosome Res* **10**(8): 707-715.

Boeva V, Jouannet S, Daveau R, Combaret V, Pierre-Eugene C, Cazes A, Louis-Brennetot C, Schleiermacher G, Ferrand S, Pierron G et al. 2013. Breakpoint features of genomic rearrangements in neuroblastoma with unbalanced translocations and chromothripsis. *PLoS One* **8**(8): e72182.

Boone PM, Yuan B, Campbell IM, Scull JC, Withers MA, Baggett BC, Beck CR, Shaw CJ, Stankiewicz P, Moretti P et al. 2014. The Alu-rich genomic architecture of SPAST predisposes to diverse and functionally distinct disease-associated CNV alleles. *Am J Hum Genet* **95**(2): 143-161.

Borg K, Stankiewicz P, Bocian E, Kruczek A, Obersztyn E, Lupski JR, Mazurczak T. 2005. Molecular analysis of a constitutional complex genome rearrangement with 11 breakpoints involving chromosomes 3, 11, 12, and 21 and a approximately 0.5-Mb submicroscopic deletion in a patient with mild mental retardation. *Hum Genet* **118**(2): 267-275.

Bose P, Hermetz KE, Conneely KN, Rudd MK. 2014. Tandem repeats and G-rich sequences are enriched at human CNV breakpoints. *PLoS One* **9**(7): e101607.

Burrow AA, Williams LE, Pierce LC, Wang YH. 2009. Over half of breakpoints in gene pairs involved in cancer-specific recurrent translocations are mapped to human chromosomal fragile sites. *BMC Genomics* **10**: 59.

Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, Schulz MH, Erdogan F, Li N, Kijas Z, Arkesteijn G et al. 2008. Mapping translocation breakpoints by next-generation sequencing. *Genome Res* **18**(7): 1143-1149.

Chiang C, Jacobsen JC, Ernst C, Hanscom C, Heilbut A, Blumenthal I, Mills RE, Kirby A, Lindgren AM, Rudiger SR et al. 2012. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat Genet* **44**(4): 390-397, S391.

Chiarle R, Zhang Y, Frock RL, Lewis SM, Molinie B, Ho YJ, Myers DR, Choi VW, Compagno M, Malkin DJ et al. 2011. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**(1): 107-119.

Crasta K, Ganem NJ, Dagher R, Lantermann AB, Ivanova EV, Pan Y, Nezi L, Protopopov A, Chowdhury D, Pellman D. 2012. DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482**(7383): 53-58.

Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**(1): 123-131.

de Pagter MS, van Roosmalen MJ, Baas AF, Renkens I, Duran KJ, van Binsbergen E, Tavakoli-Yaraki M, Hochstenbach R, van der Veken LT, Cuppen E et al. 2015a.

Chromothripsis in healthy individuals affects multiple protein-coding genes and can result in severe congenital abnormalities in offspring. *Am J Hum Genet* **96**(4): 651-656.

De S, Michor F. 2011. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat Struct Mol Biol* **18**(8): 950-955.

Dekker J, Marti-Renom MA, Mirny LA. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**(6): 390-403.

Dillon LW, Burrow AA, Wang YH. 2010. DNA instability at chromosomal fragile sites in cancer. *Curr Genomics* **11**(5): 326-337.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398): 376-380.

Engreitz JM, Agarwala V, Mirny LA. 2012. Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. *PLoS One* **7**(9): e44196.

Fanucchi S, Shibayama Y, Burd S, Weinberg MS, Mhlanga MM. 2013. Chromosomal contact permits transcription between coregulated genes. *Cell* **155**(3): 606-620.

Finelli P, Sirchia SM, Masciadri M, Crippa M, Recalcati MP, Rusconi D, Giardino D, Monti L, Cogliati F, Faravelli F et al. 2012. Juxtaposition of heterochromatic and euchromatic regions by chromosomal translocation mediates a heterochromatic long-range position effect associated with a severe neurological phenotype. *Mol Cytogenet* **5**: 16.

Giardino D, Corti C, Ballarati L, Finelli P, Valtorta C, Botta G, Giudici M, Grosso E, Larizza L. 2006. Prenatal diagnosis of a de novo complex chromosome rearrangement (CCR) mediated by six breakpoints, and a review of 20 prenatally ascertained CCRs. *Prenat Diagn* **26**(6): 565-570.

Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E. 2014. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**(4): 950-963.

Grzeda KR, Royer-Bertrand B, Inaki K, Kim H, Hillmer AM, Liu ET, Chuang JH. 2014. Functional chromatin features are associated with structural mutations in cancer. *BMC Genomics* **15**: 1013.

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**(5): 554-564.

Hastings PJ, Ira G, Lupski JR. 2009. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet* **5**(1): e1000327.

Higgins AW, Alkuraya FS, Bosco AF, Brown KK, Bruns GA, Donovan DJ, Eisenman R, Fan Y, Farra CG, Ferguson HL et al. 2008. Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. *Am J Hum Genet* **82**(3): 712-722.

Huebner K, Croce CM. 2001. FRA3B and other common fragile sites: the weakest links. *Nat Rev Cancer* **1**(3): 214-221.

Jiang Y, Lucas I, Young DJ, Davis EM, Karrison T, Rest JS, Le Beau MM. 2009. Common fragile sites are characterized by histone hypoacetylation. *Hum Mol Genet* **18**(23): 4501-4512.

Klein IA, Resch W, Jankovic M, Oliveira T, Yamane A, Nakahashi H, Di Virgilio M, Bothmer A, Nussenzweig A, Robbiani DF et al. 2011. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* **147**(1): 95-106.

Kloosterman WP, Cuppen E. 2013. Chromothripsis in congenital disorders and cancer: similarities and differences. *Curr Opin Cell Biol* **25**(3): 341-348.

Kloosterman WP, Guryev V, van Roosmalen M, Duran KJ, de Bruijn E, Bakker SC, Letteboer T, van Nesselrooij B, Hochstenbach R, Poot M et al. 2011. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum Mol Genet* **20**(10): 1916-1924.

Kloosterman WP, Tavakoli-Yaraki M, van Roosmalen MJ, van Binsbergen E, Renkens I, Duran K, Ballarati L, Vergult S, Giardino D, Hansson K et al. 2012. Constitutional chromothripsis rearrangements involve clustered double-stranded DNA breaks and nonhomologous repair mechanisms. *Cell Rep* **1**(6): 648-655.

Lee JA, Carvalho CM, Lupski JR. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**(7): 1235-1247.

Lu Z, Lieber MR, Tsai AG, Pardo CE, Muschen M, Kladde MP, Hsieh CL. 2015. Human lymphoid translocation fragile zones are hypomethylated and have accessible chromatin. *Mol Cell Biol* **35**(7): 1209-1222.

Luo Y, Hermetz KE, Jackson JM, Mulle JG, Dodd A, Tsuchiya KD, Ballif BC, Shaffer LG, Cody JD, Ledbetter DH et al. 2011. Diverse mutational mechanisms cause pathogenic subtelomeric rearrangements. *Hum Mol Genet* **20**(19): 3769-3778.

Macera MJ, Sobrino A, Levy B, Jobanputra V, Aggarwal V, Mills A, Esteves C, Hanscom C, Pereira S, Pillalamarri V et al. 2014. Prenatal diagnosis of chromothripsis, with nine breaks characterized by karyotyping, FISH, microarray and whole-genome sequencing. *Prenat Diagn*.

Mattei MG, Mattei JF, Ayme S, Giraud F. 1982. X-autosome translocations: cytogenetic characteristics and their consequences. *Hum Genet* **61**(4): 295-309.

Mirkin SM. 2006. DNA structures, repeat expansions and human hereditary disorders. *Curr Opin Struct Biol* **16**(3): 351-358.

Molenaar JJ, Koster J, Zwijnenburg DA, van Sluis P, Valentijn LJ, van der Ploeg I, Hamdi M, van Nes J, Westerman BA, van Arkel J et al. 2012. Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes. *Nature* **483**(7391): 589-593.

Newman S, Hermetz KE, Weckselblatt B, Rudd MK. 2015. Next-Generation Sequencing of Duplication CNVs Reveals that Most Are Tandem and Some Create Fusion Genes at Breakpoints. *Am J Hum Genet* **96**(2): 208-220.

Pellestor F, Anahory T, Lefort G, Puechberty J, Liehr T, Hedon B, Sarda P. 2011. Complex chromosomal rearrangements: origin and meiotic behavior. *Hum Reprod Update* **17**(4): 476-494.

Pellestor F, Gatinois V, Puechberty J, Genevieve D, Lefort G. 2014. Chromothripsis: potential origin in gametogenesis and preimplantation cell divisions. A review. *Fertil Steril* **102**(6): 1785-1796.

Peric-Hupkes D, Meuleman W, Pagie L, Bruggeman SW, Solovei I, Brugman W, Graf S, Flicek P, Kerkhoven RM, van Lohuizen M et al. 2010. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* **38**(4): 603-613.

Rippey C, Walsh T, Gulsuner S, Brodsky M, Nord AS, Gasperini M, Pierce S, Spurrell C, Coe BP, Krumm N et al. 2013. Formation of chimeric genes by copy-number variation as a mutational mechanism in schizophrenia. *Am J Hum Genet* **93**(4): 697-710.

Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, Vermeesch JR. 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res* **23**(3): 411-418.

Roix JJ, McQueen PG, Munson PJ, Parada LA, Misteli T. 2003. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet* **34**(3): 287-291.

Roukos V, Voss TC, Schmidt CK, Lee S, Wangsa D, Misteli T. 2013. Spatial dynamics of chromosome translocations in living cells. *Science* **341**(6146): 660-664.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**(3): 458-472.

Shapiro P. 2002. Ras-MAP kinase signaling pathways and control of cell proliferation: relevance to cancer therapy. *Crit Rev Clin Lab Sci* **39**(4-5): 285-330.

Sharp AJ, Spotswood HT, Robinson DO, Turner BM, Jacobs PA. 2002. Molecular and cytogenetic analysis of the spreading of X inactivation in X;autosome translocations. *Hum Mol Genet* **11**(25): 3145-3156.

Sobreira NL, Gnanakkan V, Walsh M, Marosy B, Wohler E, Thomas G, Hoover-Fong JE, Hamosh A, Wheelan SJ, Valle D. 2011. Characterization of complex chromosomal rearrangements by targeted capture and next-generation sequencing. *Genome Res* **21**(10): 1720-1727.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA et al. 2011. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**(1): 27-40.

Tartaglia M, Pennacchio LA, Zhao C, Yadav KK, Fodale V, Sarkozy A, Pandit B, Oishi K, Martinelli S, Schackwitz W et al. 2007. Gain-of-function SOS1 mutations cause a distinctive form of Noonan syndrome. *Nat Genet* **39**(1): 75-79.

Thomas NS, Durkie M, Van Zyl B, Sanford R, Potts G, Youings S, Dennis N, Jacobs P. 2006. Parental and chromosomal origin of unbalanced de novo structural chromosome abnormalities in man. *Hum Genet* **119**(4): 444-450.

Thomas NS, Morris JK, Baptista J, Ng BL, Crolla JA, Jacobs PA. 2010. De novo apparently balanced translocations in man are predominantly paternal in origin and associated with a significant increase in paternal age. *J Med Genet* **47**(2): 112-115.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**(7414): 75-82.

van Heesch S, Simonis M, van Roosmalen MJ, Pillalamarri V, Brand H, Kuijk EW, de Luca KL, Lansu N, Braat AK, Menelaou A et al. 2014. Genomic and Functional Overlap between Somatic and Germline Chromosomal Rearrangements. *Cell Rep* **9**(6): 2001-2010.

Verdin H, D'Haene B, Beysen D, Novikova Y, Menten B, Sante T, Lapunzina P, Nevado J, Carvalho CM, Lupski JR et al. 2013. Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. *PLoS Genet* **9**(3): e1003358.

Vissers LE, Bhatt SS, Janssen IM, Xia Z, Lalani SR, Pfundt R, Derwinska K, de Vries BB, Gilissen C, Hoischen A et al. 2009. Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture. *Hum Mol Genet* **18**(19): 3579-3593.

Wang YH. 2006. Chromatin structure of human chromosomal fragile sites. *Cancer Lett* **232**(1): 70-78.

Watson CT, Marques-Bonet T, Sharp AJ, Mefford HC. 2014. The genetics of microdeletion and microduplication syndromes: an update. *Annu Rev Genomics Hum Genet* **15**: 215-244.

Weckselblatt B, Hermetz KE, Rudd MK. 2015. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res*.

Zenker M, Horn D, Wieczorek D, Allanson J, Pauli S, van der Burgt I, Doerr HG, Gaspar H, Hofbeck M, Gillessen-Kaesbach G et al. 2007. SOS1 is the second most common Noonan gene but plays no major role in cardio-facio-cutaneous syndrome. *J Med Genet* **44**(10): 651-656.

Zhang CZ, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, Meyerson M, Pellman D. 2015. Chromothripsis from DNA damage in micronuclei. *Nature* **522**(7555): 179-184.

Zhang F, Carvalho CM, Lupski JR. 2009. Complex human chromosomal and genomic rearrangements. *Trends Genet* **25**(7): 298-307.

Zhang H, Freudenreich CH. 2007. An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae. *Mol Cell* **27**(3): 367-379.

Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, Simon AC, Becker MS, Alt FW, Dekker J. 2012. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**(5): 908-921.

Zhang Y, Rowley JD. 2006. Chromatin structural elements and chromosomal translocations in leukemia. *DNA Repair (Amst)* **5**(9-10): 1282-1297.