Consideration of the haplotype diversity at nonallelic homologous recombination hotspots improves the precision of rearrangement breakpoint identification

Morten Hillmer[1], Anna Summerer[1], Victor-Felix Mautner[2], Josef Högel[1], David N. Cooper[3], Hildegard Kehrer-Sawatzki[1]

[1]Institute of Human Genetics, University of Ulm, 89081 Ulm, Germany
[2]Department of Neurology, University Hospital Hamburg Eppendorf, 20246 Hamburg, Germany
[3]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK

Abbreviations:
AHR: allelic homologous recombination; BSP: breakpoint-spanning PCR product; DSB: double strand break; dHj: double Holliday junction; hDNA: heteroduplex DNA; NAHR: nonallelic homologous recombination; NAHGC: nonallelic homologous gene conversion without crossover; PRS1: paralogous recombination site 1; PRS2: paralogous recombination site 2; PSV: paralogous sequence variant; SER: strand exchange region; SDSA: synthesis-dependent strand annealing.

Grants: Deutsche Forschungsgemeinschaft (DFG) KE 724/12-2

Corresponding author:
Prof. Dr. Hildegard Kehrer-Sawatzki, PhD
Institute of Human Genetics
University of Ulm
Albert-Einstein-Allee 11
89081 Ulm, Germany
Phone: 0049 731 50065421
hildegard.kehrer-sawatzki@uni-ulm.de

Abstract

Nonallelic homologous recombination (NAHR) is the major mutational mechanism underlying recurrent copy number variants in humans. ~~The p~~Precise characterization of the associated breakpoints is key to identifying those features that influence NAHR frequency. Until now, high-resolution breakpoint analysis of NAHR-mediated rearrangements has generally been performed by comparison of the breakpoint-spanning sequences with the human genome reference sequence. However, we show here that the haplotype diversity of individual NAHR hotspots may interfere with breakpoint-mapping. We studied the transmitting parents of individuals with germline type-1 *NF1* deletions mediated by NAHR within the PRS1 or PRS2 hotspots. Several parental wildtype PRS1 and PRS2 haplotypes were identified that exhibited considerable sequence diversity with respect to the reference sequence, and these haplotypes also affected the number of predicted PRDM9-binding sites. Sequence comparisons between the parental wildtype PRS1 or PRS2 haplotypes and the deletion breakpoint-spanning sequences from the patients turned out to be an accurate means to assign *NF1* deletion breakpoints and proved superior to crude reference sequence comparisons that neglect to consider haplotype diversity. Our findings imply that both paralog-specific haplotype diversity patterns of NAHR hotspots (such as PRS2) and population-specific haplotype diversity must be taken into account in order to accurately ascertain NAHR-mediated rearrangement breakpoints.

Introduction

Nonallelic homologous recombination with crossover (NAHR) gives rise to numerous (and sometimes recurrent) disease-associated copy number variants (CNVs) [Mefford and Eichler, 2009; Stankiewicz and Lupski, 2010; Watson et al., 2014; Carvalho and Lupski, 2016]. The sequence substrates for NAHR include segmental duplications, low-copy repeats (LCRs) or L1- and HERV-elements that all exhibit high inter-paralogous sequence similarity (>97% identity) [Kamp et al., 2000; Sun et al., 2000; Sharp et al., 2005; Liu et al., 2011; Shuvarikov et al., 2013; Campbell et al., 2014; Startek et al., 2015]. Meiotic NAHR between intrachromosomal paralogs is thought to be very similar mechanistically to allelic homologous recombination (AHR) during meiosis, a process that is also associated with crossover but which does not generate CNVs [Lopes et al., 1999; Lupski, 2004; Sasaki et al., 2010]. As with AHR, NAHR also occurs within recombination hotspots of a few kilobases (kb) as determined by high resolution breakpoint analysis of CNVs [Reiter et al., 1996, 1998; Lopez-Correa et al., 2001; Bi et al., 2003; Bayés et al., 2003; Bosch and Jobling, 2003; Visser et al., 2005; De Raedt et al., 2006; Lindsay et al., 2006; Turner et al., 2008; Shinawi et al., 2009; Szafranski et al., 2010; Koscinski et al., 2011; Elinati et al., 2012; Coutton et al., 2013; Dittwald et al., 2013; Bengesser et al., 2014]. In these studies, NAHR breakpoints were identified by sequence analysis of breakpoint-spanning PCR products (BSPs) followed by comparison of these BSP sequences with the human genome reference sequence. By these means were determined the strand exchange regions (SERs) between the recombining paralogs, which constitute the rearrangement breakpoints. This method that was used to identify NAHR-mediated breakpoints is however dependent upon the presence of non-polymorphic sequence differences (termed *cis*-morphisms or paralogous sequence variants, PSVs) that serve to distinguish the recombining paralogs. The accurate determination of the SERs of NAHR-mediated rearrangements is important in terms of being able to identify possible correlations between certain DNA sequence features, such as PRDM9-binding sites, and NAHR frequency. PRDM9, a DNA-binding histone methyltransferase, regulates the initiation of AHR (and probably NAHR as well) by ensuring ~~the proper~~ an appropriate [HILDE: suitable?] chromatin and spatial environment for subsequent recombination events [Berg et al., 2010; Paranov et al., 2017].

As mentioned above, in most studies performed to date, high-resolution breakpoint analysis of CNVs has involved the direct comparison of BSP sequences with the reference sequence of the human genome (henceforth termed method #1). Previously, we also employed this method to identify the SERs of 68 germline type-1 *NF1* deletions at 17q11.2 which were mediated by NAHR between the LCRs NF1-REPa and NF1-REPc and exhibited breakpoints located within the NAHR hotspots PRS1 and PRS2 [Hillmer et al., 2016]. Large *NF1* deletions can be of different types (types 1–3 or atypical) and together ~~are responsible~~underlie ~~for~~ chromosome 17q11.2 deletion syndrome (MIM# 613675) [reviewed by Kehrer-Sawatzki et al., 2017]. Type-1 *NF1* deletions are predominantly of meiotic origin and occur in the germlines of healthy parents who then transmit the deletions to their offspring [Lopez-Correa et al., 2000; Messiaen et al., 2011]. In our previous study, we analysed the wildtype (non-recombinant) PRS1 and PRS2 sequences of the transmitting parents in ~~the case of~~ 8 of the 68 type-1 *NF1* deletions to investigate NAHR-associated gene conversion [Hillmer et al., 2016]. During this earlier analysis, we observed differences between some of the parental haplotypes (defined here as segments of DNA of specific length harbouring multiple variants that serve to distinguish between the DNA segments) and the reference sequences of NF1-REPa and NF1-REPc. In the present study ~~presented here~~, we have investigated the haplotype diversity of non-recombinant PRS1 and PRS2 sequences in great detail in order to ascertain whether haplotype-specific sequence diversity might influence SER determination within these NAHR hotspots. Accurate SER determination is important for any attempt to correlate the breakpoint location with specific DNA sequence features ~~and to~~in the hope of obtaining novel insights

into the sequence determinants of NAHR. To ~~do so~~this end, we fully sequenced the wildtype PRS1 and PRS2 sequences of the transmitting parents of 25 type-1 *NF1* deletions. Subsequently, we compared the non-recombinant parental haplotypes with the BSP sequences in their children, the *NF1* deletion patients, in order to determine the SERs of the respective deletions. Our aim was to investigate whether this SER identification method (henceforth termed method #2) was more accurate than method #1 which relies upon ~~the~~ a crude comparison of BSP and reference sequence without taking potential haplotype sequence diversity into account. By contrast, method #2 is personalized by virtue of a direct comparison with the non-recombinant parental haplotypes rather than with the standard human genome reference sequence which was assembled from multiple individuals and hence represents a mosaic haploid genome. Further, we investigated whether there might be a correlation between the number of nucleotide differences exhibited by the PRS1 and PRS2 haplotypes of the transmitting parents and whether these haplotypes were involved in the deletion-causing NAHR events. We also compared the BSP sequences from the patients with the parental haplotypes involved in the deletion-causing NAHR events in order to investigate the occurrence of NAHR-associated mutations of single nucleotides within the breakpoint-flanking sequences.

Materials and Methods

Patients and transmitting parents

We analysed the breakpoint-spanning PCR products (BSPs) of 25 patients with type-1 *NF1* deletions who were all of White European origin. Nineteen patients exhibited breakpoints (strand exchange regions; SERs) within the NAHR hotspot PRS2 and six patients had deletion breakpoints located within PRS1. The SERs of 19 of the 25 *NF1* deletions have been previously analysed by means of method #1 [Hillmer et al., 2016] and were reinvestigated by method #2 during the course of the present study ~~presented here~~ (Supp. Table S1). In addition, the SERs of six patients were newly analysed by both methods #1 and #2 during the course of this study. The 25 type-1 *NF1* deletions were initially identified by FISH of blood-derived cells and MLPA using DNA samples prepared from blood (P122 NF1 area probemix, version C2, MRC Holland, The Netherlands). Somatic mosaicism with normal cells was not detected using these methods. The parental origin of the *NF1* deletions was determined by microsatellite marker analysis using blood-derived DNA from the parents (data available upon request).

African DNA samples

The African DNA samples investigated in this study, in order to ascertain PRS2 haplotype diversity, are listed in Supp. Table S2.

PCR amplification and sequence analysis of BSPs and wildtype PRS1 and PRS2 fragments

Breakpoint-spanning PCR products (BSPs) from the patients, and PCR products from the wildtype (non-recombinant) PRS1 and PRS2 fragments spanning the PRS1 and PRS2 NAHR hotspots of the transmitting parents, were amplified from blood-derived DNA samples using the primers listed in Supp. Tables S3 and S4. PCR products from the wildtype PRS2 fragments were amplified from the African DNA samples using the primers listed in Supp. Table S4. PCRs were performed by means of the Expand Long Range dNTPack (Roche, Sigma Aldrich, Munich, Germany) and 400 ng genomic DNA as template. The PCR products were sequenced by Sanger sequencing with the primers listed in Supp. Table S5.

Phase determination of the PRS1 and PRS2 haplotypes

The phase of heterozygous SNPs was determined in all 25 transmitting parents by cloning the wildtype PRS1 and PRS2 PCR fragments using the StrataClone PCR Cloning Kit (Agilent Technologies, Santa Clara, CA, USA) followed by sequence analysis of at least three cloned PCR products. Additionally, nested PCRs were performed in order to determine the wildtype PRS1 and PRS2 haplotypes, employing the paralog-specific and allele-specific primers listed in Supp. Table S6. The nested PCRs were performed using the wildtype PRS1 and PRS2 PCR fragments amplified with the primers listed in Supp. Table S4 as PCR template. Sequence analysis of the nested PCR products using the primers given in Supp. Table S5 revealed the phase of the heterozygous sequence variants.

SER determination of type-1 *NF1* deletions according to method #1 and method #2

The assignment of the strand exchange regions (SERs) of type-1 *NF1* deletions by means of method #1 ~~comprised~~ involved the sequence analysis of BSPs amplified from blood-derived DNA samples of the patients with subsequent comparison of the BSP sequences with the reference sequences of PRS1 or PRS2 in NF1-REPa and NF1-REPc according to the human genome assembly 19 (GRCh 37; hg19). During this analysis, we considered only those sequence differences between the BSPs and the reference sequence ~~could be considered~~ that occurred at sites of (i) paralogous sequence variants (PSVs), which are non-polymorphic sequence differences between NF1-REPa and NF1-REPc, and (ii) SNPs [HILDE: rare variants!] with a minor allele frequency (MAF) ≤1%. Method #2 involved precisely the same analysis but additionally, the wildtype (non-recombinant) PRS1 or PRS2 sequences of the transmitting parents of the *NF1* deletion patients were compared with the BSP sequences derived from their offspring, i.e. the patients with germline type-1 *NF1* deletions. By means of method #2, all nucleotide differences between the aligned sequences could be evaluated, including SNPs [HILDE: rare variants!] with an MAF >1%.

Predicted PRDM9 binding sites and recombination- as well as replication-associated sequence motifs

All PRS1 and PRS2 haplotypes identified in the transmitting parents of patients with type-1 *NF1* deletions, and also in the African DNA samples, were investigated for the presence of potential PRDM9 A-variant binding sites with the consensus sequence 5'-CCNCCNTNNCCNC-3' [Myers et al., 2008] by means of the sequence motif search tool ‘Find Individual Motif Occurrences’ (FIMO) (http://meme-suite.org/tools/fimo). This software was also used to search for recombination- and replication-associated sequence motifs as described by Abeysinghe et al. [2003], Badge et al. [2000] and Visser et al. [2005].

Statistical analysis

A potential correlation between the numbers of nucleotide differences exhibited by the PRS1 or PRS2 haplotypes and whether the haplotypes were involved in the deletion-causing NAHR event was evaluated by employing the Kendall's tau-b correlation coefficient as well as the Wilcoxon rank-sum test using the SAS 9.3 software. The same statistical methods were used to ~~ascertain~~ assess the~~a~~ potential correlation between the number of nucleotide differences per PRS2 haplotype and the presence of PRDM9-binding motif A4.

Results

The transmitting parents of the 25 patients with germline type-1 *NF1* deletions were determined by microsatellite marker analysis. Twenty-two of the 25 deletions analysed were of maternal origin. The analysis of siblings of the deletion patients was possible in five cases and suggested that the deletions occurred by interchromosomal NAHR during maternal meiosis (data not shown).

SER determination by means of methods #1 and #2

The strand exchange regions (SERs) of 25 type-1 *NF1* deletions were comparatively analysed by both methods #1 and #2. Whereas method #1 relies upon the comparison between the breakpoint-spanning PCR product (BSP) sequence and the human genome reference sequence, method #2 is individualized in that it also compares the BSPs with the wildtype (non-recombinant) PRS1 or PRS2 sequences from the transmitting parents of the patients with type-1 *NF1* deletions. Employing method #2, all nucleotide differences could be evaluated; not only PSVs and rare variants [HILDE: you have changed this here but not elsewhere in the manuscript!] with a minor allele frequency (MAF) ≤1%, but also informative SNPs and indels with an MAF >1%. The latter are in most instances 'shared SNPs' that exhibit the same alternative alleles at paralogous sites in NF1-REPa and NF1-REPc. The evaluation of shared SNPs in order to localize SERs within the BSP sequences is only made possible through comparison with the wildtype PRS1 or PRS2 sequences of the transmitting parents amplified from either NF1-REPa or NF1-REPc. These paralog-specific PRS1 or PRS2 haplotypes indicate the origin of nucleotides within the BSP sequences, i.e. whether they were derived from NF1-REPa or NF1-REPc.

Since all sequence differences can be evaluated by method #2, higher resolution and greater accuracy of SER mapping ~~was~~ were to be expected. Indeed, method #2 proved to be more accurate in detecting SERs than method #1. The results of the SER demarcation for the 25 type-1 *NF1* deletions (19 PRS2-mediated and 6 PRS1-mediated) by means of method #2 are summarized in Supp. Tables S7 and S8. Only in one deletion case (patient ID 1598; Table 1) was the ascertained SER location ~~was~~ the same when applying both method #1 and method #2. By contrast, in 22 of the 25 *NF1* deletions analysed, the SERs assigned by method #2 were more precise and shorter that the SERs determined by method #1. In these 22 deletions, the mean SER length was 502-bp according to method #1 and 269-bp according to method #2. Thus, a substantial ~~refinement~~ reduction in SER length (of 233-bp) [HILDE: Worth also expressing as a percentage improvement?] was achieved using method #2 (Table 1). The superiority of method #2 as compared with method #1 in terms of the precision of SER mapping became particularly apparent during the analysis of the deletions of patients LL-2476 and 1547. In these patients, the SER location was incorrectly assigned by method #1 since the transmitting parents harboured PRS2 haplotypes which exhibited several nucleotide differences as compared to the reference sequence. Hence, only method #2 enabled the correct assignment of the *NF1* deletion-associated SERs in patients LL-2476 and 1547 (Figure 1).

NAHR-associated mutations and gene conversion in BSPs

The comparison of the patient-derived BSP sequence with the parental PRS1 or PRS2 haplotypes involved in the NAHR events did not reveal any sequence variants that were present exclusively in the BSP and hence absent from the parental haplotypes. Thus, no NAHR-associated *de novo* mutations were detected in the breakpoint-flanking sequences of the patients. In the BSP of patient 3662, we observed a single nucleotide difference by comparison with the PRS1 haplotype from the transmitting parent. However, this nucleotide difference was most likely caused by NAHR-associated gene conversion templated by its paralogue (Supp. Table S8).

Diversity of parental PRS1 and PRS2 haplotypes and NAHGC

None of the wildtype PRS2 haplotypes of the transmitting parents were identical to the reference sequence of PRS2. Indeed, some of the PRS2 haplotypes exhibited very considerable sequence diversity by comparison with the reference sequence (Supp. Tables S9-S14). Thus, of the 38 PRS2 haplotypes derived from NF1-REPa, ten exhibited 14-19 nucleotide differences compared to the reference sequence within a region of 3663-bp. In similar vein, 36 of the 38 PRS2 haplotypes derived from NF1-REPc exhibited ≥ 5 nucleotide

differences compared to the reference sequence within a stretch of 3663-bp (Supp. Table S9). Nonallelic homologous gene conversion without crossover (NAHGC) is known to be responsible for frequent sequence exchange between recombinationally-active paralogous sequences and for the occurrence of shared SNPs [Fredman et al., 2004; Hallast et al., 2005; Pavlicek et al., 2005; Dumont, 2015]. In the study presented here, 75% of all SNPs within PRS1 and 88% of all SNPs within PRS2 were shared between NF1-REPa and NF1-REPc (Supp. Table S15). The high numbers of shared SNPs as well as the pattern of nucleotide differences between the haplotypes, indicate that NAHGC between NF1-REPa and NF1-REPc must have contributed strongly to the haplotype diversity evident in both PRS1 and PRS2 (Supp. Table S16).

The reference PRS2 haplotypes from NF1 REPa and NF1-REPc exhibit 98.75% sequence identity. Pairwise comparisons between the parental PRS2 haplotypes from NF1 REPa and NF1-REPc involved in the deletion-causing NAHR events are indicative of 98.57%–98.93% sequence identity (Supp. Table S17). Thus, the sequence diversity of the parental PRS2 haplotypes observed did not reduce their overall sequence homology by more than 0.18%. These pairwise comparisons between parental PRS2 haplotypes from NF1-REPa and NF1-REPc indicated that in five of the 11 pairs, the identity between the paralogs was even higher than between the reference PRS2 haplotypes of NF1-REPa and NF1-REPc (Supp. Table S17).

PRS1 and PRS2 haplotypes involved in NAHR with crossover

In 14 of the 19 deletions analysed, the comparison of the parental haplotypes with the BSP sequences of their offspring (i.e. the patients harbouring the type-1 *NF1* deletions) yielded unambiguously the parental haplotypes that had been involved in the NAHR events causing the *NF1* deletions (Supp. Tables S18-S21). However, no correlation was observed between the numbers of nucleotide differences exhibited by the parental PRS1 and PRS2 haplotypes relative to the reference sequences and whether the parental haplotypes had been involved in the deletion-causing NAHR events. In other words, haplotypes with both high ~~or~~ and low sequence diversity as compared with the reference sequence were involved in the deletion-causing NAHR events (Supp. Tables S22-S25). Moreover, no nucleotide differences were detected that were present exclusively in all haplotypes involved in the deletion-causing NAHR events while being absent from those haplotypes that were not involved in the NAHR events causing the *NF1* deletions in the parental germlines.

PRS2 haplotype diversity in Europeans and Africans

Two distinct groups of PRS2 haplotypes from NF1-REPa were identified in the transmitting parents. These groups were distinguishable by virtue of their harbouring either high or low numbers of nucleotide differences as compared with the reference sequence. Whilst 22 of the 38 parental PRS2 haplotypes from NF1-REPa exhibited only two or three nucleotide differences by comparison with the NF1-REPa reference sequence, ten of the 38 parental PRS2 haplotypes harboured 14-19 nucleotide differences (Supp. Table S13). By contrast, the PRS2 haplotypes from NF1-REPc did not exhibit a comparable bimodal distribution of haplotypes into two distinct groups characterized by low and high sequence diversity (Figure 2).

All the parents of patients investigated during the course of this study were of white European descent. In order to investigate the haplotype diversity of PRS2 in another population, we analysed the PRS2 haplotypes from NF1-REPa and NF1-REPc in 11 black Africans (Supp. Tables S26 and S27). Those PRS2 haplotypes from NF1-REPa that were characterized by low numbers of nucleotide differences in Europeans were rare in Africans whereas those haplotypes exhibiting high numbers of nucleotide differences in Europeans were much more prevalent in Africans (Supp. Table S26). The NF1-REPa-derived PRS2 haplotype with the highest number of nucleotide differences as compared with the human genome reference

sequence was detected in the African DNA sample YRI 11 (haplotype HP2; Supp. Table S26). The nucleotide differences characteristic of this haplotype affect the binding site of primer 2290for which is paralog-specific for NF1-REPa and used for PRS2 breakpoint-spanning PCRs. Even if the PRS2 haplotype of individual YRI 11 is not frequent (it was observed in only one of the 11 Africans investigated), it should be appreciated that haplotype diversity may also affect paralog-specific primer binding sites thereby interfering with breakpoint-spanning PCRs. The haplotype HP2 observed in African individual YRI 11 was also detected in two patients of white European descent with type-1 *NF1* deletions exhibiting deletion breakpoints located proximal to PRS2 (unpublished results).

PRDM9-binding sites

PRDM9 is an important regulator of meiotic allelic recombination and is probably also involved in regulating NAHR [Berg et al., 2010]. Studies in mice ~~imply~~ have suggested that PRDM9 initiates meiotic recombination in a haplotype-specific manner, organizes hotspot nucleosomes and limits Holliday junction migration [Baker et al., 2014, 2015]. The presence of predicted PRDM9-binding sites within NAHR hotspots is instructive since it suggests that PRDM9 may indeed be involved in this mutational mechanism [Dittwald et al., 2013]. The *PRDM9* A-allele has been shown to be the most frequent *PRDM9* allele in Europeans [Berg et al., 2010] and also in the parents of patients with type-1 *NF1* deletions [Hillmer et al., 2016]. The *PRDM9* A-allele encodes the PRDM9 protein A-variant which binds to a specific DNA motif with the consensus sequence 5'-CCNCCNTNNCCNC-3' [Myers et al., 2008]. We investigated the parental wildtype PRS1 and PRS2 haplotypes with regard to the number of predicted PRDM9 A-variant binding sites. In total, 76 PRS2 haplotypes (38 from NF1-REPa and 38 from NF1-REPc) and 24 PRS1 haplotypes (12 from NF1-REPa and 12 NF1-REPc) were analysed. Comparing the various parental PRS1 haplotypes identified in our study, we did not detect any differences in terms of the numbers of predicted PRDM9 A-variant binding sites. By contrast, eight of the 38 parental PRS2 haplotypes derived from NF1-REPa exhibited two additional PRDM9 A-variant binding sites as compared with the PRS2 reference haplotype from NF1-REPa which is characterized by only two such binding sites. In the other 30 parental PRS2 haplotypes from NF1-REPa, two PRDM9 A-variant binding sites were predicted as in the reference sequence of NF1-REPa (Supp. Tables S28 and S29).
Differences in the number of predicted PRDM9 A-variant binding sites were also detected when comparing the PRS2 haplotypes from NF1-REPc (Supp. Table S30). Whereas three PRDM9 A-variant binding motifs are predicted in the reference sequence of PRS2 from NF1-REPc, two parental PRS2 haplotypes from NF1-REPc were identified that had lost the perfect match to two binding motifs (~~motifs~~ A2 and A4, Supp. Table S30). By contrast, four parental NF1-REPc PRS2 haplotypes were found to possess an additional PRDM9 A-variant binding motif as compared with the reference sequence. However, no correlation was observed between the number of predicted PRDM9 A-variant binding motifs per PRS2 haplotype and whether the haplotype had been involved in the deletion-causing NAHR event (Supp. Table S31 and S32).
Differences in the number of PRDM9 A-variant binding motifs were also observed between the African PRS2 haplotypes (Supp. Tables S29, S30, S33). In the Africans, 13 of the 22 PRS2 haplotypes from NF-REPa and four of the 22 PRS2 haplotypes from NF1-REPc did not exhibit the same number of predicted PRDM9 A-variant binding motifs as observed in the reference sequence. Hence, PRS2 haplotype diversity can be seen to impact upon the number of predicted PRDM9 A-variant binding sites.

Differences in PRS2 haplotype diversity patterns between NF1-REPa and NF1-REPc

The proportions of low-diversity and high-diversity PRS2 haplotypes were markedly different between NF1-REPa and NF1-REPc in the European parents (Figure 2). Indeed, 22 of 38

PRS2 haplotypes (57%) from NF-REPa exhibited low diversity as defined by only two or three nucleotide differences compared with the reference sequence (Supp. Tables S9 and S13). By contrast, PRS2 haplotypes from NF1-REPc exhibiting only two or three nucleotide differences as compared with the reference sequence were not observed. Instead, PRS2 haplotypes from NF1-REPc exhibited 4-14 nucleotide difference relative to the reference sequence (Supp. Table S10 and S13). The predominance of PRS2 haplotypes with low-diversity characteristic ~~for~~ of NF1-REPa is linked to the absence of the PRDM9-binding motif A4 since a strong correlation between the absence of this motif and a low number of nucleotide differences per haplotype was observed (Kendall's Tau-b correlation coefficient: 0.618, p = <.0001; Wilcoxon rank-sum test, two-sided, p = 0.0002; Supp. Table S34). The absence of PRDM9-binding site A4 from the low-diversity PRS2 haplotypes from NF1-REPa may reduce the number of NAHGC-mediated nucleotide changes in this paralog. According to the two current models of homologous recombination without crossover, namely synthesis-dependent strand annealing (SDSA) and the double Holliday junction (dHj) model, mismatched nucleotides within heteroduplex regions of the recombining sequences are corrected by mismatch repair systems using the unbroken DNA strand as template [McMahill et al., 2007]. In the context of NAHGC between NF1-REPa and NF1-REPc within PRS2, the recombination-associated DNA double strand breaks should be initiated by PRDM9 binding to NF1-REPc which is repaired by the unbroken paralogous sequence from NF1-REPa. Hence NF1-REPa serves as a template and its sequence remains unchanged by NAHGC (Supp. Figure S1).

Recombination- and replication-associated sequence motifs and haplotype diversity

PRS2 haplotype diversity not only affects the number of predicted PRDM9 A-variant binding sites but also the number of other predicted recombination- and replication-associated sequence motifs (Supp. Table S35) with potential consequences for NAHR.

Discussion

Haplotype diversity and SER determination

The high resolution sequence analysis of NAHR-mediated rearrangement breakpoints has been reported in several studies in order to identify the sequence determinants of NAHR [Conrad et al., 2010; Liu et al., 2011, 2012; Luo et al., 2011; Dittwald et al., 2013]. These high resolution breakpoint studies were performed by comparative analysis of long-range breakpoint-spanning PCR product (BSP) sequences and the reference sequence of the human genome, originally derived from the pooled DNA of several individuals [International Human Genome Sequencing Consortium, 2004]. Employing this approach (method #1), the NAHR-associated strand exchange regions (SERs) between the recombining paralogs were narrowed down. However, this approach fails to take into account differences between the reference sequence and the parental non-recombinant sequences that were involved in the original rearrangement-causing NAHR events. NAHR hotspots are known to exhibit complex patterns of sequence variation but NAHR hotspot haplotype diversity has not yet been systematically analysed. Failure to take the parental non-recombinant sequences into consideration may result in the loss of important information that could be used to further refine the locations of the NAHR breakpoints. In order to assess the haplotype diversity of the two NAHR hotspots PRS1 and PRS2 which mediate the type-1 *NF1* deletions, we sequenced the wildtype (non-recombinant) PRS1 or PRS2 sequences of 25 transmitting parents of patients with type-1 *NF1* deletions. Haplotypes with considerable sequence diversity as compared with the reference sequence of the human genome were identified (Supp. Tables S9-S12). In order to ascertain whether the observed haplotype diversity can influence SER determination of type-1 *NF1* deletions, we performed sequence comparison between the wildtype PRS1 or PRS2

haplotypes from the transmitting parents and the deletion breakpoint-spanning sequences from the patients. This approach, which we termed method #2, turned out to be much more accurate than method #1 in terms of SER determination. In 22 of the 25 *NF1* deletions analysed, the SERs assigned by method #2 were shorter and more precise ~~and shorter~~ than the SERs determined by method #1. In these 22 deletions, the mean SER length was 502-bp as determined by method #1 and 269-bp according to method #2. Consequently, a very significant refinement in the mean SER length of 233-bp was achieved using method #2 (Table 1). The SERs of two deletions could not be determined correctly by method #1 because the transmitting parents harboured haplotypes with several nucleotide differences, as compared with the reference sequence, in breakpoint-flanking regions (Figure 1). Our findings imply that the haplotype diversity of the recombination hotspots PRS1 and PRS2 must be taken into account in order to precisely map type-1 *NF1* deletion breakpoints.

SERs and sequence features influencing NAHR frequency

The SER of an NAHR-mediated rearrangement indicates the location of the double Holliday junction (dHj) resolution by an endonuclease according to the DNA double-strand break (DSB) repair model [Szostak et al., 1983]. This model implies that the SER demarcates one end of the dHj migration and the site of dHj resolution (Supp. Figure S2). However, the location of the NAHR-initiating DSB remains unknown since it cannot be inferred from the analysis of breakpoint-spanning sequences. Hence, the SER of an NAHR-mediated rearrangement is the only indication available as to the actual region of crossover between the recombining paralogous sequences. The precise assignment of SERs is therefore of critical importance if we are to confidently assess the relevance of specific DNA sequence features within recombination hotspots to the regulation of NAHR frequency. These sequence features may include the distance between the recombining paralogs, the extent of DNA sequence identity between the paralogs, GC content, and the frequency of the PRDM9-binding motifs within NAHR hotspots [Myers et al., 2008; Dittwald et al., 2013; Pratto et al., 2014; Peng et al., 2015; Guo et al., 2016]. PRDM9 is a meiosis-specific histone methyltransferase with a zinc-finger protein domain that binds to the sequence motif 5' CCNCCNTNNCCNC 3' thereby regulating the genome-wide positioning of AHR hotspots in humans via sequence-specific DNA binding of its zinc finger array [Baudat et al., 2010; Berg et al., 2010; Myers et al., 2008; Parvanov et al., 2010]. In view of the similarities between AHR and NAHR [Lupski, 2004; De Raedt et al., 2006; Lindsay et al., 2006], it is not unreasonable to assume that NAHR may also be induced by PRDM9 binding to NAHR hotspots and hence PRDM9 may regulate NAHR frequency. Our findings indicate that the haplotype diversity of the PRS1 and PRS2 NAHR hotspots not only impacts upon the accuracy of SER determination but also affects the number of predicted PRDM9-binding sites per haplotype (Supp. Tables S29, S30, S33). Differences in the numbers of other recombination- and replication-associated sequence motifs known to cause recurrent DNA double strand breaks were also detected between PRS2 haplotypes as a consequence of nucleotide diversity (Supp. Table S35) and it is conceivable that these could also influence NAHR frequency.

The majority (70-80%) of type-1 *NF1* deletions exhibit breakpoints within PRS2, which is thus a much more active NAHR hotspot than PRS1 [De Raedt et al., 2006; Hillmer et al., 2016]. A correlation between the number of predicted PRDM9 A-variant binding motifs per PRS2 haplotype and whether the haplotype was involved (or not) in the deletion causing NAHR event was not however observed (Supp. Table S31 and S32). It is reasonable to assume that features present in addition to PRDM9-binding sites, such as structural variants (CNVs) of the recombining paralogs or chromatin accessibility, can also influence NAHR frequency [Carvalho and Lupski, 2008; Cuscó et al., 2008; Antonacci et al., 2010; Vergés et al., 2014, 2017]. Polymorphic large inversions present in the transmitting parents have been identified that predispose to NAHR-mediated rearrangements involving a number of different

human genes [Small et al., 1997; Osborne et al., 2001; Bayés et al., 2003; Gimelli et al., 2003; Scherer et al., 2005; Visser et al., 2005; Koolen et al., 2006; Sharp et al., 2008; Antonacci et al., 2009; Hobart et al., 2010; Molina et al., 2012]. Further studies will be required to investigate whether polymorphic CNVs within the paralogs or inversions of the regions located between the paralogs involved in NAHR occur disproportionately more often in the transmitting parents of patients with type-1 *NF1* deletions. In conjunction with such structural features, it may be that specific PRS2 haplotypes could have predisposed to recurrent NAHR events in the germlines of the transmitting parents of patients with type-1 *NF1* deletions.

Paralog-specific differences in PRS2 haplotype diversity

NAHR hotspots are known to exhibit complex patterns of sequence variation involving large numbers of shared SNPs that have been generated by frequent historical sequence exchanges between the recombining paralogs mediated by nonallelic homologous gene conversion without crossover (NAHGC) [Rozen et al., 2003; Pavlicek et al., 2005; De Raedt et al., 2006; Lindsay et al., 2006; Guo et al., 2016]. Our findings indicate that frequent NAHGC is also responsible for the haplotype diversity at the NAHR hotspots PRS1 and PRS2 (Supp. Tables S15 and S16). Further, our results also imply NAHGC-mediated differences in the haplotype diversity patterns of the recombining paralogs, NF1-REPa and NF1-REPc. In the transmitting parents of patients with type-1 *NF1* deletions investigated here, who were all of white European descent, PRS2 haplotypes from NF1-REPa could be clearly separated into two groups: haplotypes exhibiting either low or high sequence diversity as compared with the reference sequence. By contrast, African individuals exhibited predominantly PRS2 haplotypes from NF1-REPa with high numbers of nucleotide differences relative to the reference (Supp. Table S26). This finding is in accordance with the genome-wide higher sequence diversity in Africans as compared with non-African populations [reviewed by Campbell and Tishkoff, 2008]. However, the haplotype diversity of PRS2 from NF1-REPc was similar in Africans to that in Europeans (Supp. Table S27). The predominance of low-diversity PRS2 haplotypes from NF1-REPa but not from NF1-REPc was found to be correlated with the absence of a predicted PRDM9-binding motif (A4) (Supp. Table S34, Figure S3). The absence of binding motif A4 from the majority of NF1-REPa haplotypes but its presence in most NF1-REPc-derived PRS2 haplotypes may affect the direction of sequence exchange by NAHGC between the paralogs. According to this ~~hypothesis~~postulate, binding of PRDM9 to motif A4 within NF1-REPc promotes sequence transfer from NF1-REPa to NF1-REPc because the NF1-REPa-derived sequence serves as a template to repair mismatches in heteroduplex regions while NF1-REPc is the recipient strand as explained in the model depicted in Supp. Figure S1. Polarity in the direction of sequence exchange between recombination intermediates has been observed in yeast and human recombination studies. During the repair of mismatches within heteroduplex DNA regions of recombinants, the unbroken DNA strand has been found to be preferentially used as a donor template to repair the broken strand [Mancera et al., 2008; Webb et al., 2008].
Alternative explanations for the paralog-specific predominance of PRS2 low-diversity haplotypes in the European parents, which was not observed in Africans, may include positive selection or genetic drift. In any case, our findings indicate paralog-specific haplotype diversity patterns at the NAHR hotspot PRS2, as well as population-specific haplotype diversity patterns; both should be taken into account in order to accurately demarcate the SERs of NAHR-mediated rearrangements.

NAHR-associated mutations in breakpoint-flanking regions

Comparison of BSP sequences from the patients with the PRS1 or PRS2 haplotypes from the transmitting parents, performed in order to determine the SERs as precisely as possible, also enabled us to investigate the occurrence of NAHR-associated mutations in the breakpoint-

flanking regions. We did not identify any sequence variants present exclusively in the deletion breakpoint-flanking sequences from the patient but absent from the parental PRS1 or PRS2 haplotypes. Consequently, no evidence for *de novo* NAHR-associated mutations in the breakpoint-flanking sequences of the patients was detected. Arbeithuber et al. [2015] reported that AHR in human sperm is associated with an increased mutation rate. These authors analysed 5796 crossover products (13,221,000 nucleotides evaluated) and observed 17 crossover-associated mutations translating into a mutation rate of 1.29×10^{-6}/bp, some 3.6-fold higher than the mutation rate observed in non-recombinant PCR products amplified from recombination hotspots HSI and HSII [Arbeithuber et al., 2015]. Since in this study we only analysed 25 NAHR-associated crossover products (100-kb evaluated), precise conclusions regarding the NAHR-associated mutation rate at the PRS1 and PRS2 hotspots cannot be drawn from our data. This notwithstanding, our findings nevertheless imply that NAHR-mediated type-1 *NF1* deletions with breakpoints in PRS1 and PRS2 do not exhibit a mutation rate that would be as high as that observed in regions flanking complex rearrangements on different chromosomes caused by replicative mechanisms [Carvalho et al., 2013; Wang et al., 2015] or microhomology-mediated end joining [Sinha et al., 2017]. Carvalho et al. (2013) observed five single nucleotide variants in a total of 23-kb of breakpoint-flanking sequences analysed, which corresponds to a *de novo* point mutation rate of ~ 2.1×10^{-4} mutations/bp. Although we were unable to obtain any accurate measure of the NAHR-associated mutation rate in the breakpoint-flanking sequences of type-1 *NF1* deletions, our findings do not indicate that NAHR-associated mutations would occur frequently enough to interfere strongly with SER determination.

Conclusion

NAHR breakpoints are known to be characterized by complex sequence patterns mediated by NAHGC. Our findings indicated that frequent NAHGC is responsible for the haplotype diversity at the NAHR hotspots PRS1 and PRS2 which in turn influences the identification of NAHR-mediated breakpoints. However, the novel method employed here, which compares breakpoint-spanning sequences with the wildtype non-recombinant haplotypes at the NAHR hotspot from the transmitting parent, proved itself to be a valuable tool to identify the SER within the NAHR hotspot with a much higher degree of accuracy than has hitherto been possible.

References

Abeysinghe SS, Chuzhanova N, Krawczak M, Ball EV, Cooper DN. 2003. Translocation and gross deletion breakpoints in human inherited disease and cancer. I: Nucleotide composition and recombination-associated motifs. Hum Mutat 22: 229-244.

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 18:2555-2566.

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, Rosenfeld JA, Ballif BC, Shaffer LG, Graves TA, Wilson RK, Schwartz DC, Eichler EE. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. Nat Genet 42:745-750.

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proc Natl Acad Sci USA 112:2109-2114.

Badge RM, Yardley J, Jeffreys AJ, Armour JA. 2000. Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. Hum Mol Genet 9:1239-1244.

Baker CL, Walker M, Kajita S, Petkov PM, Paigen K. 2014. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. Genome Res 24:724-732.

Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, Petkov PM, Paigen K. 2015. PRDM9 drives evolutionary erosion of hotspots in *Mus musculus* through haplotype-specific initiation of meiotic recombination. PLoS Genet 11:e1004916.

Baudat F, Imai Y, de Massy B. 2013. Meiotic recombination in mammals: Localization and regulation. Nat Rev Genet 14:794-806.

Bayés M, Magano LF, Rivera N, Flores R, Pérez Jurado LA. 2003. Mutational mechanisms of Williams-Beuren syndrome deletions. Am J Hum Genet 73:131-151.

Bengesser K, Vogt J, Mussotter T, Mautner VF, Messiaen L, Cooper DN, Kehrer-Sawatzki H. 2014. Analysis of crossover breakpoints yields new insights into the nature of the gene conversion events associated with large *NF1* deletions mediated by nonallelic homologous recombination. Hum Mutat 35:215-226.

Berg IL, Neumann R, Lam KW, Sarbajna S, Odenthal-Hesse L, May CA, Jeffreys AJ. 2010. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42:859-863.

Bi W, Park SS, Shaw CJ, Withers MA, Patel PI, Lupski JR. 2003. Reciprocal crossovers and a positional preference for strand exchange in recombination events resulting in deletion or duplication of chromosome 17p11.2. Am J Hum Genet 73:1302-1315.

Bosch E, Jobling MA. 2003. Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. Hum Mol Genet 12:341-347.

Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403-433.

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

Carvalho CM, Lupski JR. 2008. Copy number variation at the breakpoint region of isochromosome 17q. Genome Res 18:1724-1732.

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 45:1319-1326.

Carvalho CM, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17:224-238.

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 42:385-391.

Coutton C, Abada F, Karaouzene T, Sanlaville D, Satre V, Lunardi J, Jouk PS, Arnoult C, Thierry-Mieg N, Ray PF. 2013. Fine characterisation of a recombination hotspot at the *DPY19L2* locus and resolution of the paradoxical excess of duplications over deletions in the general population. PLoS Genet 9:e1003363.

Cuscó I, Corominas R, Bayés M, Flores R, Rivera-Brugués N, Campuzano V, Pérez-Jurado LA. 2008. Copy number variation at the 7q11.23 segmental duplications is a susceptibility factor for the Williams-Beuren syndrome deletion. Genome Res 18:683-694.

De Raedt T, Stephens M, Heyns I, Brems H, Thijs D, Messiaen L, Stephens K, Lazaro C, Wimmer K, Kehrer-Sawatzki H, Vidaud D, Kluwe L, Marynen P, Legius E. 2006. Conservation of hotspots for recombination in low-copy repeats associated with the NF1 microdeletion. Nat Genet 38:1419-1423.

Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, Rodríguez Rojas LX, Elton LE, Scott DA, Schaaf CP, Torres-Martinez W, Stevens AK, Rosenfeld JA, Agadi S, Francis D, Kang SH, Breman A, Lalani SR, Bacino CA, Bi W, Milosavljevic A, Beaudet AL, Patel A, Shaw CA, Lupski JR, Gambin A, Cheung SW, Stankiewicz P. 2013. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. Genome Res 23:1395-1409.

Dumont BL. 2015. Interlocus gene conversion explains at least 2.7% of single nucleotide variants in human segmental duplications. BMC Genomics 16:456.

Elinati E, Kuentz P, Redin C, Jaber S, Vanden Meerschaut F, Makarian J, Koscinski I, Nasr-Esfahani MH, Demirol A, Gurgan T, Louanjli N, Iqbal N, Bisharah M, Pigeon FC, Gourabi H, De Briel D, Brugnon F, Gitlin SA, Grillo JM, Ghaedi K, Deemeh MR, Tanhaei S, Modarres P, Heindryckx B, Benkhalifa M, Nikiforaki D, Oehninger SC, De Sutter P, Muller J, Viville S. 2012. Globozoospermia is mainly due to *DPY19L2* deletion via non-allelic homologous recombination involving two recombination hotspots. Hum Mol Genet 21:3695-3702.

Fredman D, White SJ, Potter S, Eichler EE, den Dunnen JT, Brookes AJ. 2004. Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36:861-866.

Gimelli G, Pujana MA, Patricelli MG, Russo S, Giardino D, Larizza L, Cheung J, Armengol L, Schinzel A, Estivill X, Zuffardi O. 2003. Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. Hum Mol Genet 12:849-858.

Guo X, Delio M, Haque N, Castellanos R, Hestand MS, Vermeesch JR, Morrow BE, Zheng D. 2016. Variant discovery and breakpoint region prediction for studying the human 22q11.2 deletion using BAC clone and whole genome sequencing analysis. Hum Mol Genet 25:3754-3767.

Hallast P, Nagirnaja L, Margus T, Laan M. 2005. Segmental duplications and gene conversion: human luteinizing hormone/chorionic gonadotropin beta gene cluster. Genome Res 15:1535-1546.

Hobart HH, Morris CA, Mervis CB, Pani AM, Kistler DJ, Rios CM, Kimberley KW, Gregg RG, Bray-Ward P. 2010. Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. Am J Med Genet C Semin Med Genet 154C:220-228.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

Kamp C, Hirschmann P, Voss H, Huellen K, Vogt PH. 2000. Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events. Hum Mol Genet 9:2563-2572.

Kehrer-Sawatzki H, Mautner VF, Cooper DN. 2017. Emerging genotype-phenotype relationships in patients with large *NF1* deletions. Hum Genet 136:349-376.

Koscinski I, Elinati E, Fossard C, Redin C, Muller J, Velez de la Calle J, Schmitt F, Ben Khelifa M, Ray PF, Kilani Z, Barratt CL, Viville S. 2011. *DPY19L2* deletion as a major cause of globozoospermia. Am J Hum Genet 88:344-350.

Koolen DA, Vissers LE, Pfundt R, de Leeuw N, Knight SJ, Regan R, Kooy RF, Reyniers E, Romano C, Fichera M, Schinzel A, Baumer A, Anderlid BM, Schoumans J, Knoers NV, van Kessel AG, Sistermans EA, Veltman JA, Brunner HG, de Vries BB. 2006. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nat Genet 38:999-1001.

Lindsay SJ, Khajavi M, Lupski JR, Hurles ME. 2006. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. Am J Hum Genet 79:890-902.

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

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

Lopes J, Tardieu S, Silander K, Blair I, Vandenberghe A, Palau F, Ruberg M, Brice A, LeGuern E. 1999. Homologous DNA exchanges in humans can be explained by the yeast double-strand break repair model: a study of 17p11.2 rearrangements associated with CMT1A and HNPP. Hum Mol Genet 8:2285-2292.

López Correa C, Brems H, Lázaro C, Marynen P, Legius E. 2000. Unequal meiotic crossover: a frequent cause of *NF1* microdeletions. Am J Hum Genet 66:1969-1674.

López-Correa C, Dorschner M, Brems H, Lázaro C, Clementi M, Upadhyaya M, Dooijes D, Moog U, Kehrer-Sawatzki H, Rutkowski JL, Fryns JP, Marynen P, Stephens K, Legius E. 2001. Recombination hotspot in *NF1* microdeletion patients. Hum Mol Genet 10:1387-1392.

Luo Y, Hermetz KE, Jackson JM, Mulle JG, Dodd A, Tsuchiya KD, Ballif BC, Shaffer LG, Cody JD, Ledbetter DH, Martin CL, Rudd MK. 2011. Diverse mutational mechanisms cause pathogenic subtelomeric rearrangements. Hum Mol Genet 20:3769-3778.

Lupski JR. 2004. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. Genome Biol 5:242.

Mancera E, Bourgon R, Huber W, Steinmetz LM. 2011. Genome-wide survey of post-meiotic segregation during yeast recombination. Genome Biol 12:R36.

McMahill MS, Sham CW, Bishop DK. 2007. Synthesis-dependent strand annealing in meiosis. PLoS Biol 5:e299.

Mefford HC, Eichler EE. 2009. Duplication hotspots, rare genomic disorders, and common disease. Curr Opin Genet Dev 19:196-204.

Messiaen L, Vogt J, Bengesser K, Fu C, Mikhail F, Serra E, Garcia-Linares C, Cooper DN, Lazaro C, Kehrer-Sawatzki H. 2011. Mosaic type-1 *NF1* microdeletions as a cause of both generalized and segmental neurofibromatosis type-1 (NF1). Hum Mutat 32:213-219.

Molina O, Anton E, Vidal F, Blanco J. 2012. High rates of *de novo* 15q11q13 inversions in human spermatozoa. Mol Cytogenet 5:11.

Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet 40: 1124-1129.

Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, Costa T, Grebe T, Cox S, Tsui LC, Scherer SW. 2001. A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet 29:321-325.

Parvanov ED, Petkov PM, Paigen K. 2010. Prdm9 controls activation of mammalian recombination hotspots. Science 327:835.

Parvanov ED, Tian H, Billings T, Saxl RL, Spruce C, Aithal R, Krejci L, Paigen K, Petkov PM. 2017. PRDM9 interactions with other proteins provide a link between recombination hotspots and the chromosomal axis in meiosis. Mol Biol Cell 28:488-499.

Pavlicek A, House R, Gentles AJ, Jurka J, Morrow BE. 2005. Traffic of genetic information between segmental duplications flanking the typical 22q11.2 deletion in velo-cardio-facial syndrome/DiGeorge syndrome. Genome Res 15:1487-1495.

Peng Z, Zhou W, Fu W, Du R, Jin L, Zhang F. 2015. Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome. Hum Mol Genet 24:1225-1233.

Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. DNA recombination. Recombination initiation maps of individual human genomes. Science 346:1256442.

Reiter LT, Murakami T, Koeuth T, Pentao L, Muzny DM, Gibbs RA, Lupski JR. 1996. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. Nat Genet 12:288-297.

Reiter LT, Hastings PJ, Nelis E, De Jonghe P, Van Broeckhoven C, Lupski JR. 1998. Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. Am J Hum Genet 62:1023-1033.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423:873-876.

Sasaki M, Lange J, Keeney S. 2010. Genome destabilization by homologous recombination in the germ line. Nat Rev Mol Cell Biol 11:182-195.

Scherer SW, Gripp KW, Lucena J, Nicholson L, Bonnefont JP, Pérez-Jurado LA, Osborne LR. 2005. Observation of a parental inversion variant in a rare Williams-Beuren syndrome family with two affected children. Hum Genet 117:383-388.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE. 2005. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78-88.

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, Schroer RJ, Novara F, De Gregori M, Ciccone R, Broomer A, Casuga I, Wang Y, Xiao C, Barbacioru C, Gimelli G, Bernardina BD, Torniero C, Giorda R, Regan R, Murday V, Mansour S, Fichera M, Castiglia L, Failla P, Ventura M, Jiang Z, Cooper GM, Knight SJ, Romano C, Zuffardi O, Chen C, Schwartz CE, Eichler EE. 2008. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat Genet 40:322-328.

Shinawi M, Schaaf CP, Bhatt SS, Xia Z, Patel A, Cheung SW, Lanpher B, Nagl S, Herding HS, Nevinny-Stickel C, Immken LL, Patel GS, German JR, Beaudet AL, Stankiewicz P. 2009. A small recurrent deletion within 15q13.3 is associated with a range of neurodevelopmental phenotypes. Nat Genet 41:1269-1271.

Shuvarikov A, Campbell IM, Dittwald P, Neill NJ, Bialer MG, Moore C, Wheeler PG, Wallace SE, Hannibal MC, Murray MF, Giovanni MA, Terespolsky D, Sodhi S, Cassina M, Viskochil D, Moghaddam B, Herman K, Brown CW, Beck CR, Gambin A, Cheung SW, Patel A, Lamb AN, Shaffer LG, Ellison JW, Ravnan JB, Stankiewicz P, Rosenfeld JA. 2013. Recurrent HERV-H-mediated 3q13.2-q13.31 deletions cause a syndrome of hypotonia and motor, language, and cognitive delays. Hum Mutat 34:1415-1423.

Sinha S, Li F, Villarreal D, Shim JH, Yoon S, Myung K, Shim EY, Lee SE. 2017. Microhomology-mediated end joining induces hypermutagenesis at breakpoint junctions. PLoS Genet 13:e1006714.

Small K, Iber J, Warren ST. 1997. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. Nat Genet 16:96-99.

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 43:2188-2198.

Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW (1983) The doublestrand-break repair model for recombination. Cell 33: 25-35.

Sun C, Skaletsky H, Rozen S, Gromoll J, Nieschlag E, Oates R, Page DC. 2000. Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. Hum Mol Genet 9:2291-2296.

Szafranski P, Schaaf CP, Person RE, Gibson IB, Xia Z, Mahadevan S, Wiszniewska J, Bacino CA, Lalani S, Potocki L, Kang SH, Patel A, Cheung SW, Probst FJ, Graham BH, Shinawi M, Beaudet AL, Stankiewicz P. 2010. Structures and molecular mechanisms for common 15q13.3 microduplications involving *CHRNA7*: benign or pathological? Hum Mutat 31:840-850.

Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP, Blayney ML, Beck S, Hurles ME. 2008. Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders. Nat Genet 40:90-95.

Vergés L, Molina O, Geán E, Vidal F, Blanco J. 2014. Deletions and duplications of the 22q11.2 region in spermatozoa from DiGeorge/velocardiofacial fathers. Mol Cytogenet 7:86.

Vergés L, Vidal F, Geán E, Alemany-Schmidt A, Oliver-Bonet M, Blanco J. 2017. An exploratory study of predisposing genetic factors for DiGeorge/velocardiofacial syndrome. Sci Rep 7:40031.

Visser R, Shimokawa O, Harada N, Kinoshita A, Ohta T, Niikawa N, Matsumoto N. 2005. Identification of a 3.0-kb major recombination hotspot in patients with Sotos syndrome who carry a common 1.9-Mb microdeletion. Am J Hum Genet 76:52-67.

Wang Y, Su P, Hu B, Zhu W, Li Q, Yuan P, Li J, Guan X, Li F, Jing X, Li R, Zhang Y, Férec C, Cooper DN, Wang J, Huang D, Chen JM, Wang Y. 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 134:589-603.

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.

Webb AJ, Berg IL, Jeffreys A. 2008. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proc Natl Acad Sci USA 105:10471-10476.

Figure Legend

Figure 1: Comparative assignment of the type-1 *NF1* deletion-associated strand exchange regions (SERs) in patients 1547 and LL-2476 using either method #1 or method #2. The reference sequence of PRS2 from NF1-REPa and NF1-REPc according to hg19 is indicated on the left. Analytical method #1 entails the comparison of breakpoint-spanning PCR product (BSP) sequences from the patients with the reference sequence of PRS2 at sites of ~~SNPs~~ rare variants (with an MAF ≤ 1%) and PSVs. Method #2 also includes the comparison of the BSP sequences with the wildtype PRS2 haplotypes from the transmitting parents. By these means, shared SNPs with an MAF > 1% can also be evaluated. Green bars indicate the borders of the SERs assigned by method #1 whereas the borders of the SERs determined by method #2 are marked by red lines. Since the haplotypes of the transmitting parents exhibited nucleotide differences compared with the reference sequence of PRS2, an accurate assignment of the SERs of the deletions in patients 1547 and LL-2476 was only possible by means of method #2.

Figure 2: Paralog-specific differences in PRS2 haplotype diversity. The columns indicate the number of PRS2 haplotypes exhibiting two to 19 nucleotide differences compared with the reference sequence of the human genome (hg19). PRS2 haplotypes derived from NF1-REPa are represented as red columns whereas haplotypes from NF1-REPc are shown as blue columns.