Extreme clustering of type-1 *NF1* deletions breakpoints co-locating with G-quadruplex forming sequences

Anna Summerer[1], Victor-Felix Mautner[2], Meena Upadhyaya[3], Kathleen Claes[4], Josef Högel[1], David N. Cooper[3], Ludwine Messiaen[5], 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

4: Center for Medical Genetics Ghent, Ghent University Hospital, Ghent, Belgium

5: Department of Genetics, University of Alabama at Birmingham, Birmingham, USA

Grant: 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

The breakpoints of type-1 *NF1* deletions encompassing 1.4-Mb are located within NF1-REPa and NF1-REPc, which exhibit a complex structure comprising different segmental duplications in direct and inverted orientation. Here, we systematically assessed the proportion of type-1 *NF1* deletions caused by nonallelic homologous recombination (NAHR) and those mediated by other mutational mechanisms. To this end, we analysed 236 unselected type-1 *NF1* deletions and observed that 179 of them (75.8%) had breakpoints located within the NAHR hotspot PRS2 whereas 39 deletions (16.5%) had breakpoints located within PRS1. Sixteen deletions exhibited breakpoints located outside of these NAHR hotspots but were also mediated by NAHR. Taken together, the breakpoints of 234 (99.2%) of the 236 type-1 *NF1* deletions were mediated by NAHR. Thus, NF1-REPa and NF1-REPc are strongly predisposed to recurrent NAHR, the main mechanism underlying type-1 *NF1* deletions. We also observed a non-random overlap between type-1 *NF1* deletion breakpoints and G-quadruplex-forming sequences (GQs) as well as regions flanking $PRDM9_A$-binding sites. These findings imply that GQs and $PRDM9_A$ binding-sites contribute to the clustering of type-1 deletion breakpoints. The co-location of both types of sequence was at its highest within PRS2, indicative of their synergistic contribution to the greatly increased NAHR activity within this hotspot.

Introduction

Large *NF1* deletions (also termed *NF1* microdeletions) cause the chromosome 17q11.2 deletion syndrome (MIM# 613675) which is characterized by a severe manifestation of neurofibromatosis type 1 (MIM# 162200) (Mautner et al., 2010; reviewed by Kehrer-Sawatzki et al., 2017). Different types of *NF1* microdeletion have been identified (type 1, 2 and 3 or atypical) which are distinguishable by the size and location of the breakpoints and their underlying mutational mechanisms (Kehrer-Sawatzki et al., 2004; Steinmann et al., 2007; Bengesser et al., 2010; Pasmant et al., 2010; Roehl et al., 2010; Messiaen et al., 2011; Vogt et al., 2012, 2014; Zickler et al., 2012). Most common are the type-1 *NF1* deletions which have breakpoints located within the low-copy repeats NF1-REPa and NF1-REPc (Dorschner et al., 2000; Jenne et al., 2001; López-Correa et al., 2001; Forbes et al., 2004). Multiplex ligation-dependent probe amplification (MLPA) and commercially available arrays have been deployed to identify *NF1* microdeletions. However, these techniques are insufficiently precise to determine the exact location of the deletion breakpoints and consequently, the molecular mechanism causing the deletions. Only breakpoint-spanning PCRs, and the subsequent sequence analysis of the breakpoint-spanning PCR products, potentiate the identification of the breakpoints at the highest level of resolution and hence the identification of the underlying mutational mechanisms. Previously performed studies have suggested that most type-1 *NF1* deletions are mediated by nonallelic homologous recombination (NAHR) within the PRS1 and PRS2 hotspots (De Raedt et al., 2006; Messiaen et al., 2011; Bengesser et al., 2014; Hillmer et al., 2016, 2017). However, as yet, no systematic analysis of an unselected cohort of type-1 *NF1* deletions has been performed in order to determine the proportion of those deletions which exhibit breakpoints that are located outside of the known NAHR hotspots PRS1 and PRS2. Further, it remains unclear what proportion of type-1 *NF1* deletions are not mediated by NAHR but originate instead by another mutational mechanism.

In the study presented here, we analysed a total of 236 type-1 *NF1* deletions in order to address these open questions. These deletions were initially identified by MLPA and represent the largest cohort of type-1 *NF1* deletions analysed to date. We determined the location of the deletion breakpoints at the highest possible resolution by means of a combination of long-range paralog-specific PCRs, array analysis and breakpoint-spanning PCRs. Our findings demonstrate unequivocally that NAHR is indeed the mechanism underlying the vast majority of type-1 *NF1* deletions. *NF1* microdeletions initially considered to be of type-1 according to MLPA, but which turn out to be caused by mutational mechanisms other than NAHR, are very rare. Thus, only two of the 236 deletions analysed were found not to be mediated by NAHR. Further, we observed pronounced clustering of NAHR-mediated breakpoints, a finding which we believe is likely to be causally determined by synergy between G-quadruplex forming sequences and PRDM9$_A$ binding sites.

Materials and Methods

Patients

In this study, we analysed genomic DNA derived from the blood of 152 NF1 patients with type-1 *NF1* deletions, which were initially identified by MLPA (P122 NF1 area probemix, version C2, MRC Holland, The Netherlands). The deletion breakpoints of these 152 patients have not previously been published. Here, we determined the precise position of the breakpoints of these 152 type-1 *NF1* deletions at the highest possible resolution. Subsequently, we combined these data with the breakpoint location of 84 type-1 *NF1* deletions characterized in our previous studies (Jenne et al., 2001; Mautner et al., 2010; Hillmer et al., 2016, 2017). In total, we evaluated the breakpoint locations of 236 type-1 *NF1* deletions originally identified by different centres as summarized in Supp. Table S1. The patients provided written informed consent and the study was approved by the respective institutional review boards.

Breakpoint-spanning PCRs (BS-PCRs)

BS-PCRs were performed by means of the Expand™ Long Range dNTPack (Merck, Darmstadt, Germany) with primers listed in Supp. Tables S2-4 and 400 ng genomic blood-derived DNA from the patients as template. The genomic locations of the BS-PCR products pertaining to the NAHR hotspots PRS1 and PRS2 are schematically indicated in Supp. Figure S1. The PCR products were analysed by Sanger sequencing using the primers listed in Supp. Table S5 and the BigDye™ Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific, Schwerte, Germany) on an ABI 3130*xl* genetic analyzer (Applied Biosystems, Waltham, Massachusetts, USA).

Breakpoint identification

The assignment of the deletion breakpoints, represented as NAHR-associated strand exchange regions (SERs), involved the sequence analysis of breakpoint-spanning PCR products (BSPs) amplified from blood-derived DNA samples from the patients. The sequences of the BSPs were compared with the reference sequences of NF1-REPa and NF1-REPc according to human genome assembly 19 (GRCh 37; hg19). In order to distinguish between NF1-REPa and NF1-REPc derived sequences within the BSPs, we considered only those sequence differences between the BSPs and the reference sequence that occurred at sites of (i) paralogous sequence variants (PSVs), which are non-polymorphic sequence differences between NF1-REPa and NF1-REPc, and (ii) rare single nucleotide variants (SNVs) with a minor allele frequency (MAF) ≤ 1% as described in our previous study (Hillmer et al., 2017).

Paralog-specific PCRs

These long-range PCRs were performed with a paralog-specific primer annealing either to NF1-REPa or NF1-REPc. The primers used for these PCRs are listed in Supp. Table S6 alongside the annealing temperatures that allowed for paralog-specific PCR assays. Sequence analysis of the PCR products indicated the copy-number of the amplified regions by evaluating homozygosity or heterozygosity at the sites of SNVs.

Array analysis

Custom-designed array CGH was performed using an 8x15K array (Agilent SurePrint G3 human CGH microarray) in order to improve the breakpoint prediction of type-1 *NF1* deletions exhibiting breakpoints located within the region of high sequence homology between NF1-REPa and NF1-REPc (genomic regions R2 and R4; Supp. Table S7 and Figure S2). Commercially available high-resolution arrays include only a very limited number of array probes located within the regions of high sequence homology between NF1-REPa and NF1-REPc. By contrast, our custom-designed array comprises 90 oligonucleotide probes suitable for copy number evaluation located within these regions. The oligonucleotides were designed using the SureDesign platform (https://earray.chem.agilent.com/suredesign/) and encompass several paralogous sequence variants (PSVs) and SNVs with a MAF ≤ 1% that distinguish between NF1-REPa and NF1-REPc. Printing of the arrays, hybridization, normalization, processing of the raw data and quality control were performed by IMGM Laboratories (Martinsried, Germany). The deletion in patient R003150/2 was also characterized by means of CytoScan™ HD array analysis (Affymetrix).

Identification of G-quadruplex forming repeats

We screened the 51-kb region of high sequence homology between NF1-REPa and NF1-REPc, located in direct orientation to each other, for the presence of the G-quadruplex forming consensus sequence, ($G_{\geq3}N_{1\text{-}7}$ $G_{\geq3}N_{1\text{-}7}$ $G_{\geq3}N_{1\text{-}7}$ $G_{\geq3}$) (Huppert and Balasubramanian, 2005) using the programs QuadBase2 (http://quadbase.igib.res.in/TetraPlexFinder) (configuration: medium stringency) (Dhapola and Chowdhury, 2016) and the non-B DNA database (https://nonb-abcc.ncifcrf.gov/apps/nBMST/default/) (Cer et al., 2013). The 51-kb region investigated ranged from nucleotide position 28,948,896 to 28,999,883 in NF1-REPa and from 30,364,989 to 30,415,983 in NF1-REPc according to hg19.

Using these analytical tools, we also screened the 24-kb spanning CMT1A-REPs at 17p12 for the presence of G-quadruplex forming sequences. The chromosome 17 regions investigated ranged from nucleotide positions 14,074,029 to 14,098,042 and 15,470,903 to 15,494,902. NAHR between the CMT1A-REPs causes 1.4-Mb deletions associated with hereditary neuropathy with liability to pressure palsies (HNPP; MIM# 162500) and the reciprocal duplications causing Charcot-Marie Tooth disease type-1 (CMT1A; MIM# 118220) (Chance et al., 1994; Lopes et al., 1998). The HNPP-associated deletion breakpoints representing the strand exchange regions (SERs) of these NAHR-mediated rearrangements have been precisely determined by Turner et al. (2008) and were evaluated by us in order to assess the overlap between SERs and G-quadruplex forming sequences as well as $PRDM9_A$ variant binding-sites as described in the following paragraph.

$PRDM9_A$ binding-site datasets

The most common variant of the human PRDM9 protein, termed $PRDM9_{A,}$ binds to the consensus sequence 5′-CCNCCNTNNCCNC-3′ which is enriched in ~40% of recombination hotspots (Myers et al., 2008; Baudat et al., 2010). It has been shown that $PRDM9_A$ also recognizes variants of this consensus motif with the sequence 5'-CCNCCNCNNCCNC-3' and 5'-

CCNCCNCNNCANC-3' (Patel et al., 2016). We searched the 51-kb region of high sequence homology between NF1-REPa and NF1-REPc for the presence of all three predicted PRDM9$_A$ binding motifs by means of the software tool "Find Individual Motif Occurrences (FIMO)" (Grant et al., 2011). Next, we determined 250-bp windows flanking the PRDM9$_A$ binding-sites to create the PRDM9$_A$ binding-site dataset. We selected this window-size since this is the size-range of the nucleosome-depleted regions mediated by PRDM9 as determined by Baker et al. (2014) who analysed the hotspots of allelic homologous recombination in the mouse. Upon PRDM9 binding, it locally trimethylates histone H3 at lysine 4 and the nucleosomes methylated by PRDM9 become laterally displaced thereby creating a nucleosome-depleted region (Baker et al. 2014). The recombination-initiating DNA double-strand break (DSB) is then induced by the topoisomerase SPO11 within the nucleosome-depleted region followed by the reciprocal exchange of DNA between chromatids, forming Holliday junctions according to the DNA double-strand break (DSB) repair model involving a double Holliday junction (dHj) (model shown in Supp. Figure S3).

A comparable analysis was performed to create a PRDM9$_A$ binding-site dataset for the regions of high sequence homology between the CMT1A REPs.

Simulation analysis

Simulation analysis was performed by means of the software Statistical Analysis System (SAS). Each simulation analysis included 2000 independent investigation steps (iterations) performed to assess the overlap defined as a minimum of 1-bp between randomly chosen strand exchange regions (SERs) with G-quadruplex forming sequences (GQs) or between SERs and the PRDM9$_A$ binding-site dataset. By means of these simulations, we tested the null hypothesis that the observed overlap of the SERs with GQs and SERs with the PRDM9$_A$ binding-site dataset was no greater than would be expected by chance alone (Supp. Text S1).

Results

Breakpoint identification of type-1 *NF1* deletions

In the study presented here, we analysed the largest cohort of type-1 *NF1* deletions reported to date. This cohort includes 236 type-1 *NF1* deletions identified in unrelated patients at different centres (Supp. Table S1). The breakpoints of 84 of these 236 deletions have been characterized in our previous studies by means of four classical breakpoint-spanning PCRs (BS-PCRs) designed to detect breakpoints located within the NAHR hotspots PRS1 and PRS2 (Jenne et al., 2001; López-Correa et al., 2001; Mautner et al., 2010; Hillmer et al. 2016, 2017). In order to identify the breakpoints of the 152 additional type-1 *NF1* deletions in this cohort, we also performed these classical BS-PCRs with primers listed in Supp. Table S2. However, only 117 of these 152 type-1 *NF1* deletions were positive for these classical BS-PCRs whereas the breakpoints of 35 deletions could not be identified by these means (Figure 1). To identify the breakpoints of these 35 deletions, we performed additional BS-PCRs with primers located between PRS1 and PRS2 (Supp. Figure S1, Table S3). Additionally, we performed custom-designed CGH array analysis with a high density of probes located within the regions of high sequence similarity between

NF1-REPa and NF1-REPc (Supp. Table S7 and Figure S2). To verify the array results, we applied paralog-specific long-range PCRs. Sequence analysis of the paralog-specific PCR products indicated their copy number by evaluation of homozygosity or heterozygosity of SNVs. By means of these methods, the breakpoints of 33 of the 35 type-1 deletions could be identified (Figure 1; Supp. Table S8).

Surprisingly, 12 of the 33 deletions turned out to exhibit breakpoints that were located within PRS1 and PRS2, even though they had not been positive for the classical BS-PCRs as previously performed (Hillmer et al., 2017). However, by means of optimized PCR conditions or different primer sets used for the BS-PCRs, positive breakpoint-spanning PCR products (BSPs) were successfully obtained (Supp. Text S2 and Table S4). We noted that the breakpoints of four of these 12 deletions were located within PRS1 and the remaining eight had breakpoints located within PRS2 (Supp. Table S8). The breakpoints of two of these PRS2-mediated deletions, those of patients SB94 and R131070/18, could not be amplified with the primers used for the classical BS-PCRs because of the presence of rare sequence variants at the primer binding site of primer 2290for. However, optimized primer sets were established that allowed for the amplification of positive BSPs in these cases (Supp. Text S2 and Table S4). Sequence analysis of these products indicated that these two patients (SB94 and R131070/18) exhibited high sequence diversity within PRS2, as determined by comparison of the BSP sequence with the human reference sequence. Both patients had a rare PRS2 haplotype that was also detected in three other individuals by means of the analysis of the wild-type sequence of PRS2 (Supp. Figure S4). The high sequence diversity of this haplotype compared with the reference sequence probably resulted from numerous historical gene conversion events between NF1-REPa and NF1-REPc.

We also identified five deletions with breakpoints located centromeric to the proximal boundary of PRS2 as previously defined (Supp. Table S9 and Figure S1). Owing to the close proximity of the deletion breakpoints to PRS2, these five deletions were considered to be PRS2-mediated and the extent of PRS2 had to be corrected accordingly (Supp. Figures S1 and S5). Breakpoint-spanning PCR products were amplified from DNA of these five patients by means of newly established primer sets (listed in Supp. Table S4).

Among the 33 type-1 deletions, we identified 13 deletions with breakpoints located in the region between PRS1 and PRS2, a 13,858-bp region exhibiting high sequence similarity between NF1-REPa and NF1-REPc (Supp. Figure S1, Figures 1 and 2C). These 13 deletions were also mediated by NAHR as determined by sequence analysis of the breakpoint-spanning PCR products. Remarkably, eight of these 13 deletions had breakpoints located within a 476-bp region between PRS1 and PRS2. In total, eight (3.4%) of the 236 deletions exhibited breakpoints within this 476-bp region. Three NAHR-mediated type-1 deletions were identified which had breakpoints located centromeric to PRS1 but still within the region of high-sequence homology between NF1-REPa and NF1-REPc (Supp. Tables S8 and S10, Figures 1 and 2).

Of the 236 *NF1* deletions analysed, only two (0.8%) were not mediated by NAHR [R003150/2 and R282241/39]. This conclusion was based upon the positions of the deletion breakpoints as determined by paralog-specific PCRs and array analysis. The proximal deletion breakpoint of patient R003150/2 was found to be located within the *SMURF2-P* pseudogene of NF1-REPa whereas the distal deletion breakpoint was identified within the *LRRC37B* gene of NF1-REPc (Supp. Figure S6). In patient R282241/39, the proximal deletion breakpoint was located telomeric to PRS2 within NF1-REPa whereas the distal deletion breakpoint within NF1-REPc was located centromeric to PRS1 (Supp. Figure S7 and Table S11). Even though we were unable to narrow down the deletion breakpoints any further by means of breakpoint-spanning PCRs, our findings clearly indicated that the breakpoints were not located within those regions exhibiting high sequence similarity between NF1-REPa and NF1-REPc. Since high sequence homology within extended genomic regions is a prerequisite for NAHR to occur, we conclude that these deletions were not mediated by NAHR.

Taken together, the analysis of the 236 *NF1* deletions indicated that 179 (75.8%) had breakpoints located within the NAHR hotspot PRS2, and 39 (16.5%) within PRS1 (Figure 1). Only 18 (7.6%) of the 236 deletions exhibited breakpoints that were not located within the known NAHR hotspots PRS1 and PRS2. Remarkably, 16 of these 18 deletions were also caused by NAHR. Taken together, 234 (99.2%) of the 236 type-1 deletion breakpoints investigated were found to have been mediated by NAHR which therefore constitutes the main mutational mechanism underlying type-1 *NF1* deletions.

G-quadruplex forming sequences overlap with SERs of type-1 *NF1* deletions

G-quadruplex forming sequences (GQs) have been reported to be enriched in genomic regions flanking the breakpoints of non-recurrent CNVs in humans (Bose et al., 2014). Here, we investigated whether GQs and the breakpoints of type-1 *NF1* deletions would coincide more often than might be expected assuming a random distribution of both sequences within the 51-kb region of high sequence homology between NF1-REPa and NF1-REPc. The breakpoints of type-1 *NF1* deletions are represented as strand exchange regions (SERs) between NF1-REPa and NF1-REPc. The SERs indicate the locations of double Holliday junction (dHj) resolution (Supp. Figure S3) and exhibit 100% sequence homology between NF1-REPa and NF1-REPc. The SERs are flanked by paralogous sequence variants (PSVs) which serve to distinguish NF1-REPa from NF1-REPc. In total, we analysed 129 SERs, encompassing 23-bp up to 670-bp, located within a 51-kb stretch of direct sequence homology between NF1-REPa and NF1-REPc (Supp. Table S12). Within this 51-kb region, 21 GQs were identified, ranging in length from 15-bp to 39-bp (Supp. Table S13). An overlap of at least 1-bp between the 21 GQs and the 129 SERs of type-1 *NF1* deletions was observed in a total of 64 cases. In order to investigate whether this overlap could have occurred by chance alone (null hypothesis), we performed a simulation analysis. The simulations included 2000 investigation steps (iterations). In each step, the number of overlaps between 129 randomly chosen SERs located within the 51-kb region of high sequence homology between NF1-REPa and NF1-REPc and the 21 GQs was determined. However, in none of these simulations did the number of overlaps between SERs and GQs exceed the observed number of overlaps (N= 64). These simulations therefore indicated that the null hypothesis had to be rejected (empirical $p < 0.00001$; Supp. Text S1). Consequently, the overlap between the observed SERs and the 21 GQs may be interpreted as being non-random and likely to be causally determined (Supp. Figure S8; Figure 3).

PRDM9$_A$ binding-sites and SERs of type-1 *NF1* deletions

By means of simulation analysis, we also investigated whether a non-random overlap might exist between the 129 SERs of type-1 *NF1* deletions and 250-bp regions [HILDE: Do you mean ±250bp or ±125bp? Correct also in Suppl. Figure legends] flanking 26 predicted PRDM9$_A$ binding-sites (PRDM9$_A$ binding-site dataset) located within the 51-kb region of high sequence homology between NF1-REPa and NF1-REPc (Supp. Table S14). According to these simulations, the null hypothesis, namely that the overlap between the SERs of type-1 *NF1* deletions and the PRDM9$_A$ binding-site dataset could have occurred by chance alone, had to be rejected (empirical $p < 0.00001$; Supp. Text S1). Instead, the simulations imply a non-random overlap between the PRDM9$_A$ binding-site dataset and the SERs of type-1 *NF1* deletions (Supp. Figure S9; Figure 3).

Co-location of SERs, GQs and regions flanking PRDM9$_A$ binding-sites in PRS2

The co-location of GQs, the PRDM9$_A$ binding site data set and SERs was observed significantly more often in PRS2 than elsewhere in the region of high sequence homology between NF1-REPa and NF1-REPc including also PRS1 (two-tailed $p = 0.0005$; Fisher's exact test) (Supp. Table S15). This finding implies that the co-location of GQs and PRDM9$_A$ binding sites within PRS2 makes a significant contribution to the increased NAHR frequency within this hotspot.

Overlap between other NAHR-mediated deletion breakpoints and GQs or PRDM9$_A$ binding-sites

By means of comparable simulation analyses to those described above, we investigated whether the SERs of other NAHR-mediated deletions such as those associated with hereditary neuropathy with liability to pressure palsies (HNPP) would also overlap in a non-random fashion with GQs or regions harbouring PRDM9$_A$ binding-sites (Supp. Figures S10 and S11). We selected these SERs for analysis since they have been very precisely mapped (Turner et al., 2008). A non-random overlap was indeed noted between SERs of CMT1A-REP-mediated rearrangements and GQs as well as regions flanking PRDM9$_A$ binding-sites (empirical $p < 0.00001$; Supp. Figures S10 and S11).

Discussion

Type-1 *NF1* deletions are the most common among all types of large *NF1* deletion (Messiaen et al., 2011). Hence, their characterization is important not only in the context of the determination of the deletion-causing mutational mechanisms but also in relation to the genotype/phenotype relationship in patients with large *NF1* deletions. In the study presented here, we analysed 236 type-1 *NF1* deletions, initially identified by MLPA, in order to determine the precise proportion of type-1 deletions with breakpoints located within the known NAHR hotspots PRS1 and PRS2. One further aim of our study was to determine the proportion of type-1 *NF1* deletions not mediated by NAHR but instead by other mutational mechanisms. Our study has not been biased in that we did not omit any deletions from further consideration when their breakpoints could not be narrowed down to specific genomic regions. Instead, all 236 deletions were analysed at the highest possible resolution. Taken together, the breakpoints of 234 (99.2%) of the 236 type-1 deletions investigated were mediated by NAHR. Only two deletions exhibited breakpoints that were not located within regions of high sequence homology between NF1-REPa and NF1-REPc and hence were most unlikely to have been mediated by NAHR (Figure 1, Supp. Figures S6 and S7). We conclude that NAHR is the preponderant mutational mechanism causing type-1 *NF1* deletions. Previous analyses have indicated that the majority of type-1 *NF1* deletions are of maternal origin and caused by an interchromosomal exchange during meiosis I (López-Correa et al., 2000; Neuhäusler et al., 2018). Taken together with the data presented here, we may infer that NAHR during meiosis is the predominant mechanism underlying type-1 *NF1* deletions.

Our analysis demonstrates a pronounced clustering of breakpoints within the known NAHR hotspot PRS2; although this has been suggested by the authors of previous studies, these were not performed systematically enough to allow the determination of the precise proportion of PRS2-mediated deletions among an unselected group of type-1 *NF1* deletions (De Raedt et al., 2006; Hillmer et al., 2016). Here, we demonstrate that 179 of 236 deletion breakpoints (75.8%) were located within PRS2 whereas 39 (16.5%) were located within PRS1, which therefore represents a much weaker NAHR hotspot than PRS2. Only 18 deletions (7.6%) exhibited breakpoints that were not located within these known NAHR hotspots. The breakpoints of 16 of these 18 deletions were either located within the 14-kb region of high sequence homology between the NAHR hotspots PRS1 and PRS2 (N=13) or centromeric to PRS1 but still within the region of high sequence similarity between NF1-REPa and NF1-REPc (N=3) (Figure 2). These 16 deletions were all mediated by NAHR. Remarkably, eight of these deletions exhibited breakpoints located within a short 476-bp region located between PRS1 and PRS2, which may represent another preferred region of DNA double strand breaks within NF1-REPa and NF1-REPc (Supp. Figure S1, Figure 2).

The clustering of type-1 *NF1* deletions is remarkable and suggests that specific sequence features are causally responsible for it. A high GC-content has been identified in the PRS2 hotspot and also in PRS1, although less pronounced than in PRS2 (Hillmer et al., 2016). A high GC-content has also been noted in other genomic regions harbouring NAHR breakpoints (Dittwald et al., 2013). Further, allelic homologous recombination (AHR) hotspots active during meiosis have also been shown to be characterized by a high GC-content (Gerton et al., 2000; Fullerton et al., 2001; Bagshaw et al., 2006; Hansen et al., 2011). Genomic regions rich in GC may harbour G-quadruplex forming DNA sequences (GQs) which have the capacity to adopt non-B DNA conformations and cause chromosome breakage *in vivo* (reviewed by Bochman et al., 2012; van Kregten and Tijsterman, 2014). Failure to resolve non-canonical DNA structures such as G-quadruplexes has been associated with genomic instability (Kruisselbrink et al., 2008; Ribeyre et al., 2009; Mac Donald et al., 2016; reviewed by Rhodes and Lipps; Wanzek et al., 2017).

GQs have been suggested to promote meiotic homologous recombination since genome-wide computational studies in yeast demonstrated the co-location of GQs and meiotic double strand breaks (Capra et al., 2010). A role for GQs during meiotic AHR is also supported by the finding that Hop1, a protein that is critical for the synapsis of homologous chromosomes during meiosis, binds to G-quadruplex forming DNA sequence structures *in vitro* (Muniyappa et al., 2000; Anuradha and Muniyappa, 2004). Hop1 promotes intermolecular pairing between GQs at sites of meiosis-specific DNA double strand breaks suggesting that GQs could mediate the pairing between homologous chromosomes during meiotic prophase I (Kshirsagar et al., 2017). Since meiotic AHR and NAHR events are assumed to be mechanistically similar processes (Lupski et al., 2004; Liu et al., 2011), GQs are likely to play an important role during NAHR as well as AHR. This postulate is corroborated by our observation of the co-location of GQs and breakpoints (strand exchange regions, SERs) of type-1 *NF1* deletions (Supp. Figure S8; Figure 3). A comparable co-location of SERs and GQs was also observed for NAHR-mediated rearrangements between the CMT1A-REPs at 17p12 causing CMT1A or HNPP (Supp. Figure S10). These findings may be indicative of a general causal relationship between GQs and NAHR breakpoints which may extend beyond type-1 *NF1* deletions so as to include other NAHR breakpoints at other loci.

In many species including human, the DNA-binding histone methyltransferase PRDM9 has been shown to be an important regulator of AHR during meiosis. PRDM9 determines the location of recombination hotspots and facilitates the association of hotspots with the chromosome axis at sites of programmed DNA double-strand breaks (DSBs) thereby initiating the genetic exchange between chromosomes (reviewed by Paigen and Petkov, 2019). In addition to regulating AHR activity at hotspots, PRDM9 is also likely to be an important regulator of NAHR activity (Berg et al., 2010; Pratto et al., 2014). In the current study, a significant co-location of regions flanking PRDM9$_A$ binding-sites and the SERs of type-1 *NF1* deletions (Supp. Figure S9). We also observed a significant co-location of regions flanking PRDM9$_A$ binding-sites and the SERs of deletions mediated by the CMT1A-REPs at 17p12 (Supp. Figure S11). These findings thus appear to be generalizable, and as with GQs, PRDM9$_A$ binding-sites constitute important sequence features that not only contribute to determining the location of NAHR-mediated breakpoints but which may also be responsible for the clustering of these breakpoints. The significant co-location of GQs and PRDM9$_A$ binding-sites within PRS2 further implies that these elements may functionally synergize so as to promote the greatly increased NAHR activity within this hotspot.

References

Anuradha S, Muniyappa K. 2004. Meiosis-specific yeast Hop1 protein promotes synapsis of double-stranded DNA helices via the formation of guanine quartets. Nucleic Acids Res 32:2378-2385.

Bagshaw AT, Pitt JP, Gemmell NJ. 2006. Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots. BMC Genomics 7:179.

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.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327:836−840.

Bengesser K, Cooper DN, Steinmann K, Kluwe L, Chuzhanova NA, Wimmer K, Tatagiba M, Tinschert S, Mautner VF, Kehrer-Sawatzki H. 2010. A novel third type of recurrent *NF1* microdeletion mediated by nonallelic homologous recombination between *LRRC37B*-containing low-copy repeats in 17q11.2. Hum Mutat 31:742-751.

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.

Bochman ML, Paeschke K, Zakian VA. 2012. DNA secondary structures: stability and function of G-quadruplex structures. Nat Rev Genet 13:770-780.

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

Capra JA, Paeschke K, Singh M, Zakian VA. 2010. G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in *Saccharomyces cerevisiae*. PLoS Comput Biol 6:e1000861.

Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT, Bacolla A, Collins JR, Stephens RM. 2013. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res 41:D94-D100.

Chance PF, Abbas N, Lensch MW, Pentao L, Roa BB, Patel PI, Lupski JR. 1994. Two autosomal dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome 17. Hum Mol Genet 3:223-228.

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.

Dhapola P, Chowdhury S. 2016. QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. Nucleic Acids Res 44:W277-283.

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.

Dorschner MO, Sybert VP, Weaver M, Pletcher BA, Stephens K. 2000. *NF1* microdeletion breakpoints are clustered at flanking repetitive sequences. Hum Mol Genet 9:35-46.

Forbes SH, Dorschner MO, Le R, Stephens K. 2004. Genomic context of paralogous recombination hotspots mediating recurrent *NF1* region microdeletion. Genes Chromosomes Cancer 41:12-25.

Fullerton SM, Bernardo Carvalho A, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol 18:1139-1142.

Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 97:11383-11390.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27:1017-1018.

Hansen L, Kim NK, Mariño-Ramírez L, Landsman D. 2011. Analysis of biological features associated with meiotic recombination hot and cold spots in *Saccharomyces cerevisiae*. PLoS One 6:e29711.

Hillmer M, Wagner D, Summerer A, Daiber M, Mautner VF, Messiaen L, Cooper DN, Kehrer-Sawatzki H. 2016. Fine mapping of meiotic NAHR-associated crossovers causing large *NF1* deletions. Hum Mol Genet 25:484-496.

Hillmer M, Summerer A, Mautner VF, Högel J, Cooper DN, Kehrer-Sawatzki H. 2017. Consideration of the haplotype diversity at nonallelic homologous recombination hotspots improves the precision of rearrangement breakpoint identification. Hum Mutat 38:1711-1722.

Huppert JL, Balasubramanian S. 2005. Prevalence of quadruplexes in the human genome. Nucleic Acids Res 33:2908-2916.

Jenne DE, Tinschert S, Reimann H, Lasinger W, Thiel G, Hameister H, Kehrer-Sawatzki H. 2001. Molecular characterization and gene content of breakpoint boundaries in patients with neurofibromatosis type 1 with 17q11.2 microdeletions. Am J Hum Genet 69:516-527.

Kehrer-Sawatzki H, Kluwe L, Sandig C, Kohn M, Wimmer K, Krammer U, Peyrl A, Jenne DE, Hansmann I, Mautner VF. 2004. High frequency of mosaicism among patients with neurofibromatosis type 1 (NF1) with microdeletions caused by somatic recombination of the *JJAZ1* gene. Am J Hum Genet 75:410-423.

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

Kshirsagar R, Khan K, Joshi MV, Hosur RV, Muniyappa K. 2017. Probing the potential role of non-B DNA structures at yeast meiosis-specific DNA double-strand breaks. Biophys J 112:2056-2074.

Kruisselbrink E, Guryev V, Brouwer K, Pontier DB, Cuppen E, Tijsterman M. 2008. Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective *C. elegans*. Curr Biol 18:900-905.

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.

Lopes J, Ravisé N, Vandenberghe A, Palau F, Ionasescu V, Mayer M, Lévy N, Wood N, Tachi N, Bouche P, Latour P, Ruberg M, Brice A, LeGuern E. 1998. Fine mapping of *de novo* CMT1A and HNPP rearrangements within CMT1A-REPs evidences two distinct sex-dependent mechanisms and candidate sequences involved in recombination. Hum Mol Genet 7:141-148.

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 66:1969-1974.

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.

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

Mautner VF, Kluwe L, Friedrich RE, Roehl AC, Bammert S, Högel J, Spöri H, Cooper DN, Kehrer-Sawatzki H. 2010. Clinical characterisation of 29 neurofibromatosis type-1 patients with molecularly ascertained 1.4 Mb type-1 *NF1* deletions. J Med Genet 47:623-630.

McDonald KR, Guise AJ, Pourbozorgi-Langroudi P, Cristea IM, Zakian VA, Capra JA, Sabouri N. 2016. Pfh1 is an accessory replicative helicase that interacts with the replisome to facilitate fork progression and preserve genome integrity. PLoS Genet 12:e1006238.

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.

Muniyappa K, Anuradha S, Byers B. 2000. Yeast meiosis-specific protein Hop1 binds to G4 DNA and promotes its formation. Mol Cell Biol 20: 1361-1369.

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.

Neuhäusler L, Summerer A, Cooper DN, Mautner V-F, Kehrer-Sawatzki H. 2018. Pronounced maternal parent-of-origin bias for type-1 *NF1* microdeletions. Hum Genet May 5. doi: 10.1007/s00439-018-1888-x. [Epub ahead of print]

Paigen K and Petkov PM. 2018. PRDM9 and its role in genetic recombination. Trends Genet 34:291-300.

Pasmant E, Sabbagh A, Spurlock G, Laurendeau I, Grillo E, Hamel MJ, Martin L, Barbarot S, Leheup B, Rodriguez D, Lacombe D, Dollfus H, Pasquier L, Isidor B, Ferkal S, Soulier J, Sanson M, Dieux-Coeslier A, Bièche I, Parfait B, Vidaud M, Wolkenstein P, Upadhyaya M, Vidaud D; Members of the NF France Network. 2010. *NF1* microdeletions in neurofibromatosis type 1: from genotype to phenotype. Hum Mutat 31:E1506-1518.

Patel A, Horton JR, Wilson GG, Zhang X, Cheng X. 2016. Structural basis for human PRDM9 action at recombination hot spots. Genes Dev 30:257-265.

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

Rhodes D, Lipps HJ. 2015. G-quadruplexes and their regulatory roles in biology. Nucleic Acids Res 43:8627-8637.

Ribeyre C, Lopes J, Boulé JB, Piazza A, Guédin A, Zakian VA, Mergny JL, Nicolas A. 2009. The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. PLoS Genet 5:e1000475.

Roehl AC, Vogt J, Mussotter T, Zickler AN, Spöti H, Högel J, Chuzhanova NA, Wimmer K, Kluwe L, Mautner VF, Cooper DN, Kehrer-Sawatzki H. 2010. Intrachromosomal mitotic nonallelic homologous recombination is the major molecular mechanism underlying type-2 *NF1* deletions. Hum Mutat 31:1163-1173.

Steinmann K, Cooper DN, Kluwe L, Chuzhanova NA, Senger C, Serra E, Lazaro C, Gilaberte M, Wimmer K, Mautner VF, Kehrer-Sawatzki H. 2007. Type 2 *NF1* deletions are highly unusual by virtue of the absence of nonallelic homologous recombination hotspots and an apparent preference for female mitotic recombination. Am J Hum Genet 81:1201-1220.

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.

van Kregten M, Tijsterman M. 2014. The repair of G-quadruplex-induced DNA damage. Exp Cell Res 329:178-183.

Vogt J, Mussotter T, Bengesser K, Claes K, Högel J, Chuzhanova N, Fu C, van den Ende J, Mautner VF, Cooper DN, Messiaen L, Kehrer-Sawatzki H. 2012. Identification of recurrent type-2 *NF1* microdeletions reveals a mitotic nonallelic homologous recombination hotspot underlying a human genomic disorder. Hum Mutat 33:1599-1609.

Vogt J, Bengesser K, Claes KB, Wimmer K, Mautner VF, van Minkelen R, Legius E, Brems H, Upadhyaya M, Högel J, Lazaro C, Rosenbaum T, Bammert S, Messiaen L, Cooper DN, Kehrer-Sawatzki H. 2014. SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. Genome Biol 15:R80.

Wanzek K, Schwindt E, Capra JA, Paeschke K. 2017. Mms1 binds to G-rich regions in *Saccharomyces cerevisiae* and influences replication and genome stability. Nucleic Acids Res 45:7796-7806.

Zickler AM, Hampp S, Messiaen L, Bengesser K, Mussotter T, Roehl AC, Wimmer K, Mautner VF, Kluwe L, Upadhyaya M, Pasmant E, Chuzhanova N, Kestler HA, Högel J, Legius E, Claes K, Cooper DN, Kehrer-Sawatzki H. 2012. Characterization of the nonallelic homologous recombination hotspot PRS3 associated with type-3 *NF1* deletions. Hum Mutat. 33:372-383.

Figure legends:

Figure 1: Breakpoint-analysis in 236 type-1 *NF1* deletions initially identified by MLPA. The breakpoints of 84 of these 236 deletions were characterized in previous studies by means of four classical breakpoint-spanning PCRs (BS-PCRs) which detect breakpoints located within the NAHR hotspots PRS1 and PRS2. In the present study, we analysed 152 additional type-1 *NF1* deletions in this cohort. Taken together, 201 type-1 *NF1* deletions were initially found to be positive for the classical BS-PCRs whereas the breakpoints of 35 deletions could not be identified by these means. To identify the breakpoints of these 35 remaining deletions, we improved the PCR conditions for the classical BS-PCR, established new BS-PCRs, and performed both custom-designed array CGH and paralog-specific PCRs (PS-PCRs) to determine the copy number of the corresponding genomic regions [HILDE: Is this correct? I am trying to avoid using

the word ‘performed’ twice in the same sentence]. The locations of the breakpoints are indicated in round brackets whereas the relative proportions are given in square brackets. Taken together, 234 of the 236 type-1 *NF1* deletions analysed were found to be mediated by NAHR, whereas only two deletions had breakpoints located within regions that were not homologous between NF1-REPa and NF1-REPc (Supp. Figures S5 and S6) and hence were not mediated by NAHR.

Figure 2: Schema of the low-copy repeats NF1-REPa and NF1-REPc and locations of the type-1 *NF1* deletion breakpoints analysed. (**A**) Structure of NF1-REPa and NF1-REPc. The LCRs exhibit a modular structure comprising different segmental duplications. H19 indicates the regions with high sequence similarity to chromosome 19p13.12. The 51-kb directly oriented region of high sequence homology between NF1-REPa and NF1-REPc is indicated by white horizontal arrows. (**B**) Summary of the breakpoint location of 234 NAHR-mediated type-1 deletions pertaining to NF1-REPa. (**C**) Locations of the breakpoints of the 13 deletions which exhibited strand exchange regions (SERs) between PRS1 and PRS2. The SER of the deletion in patient R608111/100 could not be identified at the highest possible resolution owing to the lack of DNA. However, paralog-specific PCRs indicated that the breakpoints were located between PRS1 and PRS2 within a 6-kb region homologous between NF1-REPa and NF1-REPc. Five patients exhibited breakpoints at the centromeric end of PRS2, thereby necessitating an increase in our estimation of the length of PRS2.

Figure 3: Locations of the 129 SERs of type-1 *NF1* deletions, 21 G-quadruplex forming sequences and 26 PRDM9_{A} binding sites within the 51-kb region of high sequence homology between NF1-REPa and NF1-REPc (marked by the grey rectangle). The most common variant of the human PRDM9 protein (termed PRDM9_{A}) recognizes the consensus sequence 5′-CCNCCNTNNCCNC-3′. These predicted PRDM9_{A} consensus binding sites are indicated by green dots. PRDM9_{A} also binds to variants of this consensus motif and these variant binding sites (5'-CCNCCNCNNCCNC-3' and 5'-CCNCCNCNNCANC-3') are indicated by lilac dots. The relative locations of the NAHR hotspots PRS1 and PRS2 are also indicated. A non-random overlap of SERs and G-quadruplex forming sequences, and between SERs and 250-bp regions flanking PRDM9_{A} binding sites, was observed.