Multiple paralogues and recombination mechanisms contribute to the high incidence of 22q11.2 Deletion Syndrome

Authors

Lisanne Vervoort[1], Nicolas Dierckxsens[1,2], Marta Sousa Santos[1], Senne Meynants[1], Erika Souche[1], Ruben Cools[1], Tracy Heung[3], Koen Devriendt[1], Hilde Peeters[1], Donna M. McDonald-McGinn[4,5,6], Ann Swillen[1], Jeroen Breckpot[1], Beverly S. Emanuel[4,5], Hilde Van Esch[1], Anne S. Bassett[3], Joris R. Vermeesch[1,*]

Affiliations

[1] Department of Human Genetics, KU Leuven, Leuven, Belgium
[2] Genomics and Regulatory Systems Unit & Marine Climate Change Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan
[3] Clinical Genetics Research Program, Centre for Addiction and Mental Health, Toronto, ON, Canada
[4] Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[5] Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, PA, USA
[6] Department of Human Biology and Medical Genetics, Sapienza University, Rome, Italy

*Correspondence to: joris.vermeesch@kuleuven.be

Running Title:

Mapping the recombination breakpoints within Chr22 low copy repeats

Keywords:

Genetics, Genomics, Chromosome 22, Low Copy Repeats, 22q11DS

Abstract: The 22q11.2 deletion syndrome (22q11.2DS) is the most common microdeletion disorder. Why the incidence of 22q11.2DS is much greater than that of other genomic disorders remains unknown. Short read sequencing cannot resolve the complex segmental duplications (SD) to provide direct confirmation of the hypothesis that the rearrangements are caused by non-allelic homologous recombination between the low copy repeats on Chromosome 22 (LCR22s). To enable haplotype-specific assembly and rearrangement mapping in LCR22 clusters, we combined fiber-FISH optical mapping with whole genome (ultra-)long read sequencing or rearrangement-specific long-range PCR on 25 families comprising several different LCR22-mediated rearrangements. Unexpectedly, we demonstrate that not only different paralogous SDs but also palindromic AT-rich repeats (PATRR) within LCR22s are driving 22q11.2 rearrangements. In addition, we show the existence of two different inversion polymorphisms preceding rearrangement, and somatic mosaicism. The existence of different recombination sites and mechanisms in paralogues and PATRRs which are copy number expanding in the human population are a likely contributors for the high 22q11.2DS incidence.

Introduction

Low copy repeats (LCRs), also referred to as segmental duplications, constitute 6.6% of the human genome (Nurk et al. 2022a) and played an important role during human evolution and adaptation (Dennis and Eichler 2016). They are defined as DNA segments with a length of at least 1 kb that share >90% of sequence identity (Bailey et al. 2001, 2002). Their high sequence homology is known to be a driver of non-allelic homologous recombination (NAHR), caused by meiotic misalignment of homologous chromosomes or sister chromatids, resulting in reciprocal deletions, duplications, and inversions (Inoue and Lupski 2002; Porubsky et al. 2022). These recurrent genomic rearrangements often cause genomic disorders, forming collectively an important cause of disabling diseases in the general population (Angelis et al. 2015).

The 22q11.2 deletion syndrome (22q11.2DS, MIM 188400) is the most common microdeletion in humans with an estimated incidence of 1 in 2148 live births (Blagojevic et al. 2021). 22q11.2DS has a heterogeneous presentation including multiple congenital and later-onset features, such as cardiac, palatal, metabolic, cognitive, and neuropsychiatric abnormalities (McDonald-McGinn et al. 2015). The presence and severity of the associated clinical expression is variable but the causes of this high variability remain largely unexplained (McDonald-McGinn et al. 2015). Extensive investigation of the recombination locus is complex due to the presence of eight LCR clusters on the chromosome, commonly named LCR22-A until -H (Shaikh et al. 2000). Rearrangements between different segments occur (Campbell et al. 2018; McDonald-McGinn et al. 2015), but deletions of the 3Mb region extending from LCR22-A to -D represent the main cause (85%) of the 22q11.2DS (Campbell et al. 2018). CNVs distal to LCR22-D represent a separate condition (MIM 611867).

Why the incidence of the 22q11.2DS is an order of magnitude higher than any other genomic disorder remains an enigma. Although it is assumed that the rearrangements are caused by NAHR between the LCR22s, direct confirmation of this hypothesis is lacking. This is because for a long time, a gapless reference of the LCR22s was missing. The gaps were a consequence of a complex intricate segmental duplication structure which could not be resolved by short read sequencing techniques and the lack of bioinformatic pipelines to create haplotype-resolved LCR22 assemblies (Vollger et al. 2019). Several attempts to map the rearrangements have provided anecdotal results: using long-range PCR and sequencing of two distal 22q11.2 deletions, the Breakpoint Cluster Region (BCR) module was suggested to be the NAHR site in the distal LCR22-D/E and LCR22-E/F deletions (Shaikh et al. 2007). By mapping shared and paralogous sequence polymorphisms of LCR22-A and -D by sequencing bacterial artificial chromosomes (BACs), this same BCR module was suggested to also drive the NAHR in a single LCR22-A/D deletion patient (Guo et al. 2016). With the help of optical mapping techniques, (Demaerel et al. 2019) were able to uncover the complex repeat structure of the LCR22s in the human population. They are characterized by a high variability in size (200kb – 3Mb) and structural organization of the LCR22-A haplotype (Demaerel et al. 2019; Pastor et al. 2020). Most recently, the telomere-to-telomere (T2T) consortium released the first gapless fully sequenced haploid genome, including complete LCR22s (Nurk et al. 2022a) (**Fig.1A**). Although the sequence represents an existing LCR22-A haplotype, the allele is rather short and some specific segmental duplications, present in part of the human population, are missing (**Fig. 1B**). Hence, this initial T2T reference cannot be used as an accurate representation in LCR22-mediated NAHR research.

It remains unknown whether NAHR alone can explain the high incidence of the 22q11.2DS. To enable LCR22 haplotype-aware assembly and rearrangement mapping in those regions, we combined optical mapping with whole genome (ultra-)long read sequencing and/or rearrangement-specific long-range PCR. A combination of these methods was applied on 25 families (patient and parent-of-origin, **Supplemental Table S1**) to identify the 22q11 breakpoint regions in the proband, the genomic elements involved and gain insights on the recombination mechanisms

Results

Fiber-FISH identification of LCR22-specific recombination clusters

To map NAHR sites, we first generated *de novo* assembly fiber-FISH or optical maps of all LCR22-A, -B, -C and -D alleles in parental and patient genomes (Demaerel et al. 2019) (**Fig. 1B-D**). Fiber-FISH provides haplotype-aware *de novo* assembly at subunit resolution with long-range structural information of the targeted LCR22 loci. In addition, the loci can be assembled without a priori structural knowledge of the LCR22s, tackling assembly problems of regions associated with incorrect reference genome representation and extreme structural variation.

Crossover regions were determined in 25 families, by comparison of the rearrangement alleles in patient and parent-of-origin **(Supplemental Fig. S1, Supplemental Fig. S2 and Supplemental Fig. S3)**. Our sample collection included five different LCR22-mediated recombinations: 15 LCR22-A/D deletions, five LCR22-A/B deletions (four patient-parent duos and one index patient), three LCR22-A/C deletions, one LCR22-B/D deletion and, one LCR22-C/D deletion **(Table S1)**. Rearrangements occurred for nine families on the maternal allele and for 15 families on the paternal allele. Fiber-FISH patterns of the patient and the parent-of-origin, in whom the rearrangement occurred, were analyzed. All parental alleles showed normal heterozygous LCR22-A, homozygous LCR22-B and -C, and hetero- or homozygous LCR22-D patterns, as described in (Demaerel et al. 2019) and (Pastor et al. 2020). In addition, an inversion between LCR22-A and -D was observed in the parent-of-origin of one family (BD001) **(Fig. 4A, Supplemental Fig. S1)**.

Based on *in silico* predictions, several of the segmental duplications identified could act as NAHR substrate, and these SDs NAHR sites cluster in a subset (**Table 1, Supplemental Fig. S1**). We chose to use reference genome hg38 for comparison and visualization of the LCR22s. For nested deletions involving LCR22-C, the rearrangements cluster in a 10kb region (Chr22: 20,688,715-20,698,995), which is copy number variable in LCR22-A and -D **(Table 1)**, from now on named RL-C (recombination locus LCR22-C). LCR22-B-mediated nested deletions can be subcategorized into two groups: (I) two out of five crossovers occurred in a 20kb unit (Chr22: 20,324,573-20,344,531; RL-B1), and (II) in three out of five the region was refined to an 8kb locus (Chr22: 20,331,987-20,339,583; RL-B2) within this 20kb unit. The 20kb unit is copy number variable in both LCR22-A and LCR22-D (Demaerel et al. 2019). Due to the size and the structural variation, rearrangements involving both LCR22-A and -D are the most complex to analyze. In 13 out of 15 families, recombination occurs within a 160kb locus (RL-AD1) whereas in the other two families, the recombination was restricted to a 20kb (RL-AD2) interval within RL-AD1. All LCR22 recombination loci identified are part of the larger RL-AD1 locus (**Table 1**).

Klik of tik om tekst in te voeren.Klik of tik om tekst in te voeren.

Long-read sequencing toolkit to map *22q11.2* recombinations

Since fiber-FISH resolution is limited, we leveraged long-read sequencing and optimized a *de novo* assembler algorithm to scrutinize these crossovers in more detail (**Fig. 2A**). Long read sequencing was performed in ten of the 25 families. Ultra-long read ONT sequencing (ULK, **Fig. 2B**) was performed in eight families. N50 values were over 50kb (maximum of 132kb) with an output range of 4-111Gb. To increase coverage, the ULK data were complemented by standard-long read ONT sequencing (SLK, **Fig. 2B**) and/or High Duplex (HD) ONT sequencing in three families. For one family, only SLK was performed. SLK and HD sequencing resulted in 36-115Gb and 96Gb of sequencing data respectively, with an N50 value above 21kb (**Supplemental Table S2**). This allowed the identification of the rearrangement breakpoint in all nine families (**Supplemental Table S1**).

In addition, we were able to infer the crossover site by long-range PCR and subsequent long-read PacBio sequencing in one index patient (AB004, **Fig. 2B**) with a LCR22-A/B deletion.

Ethnicity of all samples for which genome wide long read sequencing was available was determined based on a principal component analysis. All samples appear to be of European descent, although American influences cannot be excluded (**Supplemental Fig. S4**). Also, the relatedness score of all pairs of samples was computed. All parent-child duos' relatedness scores ranged between 0.222 and 0.225 while the relatedness scores of other duos were below 0.166, confirming the parent-child relationship of all parent-child duos (**Supplementary Fig. S5**).

Since existing assembly algorithms were not able to resolve any of the LCR22 regions, we developed a new targeted haplotype-aware assembler. Considering that we are interested in assembling relatively short regions as accurately as possible, we opted for a seed-and-extend method that assembles both haplotypes in parallel, starting from a given seed sequence. The algorithm keeps track of the variation between the haplotypes and it looks for mismatch patterns between the reads and each haplotype assembly to select the correct path during each iteration. **(Fig. 2C).**

Variability of the crossover site at nucleotide resolution

We leveraged these long-read sequencing approaches to resolve the recombination alleles for the rearrangements occurring in ten of the 25 families using genome wide and targeted approaches in nine and one families, respectively (**Fig. 3**, **Table 2**, and **Supplemental Fig. S4**). As opposed to the larger, apparently uniform, blocks identified by optical mapping, the breakpoint loci identified by this approach were scattered within these blocks (**Fig. 3**) and located in a variety of gene loci and involving various repetitive elements (**Table 2, Supplemental Fig. S4**).

In three out of the five LCR22-A/B deletions, the fiber-FISH pattern showed a clear transition from LCR22-A to -B in absence of large SDs structures of >20kb where NAHR could have taken place **(Supplemental Fig. 1, Table 1)**. Using ultra-long read sequencing (patients AB001 and AB002) and single-molecule real-time long-read sequencing (patient AB004), we revealed that the crossover occurred within a palindromic AT-rich repeat (PATRR). In addition, in family AB001, a 50bp LINE element, which was not present on either parental allele, was inserted at the recombination site between two PATRRs. The presence of a LINE insertion and the involvement of the PATRRs provides evidence for non-homologous end-joining (NHEJ) as a cause of 22q11.2 deletion.

Complex recombination patterns mediated by LCR22 inversions

In a family (BD001) with a *de novo* LCR22-B/D deletion, the fiber-FISH LCR22 structures showed (I) a normal LCR22-A, -B, -C, and -D, (II) an LCR22 block with the proximal start of LCR22-A combined with the proximal part of LCR22-B in an inverted orientation (LCR22-A/Binv), and (III) the distal end of LCR22-A in an inverted orientation coupled to the distal end of LCR22-D (LCR22-Ainv/D) **(Fig. 4A, Supplemental Fig.S1)**. In the parent-of-origin, the normal LCR22-A, -B, -C, and -D alleles were observed, as well as an allele indicative of an LCR22-A/D inversion **(Fig. 4A, Supplemental Fig. S1)**. Interphase-FISH using probes proximal and distal from LCR22-A and proximal from LCR22-B show the presence of an inversion in 100% of the cells of both the patient and the parent-of-origin **(Fig. 4B, Supplemental Table S3)**. Indeed, the presence of a parental LCR22-A/D inversion in combination with a NAHR recombination between LCR22-A and -B does explain the observed fiber-FISH structures in the patient. Using long-read sequencing, this recombination was shown to be intronic in the *FAM230* paralogue in both LCR22-A and -B. As a consequence, the locus between LCR22-B and -D is deleted.

In a family (AB002) with a *de novo* LCR22-A/B deletion, a mosaic LCR22-A/B deletion was identified via low-pass sequencing, and validated by arrayCGH, SNP array, and dual color interphase-FISH (TUPLE1/Arsa). Fiber-FISH uncovered the presence of three different alleles in the patient: (I) a normal LCR22-A and -B haplotype, (II) the LCR22-A/B deletion haplotype, and (III) a haplotype carrying an inversion between LCR22-A and -B **(Fig. 4C, Supplemental Fig. S1)**. The fiber-FISH patterns in the parent-of-origin showed two normal alleles **(Fig. 4C, Supplemental Fig. S1)**. In addition, interphase-FISH validated the presence of the three haplotypes and uncovered that each cell carried a wild type 22q11.2 locus and either the LCR22-A/B deletion or inversion **(Fig. 4D, Supplemental Table S3)**. We hypothesize the LCR22-A/B deletion was created from the LCR22-A/B inversion allele in an early stage during embryogenesis. Based on whole-genome ultra-long read sequencing data of the family, the deletion crossover was pinpointed in a PATRR of LCR22-A and -B of the inversion allele. Similarly, these PATRRs seem to be responsible for the creation of the inversion allele as well **(Fig. 4C)**. The most parsimonious explanation is that two consecutive PATRR-mediated events created an LCR22-A/B inversion and a deletion allele in the patient from family AB002 **(Fig. 4C)**.

Discussion

Due to the limitations of short- and standard long-read sequencing methods and the absence of an accurate reference genome, the exact positions of the recombinations of the 22q11.2DS remained uncharted. As a consequence, the mechanisms involved in the recombination and the reason for the high incidence of 22q11.2 rearrangements remained unknown. The recent efforts of the T2T consortium have proven that assembling even a relatively short LCR22-A is not a trivial task and requires combining both HiFi and ONT reads at sequencing depths several magnitudes higher than regular datasets were able to resolve the LCR22-A (Rautiainen et al. 2023). In this study, we mapped the recombination sites in 25 families with *de novo* 22q11.2 CNVs using fiber-FISH and optical mapping approaches. By developing and implementing a method that is capable of producing haplotype-aware assemblies of the 22q11 region with only ONT based long reads, we were able to map the recombination sites at the nucleotide resolution in ten families using long-read sequencing. Since the different LCRs22 -A and -D haplotypes were mapped in the parents but not phased, it is not possible to distinguish between interchromosomal and interchromatidal events. Here, we demonstrate that different paralogues in the LCR22s can drive recombination and that most paralogous are copy variable in the population. Furthermore, we reveal that different rearrangements occur in the PATRR loci within the LCR22s, suggesting that not only NAHR but also breakage-mediated repair mechanisms occur. In addition, the involvement of LINE mediated rearrangements could also be caused by replication or transcription-based mechanisms (Song et al. 2018; Robberecht et al. 2013). We hypothesize that the occurrence of different mechanisms and the variability in the number of paralogues regions strongly contribute to the higher incidence of 22q11.2DS as compared to other deletion syndromes. However, also other selective pressures such as f.e. variability in embryonic lethality may contribute.

PATRRs are palindromic sequences that create genomic instability via the formation of single-stranded hairpin or double-stranded cruciform secondary structures. Those cruciforms are sensitive to the generation of double-strand breaks, which are repaired via NHEJ (Kato et al. 2012a). The LCR22-B PATRR is known to drive recurrent 22q11.2 translocations with the 1p21.2, 8q23.1, 11q23 and 17q11.2loci (Kurahashi et al. 2006; Kato et al. 2012b), but has not been reported to be involved in 22q11.2 deletions. Here, we show involvement of PATRR sequence located within LCR22-A and -B to create LCR22-A/B deletions. It is known that PATRR size polymorphisms influence the rearrangement frequency of *de novo* t(11;22) translocations (Kato et al. 2006; Tong et al. 2010). For example, larger and symmetric PATRRs on chromosomes 11 and 22 are more prone to t(11;22) translocations (Kato et al. 2006; Tong et al. 2010). Exploring publicly available assemblies using RepeatMasker (Smit et al. 2013), we observed copy number variations of PATRR-HSATI-AluY triplets ranging between 1 and 26 copies in ten investigated alleles **(Supplemental Table S4**). The biggest variation was observed in LCR22-B with fragment lengths going from approximately 3kb to

25 kb. Moreover, since all the deletion breakpoints mapped within PATRR sites were found in rearrangements involving LCR22-B it is likely that this repeat expansion might affect the rearrangement frequency. Population scaled mapping of this variation will provide more insights about their role. Also, it has been shown that secondary structure formation and double-strand breaks can occur both during meiosis and mitosis (Correll-Tash et al. 2021). Here, we identified an individual with a possible meiotic, followed by a mitotic progression of PATRR-driven rearrangements (AB002, **Fig. 4C**). Alternatively, both the inversion and the deletion could have occurred post-zygotically, but this possibility appears less likely given the absence of a non-rearranged LCR22 haplotype inherited from the parent-of-origin.

Different SDs contribute to NAHR events at the 22q11.2 locus (**Table 2**). Hence, various recombination loci may be present in the shared modules between the two involved LCR22s (proximal and distal) and these create variability of the crossover locus. The identification of multiple subunits driving NAHR is also observed in other genomic disorders, for example neurofibromatosis type I deletions (Summerer et al. 2018), and Sotos syndrome 5q35 CNVs (Visser et al. 2005a, 2005b). In 22q11.2DS, in addition, many of the NAHR and PATRR sites appear to be copy number variable. It thus seems likely that expanded haplotypes may be more likely to predispose to rearrangement events. This hypothesis can now be tested by analyzing LCR22s in a larger (22q11.2DS) population. Interestingly, the genes where the rearrangements took place appear to involve the core 22q11 duplicons. Core duplicons were defined structurally as ancestral duplicons that were represented in more than 67% of the blocks within a given clade (Jiang et al. 2007). In this paper, no core duplicons for 22q11 were identified, likely because it remained impossible to sequence the segmental duplications on 22q11. When looking at the individual SDs driving the SD22 expansion it seems likely the *GGT3P* gene module is a core duplicon. Also the *FAM230*, *POM121* and *BRC* modules are very recurrent. Interesting, polymorphisms of the core duplicons in Chromosome 15 (*GOLGA8*), Chromosome 16 *(BOLA2*) have been shown to drive the rearrangements (Paparella et al. 2023; Maggiolini et al. 2019; Antonacci et al. 2014; Nuttle et al. 2016).

Among the families studied, we identified two individuals with LCR22-mediated inversions, one a parent-of-origin with a LCR22-A/D inversion that led to a LCR22-B/D deletion offspring, and the other a patient with a mosaic genotype that included a LCR22-A/B inversion. Previous studies had failed to identify such inversions in 22q11.2DS (Gebhardt et al. 2003; Vergés et al. 2017). However, availability of a gapless telomere-to-telomere assembly of the human genome (T2T-CHM13) has enabled a more complete analysis of genomic variation (Nurk et al. 2022a). Using long read sequencing and Strand-seq data of 52 samples from the 1000 Genomes Project and mapping these against the T2T-CHM13 reference has provided a more complete view of the genome-wide inversion polymorphism (Porubsky et al. 2022; Hanlon et al. 2022; Porubsky et al. 2023). Although not directly related to disease, LCR-mediated inversion polymorphisms drive NAHR, thus can lead to genomic disorders (Shaw and Lupski 2004). Examples of this phenomenon can be found in the high incidence of inversion-carrying parents-of-origin in Williams-Beuren syndrome (Osborne et al. 2001), Angelman syndrome, Sotos syndrome, 8p23.1 microdeletion, and 15q23 and 15q24 microdeletion syndromes (Puig et al. 2015). In some cases, these disease-predisposing inversion polymorphisms can be linked to phenotypic consequences (Boettger et al. 2012; Steinberg et al. 2012). To address the incidence and eventual phenotypic consequences in 22q11.2DS, it will be critical to survey many more human genomes and to sequence and resolve the large complex LCR22s flanking the inversion polymorphisms, along with deep phenotyping data.

We observed one individual to be mosaic, with 50% white blood cells to have a normal oriented LCR22 haplotype, 25% a LCR22-A/B inversion haplotype, and 25% a LCR22-A/B deletion. Mosaicism of the 22q11.2 deletion is rare, with few cases previously reported (Consevage et al. 1996; Halder et al. 2008; Chen et al. 2019; Patel et al. 2006; Chen et al. 2004). These reports were based on standard interphase or metaphase FISH testing, and thus no recombination sites were mapped. Interestingly, one case described 22q11.2 deletion mosaicism in a miscarried fetus (85% of cells) as well as in the mother (11% of cells), suggesting an increased recombination susceptibility for the specific chromosome involved (Patel et al. 2006). It will be of interest to map the LCR22 haplotypes of other individuals with 22q11.2DS mosaicism. We hypothesize inversions and PATRRs to be drivers of postzygotic rearrangements.

In conclusion, we uncovered crossovers occur in different paralogues within the LCR22s and both NAHR and NHEJ can cause 22q11.2DS. However, the large and complex LCR22s -A and -D remain challenging to sequence which resulted in only a limited number of recombinations resolved, despite being the most frequent. Fiber FISH and Bionano optical mapping provide an orthogonal approach for studying repetitive regions in the genome by generating high-resolution maps from ultra-long DNA molecules. Those approaches are likely to support and guide sequencing based *de novo* genome assemblies for those complex regions. With improved and cheaper long read sequencing technologies, it becomes feasible to *de novo* assemble the ultra-long and complex LCR22s which, in turn, will enable to saturate the landscape of 22q11.2 rearrangements. Since haplotype variability of SDs has been shown to affect rearrangement predisposition for other genomic disorders (Steinberg et al. 2012), exploration of the size, numbers and orientation of paralogue variability can potentially uncover predisposing or protective alleles. The presence of a variable number and sequence polymorphism within the PATRR sequences in both LCR22-A and -B hint that this may be the case. In addition, with fully sequenced LCR22s it becomes possible to explore how the LCR22 repeat variability, orientation and rearrangements affect the 22q11.2DS phenotype.

Materials and Methods

Sample collection

A total of 49 Epstein-Barr virus transformed (EBV) cell lines, of which 25 were index patients with a *de novo* 22q11.2 deletion and 24 were parents-of-origin, were collected for the study (**Supplemental Table S1**). Four samples were collected from Albert Einstein College of Medicine (New York), 24 from University of Toronto, 2 from Children's Hospital of Philadelphia and 21 from University Hospital Leuven . Seven of the LCR22-A/D deletion duos (AD009-AD015, **Supplemental Table S1**) were previously used in the study of (Demaerel et al. 2019), and here their patterns were re-analyzed for crossover site delineation. All patients and parents had given written consent to participate in genetic research of the 22q11.2 CNV. The EBV cell lines were the starting material for the fiber-FISH and sequencing sample preparation. Study approval was obtained from the Medical Ethics Committee of the University Hospital/KU Leuven (S52418), and at the Institutional Review Boards of the originating sites of the participating families: Clinical Genetics Research Program at the Centre for Addiction and Mental Health (REB# 114/2001-02), the Albert Einstein College of Medicine (IRB# 1999-201-047), and Children's Hospital of Philadelphia (IRB protocol #07-005352). Additional information regarding the samples is available in **Supplemental Table S1**.

Fiber-FISH & Optical mapping

To haplotype the LCRs on Chromosome 22, we used the LCR22-specific fiber-FISH method as described in (Demaerel et al. 2019). In short, long DNA fibers were extracted from EBV cell lines from probands and the parent-of-origin using the Genomic Vision extraction kit (Genomic Vision). These long DNA molecules were combed onto slides and hybridized using a LCR22-specific customized probe set (Demaerel et al. 2019). Following automated microscopy scanning of the slides (FiberVision, Genomic Vision), the data were analyzed by manually indicating regions of interest (FiberStudio, Genomic Vision). Haplotypes were *de novo* assembled, with a coverage of at least 5X, using matching colors and distances between the probes as anchors. Patterns of recombined LCR22s of the patients were compared to the parental patterns to identify the haplotype alteration position. Optical mapping data was obtained for the AD004 Parent. DNA was extracted using the SP-G2 Blood & Cell culture DNA Isolation kit (#80060, Bionano Genomics) and labeled using the DLS-G2 DNA labeling Kit (DLE-1 labeling enzyme, #80046, Bionano Genomics). The sample was loaded onto Saphyr Chips G2.3 (Bionano Genomics, linearized and visualized using the Saphyr Instrument (Bionano Genomics), according to the System User Guide. The analysis was performed by *de novo* assembly against the hg38 reference genome in Bionano Access Software (Bionano Genomics).

(Ultra-)long read sequencing via Oxford Nanopore Technologies (ONT)

Ultra-high molecular weight (UHMW) DNA (50kb -1Mb) was extracted via the UHMW DNA extraction protocol of the Nanobind CBB Big DNA kit (Circulomics) or via the Monarch HMW DNA

Extraction Kit for Cells & Blood (New England Biolabs) and quantified using Qubit dsDNA Broad Range kit (ThermoFisher). Approximately 40µg was used as input for sequencing (SQK-ULK001 library preparation kit, ONT). The UHMW DNA was tagmented and adapters attached to the DNA ends, followed by a disk-based clean-up reaction or spermine precipitation (SQK-ULK001, ONT). One third of the library was loaded onto a Promethion flow cell (ONT). The flow cells were washed twice and reloaded with the remaining 2/3 of the library after 24h and 48h. Run statistics are presented in **Supplemental Table S2**. ULK was planned for all samples. However, ULK has a low yield and with the improvements of the *de novo* alignment software used for local assembly, the successful assembly of the 22q11 region with SLK was possible. Therefore, SLK was performed for recently sequenced samples. Since the LCR22-A/D rearrangements are the most complex to sequence, families where the Fiber-FISH suggested this rearrangement type, were prioritized for sequencing.

De novo **assembly and sequence alignment**

Ultra-long Nanopore reads were aligned to the human reference genome (hg38) with minimap2 (Li 2018). To facilitate the visualization of the alignment with the Integrative Genomics Viewer (IGV) (Robinson et al. 2011), the 22q11 region was isolated with SAMtools (Li et al. 2009). *De novo* assembly was performed with NOVOLoci, a targeted haplotype-aware assembler (available at https://github.com/ndierckx/NOVOLoci). To prove that NOVOLoci can successfully assemble the challenging 22q11 region, the sequencing data used in the Verkko assembler paper (Rautiainen et al. 2023)was used. The 22q11 region was assembled using NOVOLoci with ONT data only. The resulting assemblies, the Verkko assemblies based on ONT data only and a combination of ONT and PacBio data and the Flye assemblies (Kolmogorov et al. 2019) were then mapped to the reference genome HG002 available from https://github.com/marbl/hg002 using LAST (Kiełbasa et al. 2011). NOVOLoci managed to assemble the full 22q11 region by using ONT data only while other assemblers failed to assemble the region with ONT data only and Verkko needed both ONT and PacBio data to fully assemble the region (Supplementary figure S6). NOVOLoci requires a seed sequence to initiate the assembly and produces separate assemblies for each haplotype. For the CTLR-Seq libraries, the target sequences served as seed sequences for the assemblies. For the whole-genome libraries, non-duplicated sequences downstream from the target regions were selected. The assembly quality was verified by converting the nucleotide sequence into fiber-FISH probe patterns. Assemblies were only retained when the two haplotypes of both the patient and the parent have identical probe patterns between the Fiber-FISH and ONT assemblies. To identify the rearrangement region, a multiple alignment between shared subunits among the two parental alleles was conducted using MAFFT (Katoh et al. 2019). A custom script was used to identify unique SNPs between these shared subunits, facilitating the identification of the transition between the two LCR22s.

Deletion (and inversion) breakpoint identification in B2-B2 LCR22-A/B patterns

In patients and parents of families AB001, AB002, and AB004, reads (partly) covering the wild type and the rearranged LCR22s were manually selected based on LCR22 flanking sequence. In the parents-of-origin, two haplotypes for LCR22-A and -B could be differentiated, based on SNPs in the flanking sequence. The composition of the reads was determined using BLAT (Kent 2002), and the repeat composition between the two B2 probes in the rearranged allele was determined using RepeatMasker (Smit et al. 2013).

Long-range PCR over the LCR22-A/B recombination site and PacBio sequencing

Long-range PCR was performed using the TaKaRa LA PCR kit (TaKaRa Bio). PCR conditions were optimized for the extension-annealing phase, taking into account the presence of AT-repeats (Inagaki et al. 2005): aspecific bands are present when the temperature is below 60°C and there is no reaction above 63°C. A single primer (5’-ATACTACTGTGGCTTTGTTCCAAAG) was used as both forward and reverse primer. PCR was performed by an initial denaturation of 2 minutes at 94°C, 30 cycli of 30 seconds at 94°C followed by 7 minutes at 63°C, and the final elongation was at 60°C for 10 minutes. Fragments were analyzed on agarose gel.

A PacBio library was generated from the amplicons according to the Template Preparation and Sequencing protocol (Template Prep kit 3.0, Pacific Biosciences). Four libraries (22q11.2 patient AB004, his two children with a 22q11.2 deletion and the mother of the children) were pooled and loaded onto a single SMRT cell on a PacBio RSII using a DNA/polymerase binding kit P6 v2 (Pacific

Biosciences) loading concentration 25pM) and DNA Sequencing Reagent kit 4.0 v2 (Pacific Biosciences). The RS_Long_Amplicon_Analysis.1 pipeline was used for analysis.

Low-pass sequencing

The mosaic LCR22-A/B (AB002) deletion was detected in context of non-invasive prenatal testing of the pregnant patient JV2001. Procedures for non-invasive prenatal testing were followed as described (Bayindir et al. 2015).

Array comparative genomic hybridization (ArrayCGH)

ArrayCGH was performed using the 60k CyoSure Constitutional v3 array (Oxford Gene Technology). Data analysis and visualization of the results was done using CytoSure Interpret Software (v4.10.44) with embedded Circular Binary Segmentation algorithm for automated copy number calling. The analysis was performed using hg19/GRCh37 genome build.

SNP array

Genotyping was performed using Illumina HumanCytoSNP-12 BeadChip according to the Illumina Infinium HD Ultra protocol. Genotype, logR ratio and B-allele frequency (BAF) were extracted from the raw intensity data using the GenomeStudio software (v2.0.5) with the embedded genotype calling algorithm.

Interphase FISH

Dual-color interphase-FISH was performed using Vysis DiGeorge LSI TUPLE1 (HIRA) Spectrum Orange / LSI ARSA Spectrum Green probe set (Abbott). 100 nuclei from blood, urine, and buccal mucosa were scored, by assessing the presence or absence of the Spectrum Orange fluorescent probe targeting the 22q11.2 HIRA region.

Targeted interphase-FISH

BAC DNA was extracted from BAC clones (BacPac Resources, CHORI, Oakland) using the Nucleobond Xtra BAC kit (Macherey-Nagel) and subsequently labeled (Nick translation protocol, Abbott Molecular Inc.) (**Table 3**). To score the inversion and deletion frequency, EBV transformed cells of AB002 patient, a normal control, and a non-mosaic heterozygous LCR22-A/B deletion patient were fixed in glass slides. For family BD001, slides were prepared using EBV transformed cells from the patient and both parents. To classify a chromosome as normal, inverted or containing a deletion, combinations of blue (CH17-320A22), orange (RP11-354K13), green (CH17-222C16) and/or red (CH17-389E17) BAC probes were used (**Supplemental Table S3**).

Ethnicity determination

Data from 1000genomes phase3 were downloaded, converted to PLINK format and pruned to remove variants in linkage disequilibrium using PLINK (v1.9). The VCF files of the 22q11 samples have been generated by mapping the reads to reference genome (T2T) with minimap2 (v2.25) and variant calling with the PEPPER-Margin-DeepVariant pipeline (v0.8). The VCF files have then been lifted over to hg38 using Picard (v2.18.23), normalized and converted to BCF using BCFtools (v1.9), converted to PLINK format, merged and pruned to remove variants in linkage disequilibrium using PLINK (v1.9). Variants successfully genotyped in at least 90% of the 22q11 samples and available in 1000genomes phase3 dataset have been selected to perform a principal component analysis using PLINK (v1.9). A total of 646 variants could be used for the analysis.

Relatedness calculation

The VCF files of the 22q11 samples (obtained as described above) have been filtered to keep only phased variants on autosomes prior to merging with PLINK (v1.9). Only variants successfully genotyped in 80% of the samples were retained for relatedness calculation with VCFtools (v0.1.16). All pairs of samples were classified as either Parent-Child or Unrelated depending on the *a priori* relatedness knowledge. All duos consisting of patients only or parents only were considered as unrelated. Duos consisting of parents and patients from different families were also considered as unrelated.

Analysis of repetitive element composition in PATRR sites

Assemblies were downloaded from the UCSC Genome Browser (hg38), Telomere-to-Telomere Consortium (Nurk et al. 2022b; Raae et al. 2022) (T2T-CHM13v2.0 and HG002 (v0.6)) and the Human Pangenome Reference Consortium (Liao et al. 2023) (HG005, HG00733 and HG01109). Contigs spanning the 22q11.2 region were selected by mapping to hg38 or T2T using minimap2 (Li 2018). RepeatMasker (Smit et al. 2013) (version 4.1.6) with HMMER (version 3. 2. 1.) search engine was run to chart the repeat composition. To determine the copy number, the PATRR located between probes D3 and B2 was extracted from the RepeatMasker outputs.

Data access

The long read sequencing data generated in this study have been submitted to the European Genome-Phenome Archive (EGA; https://ega-archive.org/) under accession number EGAD50000000855. The script used to get the local assembly is available in GitHub (https://github.com/JorisVermeeschLab/NOVOLoci) and as Supplemental Code. The seeds used to get the local assemblies using the long read data and NOVOLoci as well as the resulting assemblies and the scripts used to verify the probe content and get the SNPs from the MAFFT alignment are all available at GitHub (https://github.com/JorisVermeeschLab/22q11_breakpoint/tree/main) and as Supplemental Data. All Fiber FISH data is available as Supplemental Material, and the cell lines used to map repeats are available upon request.

Competing interest statement

The authors declare to have no competing interests.

Acknowledgements

This research was funded by Fonds Wetenschappelijk Onderzoek (grant numbers G0E1117N and G0A2622N to J.R.V.), KU Leuven (C14/18/092 to J.R.V.), Stichting Marguerite-Marie Delacroix (GV/B-453 to L.V.) and by funds from the National Institute of General Medical Sciences (GM125757 to B.S.E).

Author contributions

L.V. lead conception of the work and experimental design using fiber FISH and ONT. A.S., E.V informed parents and patients during hospital follow-up and collected blood samples. Data was collected by L.V., M.S.S., S.M., R.C., A.S., H.V.E., J.B. and L.V., N.D. and E.S. performed data analysis and interpretation. J.R.V., L.V., N.D., M.S.S., and E.S. drafted the article, which was critically revised by J.R.V., B.M., B.S.E., T.H.S., A.S., M.X., M.S.H., D.M.M. Final approval of the version to be published was given by L.V., N.D., M.S.S, S.M., E.S., R.C., T.H., K.D., H.P., D.M.M., A.S., J.B., B.S.E., H.V.E., A.S.B. and J.R.V.

Figure Legends:

Fig. 1: Fiber-FISH analysis of a LCR22-mediated (A-D) rearrangement. (A) UCSC Genome Browser screenshot of T2T with tracks for segmental duplications and BLAT sequences for fiber-FISH probes in the 22q11.2 region LCR22-A to -D **(up)** and a close up of LCR22-A **(down). (B)** UCSC Genome Browser screenshot of hg38 with tracks for gaps, segmental duplications and BLAT sequences for fiber-FISH probes in LCR22-A. **(C)** LCR22-A/D rearrangement with *de novo* assembled parental LCR22-A, -D allele that contributed to the rearrangement, and 22q11.2DS patient rearrangement alleles. The parental LCR22-A haplotype is not identical to the reference allele in A, due to presence of gaps in the reference and haplotype variability observed in the human population. The sequence in the red box is shared between three alleles and therefore considered to be the locus where putative recombination had taken place. Gray lines indicate similarities between the alleles of the patient and the parent-of-origin. **(D)** UCSC Genome Browser screenshot of hg38 with tracks for fiber-FISH probes and segmental duplications in LCR22-D.

Fig. 2: Study design and data analysis of 22q11.2DS recombination loci. (A) LCR22 composition of a family trio. The parent-of-origin is the parent in whom the recombination occurred. The red-blue composition in the patient represents the recombined LCR22 structure. In this example, an interchromosomal recombination is presented, although intrachromosomal recombinations are possible as well. **(B)** The patient and the parent-of-origin were sequenced using one or a combination of whole-genome ultra-long and/or standard-long Oxford Nanopore Technologies (ONT) sequencing. In one patient (AB004) a targeted breakpoint-specific long-range PCR was designed in combination with PacBio single-molecule real-time (SMRT) sequencing of the fragment. **(C)** Following *de novo* assembly of the different proximal and distal LCR22 alleles in the patient and parent-of-origin, the recombination-involved parental haplotypes (proximal and distal) were identified via SNP comparison, followed by a delineation of the breakpoint locus by LCR22 proximal-specific (red) and distal-specific (blue) SNPs. These SNPs are shared between the patient and the parental proximal or distal LCR22, respectively, but not with the other (proximal or distal) LCR22 involved in the CNV. The recombination locus was scrutinized or its precise coordinates, elucidating the genes and repetitive elements involved.

Fig. 3: Schematic visualization of 22q11.2 recombination breakpoints. The graph represents the 22q11.2 locus, its segmental duplications and known genes. Lines between LCR22s connect the proximal and distal breakpoint of a family-specific recombination. Solid red lines represent PATRR-mediated recombinations, and solid blue lines represent NAHR-mediated recombinations. The position of genes located in the involved recombination sequences are shown via blue or green vertical dotted lines.

Fig. 4: Inversion-associated rearrangements. (A) Schematic representation of the rearrangement of family BD001: the parent-of-origin (left) carries a LCR22-A/D inversion, which is recombined between LCR22-A and -B (dotted line) causing the presence of a deletion allele in the patient, missing the sequence from LCR22-B until -D. In addition, the patient inherited one of the wild type alleles of the other parent (right). Colored dots represent the probes used in the interphase-FISH. **(B)** Interphase-FISH results using color-labeled probes CH17-320A22 (proximal LCR22-A, blue, B), CH17-222C16 (distal LCR22-A, green, G), and CH17-389E17 (proximal LCR22-B, red, R) in the parent-of-origin (left), patient BD001 (middle), and the other parent (right). **(C)** Schematic representation of the rearrangement of family AB002: both parents carry two normal LCR22 haplotypes (LCR22-A until -D). The patient carries a wild type allele (right) and an LCR22-A/B inversion allele is observed (inverted gray arrow), created by the recombination (dotted line and arrow) of the LCR22-A and -B allele from the parent-of-origin. We hypothesize the deletion is created in a consecutive manner by a new recombination between these blocks. Colored dots represent the probes used in the interphase-FISH. **(D)**

Interphase-FISH results using color-labeled probes CH17-320A22 (proximal LCR22-A, blue, B), CH17-222C16 (distal LCR22-A, green, G), CH17-389E17 (proximal LCR22-B, red, R), and RP11-354K13 (distal LCR22-D, yellow, Y) in a control individual (left), deletion-carrying cell of patient AB002 (middle), and inversion-carrying cell of patient AB002 (right).

Table Legends

Table 1: 22q11.2 LCR rearrangement loci at fiber-FISH resolution. The chromosomal locus corresponds to the genomic location in reference genome hg38. CNV in superscript indicates the presence of copy number variability of the recombination locus in the LCR22 (-A and -D), based on (Demaerel et al. 2019) and (Pastor et al. 2020).

Table 2: 22q11.2 LCR rearrangement loci (RL) at nucleotide resolution. First column shows the family identifier, based on **Supplementary_Table_1**, and diagnosed deletion type. The fiber-FISH based (**Table 1**) as well as the length of the RL is presented in the second column. The third column provides the position of the last proximal-specific SNP and the first distal-specific SNP and the length of the locus. The Exact position could not be mapped for the three PATRR-mediated LCR22-A/B deletions due to presence of several PATRRs in the reference genome. Columns four and five show the gene and repeat content, respectively. Repetitive elements and gene exons/introns can be fully or partly covered in the recombination locus. More information on the exact composition of the recombination locus can be found in **Supplementary_Fig_3**. In family BD001, the deletion is diagnosed as LCR22-B/D, however, the recombination occurred between LCR22-A and -B (corresponding to chromosomal loci in the third column), as explained in **Figure 4C**. (**RL** - Recombination Locus, **SLR** – Standard Long Read sequencing; **ULR** – Ultra-long Read sequencing; **HDR** – High Duplex Read sequencing)

Table 3: Extracted BAC clones with LCR22-related location and labeling. The two last columns indicate whether the probe was used to score inversion status in the corresponding families.

References

Angelis A, Tordrup D, Kanavos P. 2015. Socio-economic burden of rare diseases: A systematic review of cost of illness evidence. *Health Policy (New York)* **119**: 964–979.

Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, Rosenfeld JA, Miroballo M, Graves TA, Vives L, Malig M, et al. 2014. Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat Genet* **46**: 1293–1302. https://pubmed.ncbi.nlm.nih.gov/25326701/ (Accessed July 8, 2024).

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte R V., Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent Segmental Duplications in the Human Genome. *Science (1979)* **297**: 1003–1007.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. *Genome Res* **11**: 1005–1017.

Bayindir B, Dehaspe L, Brison N, Brady P, Ardui S, Kammoun M, Van Der Veken L, Lichtenbelt K, Van Den Bogaert K, Van Houdt J, et al. 2015. Noninvasive prenatal testing using a novel analysis pipeline to screen for all autosomal fetal aneuploidies improves pregnancy management. *European Journal of Human Genetics* **23**: 1286–1293.

Blagojevic C, Heung T, Theriault M, Tomita-Mitchell A, Chakraborty P, Kernohan K, Bulman DE, Bassett AS. 2021. Estimate of the contemporary live-birth prevalence of recurrent 22q11.2 deletions: a cross-sectional analysis from population-based newborn screening. *CMAJ Open* **9**: E802–E809.

Boettger LM, Handsaker RE, Zody MC, Mccarroll SA. 2012. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**: 881–885.

Campbell IM, Sheppard SE, Crowley TB, McGinn DE, Bailey A, McGinn MJ, Unolt M, Homans JF, Chen EY, Salmons HI, et al. 2018. What is new with 22q? An update from the 22q and You Center at the Children's Hospital of Philadelphia. *Am J Med Genet A* **176**: 2058–2069.

Chen CP, Chern SR, Lee CC, Lin SP, Chang TY, Wang W. 2004. Prenatal diagnosis of mosaic 22q11.2 microdeletion. *Prenat Diagn* **24**: 660–662.

Chen W, Li X, Sun L, Sheng W, Huang G. 2019. A rare mosaic 22q11.2 microdeletion identified in a Chinese family with recurrent fetal conotruncal defects. *Mol Genet Genomic Med* **7**.

Consevage MW, Seip JR, Belchis DA, Davis AT, Baylen BG, Rogan PK. 1996. Association of a mosaic chromosomal 22q11 deletion with hypoplastic left heart syndrome. *American Journal of Cardiology* **77**: 1023–1025.

Correll-Tash S, Lilley B, Salmons Iv H, Mlynarski E, Franconi CP, McNamara M, Woodbury C, Easley CA, Emanuel BS. 2021. Double strand breaks (DSBs) as indicators of genomic instability in PATRR-mediated translocations. *Hum Mol Genet* **29**: 3872–3881.

Demaerel W, Mostovoy Y, Yilmaz F, Vervoort L, Pastor S, Hestand MS, Swillen A, Vergaelen E, Geiger A, Coughlin CR, et al. 2019. The 22q11 low copy repeats are characterized by unprecedented size and structural variability. *Genome Res* **29**: 1389–1401.

Dennis MY, Eichler EE. 2016. Human adaptation and evolution by segmental duplication. *Curr Opin Genet Dev* **41**: 44–52.

Gebhardt GS, Devriendt K, Thoelen R, Swillen A, Pijkels E, Fryns J-P, Vermeesch JR, Gewillig M. 2003. No evidence for a parental inversion polymorphism predisposing to rearrangements at 22q11.2 in the DiGeorge/Velocardiofacial syndrome. *European Journal of Human Genetics* **11**: 109–111.

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.

Halder A, Jain M, Kabra M, Gupta N. 2008. Mosaic 22q11.2 microdeletion syndrome: diagnosis and clinical manifestations of two cases. *Mol Cytogenet* **1**: 18.

Hanlon VCT, Lansdorp PM, Guryev V. 2022. A survey of current methods to detect and genotype inversions. *Hum Mutat* **43**: 1576–1589.

Inagaki H, Ohye T, Kogo H, Yamada K, Kowa H, Shaikh TH, Emanuel BS, Kurahashi H. 2005. Palindromic AT-rich repeat in the NF1 gene is hypervariable in humans and evolutionarily conserved in primates. *Hum Mutat* **26**: 332–342.

Inoue K, Lupski JR. 2002. Molecular Mechanisms for Genomic Disorders. *Annu Rev Genomics Hum Genet* **3**: 199–242.

Kato T, Inagaki H, Yamada K, Kogo H, Ohye T, Kowa H, Nagaoka K, Taniguchi M, Emanuel BS, Kurahashi H. 2006. Genetic variation affects de novo translocation frequency. *Science (1979)* **311**: 971.

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

Kato T, Kurahashi H, Emanuel BS. 2012b. Chromosomal translocations and palindromic AT-rich repeats. *Curr Opin Genet Dev* **22**: 221–228. https://pubmed.ncbi.nlm.nih.gov/22402448/ (Accessed January 7, 2023).

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* **20**: 1160–1166.

Kent WJ. 2002. BLAT — The BLAST -Like Alignment Tool. *Genome Res* **12**: 656–664.

Kurahashi H, Inagaki H, Ohye T, Kogo H, Kato T, Emanuel BS. 2006. Chromosomal Translocations Mediated by Palindromic DNA. *Cell cycle* **5**: 1297–1303.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324. https://pubmed.ncbi.nlm.nih.gov/37165242/ (Accessed July 9, 2024).

Maggiolini FAM, Cantsilieris S, D'Addabbo P, Manganelli M, Coe BP, Dumont BL, Sanders AD, Pang AWC, Vollger MR, Palumbo O, et al. 2019. Genomic inversions and GOLGA core duplicons underlie disease instability at the 15q25 locus. *PLoS Genet* **15**. https://pubmed.ncbi.nlm.nih.gov/30917130/ (Accessed July 8, 2024).

McDonald-McGinn D, Sullivan K, Marino B, Philip N, Swillen A, Vorstman J, Zackai E, Emanuel B, Vermeesch J, Morrow B, et al. 2015. 22q11.2 Deletion Syndrome. *Nat Rev Dis Primers* **1**.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022a. The complete sequence of a human genome. *Science (1979)* **376**: 44–53.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022b. The complete sequence of a human genome. *Science* **376**: 44–53. https://pubmed.ncbi.nlm.nih.gov/35357919/ (Accessed September 24, 2022).

Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Narvaiza I, Sudmant PH, Penn O, Chiatante G, Malig M, Huddleston J, et al. 2016. Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**: 205–209. https://pubmed.ncbi.nlm.nih.gov/27487209/ (Accessed July 8, 2024).

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

Paparella A, L'Abbate A, Palmisano D, Chirico G, Porubsky D, Catacchio CR, Ventura M, Eichler EE, Maggiolini FAM, Antonacci F. 2023. Structural Variation Evolution at the 15q11-q13 Disease-Associated Locus. *Int J Mol Sci* **24**. https://pubmed.ncbi.nlm.nih.gov/37958807/ (Accessed July 8, 2024).

Pastor S, Tran O, Jin A, Carrado D, Silva BA, Uppuluri L, Abid HZ, Young E, Crowley TB, Bailey AG, et al. 2020. Optical mapping of the 22q11.2DS region reveals complex repeat structures and preferred locations for non-allelic homologous recombination (NAHR). *Sci Rep* **10**: 1–13.

Patel ZM, Gawde HM, Khatkhatay MI. 2006. 22q11 microdeletion studies in the heart tissue of an abortus involving a familial form of congenital heart disease. *J Clin Lab Anal* **20**: 160–163.

Porubsky D, Harvey WT, Rozanski AN, Ebler J, Höps W, Ashraf H, Hasenfeld P, Paten B, Sanders AD, Marschall T, et al. 2023. Inversion polymorphism in a complete human genome assembly. *Genome Biol* **24**.

Porubsky D, Höps W, Ashraf H, Hsieh PH, Rodriguez-Martin B, Yilmaz F, Ebler J, Hallast P, Maria Maggiolini FA, Harvey WT, et al. 2022. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**: 1986-2005.e26.

Puig M, Casillas S, Villatoro S, Cáceres M. 2015. Human inversions and their functional consequences. *Brief Funct Genomics* **14**: 369.

Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, Eichler EE, Phillippy AM, Koren S. 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**: 1474–1482.

Robberecht C, Voet T, Esteki MZ, Nowakowska BA, Vermeesch JR. 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res* **23**: 411–418. https://pubmed.ncbi.nlm.nih.gov/23212949/ (Accessed July 8, 2024).

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. *Nat Biotechnol* **29**: 24–26.

Shaikh TH, Kurahashi H, Saitta SC, Mizrahy O'Hare A, Hu P, Roe BA, Driscoll D a, McDonald-McGinn DM, Zackai EH, Budarf ML, et al. 2000. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* **9**: 489–501.

Shaikh TH, O'Connor RJ, Pierpont ME, McGrath J, Hacker AM, Nimmakayalu M, Geiger E, Emanuel BS, Saitta SC. 2007. Low copy repeats mediate distal chromosome 22q11.2 deletions: Sequence analysis predicts breakpoint mechanisms. *Genome Res* **17**: 482–491.

Shaw CJ, Lupski JR. 2004. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet* **13 Spec No**: R57–R64.

Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

Song X, Beck CR, Du R, Campbell IM, Coban-Akdemir Z, Gu S, Breman AM, Stankiewicz P, Ira G, Shaw CA, et al. 2018. Predicting human genes susceptible to genomic instability associated with Alu/Alu-mediated rearrangements. *Genome Res* **28**: 1228–1242. https://pubmed.ncbi.nlm.nih.gov/29907612/ (Accessed July 8, 2024).

Steinberg KM, Antonacci F, Sudmant PH, Kidd JM, Campbell CD, Vives L, Malig M, Scheinfeldt L, Beggs W, Ibrahim M, et al. 2012. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat Genet* **44**: 872–880.

Summerer A, Mautner VF, Upadhyaya M, Claes KBM, Högel J, Cooper DN, Messiaen L, Kehrer-Sawatzki H. 2018. Extreme clustering of type-1 NF1 deletion breakpoints co-locating with G-quadruplex forming sequences. *Hum Genet* **137**: 511–520.

Tong M, Kato T, Yamada K, Inagaki H, Kogo H, Ohye T, Tsutsumi M, Wang J, Emanuel BS, Kurahashi H. 2010. Polymorphisms of the 22q11.2 breakpoint region influence the frequency of de novo constitutional t(11;22)s in sperm. *Hum Mol Genet* **19**: 2630–2637.

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**: 1–11.

Visser R, Shimokawa O, Harada N, Kinoshita A, Ohta T, Niikawa N, Matsumoto N. 2005a. Identification of a 3.0-kb Major Recombination Hotspot in Patients with Sotos Syndrome Who Carry a Common 1.9-Mb Microdeletion. *The American Journal of Human Genetics* **76**: 52–67.

Visser R, Shimokawa O, Harada N, Niikawa N, Matsumoto N. 2005b. Non-hotspot-related breakpoints of common deletions in Sotos syndrome are located within destabilised DNA regions. *J Med Genet* **42**.

Vollger MR, Dishuck PC, Sorensen M, Welch AME, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat Methods* **16**: 88–94.

Table 1

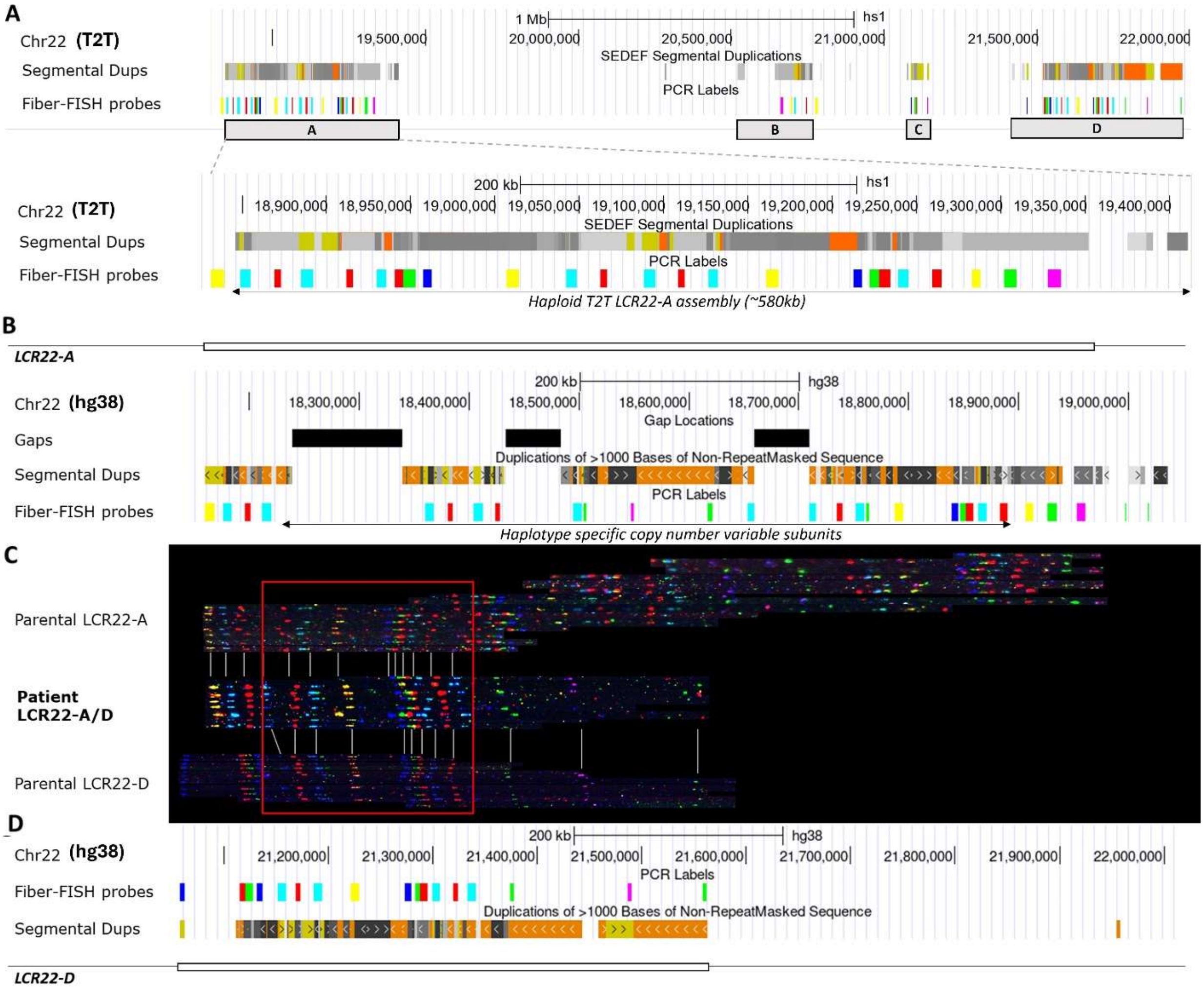
Deletion type	# of samples	Recombination locus and size	Locus in proximal LCR22	Locus in distal LCR22
LCR22-A/D	13	RL-AD1 (160kb)	chr22:18,729,876- 18,891,257^{cnv}	chr22:21,162,204- 21,325,236^{cnv}
	2	RL-AD2 (20kb)	chr22:18,838,944- 18,859,182^{cnv}	chr22:21,115,387- 21,137,124^{cnv}
LCR22-A/B	2	RL-B1 (20kb)	chr22:18,196,206- 18,218,744^{cnv}	chr22:20,324,573- 20,344,531
	3	RL-B2 (8kb)	chr22:18,414,455- 18,422,770^{cnv}	chr22:20,331,987- 20,339,583
LCR22-A/C	3	RL-C (10kb)	chr22:18,845,620- 18,855,844^{cnv}	chr22:20,688,715- 20,698,995
LCR22-B/D	1	RL-B1 (20kb)	chr22:20,324,573- 20,344,531	chr22:21,152,820- 21,173,481^{cnv}
LCR22-C/D	1	RL-C (10kb)	chr22:20,688,715- 20,698,995	chr22:21,279,982- 21,291,168^{cnv}

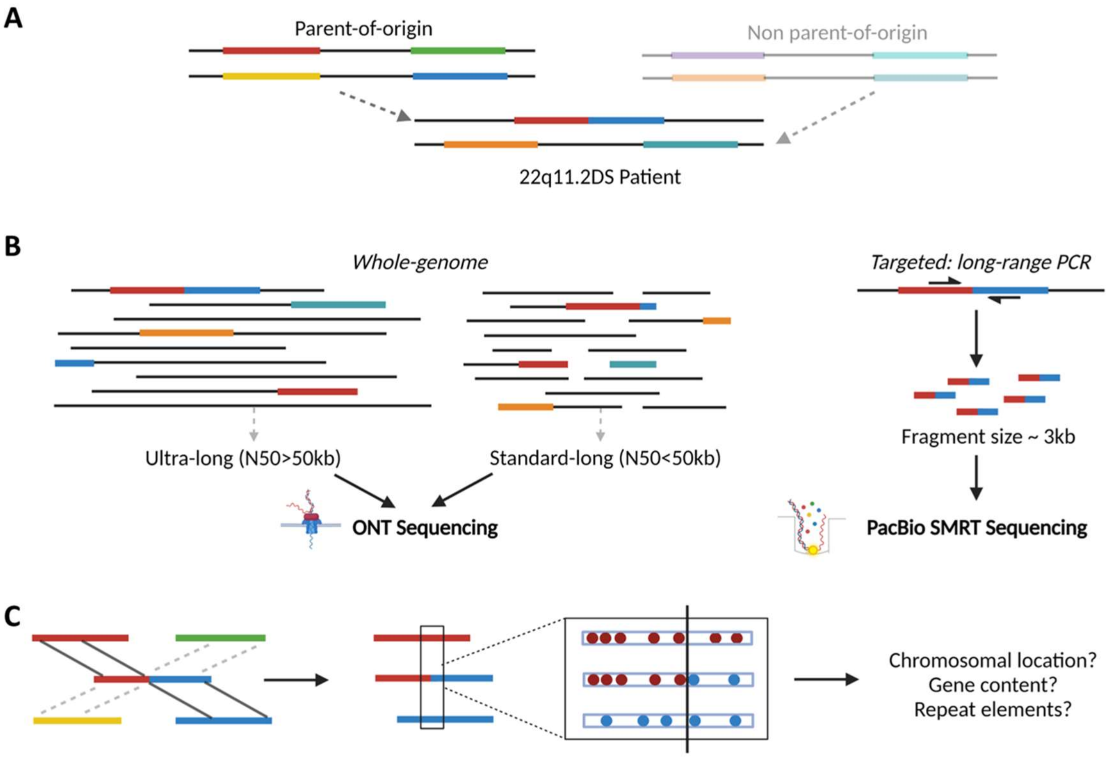
Table 2

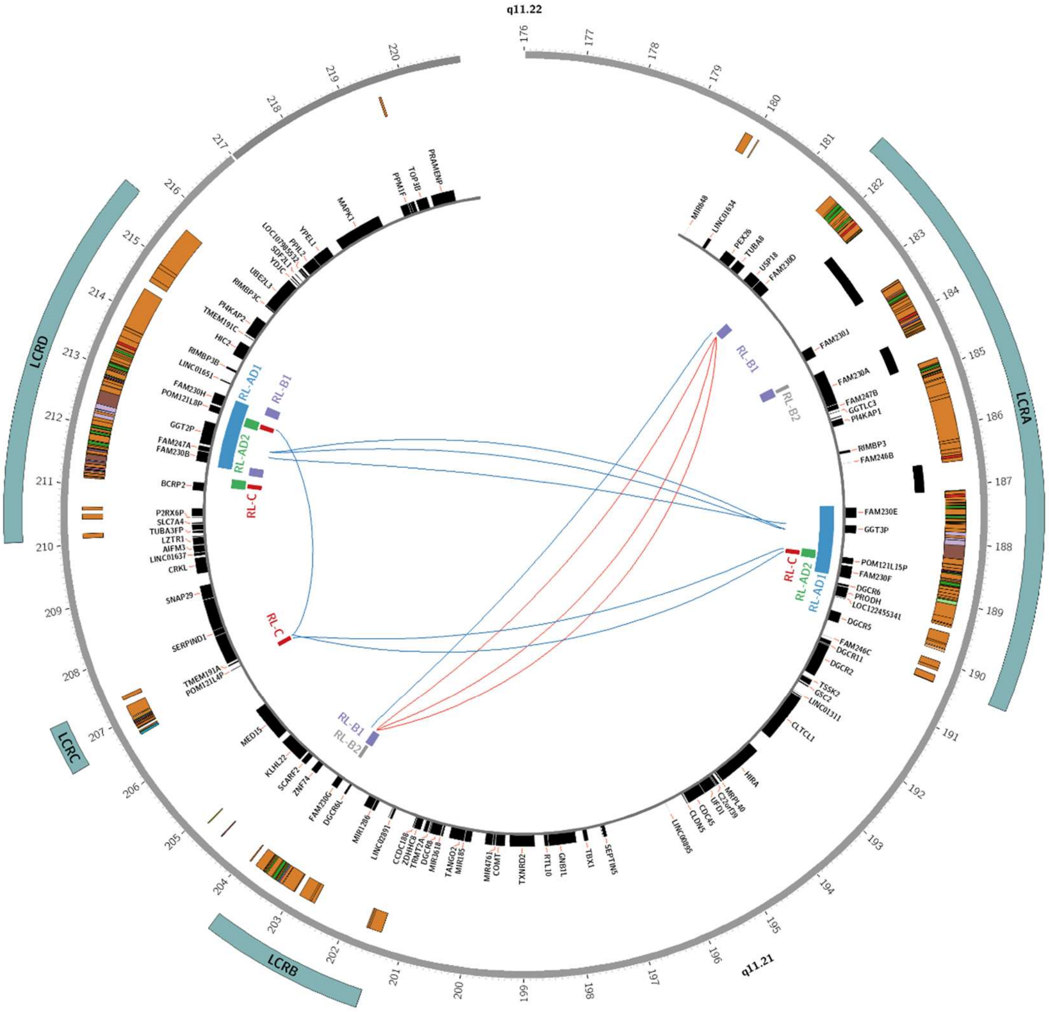
Family (Deletion)	Sequencing technology	RL	Nucleotide level rearrangement locus (hg38)	Gene content	Repeat content
AD004 (LCR22-A/D)	SLR, ULR, HDR (ONT)	RL-AD1 (160kb)	319 bp chr22:18,775,948-18,776,026 (A) chr22:21,210,278-21,210,356 (D)	intronic: *GGT3P* (A) *GGT2* (D)	L2 LINE
AD007 (LCR22-A/D)	SLR (ONT)	RL-AD1 (160kb)	421 bp chr22:18,791,335-18,791,756 (A) chr22:21,225,667-21,226,088 (D)	exonic/intronic: *GGT3P* (A) *GGT2* (D)	L2 LINE
AD008 (LCR22-A/D)	ULR (ONT)	RL-AD1 (160kb)	650 bp chr22:18,791,386-18,792,037 (A) chr22:21,225,716-21,226,367 (D)	exonic/intronic: *GGT3P* (A) *GGT2* (D)	L2 LINE
AB001 (LCR22-A/B)	ULR (ONT)	RL-B2 (8kb)	AT repeat within 8kb locus chr22:18,201,246-18,207,385 (A) chr22:20,336,265-20,339,445 (D)	intronic: *FAM230D* (A) *FAM230G* (B)	PATRR
AB002 (LCR22-A/B)	ULR (ONT)	RL-B2 (8kb)	AT repeat within 8kb locus chr22:18,201,246-18,207,385 (A) chr22:20,336,265-20,339,445 (D)	intronic: *FAM230D* (A) *FAM230G* (B)	PATRR
AB004 (LCR22-A/B)	Long-range PCR and PacBio SMRT sequencing	RL-B2 (8kb)	AT repeat within 8kb locus chr22:18,201,246-18,207,385 (A) chr22:20,336,265-20,339,445 (D)	intronic: *FAM230D* (A) *FAM230G* (B)	PATRR
AC001 (LCR22-A/C)	ULR (ONT)	RL-C (10kb)	119 bp chr22:18,855,275-18,855,394 (A) chr22:20,698,426-20,698,545 (C)	-	AluJo SINE
AC002 (LCR22-A/C)	ULR (ONT)	RL-C (10kb)	866 bp chr22:18,845,968-18,846,834 (A) chr22:20,689,067-20,689,933 (C)	exonic: *POM121L15P* (A) *POM121L4P* (D)	AluS SINE
BD001 (LCR22-B/D)	STR, ULR (ONT)	RL-B1 (20kb)	200 bp chr22:18,187,091-18,187,290 (A) chr22:20,353,454-20,353,653 (B)	intronic: *FAM230D* (A) *FAM230G* (B)	/
CD001 (LCR22-C/D)	STR, ULR (ONT)	RL-C (10kb)	200 bp chr22:20,697,310-20,697,507 (C) chr22:21,290,100-21,290,300 (D)	intronic: *POM121L8P* (D)	/

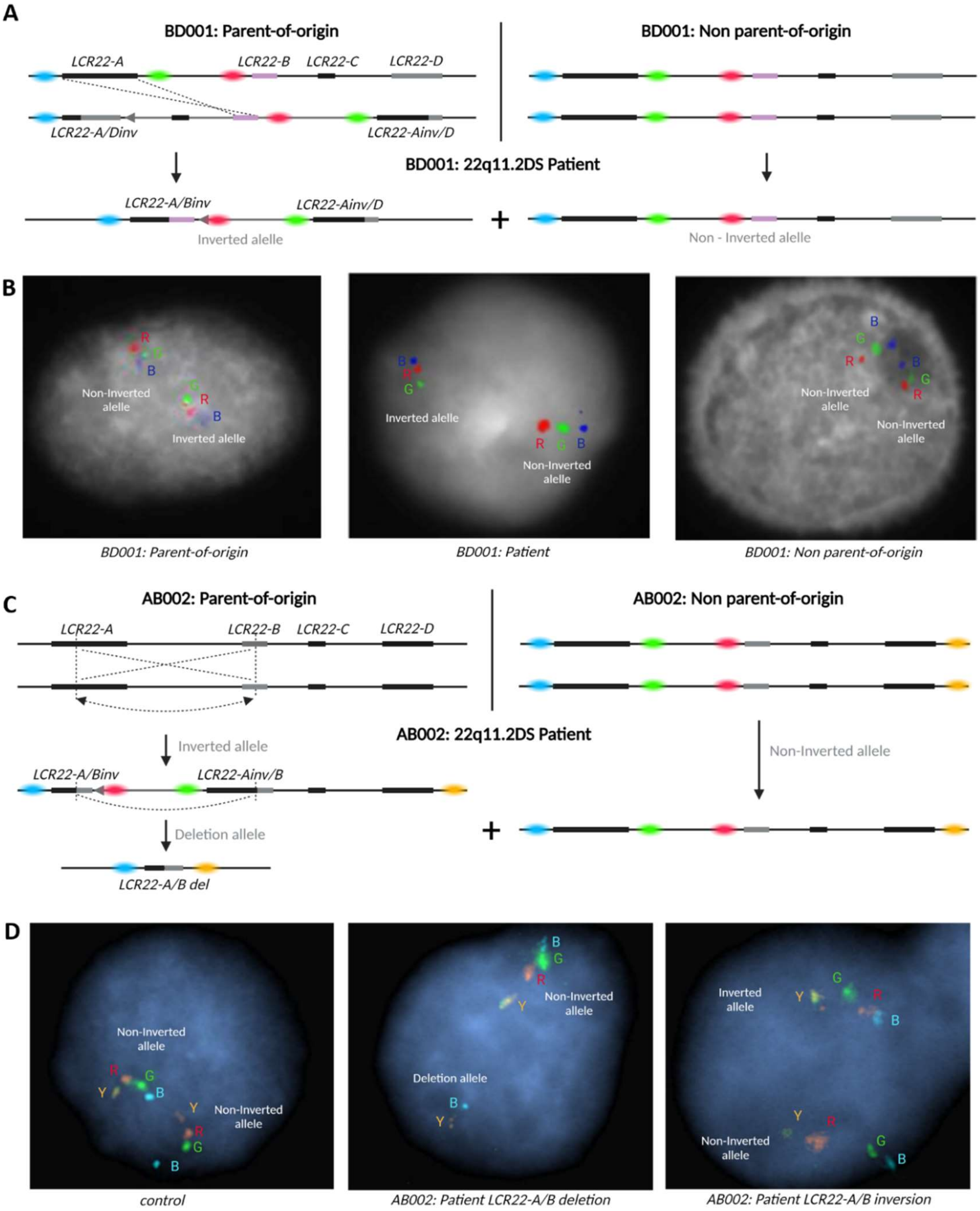
Table 3

BAC clone	**Location**	**Labeling**	**AB002**	**BD001**
CH17-320A22	Proximal LCR22-A	Aqua 431dUTP	x	x
CH17-222C16	Distal LCR22-A	Spectrum Green	x	x
CH17-389E17	Proximal LCR22-B	Spectrum Orange	x	x
RP11-354K13	Distal LCR22-D	Spectrum Green + Spectrum Orange	x	


A
Chr22 (T2T)
Segmental Dups
Fiber-FISH probes
1 Mb
hs1
19,500,000
20,000,000
20,500,000
21,000,000
21,500,000
22,000,000
SEDEF Segmental Duplications
PCR Labels
A
B
C
D
200 kb
18,900,000
18,950,000
19,000,000
19,050,000
19,100,000
19,150,000
19,200,000
19,250,000
19,300,000
19,350,000
19,400,000
Haploid T2T LCR22-A assembly (~580kb)
B
LCR22-A
Chr22 (hg38)
Gaps
Segmental Dups
Fiber-FISH probes
200 kb
hg38
18,300,000
18,400,000
18,500,000
18,600,000
18,700,000
18,800,000
18,900,000
19,000,000
Gap Locations
Duplications of >1000 Bases of Non-RepeatMasked Sequence
PCR Labels
Haplotype specific copy number variable subunits
C
Parental LCR22-A
Patient LCR22-A/D
Parental LCR22-D
D
Chr22 (hg38)
Fiber-FISH probes
Segmental Dups
200 kb
hg38
21,200,000
21,300,000
21,400,000
21,500,000
21,600,000
21,700,000
21,800,000
21,900,000
22,000,000
PCR Labels
Duplications of >1000 Bases of Non-RepeatMasked Sequence
LCR22-D


A
Parent-of-origin
Non parent-of-origin
22q11.2DS Patient
B
Whole-genome
Targeted: long-range PCR
Ultra-long (N50>50kb)
Standard-long (N50<50kb)
ONT Sequencing
Fragment size ~ 3kb
PacBio SMRT Sequencing
C
Chromosomal location?
Gene content?
Repeat elements?


q11.22
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
q11.21
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
LCRA
LCRB
LCRC
LCRD
MIR648
LINC01634
PEX26
TUBA8
USP18
FAM230D
FAM230J
FAM230A
FAM247B
GGTLC3
PI4KAP1
RIMBP3
FAM246B
FAM230E
GGT3P
POM121L15P
FAM230F
DGCR6
PRODH
LOC122455341
DGCR5
FAM246C
DGCR11
DGCR2
TSSK2
GSC2
LINC01311
CLTCL1
HIRA
MRPL40
C22orf39
UFD1
CDC45
CLDN5
LINC00895
SEPTIN5
TBX1
GNB1L
RTL10
TXNRD2
COMT
MIR4761
MIR185
TANGO2
MIR3618
DGCR8
TRMT2A
ZDHHC8
CCDC188
LINC02891
MIR1286
DGCR6L
FAM230G
ZNF74
SCARF2
KLHL22
MED15
POM121L4P
TMEM191A
SERPIND1
SNAP29
CRKL
LINC01637
AIFM3
LZTR1
TUBA3FP
SLC7A4
P2RX6P
BCRP2
FAM230B
FAM247A
GGT2P
POM121L8P
FAM230H
LINC01651
RIMBP3B
HIC2
TMEM191C
PI4KAP2
RIMBP3C
UBE2L3
YDJC
SDF2L1
LOC107985532
PPIL2
YPEL1
MAPK1
PPM1F
TOP3B
PRAMENP
RL-AD1
RL-AD2
RL-B1
RL-B2
RL-C
RL-B1
RL-B2
RL-C
RL-AD2
RL-AD1
RL-C
RL-B1
RL-B2


A
BD001: Parent-of-origin
LCR22-A
LCR22-B
LCR22-C
LCR22-D
LCR22-A/Dinv
LCR22-Ainv/D
BD001: Non parent-of-origin
BD001: 22q11.2DS Patient
LCR22-A/Binv
LCR22-Ainv/D
Inverted allele
+
Non - Inverted alelle
B
R
G
B
Non-Inverted alelle
G
R
B
Inverted allele
BD001: Parent-of-origin
B
R
G
Inverted allele
R
G
B
Non-Inverted alelle
BD001: Patient
B
G
R
B
G
R
Non-Inverted alelle
Non-Inverted alelle
BD001: Non parent-of-origin
C
AB002: Parent-of-origin
LCR22-A
LCR22-B
LCR22-C
LCR22-D
AB002: Non parent-of-origin
AB002: 22q11.2DS Patient
Inverted allele
LCR22-A/Binv
LCR22-Ainv/B
Deletion allele
LCR22-A/B del
Non-Inverted allele
+
D
Non-Inverted allele
R
G
Y
B
Y
R
G
Non-Inverted allele
B
control
B
G
R
Y
Non-Inverted allele
Deletion allele
B
Y
AB002: Patient LCR22-A/B deletion
Inverted allele
G
Y
R
B
Y
R
Non-Inverted allele
G
B
AB002: Patient LCR22-A/B inversion

CSH PRESS GENOME RESEARCH

Multiple paralogues and recombination mechanisms contribute to the high incidence of 22q11.2 Deletion Syndrome

Lisanne Vervoort, Nicolas Dierckxsens, Marta Sousa Santos, et al.

Genome Res. published online November 13, 2024
Access the most recent version at doi:10.1101/gr.279331.124

Published by Cold Spring Harbor Laboratory Press