KU Leuven
Biomedical Sciences Group
Faculty of Medicine
Department of Human Genetics

KU LEUVEN

DOCTORAL SCHOOL
BIOMEDICAL SCIENCES

LOW COPY REPEATS FLANKING CHROMOSOME 22Q11.2 DELETION SYNDROME

Lisanne VERVOORT

Jury:

Supervisor:	Prof. Dr. Joris Vermeesch
Co-supervisor:	Prof. Dr. Jeroen Breckpot
Chair examining committee:	Prof. Dr. Tassos Economou
Chair public defense:	Prof. Dr. Jan Cools
Jury members:	Prof. Dr. Koen Devriendt
	Prof. Dr. Stephan Claes
	Prof. Dr. Björn Menten
	Prof. Dr. Beverly Emanuel
	Prof. Dr. Alexander Urban

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Biomedical Sciences

September 2022

Author: Lisanne Vervoort
Cover design: Lisanne Vervoort, Tatjana Jatsenko, and Olga Tsuiko

Printed by Procopia

PERSONAL ACKNOWLEDGEMENTS

"All is possible, when you have the right people there to support you"

Vijf jaar lang heb ik mij dag in dag uit mogen amuseren met DNA, microscopen en cellen. Ik ben dan ook zeer blij en trots met het resultaat van dit boekje, maar zoals de quote hierboven aangeeft, is dit geen werk van mij alleen geweest. Daarom zou ik hier toch even een welgemeende dankuwel willen zeggen aan alle mensen die mij gedurende deze vijf jaar begeleid en gesteund hebben.

De persoon die ik hierbij als eerste moet en wil bedanken, is Joris. Ik begon zonder enige voorkennis en ervaring aan dit project, en ben hem dan ook enorm dankbaar voor de kans om te starten in zijn labo. Vele meetings en ideeën hebben over de jaren heen dit werk vormgegeven. Uiteindelijk hebben we ook samen een stukje van de wereld (en luchthavens) gezien (Canada, Zweden, Oostenrijk en Kroatië) en heeft hij mij de kans gegeven om ons werk internationaal voor te stellen. Bedankt Joris, om mij kennis te laten maken met de fantastische wereld van de genetica, mij te laten verdiepen in het onderwerp van structurele variatie en mij alle kansen te geven die hebben geleid tot enorm mooie en waardevolle herinneringen, zowel op wetenschappelijk als persoonlijk gebied.

Ook mijn co-promotor, Jeroen, stond steeds voor mij klaar om klinische zaken te verduidelijken, bloedstalen te verzamelen en advies te geven vanuit België of Canada. Bedankt Jeroen, om vijf jaar geleden deze rol op te nemen en mij zo goed mogelijk te ondersteunen. Bedankt aan de twee interne KU Leuven juryleden, Professoren Stephan Claes en Koen Devriendt, om mij bij te sturen met het nodige advies en kritisch te laten denken door de juiste opmerkingen gedurende de jaarlijkse evaluaties. Thank you to the external jury members, Professors Beverly Emanuel, Alexander Urban, and Björn Menten, for carefully reading my thesis manuscript, and providing corrections and input to improve the quality of this thesis. Ten slotte zou ik ook Professor Jan Cools willen bedanken, voor het accepteren van de rol als moderator gedurende mijn publieke verdediging.

Als doctoraatsstudent behoorde ik tot het Centrum Menselijke Erfelijkheid (CME) aan het UZ en KU Leuven. Tot deze groep behoren de klinisch genetici, die ik wil bedanken voor hun hulp bij het verzamelen van stalen van 22q11.2 deletie syndroom patiënten en hun ouders, en voor hun tijd om deze families correct te informeren over het onderzoek en de resultaten. De diagnostische teams stonden steeds klaar om mij meer informatie te geven over specifieke protocollen of de interpretatie van genetische testresultaten. Verder ook bedankt aan alle mensen van de Genomics Core, voor de hulp en het advies bij alle kleine of grote vragen omtrent library protocols en bioinformatica codetaal. De nauwe samenwerking tussen clinici, diagnostiek en onderzoekers is zeer uniek en vormt dan ook de basis voor deze stimulerende omgeving. Verder mogen ook Annemie, Katleen, Nathalie, Veerle en Erik niet ontbreken voor het afhandelen van alle logistieke en administratieve zaken. En tot slot kan ik niet genoeg bedankt zeggen aan

het weefselkweek team van Wim en zijn collega's: ik denk niet dat ik overdrijf als ik zeg dat zij de laatste vijf jaar minstens 200 cellijnen voor mij hebben moeten ontdooien, opgroeien en invriezen. Daarom nogmaals: bedankt aan elk CME-lid dat heeft bijgedragen aan het 22q11 onderzoek! Zonder jullie enthousiasme had ik dit nooit kunnen verwezenlijken.

Gedurende vijf jaar heb ik nauw mogen samenwerken met het '22q11.2 team' van Leuven, een gepassioneerde groep van onderzoekers en clinici. Hierbij zou ik graag mijn dank uitdrukken aan Wolf, om me álles te leren over dit syndroom, de low copy repeats en uiteraard de fiber-FISH methode. Verder ook aan Professor Ann Swillen, om deze groep enthousiast te leiden, input te geven aan ons onderzoek en de jaarlijkse 22q11 oudermeeting te organiseren. Several people were part of this dynamic group: Adrian, Benjamin, Elfi, Jente, Kris, Nehir, Nicolas, and Ruben. I would like to express my gratitude to each of you for the collaborations and your significant contributions to the research field of our syndrome. Internationally, we were part of the 22q11.2 IBBC International Brain Behavior Consortium. This group is outstanding in delivering high-impact scientific output, organizing both high quality and pleasant meetings, and should be praised for the establishment of a professional and still open community.

One of the most exciting experiences during my PhD, was the research visit at the University of Stanford. Thank you Alex, for the opportunity to be part of your lab for three months. Bo, thank you for guiding me through the different labs, the Stanford campus, and the CTLR-Seq method. Another thank you to the lab people, Alison, Agnes, Billie, GiWon, Sharmili, Stephanie, and Xiangqi, for making me part of the team and invite me to the pizza parties and afterwork drinks. I was extremely lucky to meet several people of whom five really need to be mentioned here: Bjort, Grace, Lauranne, Romy, and Sofie, thank you for the cozy dinner evenings and unforgettable weekend trips. Deze periode werd afgesloten met een roadtrip doorheen West-Amerika, waarbij ik het allerbeste reisgezelschap had: Hannah, Marie, Nikki, Paulien en Thomas. Met heel veel nostalgie denk ik terug aan margarita pitchers op Venice Beach, Grand Canyon *Schluchten* op TikTok en exotische ziektes in Las Vegas, wanneer 'Gloria' of Jan Smit doorheen de boxen galmt.

The office. Although some people discovered some striking similarities with the Netflix-hype, I am still more a fan of our own 5th floor group. Thank you to everyone for the nice time: the current group (Angelica, Armelle, Benjamin, Charlotte, Dhanya, Elise, Greg, Jente, Julio, Kate, Laura, Mathilde, Matt, Nicolas, Olga, Stefania, Tine, Yan, and Yoeri), former people (Alena, Ans, Aspasia, Berardo, Carlijn, Chiara, Darine, Huiwen, Lise, Maria, Molka, Nele, Romain, Simon, Vincent, Wolf, and Yasmine), and the 'Voet lab' (Elia, Elisa, Inge, Koen, Sarah, Sebastiaan, and Thomas). I wish you all the best with your PhDs/postdocs, houses, children, and life!

I want to take the time to thank a few CME people in particular. Tanja, for the scientific advice, the patient (!) help with the design of my cover and the administrative tasks of the defense. Olga, for taking care of our work-life balance and to schedule 'very important meetings' from time to time. Heleen, merci voor het uitleggen van enorm veel genetische begrippen en uiteraard

voor alle dessertjes, van de salted caramel brownie tot de gender reveal taart! Lieselot, voor de bureau-babbeltjes, en standaard of extreem chique lunches. Eline, ik vrees dat niemand ooit ons enthousiasme over de *Tomorrowland ESHG edition* zal begrijpen, inclusief *the swinging geneticists* band, maar wij weten beter. Laetitia, mijn vaste adres en knuffelcontact voor het keuren en bespreken van zalando bestellingen, feestjes en andere zaken. Altijd kon ik bij jou terecht, tenzij we thuis moesten geraken zonder gsm/licht, of wanneer ik liever ging slapen in plaats van iets te laten weten.

Tot slot mogen mijn allerliefste bureauburen absoluut niet ontbreken, Greet en Margot, want elke woensdag of thuiswerkdag voelde onze rij toch wel een beetje incompleet. Greet, teacher of labskills en co-founder van onze *still need a name* mini-company. Long reads zijn onze specialiteit (fiber-FISH, Bionano, Nanopore), initiaties en workshops beschikbaar op aanvraag, tof gezelschap krijg je er gratis bij. Merci voor al je kennis in het labo en je gastvrijheid tijdens een garden party of fietstussenstop. Margot, vijf jaar geleden zijn we samen aan dit 'avontuur' gestart en binnenkort zullen we ook beiden vaarwel zeggen aan het CME. Naast het bespreken van werk-gerelateerde issues, konden we nog veel langer babbelen over weekendjes, sporten, en cocktails. Naar het werk komen zorgde dus niet perse voor efficiëntere, maar wel absoluut voor leukere werkdagen. Ik had me niemand beter kunnen voorstellen om letterlijk naast mij te 'zitten/staan' tijdens dit doctoraat en ben er van overtuigd dat er in de toekomst nog vele aperitiefjes, Antwerpen/Mechelen dagtrips en fietstochtjes via Rillaar zullen volgen!

Naast het werk kon ik steeds rekenen op een groep van enthousiaste en geweldige vrienden, die ik hier ook even zou willen bedanken. Mijn highschool clan en hun (bíjna even leuke) aanhangsels, Hannah & Aaron, Elise & Stef, Suze & Daan, en Karen & Daan: het is steeds uitkijken naar en enorm nagenieten van onze housewarmings, festival- en campingedities, of een occasioneel chique hotelbezoek in ons geliefde Antwerpen. Hoewel onze doelen het komende jaar niet meer kunnen verschillen (huizen verbouwen, appels plukken aan de andere kant van de wereld, of goedkope tickets naar Boston scoren), kijk ik er al naar uit om terug met z'n allen samen te zijn! Ook de farma ladies mogen hier absoluut niet ontbreken: Aline, Bea, Emma, Evelyne, Hélène, Marlies, Nikki en Stefanie. Wat begon met farma galabals en financieel voordelig studentenfood, is geëvolueerd naar fancy cocktails, Ottolenghi-inspired diners, en onvergetelijke roadtrips doorheen Canada en Zuid-Afrika. Een speciale shout-out hier toch wel naar Bea, om twee jaar lang een geweldige roomie te zijn in onze Heverly Hills mansion. We hebben die quarantaines toch maar mooi samen doorstaan. Vrijdag- en zaterdagavond waren standaard gereserveerd voor 'de volleybal'. Met spijt in het hart zal ik het komende seizoen geen deel meer uitmaken van Team 5.0, maar ik ben wel enorm dankbaar voor de voorbije jaren en de speelsters/vriendinnen waarmee ik op het veld (of andere oppervlakken in 't toreke) heb gestaan. Ook dikke merci aan Pieter, voor de afgelopen tien fantastische maanden.

Als laatste mag mijn familie zeker niet ontbreken in dit dankwoord. Mijn grootouders, oma & opa en moeke & vake, voor de babbeltjes, telefoontjes, chocolaatjes en dessertjes. Mijn vake

had dit moment ongetwijfeld graag bijgewoond. Ik ben hem dan ook enorm dankbaar voor alle Amerika verhalen en zo de liefde voor dit toch wel vreemde maar fantastische land door te geven. Bij de familie groep reken ik ook het Everix trio: dankjulliewel voor de feestjes, uitstapjes en vakanties. Het is nog steeds wachten op het eerste stille (of zelfs rustige) moment in bijna 20 jaar. Mijn zussen, Paulien, Marie, Nel en Flo: ook al kunnen we bijna niet meer van elkaar verschillen, weet dat jullie ieder op zich geweldig zijn en ik me geen betere zussen kan inbeelden. Hopelijk letten jullie tijdens de presentatie wel iets beter op dan de voorbije vijf jaar en weten jullie dan eindelijk ook dat mijn doctoraat niet over zebravisjes gaat. En ten slotte, mijn ouders, mama & papa, dankjulliewel voor alles, onder andere de hulp bij mijn verhuizingen (en dat is wel een paar keer voorgevallen tijdens dit doctoraat), het advies over alles behalve wetenschappelijk zaken, maar vooral jullie onvoorwaardelijke steun!

Indien het nog niet duidelijk zou zijn: ik ben enorm dankbaar voor deze geweldige periode en de fantastische personen die naast mijn stonden. Daarnaast heb ik zeer veel goesting in de toekomst en ben ik benieuwd naar de uitdagingen die mijn pad nog zullen kruisen.

Lisanne

TABLE OF CONTENTS

SCIENTIFIC SUMMARY

Rearrangements of chromosome segments that alter the overall DNA structure and cause a phenotypic effect are classified as genomic disorders. They are collectively an important cause of disabling diseases in the population, and costly to patients, their families, and society. With an incidence of 1 in 2148 live births, the 22q11.2 deletion syndrome (22q11.2DS) is the most common microdeletion syndrome in humans.

The 22q11.2DS is caused by non-allelic homologous recombination between low copy repeats (LCRs) or segmental duplications on chromosome 22q11.2. LCRs are blocks of DNA of at least 1kb that are duplicated to inter- and intrachromosomal loci. Due to the high sequence identity between these blocks, homologous segments may misalign during meiosis resulting in deletions and duplications. The presence of eight LCRs on chromosome 22q (LCR22-A until -H) entails that this locus is structurally one of the most complex areas of the human genome. In 90% of the 22q11.2DS patients, a 3 Mb deletion occurs between LCR22-A and -D, the two largest LCR22s, but several other rearrangements exist.

Unfortunately, the large number of subunits with a high percentage of sequence identity, and clustering of these subunits, frequently produces alignment errors using short-read and even standard long-read sequencing data. As a consequence, the hg38 reference genome still comprises three unresolved sequence gaps in LCR22-A, hampering the study of the LCR22 structural organization, LCR22 genes, and exact breakpoints of the 22q11.2 rearrangements.

In this thesis, we successfully unraveled the overall architecture of the 22q11.2 locus for the first time by the development of an LCR22-specific fiber-FISH technique. The method provides long-range structural LCR22 information and therefore bypasses the biased mapping of sequencing data. We assembled a total of 44 LCR22-A alleles and uncovered a variety of over 25 haplotypes, ranging in size between 250kb and 2Mb, demonstrating the extreme level of interindividual hypervariability of this locus. Due to the high level of LCR22-A recombination in pedigree linkage analyses, we hypothesized that new haplotypes arose via the mechanism of allelic homologous recombination. However, LCR22-A fiber-FISH assembly of families with identified recombination did not show the creation of hybrid alleles. By studying the LCR22-A composition in great apes, we demonstrate the expansion and variability can be considered as human-specific.

Moreover, we used the fiber-FISH method to visualize the 22q11.2DS rearrangement loci at subunit level. To pinpoint the exact recombination sites at nucleotide level, we leveraged Nanopore long-read sequencing approaches. First, atypical 22q11.2 deletions were investigated, since one of the breakpoints resides in unique sequence between the LCR22s. We found that this subclass of deletions was created by replication-based mechanisms or by a complex two-step recombination mechanism. Second, we examined the range of

recurrent LCR22 deletions and showed variability in the rearrangement locus. In addition, a subset of deletions was mediated via palindromic AT-rich repeats, implicating the involvement of non-homologous end-joining pathways. Hence, involvement of different loci and mechanisms probably explains why the 22q11.2DS is the most common microdeletion disorder in humans.

In summary, we discovered human-specific expansion and variability of the LCR22-A haplotype, as well as variability in both the breakpoint locus and mechanisms involved in the 22q11.2DS rearrangements. These findings are fundamental for the 22q11.2DS research community and pave the way towards further investigation of this complex locus at the molecular and cellular level to unravel part of the genotype-phenotype correlation of this genomic disorder. In addition, this research will provide a paradigm for the study of other rare genetic disorders with incomplete penetrance.

WETENSCHAPPELIJKE SAMENVATTING

Genomische aandoeningen worden veroorzaakt door herschikkingen van chromosomale segmenten die de algemene DNA-structuur aantasten met een fenotypisch effect tot gevolg. Deze subgroep van genetische aandoeningen vormt een belangrijke oorzaak van invaliditeit in de populatie, met hoge kosten voor patiënten, hun families en de samenleving tot gevolg. Met een geschatte incidentie van 1 op 2148 geborenen is het 22q11.2 deletie syndroom (22q11.2DS) het meest voorkomende microdeletie syndroom in de humane populatie.

Het 22q11.2DS wordt veroorzaakt door een niet-allelische homologe recombinatie tussen low copy repeat eenheden (LCRs) of segmentele duplicaties op chromosoom 22q11.2. Deze LCR-blokken hebben een lengte van ten minste 1kb en zijn gedupliceerd naar verschillende plaatsen op hetzelfde of een ander chromosoom. Aangezien zij zeer gelijkend zijn aan elkaar, kunnen de verkeerde homologe segmenten recombineren tijdens de meiose wat resulteert in deleties en duplicaties. De 22q locus is een van de meest structureel complexe regio's in het humane genoom, omwille van de aanwezigheid van acht LCRs, genaamd LCR22-A tot -H. Verschillende herschikkingen tussen de LCR22s zijn mogelijk, maar de meerderheid van de patiënten draagt een 3Mb deletie, die ontstaan is door recombinatie tussen LCR22-A en -D, de twee grootste LCR22s.

Hoewel deze DNA structuren dus essentieel zijn voor het ontstaan van het 22q11.2DS, wordt onderzoek hiernaar belemmerd door hun complexiteit: het grote aantal eenheden, gecombineerd met de hoge sequentie gelijkheid tussen de duplicaties en de clustering van deze eenheden, veroorzaakt fouten in het analyseproces. Zelfs het gebruik van nieuwere sequencing methodes, zoals het lezen van zeer lange DNA-fragmenten, kon tot nog toe geen oplossing bieden om deze regio samen te stellen. Dit heeft tot gevolg dat er nog steeds drie onopgeloste 'gaten' zijn in LCR22-A van het hg38 referentiegenoom. Hierdoor was onderzoek naar de structurele organisatie, genen en exacte breekpunten van de 22q11.2 herschikkingen moeilijk tot onmogelijk.

In deze thesis hebben we voor de eerste keer de globale samenstelling van de 22q11.2 locus in kaart kunnen brengen door het ontwikkelen van een LCR22-specifieke fiber-FISH techniek. Deze methode genereert structurele informatie over een lange afstand en maakt het zo mogelijk om de LCR22-geassocieerde problemen te omzeilen. In totaal werden er 44 LCR22-A allelen samengesteld die meer dan 25 haplotypes vertegenwoordigen met lengtes variabel tussen 250kb en 2Mb. Dit toont de zeer grote interindividuele variabiliteit aan in de 22q11.2 locus. Eerder onderzoek naar meiotische recombinaties in grote humane families toonde aan dat recombinatie frequent plaatsvindt in de LCR22-A locus. Hierdoor ontstond de hypothese dat deze uitgebreide set aan LCR22-A haplotypes zich gevormd kon hebben via het mechanisme van allelische homologe recombinatie. Om deze hypothese te testen, stelden we de LCR22-A haplotypes samen van families waar een recombinatie geobserveerd

was, maar we konden geen hybride haplotypes identificeren. We vroegen ons ook af of de hypervariabiliteit specifiek was voor mensen. Via de analyse van deze regio's in de mensapen, kunnen we concluderen dat de LCR22-A expansie en hypervariabiliteit humaan-specifiek zijn.

Verder werd de fiber-FISH techniek gebruikt om de herschikkingen in de 22q11.2DS families te visualiseren op subunit resolutie. Bijkomend werden de exacte recombinatie posities gelokaliseerd door gebruik te maken van de nieuwste Nanopore sequencing methodes, die de DNA-samenstelling van zeer lange fragmenten kunnen bepalen. Als eerste ontrafelden we de atypische 22q11.2 deleties, waar ten minste één van de recombinatieplaatsen niet in de LCR22, maar in de unieke regio is gelegen. Zowel replicatiegebaseerde mechanismen als complexe tweestapprocessen zijn verantwoordelijk voor het creëren van deze klasse van deleties. Ten tweede exploreerden we de standaard 22q11.2 deleties, die steeds plaatsvinden tussen twee LCR22s. In deze klasse observeerden we naast variabiliteit van de breekpunt locus ook variabiliteit van het ontstaansmechanisme, aangezien een deel van de herschikkingen gemedieerd werden door palindromische AT-rijke repeatstructuren. Dit wil dus zeggen dat niet enkel niet-allelische homologe recombinatie verantwoordelijk is voor het ontstaan van de 22q11.2 deleties. De betrokkenheid van verschillende recombinatieregio's en mechanismen verklaart waarschijnlijk waarom het 22q11.2DS het meest voorkomende microdeletie syndroom is in de humane populatie.

Samengevat, tijdens dit onderzoek ontdekten we de humaanspecifieke expansie en variabiliteit van de LCR22-A locus én de variabiliteit van recombinatie regio's en mechanismen die leiden tot de 22q11.2 herschikkingen. Deze bevindingen zijn essentieel in het kader van 22q11.2 onderzoek en vormen de basis om deze complexe locus verder te ontrafelen op cellulair en moleculair level. Dit onderzoek zal uiteindelijk leiden tot het ontcijferen van de genotype-fenotype correlatie in het 22q11.2DS. Bijkomend betekent deze studie een fundament voor andere onderzoeken die focussen op de fenotypische variabiliteit bij genomische aandoeningen.

LIST OF ABBREVIATIONS

22q11.2DS	22q11.2 deletion syndrome
AHR	Allelic homologous recombination
ArrayCGH	Array comparative genomic hybridization
BAC	Bacterial artificial chromosome
BLAT	BLAST-like alignment tool
Bp	Base pair
CEPH	Centre d'Etude du Polymorphism Humain
CNV	Copy number variation
CTLR-Seq	CRISPR-targeted ultra-long read sequencing
FISH	Fluorescent *in situ* hybridization
FoSTeS	Fork Stalling and Template Switching
GWAS	Genome-wide association study
IBBC	International 22q11.2DS Brain Behavior Consortium
IGV	Integrative Genomics Viewer
iPSC	Induced pluripotent stem cell
LCR	Low copy repeat
LCR22	Low copy repeat on chromosome 22
(O)MIM	(Online) Mendelian Inheritance in Man
MLPA	Multiplex ligation-dependent probe amplification
MMBIR	Microhomology-mediated break-induced replication
NAHR	Non-allelic homologous recombination
NHEJ	Non-homologous end-joining
NIGMS	National Institute of General Medical Sciences
NIMH	National Institute of Mental Health
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PATRR	Palindromic AT-rich repeat
PCR	Polymerase chain reaction
SD	Segmental duplication
SMRT	Single-molecule real-time
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
TAD	Topologically associated domains
VEGF	Vascular endothelial growth factor

CHAPTER 1
INTRODUCTION

1 INTRODUCTION

1.1 Genomic disorders

Pathological alterations in the composition or the structure of our DNA result in genetic disorders. Different classes can be distinguished, ranging from aneuploidies, defined by a change of the total number of chromosomes, to single nucleotide mutations, in which the disease is caused by the variation of one single base pair (bp) (Gu et al. 2008; Harel and Lupski 2018). Rearrangement of a segment of the chromosome, typically encompassing 100bp or more, is classified as structural variation (Hollox et al. 2022). Genomic disorders are a subclass of genetic disorders in which the disease phenotype is caused by these DNA rearrangements, rather than single nucleotide changes (Harel and Lupski 2018).

1.1.1 Exponential increase of structural variation detection

Technological innovation has driven our ability to detect structural variation. First, G-banded karyotyping provided indications of larger-scale rearrangements. Second, the introduction of array comparative genomic hybridization (arrayCGH) and single nucleotide polymorphism (SNP) assays enabled the scanning of the genome for copy number variations (CNV) without prior assumptions. Systematic screening of patients with developmental disorders using chromosomal microarrays resulted in the identification of several hitherto unknown genomic disorders (Lupski 2009). Third, the structural variation catalogue was rapidly expanding by the use of whole-exome and whole-genome sequencing based on read depth or presence of split reads and conflicting mate pairs in short-read sequencing data (Hollox et al. 2022). However, these analyses are still limited by short-read sequencing associated problems. Nowadays, these problems are solved by long-read sequencing approaches resulting in the gap-free assemblies of whole human chromosomes (Miga et al. 2020; Logsdon et al. 2021) and even a complete human genome (Nurk et al. 2022). This started a new era of structural variation detection overload, switching the challenges from detection towards documentation, interpretation, and clinical validation of newly observed CNVs. To that aim, the Human Pangenome Reference Consortium was established (Wang et al. 2022).

1.1.2 Rearrangements via non-allelic homologous recombination

Genomic disorders can be subdivided into the nonrecurrent and recurrent rearrangements, the latest being the focus of this thesis. Nonrecurrent and complex rearrangements are characterized by locus and breakpoint variability, which are not predictable based on the genomic architecture. Molecular mechanisms responsible for these CNVs include non-homologous end-joining (NHEJ), Fork Stalling and Template Switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) (Harel and Lupski 2018). In the recurrent rearrangements, several patients are described with breakpoints clustering in

a specific locus. This is caused by the presence of duplication modules in this locus, serving as drivers for the rearrangements (Lupski 2009).

Recurrent genomic disorders are caused by meiotic misalignment of high sequence identity (>90%) blocks (**Figure 1.1A**), resulting in rearrangements of the involved segment, a mechanism known as non-allelic homologous recombination (NAHR) (Gu et al. 2008). The non-allelic recombination substrates are typically low copy repeats (LCRs) or segmental duplications (SDs), which are flanking the involved locus (Gu et al. 2008). LCRs are blocks of DNA with a length of at least 1kb and duplicated to several inter- and intrachromosomal loci in the genome (Bailey et al. 2001, 2002a).

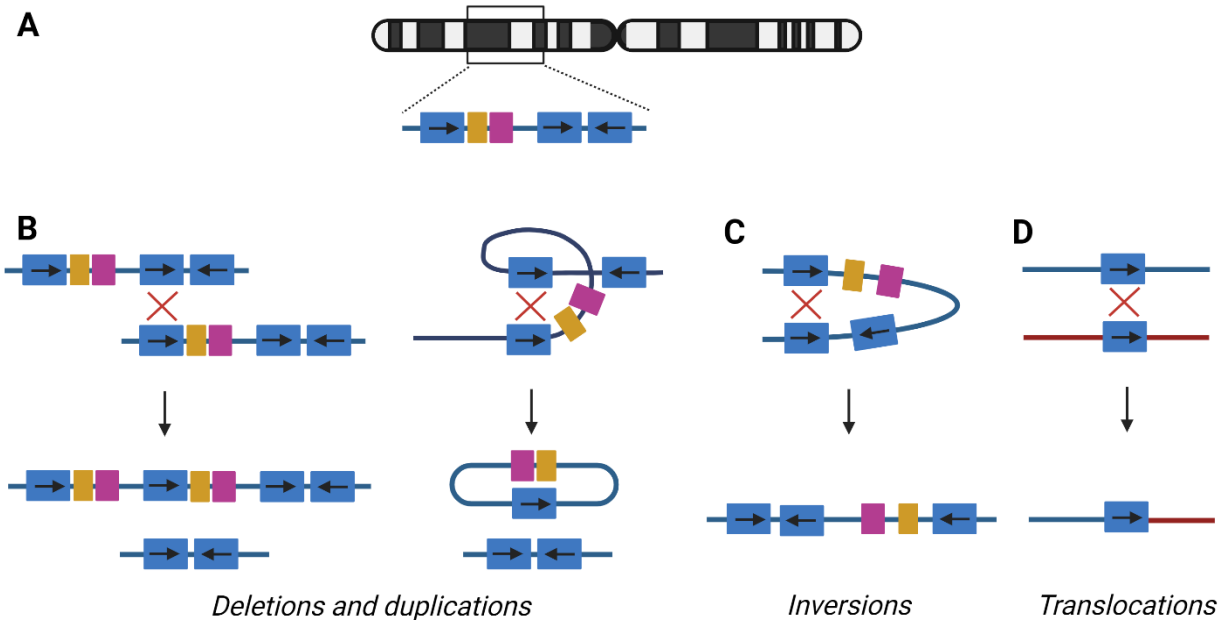


Figure 1.1: Non-allelic homologous recombination. *(A) Example of LCR structure on a random chromosome. The blue duplicons can be in direct or inverted orientation, indicated by arrows. (B) NAHR between two direct repeats on two homologous chromosomes, interchromosomal NAHR (left), will lead to the creation of a duplication and deletion allele. Intrachromosomal NAHR (right), between two repeats on the same chromosome, creates a deletion and a ring chromosomal segment. (C) The recombination between indirect repeats on the same chromosome will result in an inversion of the segment. (D) Translocations are shaped if the NAHR is between two different chromosomes.*

If both LCRs are in direct orientation, NAHR between homologous chromosomes will result in a chromatid with a deletion and another with the reciprocal duplication (**Figure 1.1B**). These are considered CNVs, since they are associated with the gain or loss of DNA (Hollox et al. 2022). Intrachromosomal NAHR can only create deletions and a ring-shaped chromosomal segment (**Figure 1.1B**). Therefore, deletions are more frequently observed compared to duplications. Both deletions and duplications can lead to a clinical phenotype via a diversity of mechanisms: gene-dosage effect (Ewart et al. 1993), expression of a new gene via gene fusion (Aigner et al. 2013), interaction with regulatory elements (Montavon et al. 2012), and interruption of chromatin structure (Gheldof et al. 2013) are a few examples. Several clinical deletion and duplication syndromes are known and a limited overview is provided in **Supplementary Table S1.1**. In general, duplication syndromes have a milder phenotype compared to the deletion syndromes, since gene deficiencies

overall have more phenotypic consequences (Lupski 2009). Therefore, duplication carriers are less represented in clinical cohorts and the population incidence was long underestimated.

NAHR between LCRs in opposite orientation will lead to inversions (**Figure 1.1C**). These are copy neutral events, since no gain or loss is associated with the rearrangement (Hollox et al. 2022). If heterochromatic sequence is inverted, the inversion will be harmless. If the inversion directly affects a gene, this can lead to disease, either by disrupting the gene or by alteration of the expression level (Puig et al. 2015). For example, the majority of severe hemophilia A cases is caused by an inversion disrupting the *F8* gene on chromosome X, mediated by two LCRs (Lakich et al. 1993). Another recurrent inversion involves the *IDS* gene on chromosome X, encoding the iduronate 2-sulfatase enzyme, responsible for breakdown of glycosaminoglycans. The inversion causes *IDS* gene disruption, leading to mucopolysaccharidosis type II, a lysosomal storage disorder (Bondeson et al. 1995). Although not causing disease, some human inversion polymorphisms are associated with an abnormal phenotype (Puig et al. 2015). The largest known human inversion polymorphism is located in the 8p23.1 locus, spanning a length of 4.5Mb. These 8p23.1 inversion carriers have a lower risk for developing systemic lupus erythematosus and rheumatoid arthrosis compared to individuals carrying the reference allele (Namjou et al. 2014).

Translocations are created by crossovers between elements on different chromosomes (**Figure 1.1D**). Examples of recurrent translocations are t(11;22)(q23;q11), t(8,22)(q24.13;q11.21), and t(4;8)(p16;p23). Carriers of the balanced translocation are phenotypically normal in most cases, but their offspring are at risk for inheriting a derivative chromosome, leading to genomic imbalance (Ou et al. 2011). For example, in the unbalanced der(4)t(4;11)(p16.2;p15.4) translocation, the 4p16.2-pter monosomy expresses as Wolf-Hirschhorn syndrome, and the imprinted 11p15.4-pter trisomy manifests as Russell-Silver syndrome or Beckwith-Wiedemann syndrome, when maternally or paternally inherited, respectively (Ou et al. 2011).

Hence, the NAHR mechanism is responsible for a range of rearrangements, involving several, but LCR-specific parts of the genome. If a rearrangement manifests as an abnormal clinical phenotype, it can be classified as a genomic disorder.

1.1.3 Genomic predisposition for deletion/duplication syndromes

Parental inversion polymorphisms between LCRs predispose the region to NAHR, resulting in offspring with genomic disorders (Shaw and Lupski 2004). Population embedded inversion polymorphisms are drivers of many genomic disorders. For example, the 1.5Mb inversion on 7q11.23 is a driver of the deletion causing Williams-Beuren syndrome (Osborne et al. 2001). The inversion has a frequency of 12.4% in deletion-transmitting parents, but is only

present in 2.9% of the control population (Puig et al. 2015). Other examples are the inversions leading to Angelman syndrome (inversions in 33% of the mothers), Sotos syndrome, 8p23.1 microdeletion, and 15q23 or 15q24 microdeletion syndrome (Puig et al. 2015). In addition, in some cases, these disease-predisposing inversion polymorphisms can be linked to phenotypic consequences as well. For example, the two haplotypes of the 1.5Mb inversion polymorphism in the 17q21.31 locus have different characteristics: the direct H1 haplotype carries mutations linked to Parkinson disease and other neurodegenerative diseases, the inverted H2 haplotype is associated with an increased risk for 17q21.31 rearrangements and positive selection in the human population (Boettger et al. 2012; Steinberg et al. 2012). Thus, although not directly related to disease, inversion polymorphisms between LCRs are an important driving cause of genomic disorders.

1.2 Segmental duplications or low copy repeats

SDs or LCRs play an important role in the origin of genomic disorders. They constitute 6.6% of the human genome (Nurk et al. 2022). Due to the multiple mapping options, reads originating from LCRs are frequently misassigned, creating errors and gaps in reference assemblies. As a consequence, they are frequently removed in standard analysis pipelines and specialized techniques are necessary to investigate their importance in genome stability and evolution (Bailey and Eichler 2006).

1.2.1 Distribution, origin and evolution

LCRs are not randomly distributed across the genome, but are primarily clustering in pericentromeric, subtelomeric, and interstitial loci. In these regions, there is up to 10-fold enrichment for LCRs with chromosome-specific differences: chromosome 3 has a low LCR density, while chromosomes 22 and Y harbor the largest LCR proportions (Bailey et al. 2001). They are composed of regular genomic architectural features as genes, repeat elements, and regulatory sequences, but differ from the standard unique sequence since they are copied to inter- and intrachromosomal loci. However, compared to other repeat elements, their copy number is limited and ranges between 2 and 50 (Carvalho and Lupski 2016).

Pericentromeric, subtelomeric, and interstitial LCRs differ in characteristics and mechanisms of origin (**Figure 1.2A**). The duplication content of pericentromeric LCRs originated mainly from interchromosomal duplication events, as proposed in the two-step model (**Figure 1.2B**). In a first 'initial seeding event', LCR sequence from different genomic loci is juxtaposed in a duplicon block. Second, these duplicon blocks are copied to other pericentromeric sites, creating a mosaic structure (Bailey and Eichler 2006). Interchromosomal duplications are enriched in subtelomeric LCRs via serial translocations (**Figure 1.2D**): consecutive events of double-strand breakage and repair in these subtelomeric regions created a mosaic pattern of LCR-containing sequences (Bailey and

Eichler 2006; Abdullaev et al. 2021). The largest LCRs in the human genome are located in interstitial regions and enriched for intrachromosomal duplications. These interstitial LCR paralogues have the highest similarity. The complex patterns are formed by serial duplications, using the LCRs themselves as homology substrates in consecutive rounds (**Figure 1.2C**) (Bailey and Eichler 2006; Abdullaev et al. 2021). *Alu* repeats are frequently observed at or in the vicinity of the boundaries of LCRs, suggesting involvement of both NAHR and replication-based mechanisms (NHEJ, MMBIR) in the creation of these complex structures (Bailey and Eichler 2006; Hastings et al. 2009; Carvalho and Lupski 2016; Abdullaev et al. 2021).

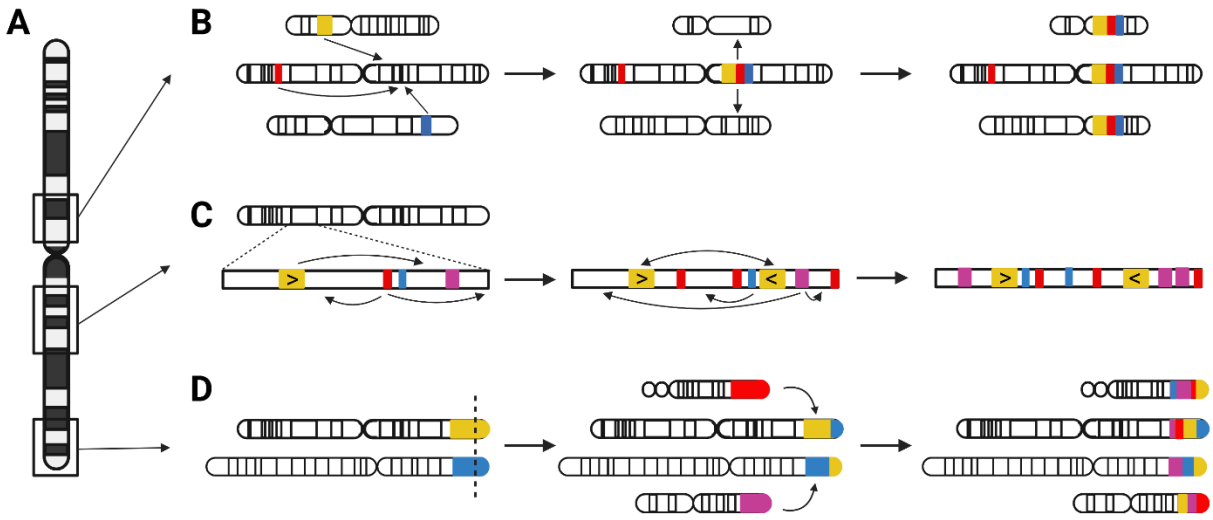


Figure 1.2: Origin of pericentromeric, interstitial, and subtelomeric LCRs. *(A) Relative chromosomal location of pericentromeric, interstitial, and subtelomeric LCRs. (B) Two consecutive events created the pericentromeric LCRs. First, during the initial seeding event, LCR sequences from different chromosomes are merged into an LCR block. Second, blocks are duplicated to other pericentromeric loci. (C) Mosaic patterns of the interstitial LCRs were formed by several rounds of serial duplication. (D) Serial translocation, by double-strand breakage and repair, is responsible for the subtelomeric LCR structure.*

The proportion of LCR sequence varies substantially between the genomes of different species. The number of LCRs is very low in fly and worm in comparison to the human genome (Bailey and Eichler 2006). Although first thought that the duplication content was lower in mammalian genomes (mouse, dog, cow) as well (Bailey and Eichler 2006), genome sequencing revealed similar levels of recent duplications (Marques-Bonet et al. 2009a). However, since they show a tandem organization, the architecture of these duplications differs radically from the human mosaic LCR structure (Marques-Bonet et al. 2009a).

The LCRs evolved into their current organization starting in primates and are therefore of relatively recent origin. The genome of the marmoset, a New World monkey, has lower LCR levels than hominids (great apes: orangutan, gorilla, chimpanzee, bonobo; and humans), suggesting an expansion of LCRs after the divergence of Old and New World monkeys, 35 million years ago (Marques-Bonet et al. 2009a; Sudmant et al. 2013). This is consistent with the LCR duplication burst during human evolution (Marques-Bonet et al. 2009b).

Investigation and comparison of great ape and human genomes revealed more genetic variation due to LCR structure than single nucleotide variants in other genomic loci (Dennis et al. 2017). Due to the presence of *Alu* elements at the boundaries and breakpoints of the LCRs, the expansion is thought to be caused by *Alu-Alu* mediated rearrangements. This hypothesis is concordant to the burst of *Alu* elements 35 million years ago (Marques-Bonet et al. 2009a; Dennis et al. 2017). Hence, LCRs are considered as fundamental components in the shaping process of primate genomes.

1.2.2 Transcriptomic innovation and phenotypic effects in evolution

Incorporation of one or more extra copies in the genome is associated with disease and evolutionary consequences. On the one hand, two identical copies on different chromosomal locations can act as NAHR substrates, leading to genetic instability, rearrangements, and eventually disease. On the other hand, natural evolutionary processes can act on the copied segment itself, creating gene segments and transcripts with a completely new (neofunctionalization) or altered function (subfunctionalization), compared to the ancestral gene (Marques-Bonet et al. 2009a; Dennis and Eichler 2016). Hence, due to their duplication potential and nature, LCRs are ideal substrates to influence gene evolution. If these duplications are specific to the human lineage, they can contribute to important adaptive traits.

One mechanism leading to a new gene product is incomplete duplication of an ancestral gene. For example, the *SRGAP2A* gene (Chr1q32.1) is important in neuronal migration in mammals and is partially duplicated in the human lineage (*SRGAP2B*, Chr1q21.1; and *SRGAP2C,* Chr1p11.2) The *SRGAP2C* paralogue is the most recent one and dimerizes with *SRGAP2A* in the human embryonic cortex. The function of *SRGAP2C* is antagonistic to the ancestral *SRGAP2A*, since it is a cortical development gene involved in dendritic spine maturation (Charrier et al. 2012; Dennis et al. 2012). Another example is the human-specific *ARHGAP11B* gene (Chr15q13.2), the product of incomplete duplication of *ARHGAP11A* (Chr15q13.3). It exerts a completely new function, by influencing progenitor cells of the radial glia neurons, leading to cortical layer expansion of the developing brain (Florio et al. 2015). The *NOTCH2NL* gene has three human-specific paralogues: *NOTCH2NLA* (Chr1q21.1), *NOTCH2NLB* (Chr1q21.2), and *NOTCH2NLC* (Chr1q21.2). They are expressed in radial glia and important in the *Notch* signaling pathway. In that way, they influenced human-specific neuronal differentiation and alterations in size and complexity of the human neocortex (Fiddes et al. 2018). Hence, the emergence of human-specific genes due to incomplete duplications have contributed to critical adaptive traits regulating brain size and function.

If the duplication juxtaposes two partial genic fragments and the necessary regulatory elements, a fusion gene with new function can be created. Indeed, there is evidence for an enrichment of gene fusions in human LCR regions (McCartney et al. 2019). The *HYDIN* gene

(Chr16q22.2), involved in cilia motility, was duplicated in the human lineage, although the promotor and polyadenylation site were missing. However, the partial duplication was juxtaposed with active regulatory elements in the locus, leading to transcription of the *HYDIN2* gene (Chr1q21.1). Interestingly, whereas the ancestral *HYDIN* gene is mainly expressed in ciliated tissues, the human-specific *HYDIN2* transcripts are specific for neuronal tissues (Dougherty et al. 2017). The incomplete duplication and consecutive fusion between *FAM7A* and *CHRNA7* created the *CHRFAM7A* fusion gene (Chr15q13.2). Although research is hampered due to the large sequence identity between the ancestral and fusion gene, the gene product is involved in ion channel function (Sinkus et al. 2015; Dennis et al. 2017). So, in addition to incomplete gene duplication, gene fusion is another mechanism contributing to gene evolution.

To conclude, the presence of genes in loci subjected to duplication have a tremendous evolutionary potential. Human-specific genes were identified with important functions in the development and maturation of the brain, differentiating humans from chimpanzees (Dennis and Eichler 2016). A next step will be to link these human-specific genetic alterations to complex brain diseases as schizophrenia, intellectual disability, and developmental delay. Due to their complex genetic architecture, targeted studies are essential and therefore, the extent of the evolutionary and adaptive impact is only starting to be discovered.

1.2.3 Structural variation of LCRs in the human population

The LCR loci are important substrates for the creation of human-specific genes via duplication, but due to their duplication potential, they can also be polymorphic within the human population (Goidts et al. 2006b). These inter-human copy number variants can give rise to phenotypic differences between individuals of different populations. The copy number of the amylase gene (*AMY1*, Chr1p21.1) differs between 2 and 15 in modern humans. The gene has an essential function in the digestion of starch and copy number is therefore correlated with the amount of starch in the diet between populations (Perry et al. 2007). The *BOLA2* unit (Chr16p11.2) is a CNV under positive selection and 3 to 8 copies are reported in the human population. This polymorphism is associated with the maturation of iron-sulfur proteins and iron homeostasis: anemia is described in individuals with lower copy number and deletions of *BOLA2*, while expansions are protective against iron deficiency (Nuttle et al. 2016; Giannuzzi et al. 2019). Another interesting CNV is *DUF1220* (Chr1q21.1), characterized by neuron-specific expression and positive selection. There is strong association with head circumference in the normal population and brain-size manifestations in the disease population (Dumas et al. 2012). *TCAF* genes (Chr7q35) are highly conserved in evolution, but show copy number variation in the human population. They have a valuable effect on regulating *TRPM8*, the ion channel for thermal sensation in somatosensory neurons. Copy numbers can be linked to demographic differences and implicate an adaptive role for changing conditions (Hsieh et al. 2021). So, human

polymorphic LCR loci are common and phenotypic consequences can be expected if genes are involved in the duplication.

Human-specific structural variation is described for several LCR loci. The 15q13.3 locus comprises five possible structural configurations, including copy number variants and inversions. The locus, and more specifically the *GOLGA* repeat, is important in evolution and a hub for the breakpoint loci of Prader-Willi/Angelman syndrome, 15q24 and 15q25.3 microdeletions (Antonacci et al. 2014). Similar amounts of structural variation were reported for the 17q11.23, 7q11.23, 15q13.3, and 16p12.2 loci (Boettger et al. 2012; Mostovoy et al. 2021), although they could not be associated (yet) to phenotypic differences in the human population.

These studies introduce a paradox: LCRs, which are associated with an increased susceptibility for genomic rearrangements, and therefore are expected to be 'selected out' during evolution, are actually expanded during human evolution (Goidts et al. 2006b). An explanation is that the genomic instability created by an increased copy number is neutralized or compensated by an evolutionary advantage, highlighting their extreme evolutionary and adaptive potential (Dumas et al. 2012).

1.3 Challenges to investigate Low copy repeats

1.3.1 Drawbacks of short-read sequencing technologies

Second generation sequencing techniques have proven to be extremely sensitive for the detection of single nucleotide variants in the euchromatic non-duplicated loci of the genome. To detect structural variants in high-throughput sequencing data, three specific approaches are used: (1) sequence read depth, based on coverage changes over the locus indicating the presence of a CNV, (2) split reads, if the read contains the structural variant breakpoint and is mapping to two separate loci in the genome, and (3) discordant mate pair, if the distance and/or orientation of two reads in a pair is different than expected (Hollox et al. 2022). However, since the length of the LCR modules surpasses the length of second generation sequencing reads, LCR research is hampered by additional assembly and alignment challenges. The inability to differentiate between two alleles on homologous chromosomes or two (or more) paralogues on the same chromosome (or a combination) is the main cause for these failures (Bailey et al. 2001). Global structural information is missing in the short reads, hampering a correct assembly of structurally complex loci. In addition, the short reads are too small to correctly align to one locus in the genome and misalignments can lead to false structural variant calls, further complicated by the absence of an accurate reference genome in this specific locus (Chan et al. 2018).

1.3.2 Solutions

To decipher the exact structural composition of the LCRs and pinpoint the breakpoints of the associated genomic disorders, long range information over the repeat is necessary. For example, the LCRs on chromosome 17 involved in the rearrangement leading to neurofibromatosis type 1 were puzzled together using long range PCR fragments. Primers were chosen to target unique paralogous sequence variants to copy one specific LCR-fragment (De Raedt et al. 2006). However, such approaches are not applicable for larger LCRs and informative DNA fibers that cover the repeat (partially) as well as the delineating unique locus are necessary. Applications to achieve this long range structural information include optical mapping techniques (**Figure 1.3**) and long-read sequencing (**Figure 1.4**).

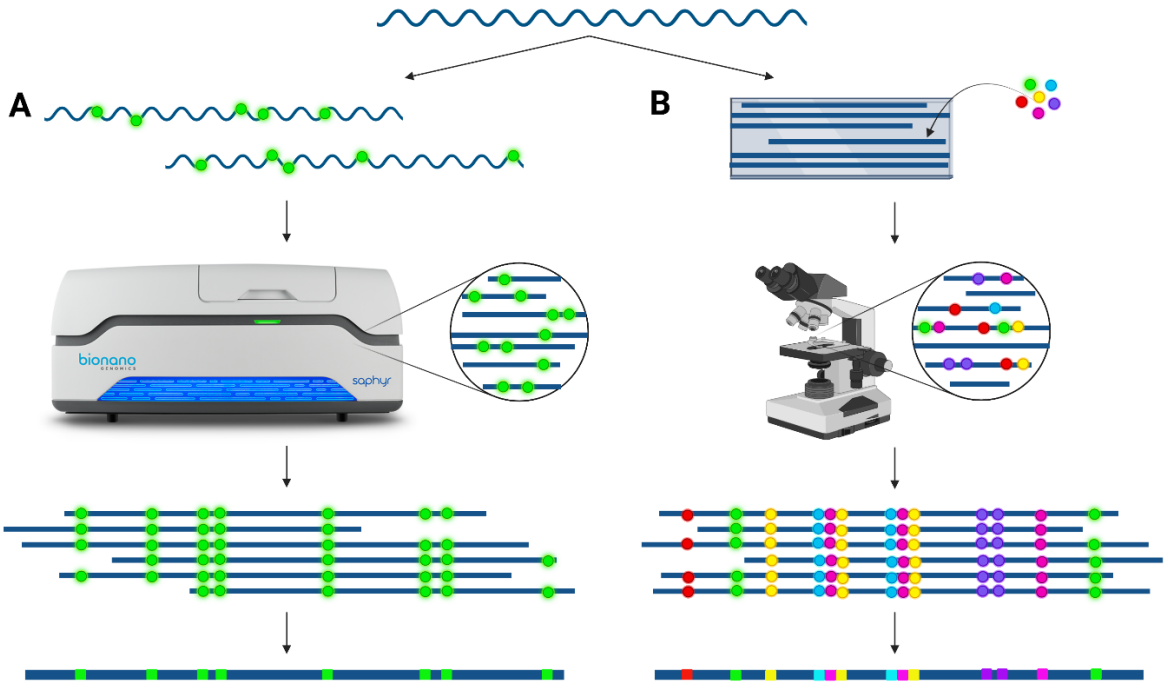


Figure 1.3: De novo assembly via optical mapping techniques. *(A) Bionano optical mapping: ultra-long (>100kb) DNA fibers are fluorescently labeled on specific locations across the whole genome and subsequently linearized and scanned. These images are converted into molecules to visualize the labeling pattern across the DNA strand. By overlapping identical patterns, a consensus pattern can be created. (B) Fiber-FISH: ultra-long (>300kb) DNA fibers are stretched onto slides and hybridized with a combination of fluorescently colored probes targeting the locus of interest. The slide is scanned using a microscope, allowing the detection of the different colors via scanning at different excitation levels. A consensus of the region of interest can be compiled based on overlapping colors and distances between the probes.*

Optical mapping techniques as fiber-FISH (Fluorescent *in situ* hybridization) and Bionano optical mapping, enable the *de novo* assembly of structurally complex loci with a resolution of 5-10kb (Chan et al. 2018). In the Bionano method, fluorescently labeled DNA fibers are scanned and the *de novo* assembly is based on the labeling pattern of the reads (**Figure 1.3A**). Due to the lengths of these fibers (N50 >150kb in general), large deletions, duplications, inversions, and translocations can be visualized. Using this technique, LCR haplotypes and additional structural variants were discovered in the 7q11.23, 15q13.3, and 16p12.2 locus (Mostovoy et al. 2021). The method is not targeted and therefore, structural

variants can be inferred at a whole-genome scale (Levy-Sakin et al. 2019). In the fiber-FISH method, one specific locus of the genome is scrutinized by using fluorescently labeled probes targeting the region of interest (**Figure 1.3B**). In short, long DNA molecules (>300kb) are stretched onto a glass surface and hybridized with probes (Louzada et al. 2017). This approach has proven to be successful for structural variation detection (Algady et al. 2018; Shi et al. 2019). Hence, long-read optical mapping approaches enable us to detect and compile structural variation haplotypes at subunit level.

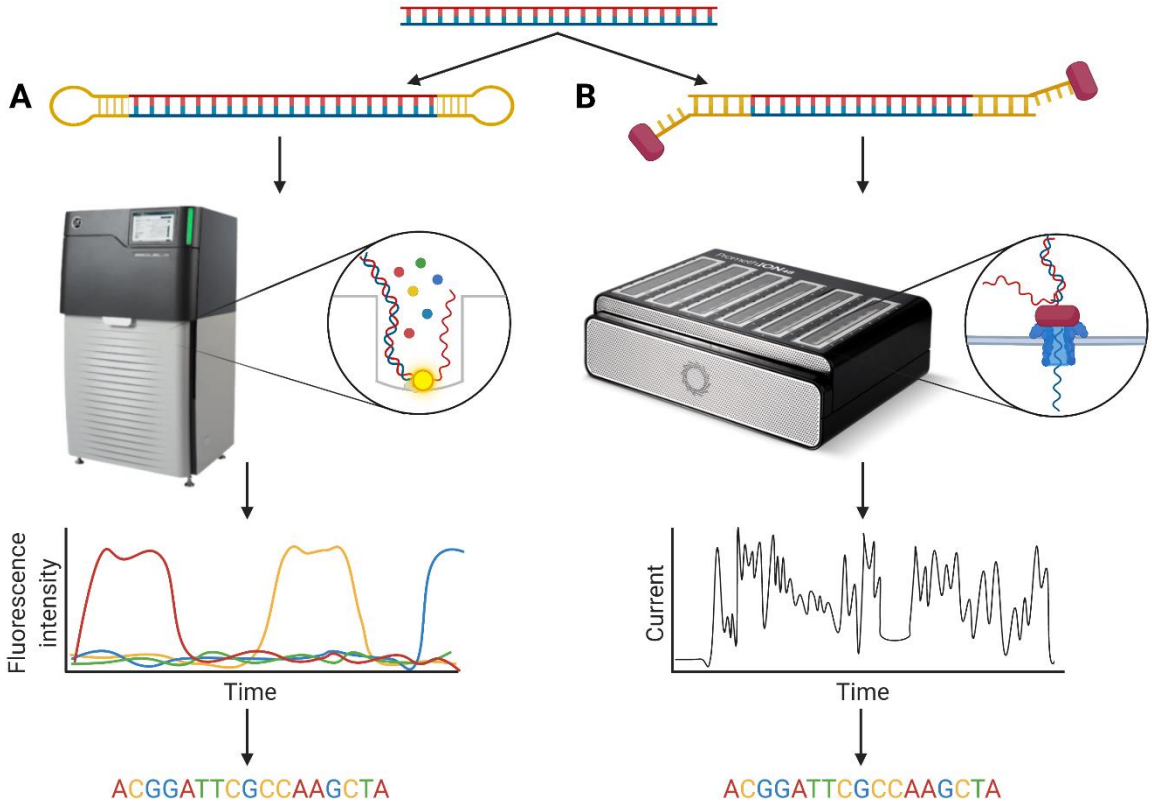


Figure 1.4: Long-read sequencing approaches. *(A) PacBio single-molecule real-time sequencing. SMRT bells, created by ligating hairpin adapters to the DNA fragments, are loaded onto the PacBio sequencing device. Fluorescent labeled nucleotides are added during the replication process. This fluorescent signal is converted to a nucleotide sequence. (B) Nanopore sequencing. During library preparation, a motor protein to control the translocation speed, is added to the DNA fragments. Passing of the DNA strand through the nanopore generates an electric current, which can be translated into a nucleotide sequence.*

Sequencing of long reads without the need for prior PCR amplification distinguishes the third generation from the second generation sequencing approaches. In 2011, Pacific Biosciences (PacBio) released their first single-molecule real-time (SMRT) sequencing platform (van Dijk et al. 2018). During library preparation, hairpin adaptors are ligated to the DNA template, creating a SMRT bell, which can enter the sequencing unit (zero-mode waveguide) (**Figure 1.4A**). The polymerase in this sequencing unit starts the replication after binding to the hairpin adaptor. The fluorescent-labeled nucleotides added during the reaction generate different light pulses, which can be interpreted as a specific sequence of bases (Rhoads and Au 2015) (**Figure 1.4A**). Read lengths can reach over 20kb with an accuracy of over 99% (Mohammadi and Bavi 2021). This methodology was successfully applied to

resolve the repeat size and sequence of the *FMR1* CGG repeat expansion in the Fragile X syndrome (Ardui et al. 2017). In 2014, Oxford Nanopore Technologies (ONT) launched their first nanopore sequencer, MinION (van Dijk et al. 2018). Here, DNA molecules to which a motor protein is attached, are loaded onto a flowcell (**Figure 1.4B**). Each flowcell consists of nanopores embedded in a membrane over which an electric current is applied. The negatively charged DNA will be threaded towards the positive pole and therefore, driven through the pore. This will result in ionic current changes corresponding to a specific nucleotide sequence (Wang et al. 2021) (**Figure 1.4B**). Average read lengths are dependent on the input material and library preparation method, but the longest read reported so far has a length of 2.27Mb (Payne et al. 2019). Error rates are lower than 5% but the accuracy is subject to continuous improvement (Wang et al. 2021). ONT sequencing of 3622 Icelanders has showed that the technique is able to accurately characterize structural variants at a population-scale (Beyter et al. 2021). Hence, the approachability and high-throughput possibility of nanopore sequencing started a revolution in CNV research.

Combinations of second and third generation sequencing tools and optical mapping techniques are golden standard for the generation of whole-genome *de novo* assemblies (Gordon et al. 2016; Mao et al. 2021). The aim of the telomere-to-telomere consortium was to generate the first complete assembly of a human genome, including unresolved loci as ribosomal rRNA, centromeric satellites, and segmental duplications. In 2020, they published the first gapless sequence of a complete human X chromosome (Miga et al. 2020). Using Ultra-long read Nanopore (39X coverage and N50 ~ 70kb) and SMRT PacBio (70X coverage) data, they performed a *de novo* assembly of the homozygous complete hydatidiform mole CHM13 genome. This initial draft was polished by (shorter) Nanopore and PacBio, linked-read Illumina (10X Genomics), and Bionano optical mapping data. Human chromosome X was then manually finished to a single contig, closing 29 gaps compared to the latest human reference genome (hg38). The assembly showed an accuracy of 99.995% based on mapped Illumina data (Miga et al. 2020). One year ago, the complete sequence of human chromosome 8 was published (Logsdon et al. 2021) and in April 2022, the first complete human genome (Nurk et al. 2022).

1.4 The 22q11.2 deletion syndrome

1.4.1 Epidemiology

The 22q11.2 deletion syndrome (22q11.2DS) is the most common microdeletion syndrome in humans, with an estimated incidence of 1 in 2148 live births (Blagojevic et al. 2021). Prenatal studies reported a higher incidence in fetuses, ranging from 1 in 1000 in anatomically normal fetuses to 1 in 100 in fetuses with abnormal ultrasonograms (Grati et al. 2015). Both males and females are equally affected (McDonald-McGinn et al. 2015). In terms of deletion origin, there is a slightly higher maternal origin (Costain et al. 2011; Delio

et al. 2013). The general incidence is estimated to be higher since not all 22q11.2 deletion carriers are ascertained, due to limited cytogenetic resolution, reduced penetrance, and difficulties in racial-specific diagnosis (McDonald-McGinn et al. 2015, 2005; Liu et al. 2014).

In 1965, Dr. Angelo DiGeorge described the DiGeorge syndrome (MIM: 188400): infants with a triad of symptoms, including thymus absence, hypoparathyroidism and congenital heart disease (DiGeorge 1965). In two familial cases of DiGeorge syndrome, a partial deletion of chromosome 22 (pter-q11) was observed, providing an argument that the 22q11 locus was involved in the causal mechanism (de la Chapelle et al. 1981; Kelley et al. 1982). In 1991, using FISH probes covering this locus, DiGeorge syndrome was associated with a deletion of the 22q11.2 locus in 11% of the investigated patients (Scambler et al. 1991). Later, 22q11.2 deletions were also found as causative in other similar or even unrelated disorders as velocardiofacial syndrome (MIM: 192430), conotruncal anomaly face syndrome, CATCH22 (cardiac defects, abnormal facial features, thymic hypoplasia, cleft palate, and hypocalcemia) Opitz G/BBB syndrome, and Cayler cardiofacial syndromes (McDonald-McGinn et al. 2015). Together, all these carriers of a 22q11.2 deletion are now diagnosed with the overlapping term of the 22q11.2DS.

1.4.2 Phenotype of 22q11.2DS

The phenotypic spectrum of 22q11.2DS typically involves multiple organ systems and is characterized by variability of both the number and severity of symptoms between patients (**Figure 1.5**). Characteristics can be absent or present in a subset of patients, range from mild to life-threatening, and differ in age of onset (McDonald-McGinn et al. 2015).

Dysfunctionalities during the fetal pharyngeal arch development cause the major subset of congenital features observed in the 22q11.2DS (**Figure 1.5**) (McDonald-McGinn et al. 2015). Congenital heart disease is the typical initial manifestation leading to the diagnosis and is present in 64% of patients. Since the outflow tract is affected, mainly conotruncal heart defects (truncus arteriosis, ventricular septal defect, tetralogy of Fallot) are observed (Campbell et al. 2018). In addition, cardiac complications are considered as the main cause of mortality in children with the 22q11.2DS (McDonald-McGinn et al. 2015). Approximately 77% of patients display immunological problems, encompassing allergies, abnormal T-cell function, and absence of the thymus (Morsheimer et al. 2017). In absence of severe medical problems or in combination with, the diagnosis is supported by the presence of dysmorphic facial features, involving all aspects of the face (abnormal ear helix, bulbous nasal tip, eye hooding, micrognathia, asymmetric crying facies) (Campbell et al. 2018). Other frequent (>50% of patients) medical issues include gastrointestinal difficulties, endocrine dysfunction, palatal anomalies (submucous cleft palate and velopharyngeal insufficiency), and hypocalcaemia due to hypoparathyroidism (McDonald-McGinn et al. 2015; Campbell et al. 2018).

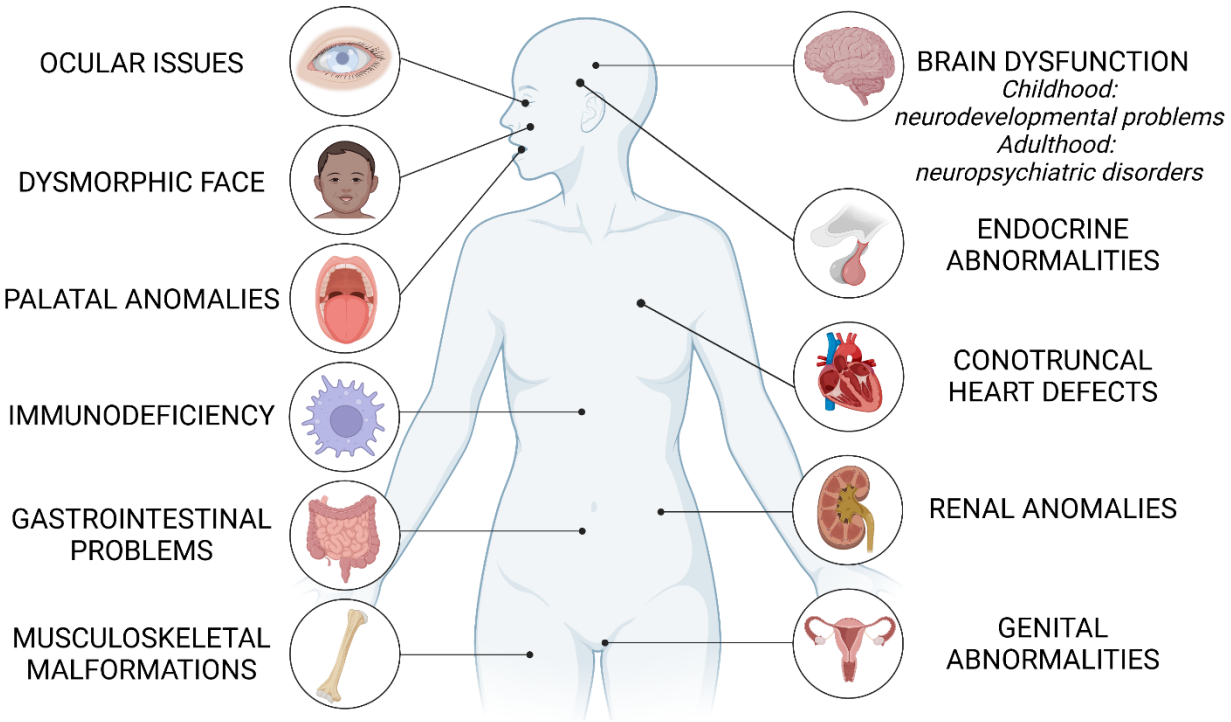


Figure 1.5: Overview of the main phenotypic features characterizing the 22q11.2DS. *Summary based on McDonald-McGinn et al. 2015, and Campbell et al. 2018.*

During childhood and adolescence, neurodevelopmental disabilities are frequently reported in patients with 22q11.2DS (McDonald-McGinn et al. 2015). In infancy, motor delays and speech and language deficits dominate, while in primary school, learning difficulties and cognitive deficits become apparent (Swillen and McDonald-Mcginn 2015). The developmental outcome is influenced by person-specific (genetic profile, prematurity, presence of peri-operative seizures) and familial and environmental (socio-economic status, IQ of the parents and siblings) risk factors (Swillen et al. 2018). The average intelligence quotient (IQ) is calculated at 70, which is low compared to the average of 100 in the normal developing adolescents. In addition, a negative correlation between age and IQ score was reported (Swillen et al. 2018), expressed as a gradual decline in cognitive development over time. Furthermore, treatment and follow-up of these developmental, cognitive, behavioral, educational, and socio-emotional concerns is complexed by the associated medical background.

Neurological and psychiatric pathologies are more prevalent in patients with the 22q11.2DS compared to the general population (Bassett and Chow 2008). Attention-deficit hyperactivity disorder, autism spectrum disorder, intellectual disability and anxiety disorders are typically diagnosed in childhood or adulthood (McDonald-McGinn et al. 2015). Later-onset manifestations include Parkinson's disease (Butcher et al. 2013) and schizophrenia (Zinkstok et al. 2019). No differences are reported in the phenotypic presentation of these neuropsychiatric disorders and their response to antipsychotic medication in comparison to the normal, non-deletion carrier population (Dori et al. 2017; Gur et al. 2017). However, the prevalence of schizophrenia in the general population is

estimated at 0.5-1%, whereas the disease is expressed in at least one in four individuals with 22q11.2DS. Hence, the presence of this deletion is one of the strongest known risk factors for schizophrenia (Zinkstok et al. 2019).

1.4.3 Genotype variability

The deletion is thought to be caused by NAHR between the SDs or LCRs on chromosome 22 (LCR22s) (Edelmann et al. 1999). A cluster of eight LCR22s is present proximal on the q-arm of this chromosome, commonly termed LCR22-A until -H. Rearrangements amongst the four proximal LCR22s, -A, -B, -C and -D, are causing the 22q11.2DS (**Figure 1.6**). In 85% of individuals with the 22q11.2DS, NAHR occurred between LCR22-A and –D, the two largest LCR22s, resulting in a 3Mb deletion (**Figure 1.6**). This region contains 46 protein-coding genes and less-characterized transcripts within the LCR22s. In 90-95% of patients the deletion occurred *de novo* (Campbell et al. 2018; McDonald-McGinn et al. 2015).

Figure 1.6: Proximal low copy repeats on chromosome 22. *Organization of the LCR22s -A until -D, the four LCR22 blocks most involved in 22q11.2 rearrangements. Deletion types and sizes are indicated below the LCR22 blocks.*

Nested deletions are defined as deletions where NAHR involved either LCR22-B or -C (McDonald-McGinn et al. 2015). 5% of 22q11.2DS patients carry the 1.5Mb LCR22-A/B deletion. The 2Mb LCR22-A/C deletion, the 1.5Mb LCR22-B/D deletion, and the 1Mb LCR22-C/D deletion account for 2%, 4%, and 1%, respectively (**Figure 1.6**) (Campbell et al. 2018). Since they are associated with a milder phenotype, they are more frequently inherited. Deletions encompassing the distal LCR22s (LCR22-E until -H) are less frequent and the phenotypes differ from the traditional 22q11.2DS (Burnside 2015). These deletions are referred to as the distal 22q11.2 deletion syndrome (MIM: 611867) (Ben-Shachar et al. 2008). Rearrangements where one of the breakpoints is located in the unique sequence between the LCR22s have been observed as well and are termed atypical deletions (Burnside 2015).

The prevalence of these non-standard (nested, distal, and atypical) deletions is probably underestimated since they are associated with a milder phenotype, and therefore harder to diagnose. In addition, not all genetic tests were able to detect these uncommon deletions (for example, standard 22q11.2del FISH using probes TUPLE or N25, are located in the unique sequence between LCR22-A and -B).

Reciprocal to the standard 22q11.2DS, the 22q11.2 duplication syndrome (MIM: 608363) was described (Portnoï 2009). Symptoms include velopharyngeal insufficiency, behavioral deficits, and dysmorphic features and therefore has a phenotypic overlap with the 22q11.2DS. Of interest, inter-patient variability is present in the 22q11.2 duplication syndrome as well (Portnoï 2009).

1.4.4 Genotype-phenotype relationships in 22q11.2DS

Whereas the discovery of the 22q11.2 deletion was a major breakthrough for our understanding of the 22q11.2DS, the focus has shifted towards understanding the causes of the individual phenotypic variation seen in the syndrome and the (genetic) drivers causing this phenotypic variability. Several genetic hypotheses were investigated to explain (part of) the phenotypic variability observed in 22q11.2DS. First, deletion length can have an impact since haploinsufficiency of more or less genes may result in a severe or mild phenotype, respectively. Second, the deletion may become pathogenic in combination with a mutation on the second, non-deleted allele from the non-parent-of-origin (Hochstenbach et al. 2012). Third, genome-wide variants can exert an additive effect on the susceptibility for specific phenotypes (Bacchelli et al. 2020).

Variation in deletion length and therefore deletion of a different number of genes, would be a straightforward explanation for part of the phenotypic variability seen in 22q11.2DS patients. In general, similar phenotypes are observed in the standard LCR22-A/D and the smaller LCR22-A/B deletion. In addition, the LCR22-B/D and LCR22-C/D deletions result in similar, but less penetrant phenotypic features. Even distal deletions express some of the features from the standard 22q11.2DS (cardiac problems, dysmorphic features, developmental delay) (Burnside 2015). However, some subtle differences can be observed. For example, by comparing IQ scores (full-scale, verbal, and performance IQ) of 1353 LCR22-A/D and 74 LCR22-A/B deletion patients, all IQ scores were lower in the LCR22-A/D deletion group, with verbal IQ the most significant difference (75.60 in LCR22-A/D compared to 82.33 in LCR22-A/B subset). This suggests that haploinsufficiency of genes within the LCR22-B/D locus contributes to IQ (Zhao et al. 2018). However, in general, the deletion type cannot accurately predict the phenotypic outcome of the 22q11.2DS.

Several genes were the focus of different studies to investigate their role and exact contribution in the 22q11.2DS (**Supplementary Table S1.2**). One of the main candidate genes to contribute to phenotypic variability is *TBX1* (T-box transcription factor 1), since heterozygous mouse models are mimicking a phenotype similar to patients with 22q11.2DS, including cardiac problems, facial dysmorphologies, thymus alterations, and cleft palate (Jerome and Papaioannou 2001; Papangeli and Scambler 2013). Nonetheless, the gene is located between LCR22-A and -B and is therefore not affected in patients with LCR22-B/D and LCR22-C/D deletions, who display a similar phenotype. An 'alternative' gene proposed for phenotypic contribution in this locus is *CRKL* (v-crk avian sarcoma virus CT10 oncogene

homologue-like). Mice models indicate *CRKL*-mediated pathways interfere with cranial and cardiac neural crest cells during development, resulting in a range of traits, including heart defects, genitourinary problems, and thymus deficits (Racedo et al. 2015; Haller et al. 2017). Another interesting gene is *DGCR8* (DiGeorge syndrome critical region gene 8), located in the unique region between LCR22-A and -B. The gene encodes an element important in microRNA biosynthesis and *microRNA-185*, expressed in the brain. Reduced expression due to haploinsufficiency was linked with neuronal alterations and results were suggesting a possible link with schizophrenia in the 22q11.2DS (Stark et al. 2008; Forstner et al. 2013).

The deletion in the 22q11.2DS population can unmask recessive variants on the remaining, non-deleted allele leading to specific phenotypes (Hestand et al. 2016) (**Supplementary Table S1.2**). For example, mutations in *SNAP29* are involved in the pathogenesis of cerebral dysgenesis, neuropathy, ichthyosis and keratoderma, a recessive disorder (McDonald-McGinn et al. 2013). Other examples are Bernard-Soulier syndrome (a bleeding disorder) caused by hemizygous *GP1BB* mutations (Kunishima et al. 2013), *CDC45* variants involved in craniosynostosis pathology (Unolt et al. 2020), and Van den Ende-Gupta syndrome, characterized by craniofacial and skeletal features, involving *SCARF2* mutations (Bedeschi et al. 2010). Hence, part of the phenotypic variability observed in the 22q11.2DS is caused by variation on the remaining allele, especially lower incidence phenotypes.

Modifiers outside the 22q11.2 locus have been identified via targeted and genome-wide approaches. First, candidate genes were tested for their phenotypic contribution. For example, the vascular endothelial growth factor (*VEGF*) was suggested as a candidate modifier for cardiovascular abnormalities in the 22q11.2 deletion syndrome (Stalmans et al. 2003). Mice studies show vascular malformations, cardiac defects, and reduced tbx1 levels in mice expressing a specific *VEGF* isoform. To associate these results with phenotypic expression in the 22q11.2DS, they compared *VEGF*-specific SNPs between healthy controls and patients with the 22q11.2DS and cardiovascular defects (Stalmans et al. 2003). However, a larger study including 122 patients (50% with cardiovascular defects) and their parents, was not able to replicate the SNP or haplotype association (Calderón et al. 2009). Second, genome-wide association studies (GWAS) can identify common genetic variants contributing to the 22q11.2DS phenotype. Such a GWAS study was performed for tetralogy of Fallot, a severe congenital heart defect. By comparing the output data from 22q11.2DS patients with (n=326) and without (n=566) tetralogy of Fallot, one SNP was significantly associated. This SNP was located in an intron in *GPR98* (G-protein coupled receptor V1) on chromosome 5q14.3. They supposed that this variant can affect transcriptional regulation of genes involved in heart development, via for example topologically associated domains (Guo et al. 2017).

To uncover the genetic factors contributing to the increased risk for the development of schizophrenia in the 22q11.2DS population, the International 22q11.2DS Brain Behavior Consortium (IBBC), funded by the National Institute of Mental Health (NIMH), was established (Gur et al. 2017). It is a collaborative effort between 22 European and American sites to recruit 22q11.2DS patients and to collect phenotype and genotype data, via psychiatric and cognitive assessments and sequencing of DNA, respectively (Gur et al. 2017). Data were used to compare common and rare variants on the non-deleted 22q11.2 allele and genome-wide, between 214 22q11.2DS patients with and 221 without schizophrenia (Cleynen et al. 2021). First, polygenic risk score was calculated for both groups and was significantly higher in the group with a psychotic illness, suggesting the contribution of genetic factors besides the deletion itself. Second, association analyses for common variants were not able to identify genome-wide or intact 22q11.2 allele associations with schizophrenia in the investigated cohort. Third, no rare variants were significantly associated. Hence, this dataset and analyses, in the largest cohort of individuals with 22q11.2DS investigated thus far, suggested the absence of variants with a large effect size on the schizophrenia risk, on the remaining allele or genome-wide (Cleynen et al. 2021).

Despite extensive research efforts, the genetic basis for the phenotypic variability of the major phenotypes in the 22q11.2DS remain elusive and new strategies are implemented to unravel this enigma (Vermeesch 2022). Davies et al. (2020) used a schizophrenia polygenic score to study the shared genetic background between schizophrenia and schizophrenia-related phenotypes. They showed that this score can be useful in risk stratification since it is associated with cognitive decline and sub – treshold psychosis as well (Davies et al. 2020). A new way of approaching the schizophrenia – genetics correlation is via the threshold model of functional low frequency single nucleotide variants (Breetvelt et al. 2022). Their results suggest that the phenotypic outcome can be determined by an interplay of structural variation and low frequency single nucleotide variant burden in the 22q11.2 locus (Breetvelt et al. 2022).

While several groups are continuing their work on investigating the whole genome and the remaining allele for phenotypic contributions, one important locus is removed from all these analyses: the low copy repeats on chromosome 22. They are responsible for and flanking the rearrangements in the 22q11.2DS, but are extremely complex and not even accurately represented in the hg38 reference genome.

1.5 Low copy repeats on chromosome 22

The acrocentric chromosome 22 is the second shortest human autosome with a length of 50Mb of which 15Mb belong to the p-arm. In 1999, chromosome 22 was claimed to be the first finished human chromosome sequence (Dunham et al. 1999). This sequence covered the euchromatic part and consisted of 12 contigs spanning 33.4Mb. Four of the 11 gaps were located in the LCR22-associated 22q11.2 locus (Dunham et al. 1999). In 2008, they

were able to close 8 of the 11 gaps by combining sequencing techniques (Cole et al. 2008). However, the 22q11.2 locus still harbored two of the remaining gaps, due to the presence of LCR22s and their associated assembly problems. Hence, the presence of eight interstitial LCR22s significantly increases the complexity of the 22q11.2 locus and causes several chromosomal conditions.

1.5.1 22q11.2 instability creating chromosomal abnormalities

The 22q11.2 locus can be considered as one of the most instable loci in the human genome. NAHR between the LCR22s results in the 22q11.2 deletion and reciprocal duplication syndrome, of which the first one is the most common microdeletion disorder in humans (McDonald-McGinn et al. 2015). The high prevalence of *de novo* cases validates the high mutation rate of the locus (Emanuel 2008). In addition to copy number variable deletions and duplications, copy number neutral events involving LCR22s are described as well.

Inversions between LCR22s are not described so far (Gebhardt et al. 2003), as opposed to non-random constitutional translocations. Recurrent translocations include t(8;22)(q24.1;q11.2), t(11;22)(q23;q11.2), and t(17;22)(q11.2;q11.2), whereas t(1;22)(p21.1;q11.2) and t(4;22)(q35.1;q11.2) belong to the non-recurrent class (Kato et al. 2012). The translocation can directly affect a gene and his function, as described for t(17;22) leading to neurofibromatosis type 1 by disrupting the *NF1* gene on chromosome 17 (Hsiao et al. 2014).

The most known non-Robertsonian translocation is t(11;22)(q23;q11.2). Since there is no gain or loss of DNA, carriers are phenotypically normal, except for reproductive problems that are often observed in this group (Shaikh et al. 1999). They can have phenotypically normal offspring when the proband inherits the two normal chromosomes or the two translocated chromosomes, as a result of balanced 2:2 segregation during meiosis. Gametes and embryos following unbalanced 2:2 segregation are conceived as well, although this will mainly lead to spontaneous abortion and explains the high miscarriage frequency in the carrier parents (Zenagui et al. 2019). However, the offspring can inherit the derivative chromosome from the translocation-carrying parent in addition to the normal non-translocation partner chromosomes. This mechanism is known as 3:1 meiotic malsegregation (Shaikh et al. 1999), and can lead to clinical symptoms since the copy number is changed. For example, carriers of the t(11;22)(q23;q11.2) translocation have a 10% risk of having a child with the supernumerary der(22)t(11;22) syndrome, also known as Emanuel syndrome (MIM: 609029), caused by the presence of a partial trisomy of the 11q23 and the 22q11 locus. Clinical features of the syndrome include ear pits, micrognathia, heart malformations, and cleft palate (Carter et al. 2009).

The breakpoints of these constitutional 22q11.2 translocations all cluster in LCR22-B, described by Emanuel in 2008 as the most rearrangement-prone site of the human genome

(Emanuel 2008). By scrutinizing these rearrangements, palindromic AT-rich repeats (PATRRs) were identified at the breakpoint locus in LCR22-B as well as in the involved partner chromosome locus, generating instability and an opportunity for rearrangement via secondary structure formation (Kato et al. 2012). However, although an important role in the generation of translocations, the contribution of palindrome-mediated pathways leading to deletions and duplications in the 22q11.2 locus is not known yet (Emanuel 2008).

Several non-constitutional translocations involving 22q were described to play a role in carcinogenesis as well. The chromosomal abnormality resulting from a translocation between chromosome 9q34 (ABL locus) and chromosome 22q11 (BCR locus) is called the Philadelphia chromosome. The *BCR-ABL1* fusion gene encodes a hybrid protein important in the disease pathway of several forms of leukemia (Groffen et al. 1984). In addition, the t(8;22) translocation is a variant of the t(8;14) translocation associated with Burkitt's lymphoma (Boerma et al. 2008), and Ewing sarcoma (MIM: 612219) is caused by the t(11;22)(q24;q12) translocation fusing the *FLI1* and *EWS* genes on chromosomes 11 and 22, respectively (Riggi et al. 2021). These oncogenic translocations arise via mitotic rearrangement and are only present in the tumor or affected tissue.

A more complex constitutional LCR22-mediated rare disorder is cat eye syndrome (MIM: 115470), characterized by iris coloboma, anal atresia, and ear malformations (McDermid and Morrow 2002). The karyotype marks the presence of a small supernumerary dicentric, bisatellited chromosome representing an inv dup(22) (q11.2) (McDermid and Morrow 2002; Emanuel and Shaikh 2001). This structure leads to a tetrasomy of the chromosome 22 p-arm and part of the 22q11.2 locus (from the centromere to the distal breakpoint). The breakpoints are mapped to LCR22-A and/or -D, and different cytogenetic cat eye syndrome types exist based on the specific LCR22 involvement (McDermid and Morrow 2002): type I and IIb marker chromosomes are symmetrical with both breakpoints clustering in LCR22-A or -D, respectively. Asymmetric cat eye syndrome marker chromosomes characterized by one LCR22-A and one LCR22-D breakpoint are classified as type IIa. The supernumerary chromosome is predicted to be created via interchromosomal misalignment or intrachromosomal recombination leading to an inversion with subsequent homologue crossover (Emanuel and Shaikh 2001). Hence, the presence of LCR sequence on chromosome 22 makes the locus prone to several types of genomic rearrangements, leading to disease and disease predisposition.

1.5.2 Complex structure of the LCR22s

The breakpoints of the 22q11.2 rearrangements coincide in the LCR22s, providing evidence for the presence of specific elements involved in homologous and non-homologous rearrangement mechanisms. Due to the difficulties associated with LCR22 research, the 22q11.2 locus can be considered as one of the most complex loci in the human genome.

Comparisons between the different reference genomes showed several inconsistencies at the level of the LCR22s (**Figure 1.7**).

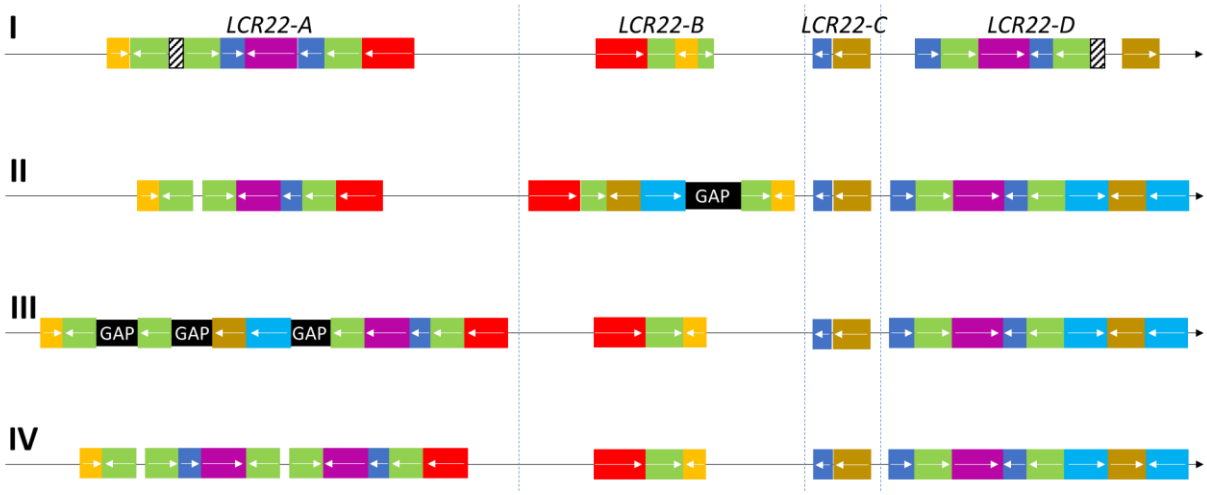


Figure 1.7: Organization of LCR22 duplicated modules in several references. *(I) First description of the LCR22 structure in Shaikh et al. (2000). These modules serve as 'essential duplicons' in the comparison and composition of the LCR22s in the other reference genomes (II-IV). The yellow duplicon contains USP18 (or a paralogue). The red duplicon encompasses (paralogues of) the genes PRODH and DGCR6. The PI4KA gene or a paralogue is present in the mustard colored blocks. The cyan duplicon always covers a RIMBP3 paralogue, the blue duplicon a BCR paralogue, and the magenta duplicon a GGT paralogue. A FAM230 paralogue can be found in the green duplicon, as well as a palindromic AT-rich repeat. (II) LCR22 organization in GRCh37/hg19. (III) LCR22 organization in GRCh38/hg38. (IV) LCR22 organization, for the first time without gaps, in the most recent reference genome (T2T-CHM13).*

Despite their complexity, an organizational overview of the LCR22 structure was provided even before the release of the first human reference genome assembly by the group of Professor Beverly Emanuel (Shaikh et al. 2000) (**Figure 1.7I**). Based on sequencing data from bacterial artificial chromosome (BAC), P1-derived artificial chromosome and cosmid insert clones, they differentiated four LCR22s, named LCR22-A, -B, -C, and -D, each represented as a mosaic patchwork of different modules. These modules are shared between one or more LCR22s via duplication and therefore have high sequence similarity (97-98%). For the two largest blocks, LCR22-A and -D, the length was estimated at 350kb and 250kb, respectively. They share long stretches of identical duplicon composition, except for the LCR22-A specific proximal USP18 (yellow) and distal DGCR6/PRODH module (red), and the LCR22-D specific PI4KA module (mustard) (**Figure 1.7**). The smaller LCR22-B and -C blocks are composed of modules that are present in LCR22-A and/or LCR22-D, but they do not share similar blocks between them, suggesting that rearrangements between the smaller LCR22s are not possible. Hence, the duplicated module structure of the LCR22s provides perfect elements for NAHR.

As opposed to the LCR22 structural representation by Shaikh et al (2000), difficulties were encountered in composing the LCR22s during the generation of the human reference genome (Chaisson et al. 2015). In hg19 (GRCh37, released February 2009), the LCR22s were delineated (**Table 1.1**) with a gap of 100kb present in LCR22-B (**Figure 1.7II**).

Surprisingly, LCR22-A was identified as the second smallest LCR22. In reference genome GRCh38, released in December 2013, relative LCR22 sizes were comparable to the first representation (Shaikh et al. 2000) with major changes compared to hg19 (**Figure 1.7III, Table 1.1**). First, a 100kb module including the genes RIMBP3, TMEM191B, and pseudogene PI4KAP1, was shifted from LCR22-B to LCR22-A. Second, no gaps were present anymore in LCR22-B, but three new gaps appeared in LCR22-A with lengths of 100kb (most proximal), 50kb, and 50kb (most distal). Hence, due to the variability and inconsistencies between the different genome assemblies, constructing the exact composition and sequence of the LCR22s is still an ongoing challenge.

Table 1.1: Exact chromosomal locations of the different LCR22s in hg19, hg38, and T2T reference genomes.

	GRCh37/hg19	**GRCh38/hg38**	**T2T-CHM13**
LCR22-A	chr22:18,639,043-19,022,986	chr22:18,156,276-19,035,473	CP068256.2:18,828,186-19,410,796
LCR22-B	chr22:20,128,537-20,731,921	chr22:20,141,014-20,377,631	CP068256.2:20,520,047-20,781,953
LCR22-C	chr22:21,021,564-21,092,560	chr22:20,667,276-20,738,272	CP068256.2:21,075,991-21,146,982
LCR22-D	chr22:21,363,668-21,916,380	chr22:21,009,379-21,562,091	CP068256.2:21,418,153-21,975,566

The lack of an accurate reference genome implies the occurrence of difficulties in standard pipelines as mapping and variant calling. This problem is even worse in the LCR22 locus due to the extreme duplication nature. As a consequence, modules involved in the rearrangement and exact nucleotide sequences were never fully charted and are hampering future research. Therefore, there is need for *de novo*, non-reference based approaches to correctly compile the LCR22 haplotype. Although the assembled chromosome 22 of the telomere-to-telomere consortium showed no gaps in the LCR22s (**Figure 1.7IV**) (Nurk et al. 2022), inter-individual genetic variation within the LCR22s has not been well characterized. In addition, exact rearrangement loci are not disentangled yet at sequence level. This will be an important step to understand the content, variation, and potential role of the LCR22s.

***Supplementary Table S1.1**: Overview of most common recurrent microdeletion and -duplication syndromes.*

Chromosomal location	Rearrangement	Main phenotypic features	MIM#/reference
1q21.1	Deletion	Microcephaly, ID, ocular anomalies, cardiac problems	612474
	Deletion (additional RBM8A SNV) (thrombocytopenia-absent radius, TAR)	Thrombocytopenia, aplasia of radii (long forearm bones), other skeletal defects, cardiac problems	274000
	Duplication	Macrocephaly, ID, autism, schizophrenia	612475
5q35.3	Deletion (Sotos)	Childhood overgrowth, mental retardation, facial dysmorphism, hyperinsulinemic hypoglycemia	117550
	Duplication	Short stature, microcephaly, facial dysmorphism, ID	Dikow et al. (2013)
7q11.23	Deletion (Williams-Beuren)	Facial dysmorphism, ID, cardiac problems, 'sociable phenotype'	194050
	Duplication	Facial dysmorphism, speech delay, cardiac problems, cryptorchidism	609757
8p23.1	Deletion	Cardiac problems, diaphragmatic hernia, ID	Wat et al. (2009)
	Duplication	Cardiac problems, ID, learning difficulties, facial dysmorphism	Barber et al. (2013)
9q34	Deletion (Kleefstra 1)	Epileptic seizures, ID, cardiac problems, facial dysmorphism	610253
	Duplication	Hyperactivity, psychomotor retardation, musculoskeletal abnormalities, facial dysmorphism	Allderdice et al. (1983)
15q11.2	Paternal deletion (Prader-Willi)	Hypotonia, ID, obesity, small hands and feet, hypogonadotropic hypogonadism	176270
	Maternal deletion (Angelman)	ID, speech and language limitations, 'happy personality'	105830
	Duplication	Autism, ID, seizures, ataxia	608636
16p11.2	Deletion	Autism, developmental delay, obesity	611913

	Duplication	Autism, attention-deficit hyperactivity disorder, facial dysmorphism	614671
17p11.2	Deletion (Smith-Magenis)	ID, facial dysmorphism, behavioral problems (anxiety, aggression, self-destructive behavior)	182290
	Duplication (Potocki-Lupski)	Hypotonia, congenital anomalies, ID	610883
17p12	Deletion (Hereditary neuropathy with liability to pressure palsies)	Neuropathy with liability to pressure palsies	600361
	Duplication (Charcot-Marie-Tooth disease type 1A)	Muscle weakness and atrophy, reduced sensation	601098
17q11.2	Deletion (NF1 microdeletion)	ID, facial dysmorphism, early-onset neurofibromas	613675
	Duplication (NF1 microduplication)	ID, facial dysmorphism, seizures	618874
17q21.31	Deletion (Koolen-De Vries)	Hypotonia, ID, 'friendly personality', facial dysmorphism, cardiac problems, seizures	610443
	Duplication	Psychomotor delay, poor social interaction, ID, facial dysmorphism, congenital malformations	613533
22q11.2	Deletion	Cardiac problems, immunodeficiency, palatal anomalies, neuropsychiatric disease, ID	192430
	Duplication	Behavioral problems, facial dysmorphism, velopharyngeal insufficiency, ID	608363
Yq11.2	Deletion (AZFa microdeletion)	Male infertility	400042

This overview of recurrent microdeletion and -duplication syndromes in the human genome is based on Harel and Lupski (2018) and the syndrome-specific pages of the OMIM website (Online Mendelian Inheritance in Man catalog, www.omim.org)

Supplementary Table S1.2: *Limited overview of mutations in genes on the 22q11.2 region and their associated phenotypes.*

Gene	Locus	Type	Phenotype	Reference
PRODH	LCR22-A	SNP	Schizophrenia susceptibility	Bender et al. (2005);
		Recessive	Hyperprolinemia type 1 (neurological deficits, seizures, ...)	Prasad, Howley, and Murphy (2008)
SLC25A1	LCR22-A/B	Recessive	Combined D2 and L2 hydroxyglutaric aciduria (encephalopathy, seizures, failed psychomotor development)	Nota et al. (2013)
HIRA	LCR22-A/B	Dominant	Impaired dendritic outgrowth, abnormal neurodevelopment	Jeanne et al. (2021)
CDC45	LCR22-A/B	Recessive	Craniosynostosis	Unolt et al. (2020)
GP1BB	LCR22-A/B	Recessive	Bernard-Soulier syndrome	Budarf et al. (1995); Ludlow et al. (1996); Kunishima et al. (2013)
TBX1	LCR22-A/B	Dominant	Congenital heart defects, thymus and parathyroid defects	Jerome and Papaioannou (2001); Papangeli and Scambler (2013)
GNB1L	LCR22-A/B	Dominant	Schizophrenia susceptibility	Ishiguro et al. (2010); Prasad, Howley, and Murphy (2008)
COMT	LCR22-A/B	Dominant	Schizophrenia susceptibility	Prasad, Howley, and Murphy (2008)
TANGO2	LCR22-A/B	Recessive	Infancy-onset metabolic crises with encephalocardiomyopathy	Kremer et al. (2016)
DGCR8	LCR22-A/B	miRNA interference	Neuronal deficits, schizophrenia	Forstner et al. (2013); Stark et al. (2008)
ZDHHC8	LCR22-A/B	Dominant	Schizophrenia susceptibility	Mukai et al. (2004)
SCARF2	LCR22-B/C	Recessive	Van den Ende–Gupta syndrome	Bedeschi et al. (2010)
PI4KA	LCR22-C/D	Recessive	Perisylvian polymicrogyria, cerebellar hypoplasia, arthrogryposis; Hypomyelinating leukodystrophy	Pagnamenta et al. (2015); Verdura et al. (2021)
SNAP29	LCR22-C/D	Recessive	Cerebral dysgenesis, neuropathy, ichthyosis, and keratoderma	McDonald-McGinn et al. (2013)
CRKL	LCR22-C/D	Dominant	Congenital heart defects, genitourinary problems	Haller et al. (2017; Racedo et al. (2015)
LZTR1	LCR22-C/D		22q11.2 deletion: reduced schwannomatosis risk	Evans et al. (2021)

CHAPTER 2

OBJECTIVES

2 OBJECTIVES

Although the 22q11.2DS is the most common microdeletion disorder, the cause of the phenotypic variability remains elusive. We hypothesize that variability in LCR22 haplotype composition and/or the exact 22q11.2 deletion breakpoints can explain (part of) the phenotypic interpatient variability in the 22q11.2DS. However, exploration of the role of the very complex LCR22s flanking the deletion is hampered. The hg38 human reference genome still contains sequence gaps in the LCR22s. Reference-based variant mapping removes those regions from standard sequence analysis pipelines. However, considering the importance of the LCRs in human evolution and knowing that their duplication nature could lead to novel transcripts, the LCR22s warrant further study. Obtaining the full sequence information will enable full mapping of the transcripts in this locus and investigate the sequence variability. In addition, we hypothesize that by studying the LCR22s we would gain better insights in the mechanistic drivers of the rearrangement of the 22q11.2DS.

The main objective of this thesis was to map LCR22s and the 22q11.2DS rearrangement breakpoints. First, the complex LCR22 structures were constructed for the first time at subunit level for individuals with the 22q11.2DS, their parents, and controls, using fiber-FISH and Bionano optical mapping techniques (chapter 3). By mapping the LCR22s, we **uncovered an extreme variability of the LCR22s with LCR22-A sizes ranging from 250 - 2000kb in more than 25 haplotypes.** To determine whether this variability would be evolutionary conserved or is human-specific, we charted the evolutionary history of the 22q11.2 locus in Great Apes (chapter 4). We hypothesized that variability in LCR22 sizes might be driven by allelic homologous recombination. To test this hypothesis we scrutinized families for recombination over LCR22-A (chapter 5). In order to unravel the mechanisms causing the 22q11.2DS rearrangement we mapped the 22q11.2 rearrangement crossover sites at nucleotide level. On the one hand, we examined atypical deletions, where at least one of the breakpoints is not located in an LCR22, by whole-genome sequencing, fiber-FISH and PCR breakpoint cloning validation (chapter 6). On the other hand, recurrent deletions between two LCR22s were scrutinized by fiber-FISH mapping of the patient/parent-of-origin and a combination of (ultra-)long read sequencing methods (chapter 7).

CHAPTER 3

The 22q11.2 low copy repeats are characterized by unprecedented size and structural variability

Wolfram Demaerel[1,10], Yulia Mostovoy[2,10], Feyza Yilmaz[3,4,10], Lisanne Vervoort[1], Steven Pastor[5], Matthew S. Hestand[1,6,7], Ann Swillen[1], Elfi Vergaelen[1], Elizabeth A. Geiger[4], Curtis R. Coughlin[4], Stephen K. Chow[2], Donna McDonald-McGinn[5], Bernice Morrow[8], Pui-Yan Kwok[2], Ming Xiao[9], Beverly S. Emanuel[5], Tamim H. Shaikh[4], and Joris R. Vermeesch[1]

[1] *Department of Human Genetics, KU Leuven, Leuven, Belgium*

[2] *Cardiovascular Research Institute, UCSF School of Medicine, San Francisco, CA, USA*

[3] *Department of Integrative Biology, University of Colorado Denver, Denver, CO, USA*

[4] *Department of Pediatrics, Section of Clinical Genetics and Metabolism, University of Colorado Denver, Aurora, CO, USA*

[5] *Division of Human Genetics, Children's Hospital of Philadelphia and Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

[6] *Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA*

[7] *Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA*

[8] *Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA*

[9] *School of Biomedical Engineering, Drexel University, Philadelphia, PA, USA*

Status: Published in *Genome Research*

Abstract

Low copy repeats are recognized as a significant source of genomic instability, driving genome variability and evolution. The LCR22s on chromosome 22 mediate non-allelic homologous recombination leading to the 22q11.2 deletion syndrome. However, LCR22s are among the most complex regions in the genome, and their structure remains unresolved. Using fiber-FISH and Bionano optical mapping, we assembled LCR22 alleles in 187 cell lines. Our analysis uncovered an unprecedented level of variation in LCR22s, including LCR22-A alleles ranging in size from 250kb to 2000kb. Further, the incidence of various LCR22 alleles varied within different populations. Thus, we present the most comprehensive map of LCR22 variation to date. This will greatly facilitate the investigation of the role of LCR22 variation as a driver of 22q11.2 rearrangements and the phenotypic variability among 22q11.2DS patients.

3 The 22q11.2 low copy repeats are characterized by unprecedented size and structural variability

3.1 Introduction

Low copy repeats, also referred to as segmental duplications, are a driving force in genome evolution, adaptation, and instability. In the most recent T2T-CHM13 reference genome, 6.6% of the reference assembly consists of LCRs (Nurk et al. 2022). Duplications have long been recognized as a potential source for the rapid evolution of new genes with novel functions (Bailey et al. 2001; Jiang et al. 2007; Dennis et al. 2017). Studies have suggested potential functional roles for genes within LCRs in synaptogenesis, neuronal migration, and neocortical expansion within the human lineage (Charrier et al. 2012; Dennis et al. 2012; Boyd et al. 2015; Florio et al. 2015). However, these regions are highly enriched for gaps and assembly errors even within recent versions of the human reference genome (Bovee et al. 2008; Genovese et al. 2013; Chaisson et al. 2015). This is because LCRs are both highly sequence identical and copy number polymorphic. These features strongly hamper the study of the precise role of LCRs as drivers of human disease or evolution.

High sequence homology between LCR copies is a driver of recurrent genomic rearrangements. Misalignment of homologous chromosomes or sister chromatids can lead to nonallelic homologous recombination (Inoue and Lupski 2002). NAHR between LCRs results in reciprocal deletions, duplications, or inversions, often referred to as genomic disorders (Inoue and Lupski 2002). The 22q11.2 deletion syndrome (22q11.2DS) is the most common genomic disorder. Chromosome 22q11.21 contains four LCR22s, termed LCR22-A until -D. All LCR22s are composed of different repeat subunits which are present in variable composition, copy number, and orientation (Shaikh et al. 2007). LCR22-A and -D are the largest and, in genome build hg38, estimated to span 1Mb and 400kb, respectively. Copy number variations (CNVs) exist within the LCR22s (Guo et al. 2011). In addition, a CNV encompassing PRODH, DGCR6, and DGCR5 (referred to as LCR22-A+) was recently mapped within LCR22-A (Guo et al. 2018). Nonetheless, the overall architecture of several LCR22s remains unresolved. Because the 22q11.2DS breakpoints are embedded within these unresolved LCR22s, their exact locations have remained elusive. We set out to map these repeats to elucidate the LCR22 structures and their variability.

3.2 Results

3.2.1 Subunit-resolution LCR22 assemblies using fiber-FISH

To resolve the LCR22 subunit organization, we first redefined repeat subunits that are present in the LCR22s of human reference hg38 (**Figure 3.1A**). We aligned the LCR22 sequences to each other, revealing all segments with a sequence similarity >99%. Based

on this LCR22 decomposition, we identified distinct repeat subunits that have a copy number of at least two on chromosome 22q11.21 (**Figure 3.1**).

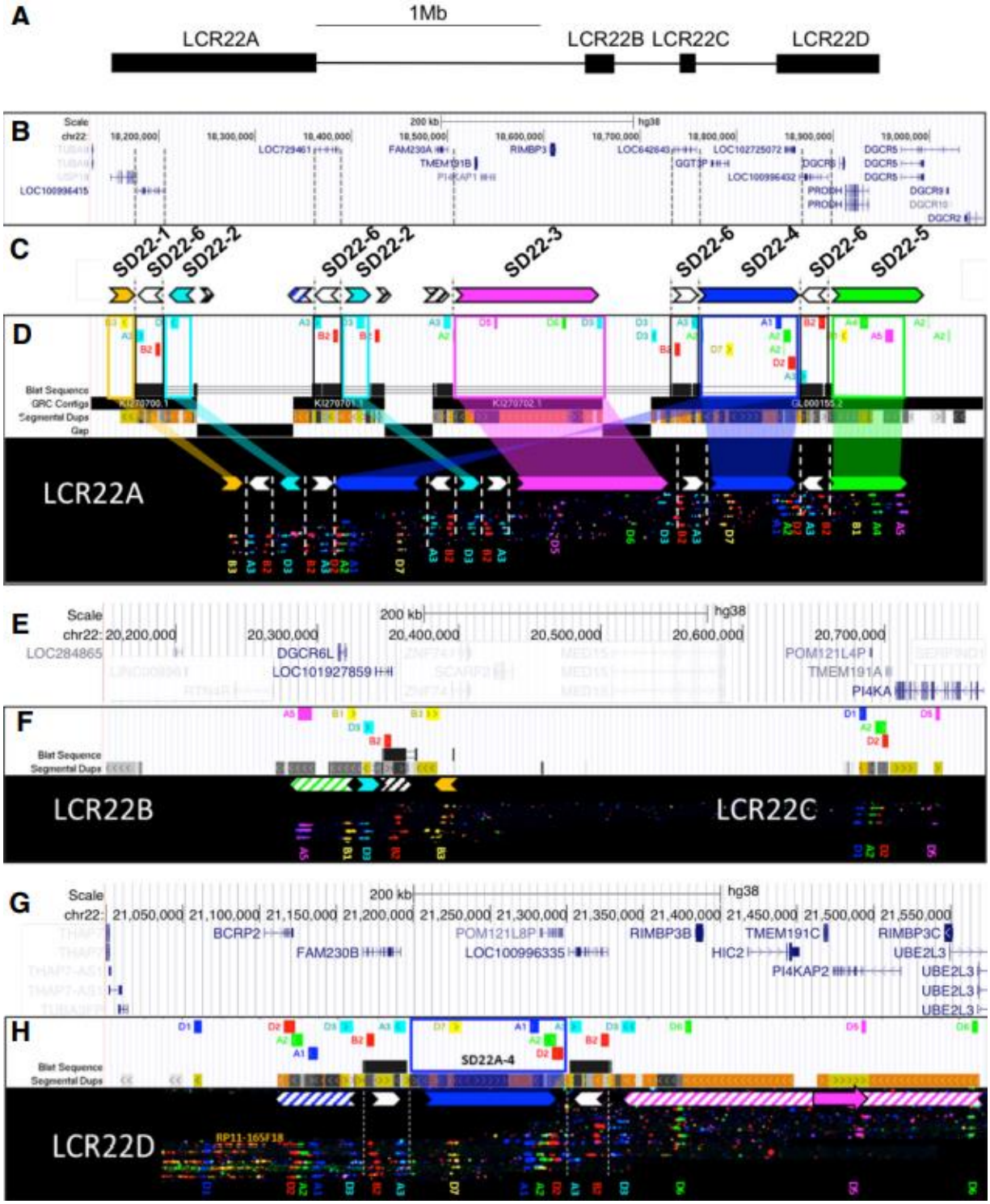


Figure 3.1: In silico hg38 fiber-FISH probe positions compared to duplicon composition of LCR22s. *(A) Schematic overview of the LCR22s in chromosome 22q11.2. (B) RefSeq-curated gene set overlapping with the LCR22s. (C) Duplicon decomposition of the hg38 structure of LCR22-A. Duplicons were deduced from mapped haplotypes. Filled, colored arrows represent copies of duplicons and hatched arrows represent partial copies of duplicons of the same color. (D) UCSC Genome Browser hg38 reference assembly tracks of Segmental Dups, GRC contigs, gap positions, and fiber-FISH probe BLAT positions (white panel). Positions of the latter are aligned with recordings of fiber-FISH patterns in LCR22-A (black bar). Decomposition of one LCR22-A haplotype to duplicons is illustrated using colored arrows. The arrow direction represents inverted or direct orientation. Larger duplicons are flanked by copies of SD22-6 (white arrows). Probe identifiers are indicated below the fiber pattern.*

(E) RefSeq annotated genes overlapping with LCR22-B and -C. (F) LCR22-B and -C, fiber-FISH patterns have the same order and distances as those predicted in hg38 and contain partial duplications of LCR22-A duplicons (hatched arrows). (G) RefSeq annotated genes overlapping with LCR22-D. (H) All LCR22-D molecules present at the same centromeric start, overlapping with predicted hg38 probe positions. The first duplicon displays a partial SD22-4 and SD22-2 (hatched blue arrow), followed by a complete SD22-4 flanked by SD22-6 copies (white arrow). The distal end of LCR22-D consists of partial duplications of SD22-3 (hatched magenta arrow). Nested, solid magenta arrow represents probe D5 position variant.

We next used fiber-FISH to further resolve the structure of LCR22s and to obtain a more accurate map of these regions than what is available in the reference genome. We designed fluorescent probes for 14 repeat subunits to visualize their order and the distance between them (**Figure 3.1, Supplementary Table S3.1**). We used long-range PCR to generate the probes, which were labeled with different colors to obtain distinct signals from adjacent subunits (**Figure 3.1D**). We used BAC probes flanking each of the LCR22 repeat clusters as probes to anchor them within unique sequence (**Supplementary Figure S3.1**).

We first assayed the LCR22-specific fiber-FISH probe pattern on DNA fibers generated from haploid cell line CHM1 and HapMap cell line GM12878. These genomes have been well characterized and were included in the Platinum Genome Project (Eberle et al. 2017). The haploid state of CHM1 significantly reduced mapping complexity of the repeat clusters. We hybridized our custom probe set on fibers of CHM1 and GM12878 and detected more than 100 informative fibers, each >200kb in size. We then tiled the clustered fibers, enabling the *de novo* assembly of the subunit order over more than 1Mb.

We compared the assembled subunit patterns to the *in silico* determined subunit positions in hg38. Probe patterns of LCR22-B and LCR22-C were in agreement with those in hg38 for both cell lines (**Figure 3.1E-F**). In contrast, the observed LCR22-A and -D patterns diverged from hg38 to different extents. LCR22-D structure was identical in CHM1 and GM12878. This structure mostly matched hg38 (**Figure 3.1G-H**), except for the position of a single probe, D5. The *de novo* assembled LCR22-A allele in CHM1 was larger than the one predicted by hg38 (**Figure 3.1B-D, Supplementary Figure S3.2**). Similarly, GM12878 also had two distinct LCR22-A alleles, both of which were different from the hg38 predicted allele (**Supplementary Figure S3.3**). Based on the distance between probe signals, LCR22-A alleles in GM12878 were estimated to be ~1.20Mb and ~0.65Mb, respectively. The CHM1 allele was also estimated to be ~1.20Mb, however, it differed in its composition from the GM12878 allele of the same size.

Because the first three observed LCR22-A assemblies differed substantially from the reference and from one another, we wondered whether those alleles were exceptional. We assembled fiber-FISH patterns in 33 additional cell lines. As observed previously, the LCR22-B and -C patterns were identical and in agreement with hg38 in all individuals tested. However, we assembled 44 additional LCR22-A alleles that showed a surprising level of variation (**Figure 3.2A**). Based on the fiber-FISH assembly of a total of 44 LCR22-A alleles in 33 samples, we observed 26 distinct haplotypes varying in length from ~300kb to >2Mb

(**Figure 3.2A**). No individual in this subset was homozygous for the structure of LCR22-A. In contrast, LCR22-D displayed less variability. We observed three haplotypes for LCR22-D (**Figure 3.2B**), with allelic variation including different positions of probe D5 and a partial duplication of the blue duplicon (**Figure 3.2B**).

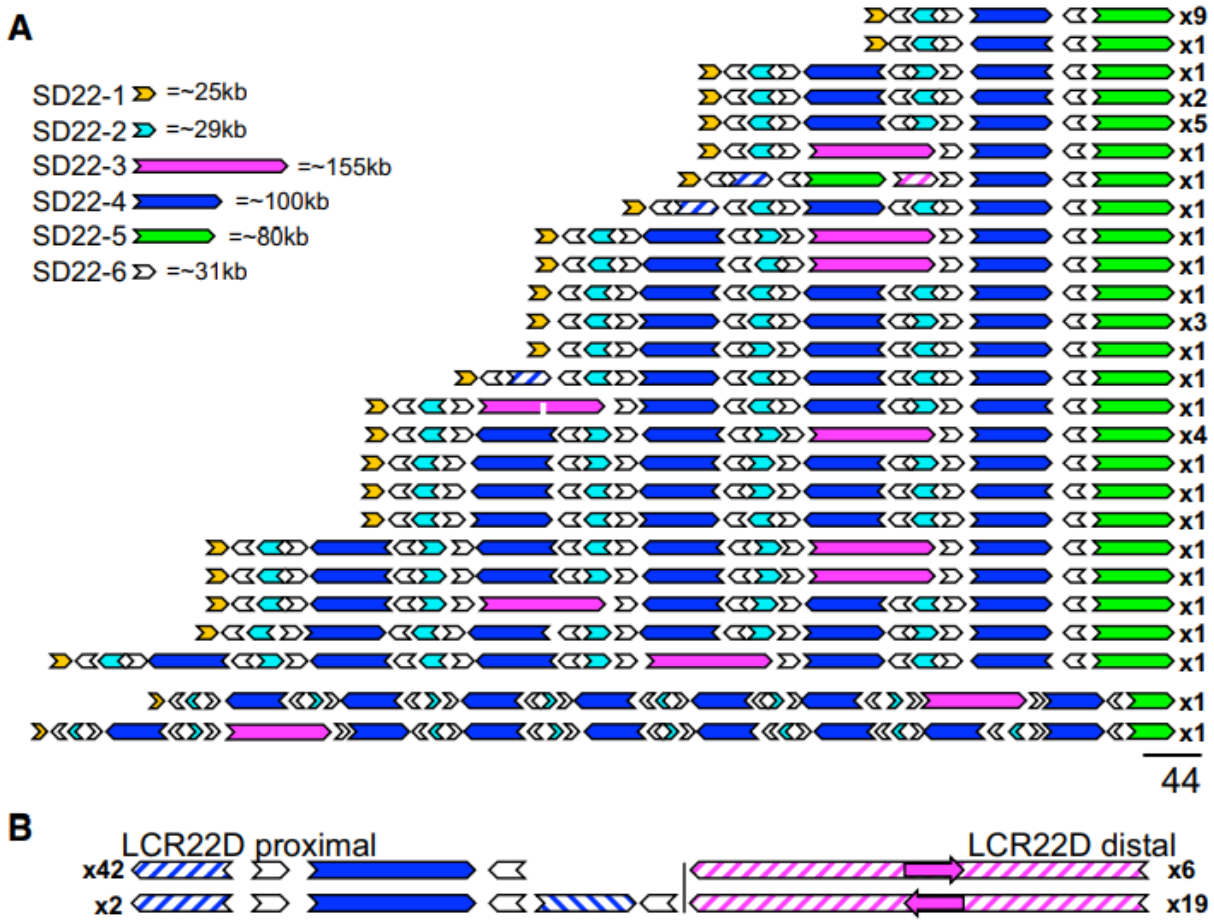


Figure 3.2: Fiber-FISH mapped haplotypes of LCR22-A and -D observed in a cohort of 33 cell lines. *(A) Twenty-six haplotypes observed for LCR22-A. Haplotypes are aligned at the distal unique anchor of LCR22-A. (B) Proximal and distal haplotypes observed for LCR22-D. Filled, colored arrows represent copies of duplicons, and hatched arrows represent partial copies of duplicons of the same color. Size estimates of individual SD22s are shown (upper left). Frequencies of haplotypes are depicted on the right.*

3.2.2 LCR22-A fiber patterns identify core duplicons

Despite the observed scale of variation within LCR22-A, we observed a nonrandom pattern of probe clusters within the mapped haplotypes. We predicted that these probe clusters represent segmental duplications or duplicons within the LCR22s, and the observed differences in LCR22 architecture is driven by the copy number variation of a small set of such duplicons. We visually deduced a minimal set of different probe clusters, which were designated as SD22s and are henceforth referred to as duplicons (**Figure 3.1**).

We identified six probe clusters to define six LCR22 duplicons, designated SD22-1 to SD22-6 (**Figure 3.1C**). All 44 fiber-FISH mapped alleles of LCR22-A presented a conserved proximal and distal end, represented by SD22-1 and SD22-5, respectively (**Figure 3.2A**). In contrast

to the proximal and distal anchors, SD22-2, -3, -4, and -6 were copy number variable among alleles, with SD22-3 being absent in some (**Figure 3.2A**). A majority of the duplicons maintain their structural integrity when comparing various LCR22-A alleles. We rarely observed partial copies of any given duplicon, except in three LCR22-A haplotypes, which had partial copies of SD22-3 and -4 (**Figure 3.2A**).

This analysis revealed that every LCR22-A allele mapped by fiber-FISH was composed of SD22-1 to SD22-6 in different orders, copy numbers, and orientations. We never observed a tandem array of any single duplicon as any two subsequent copies of SD22-2, -3, -4, or -5 were always flanked by a paralog of SD22-6. Moreover, the orientation of SD22-6 relative to its surrounding duplicons was conserved. Although, most alleles have a single copy of SD22-1 and -5 at the proximal and distal end of assembled LCR22-A haplotypes, respectively, we did occasionally observe copy number variants of SD22-5 (**Figure 3.2A**). In a previous study on a cohort of 15,579 normal individuals, Guo et al. (2018) identified a deletion (0.3%) and reciprocal duplication (1.4%) embedded in LCR22-A. Of the 33 cell lines we analyzed by fiber-FISH, one was from an individual carrying the duplication embedded within LCR22-A (**Figure 3.2A**).

3.2.3 Sequence and gene content of LCR22-A duplicons

To determine the sequence content of each of the duplicons within LCR22-A, we compared the observed probe patterns of the duplicons to the expected positions in the reference genome (**Figure 3.1B-D**). *In silico* fiber-FISH probe patterns of reference contigs KI270701.1, KI270702.1, and the proximal 150kb of reference contig GL000155.2 individually matched duplicons SD22-2, -3, and -4, respectively (**Figure 3.1C-D**). SD22-2 did not contain any genes whereas SD22-3 contained TMEM191B, RIMBP3, and PI4KAP1, and SD22-4 contained GGT3P. Further, SD22-1 contained USP18 and SD22-5 contained PRODH, DGCR5, and DGCR6. SD22-6 corresponded to a ~31kb repeat in hg38, which was present five times in the reference LCR22-A with sequence similarities of 97% and higher (**Figure 3.1C-D**, BLAT track and white arrows). The SD22-6 hg38 sequence contains paralogs of a lincRNA with sequence similarity to FAM230C. Each of these paralogs contains copies of the translocation breakpoint type A (TBTA, AB261997.1), which consists of an unstable palindromic AT-rich repeat (PATRR). Thus, the different alleles of LCR22-A contained a different copy number of the genes and other sequences based on the respective copy number of the duplicons.

3.2.4 Bionano optical mapping confirms fiber-FISH assemblies

To evaluate the fiber-FISH assemblies with an orthogonal technology, we performed Bionano optical mapping assays on a total of eight cell lines: the haploid cell line CHM1, GM12878, and two trios, containing 22q11.2DS patients and their parents. Because a certain degree of paralogous variation between segmental duplications is missed by fiber-

FISH, we expected some mismatch when comparing individual label sites. However, other than this expected paralogous variation, *de novo* assembled LCR22 duplicon order and orientation should be consistent between both data sets. We compared the fiber-FISH and Bionano results by first converting the fiber-FISH duplicon and orientation information into sequences, stitching together the duplicon sequences from the reference genome. We then *in silico* labeled these sequences to convert them to Bionano optical map format and then compared them to the observed Bionano assemblies in the same individual (**Supplementary Figures S3.2-3**, blue bar). For CHM1 and GM12878, we generated Bionano data using the Direct Label and Stain (DLS) enzyme. After examination of single molecules from these samples at both LCR22-A and -D, we generated a list of partial haplotypes with strong single molecule support for each sample. We observed that the cluster of SD22-4 and its flanking SD22-6 duplicons contained five DLS labels that were polymorphic between paralogs. These polymorphisms allowed us to stitch the partial haplotypes into end-to-end haplotypes of LCR22-A and -D.

We aligned the observed optical maps from CHM1 and GM12878 to those converted from the fiber-FISH results (**Supplementary Figures S3.2-3**). All alignments showed strong agreement between *in silico* and observed optical maps after accounting for the expected paralogous variation. The copy number, order, and orientation of detected duplicons were identical between the two techniques. Overall, the Bionano optical maps confirmed the fiber-FISH assemblies with minimal discrepancies.

3.2.5 Bionano optical mapping reveals population-specific LCR22 variation

To determine the prevalence of different variants in LCR22-A and -D, both within and between populations, we mapped the variability, by Bionano optical mapping using the Nt.BspQI enzyme, in a cohort of 154 phenotypically normal individuals from 26 populations spanning five superpopulations: African , American , East Asian , European , and South Asian (Levy-Sakin et al. 2019).

Structural variation at LCR22-A

We generated distinct LCR22-A configurations by collapsing optical map assembled contigs that overlapped LCR22-A and had single molecule support (**Figure 3.3A-D**). Not all assembled contigs were able to span the entire LCR22-A region from end to end, due to factors such as insufficient molecular coverage, short molecule lengths, and/or longer LCR22-A haplotype lengths. Additionally, the presence of two Nt.BspQI nicking sites, on opposite strands, in close proximity to one another near the beginning of SD22-3 created a fragile site on the DNA that consistently interrupted molecules in that location. Although this catalog of LCR22-A haplotypes is likely not comprehensive, it captures a substantial amount of large-scale structural variation at this locus.

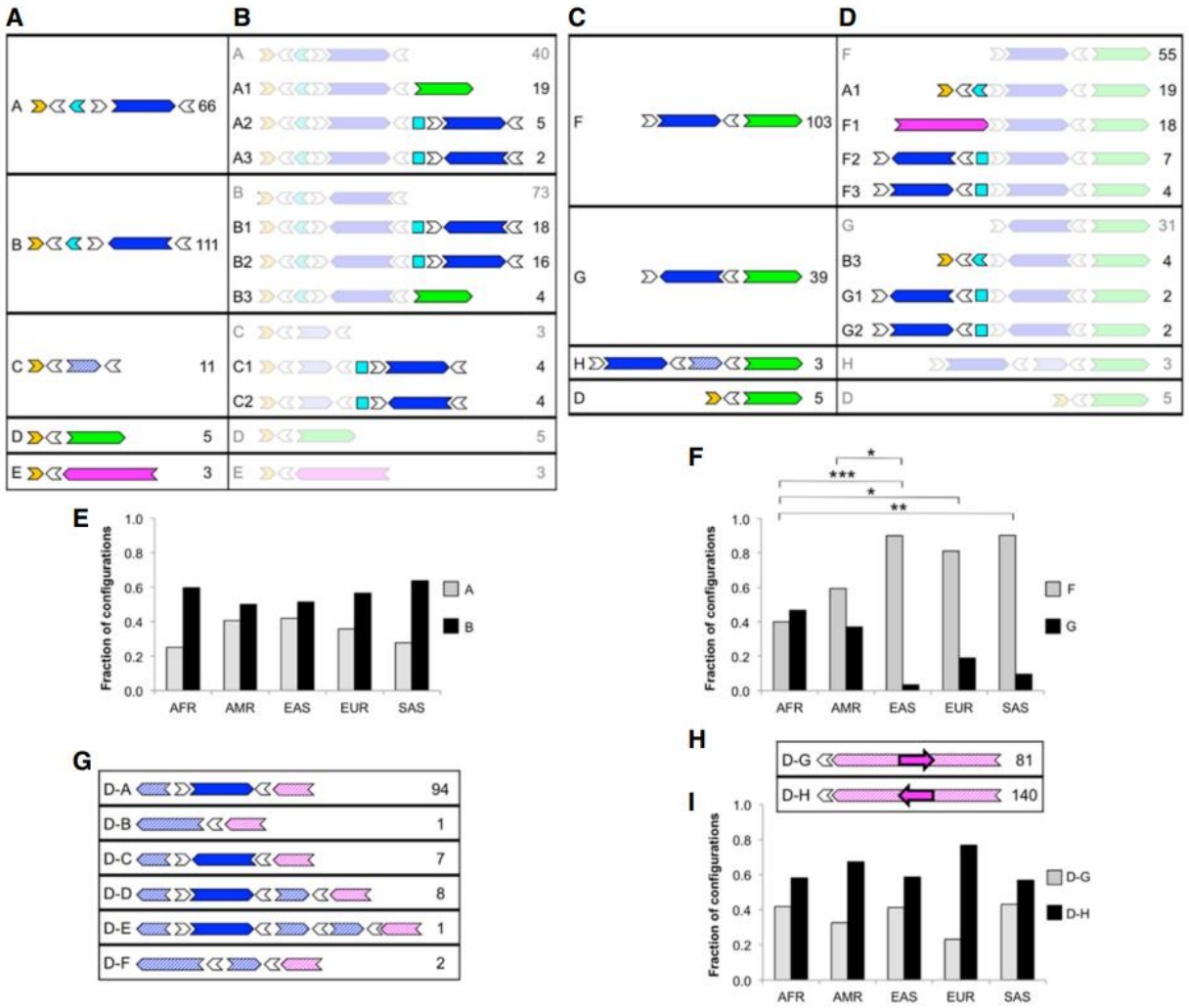


Figure 3.3: LCR22-A and -D configurations across a diverse control data set observed using Bionano optical mapping*. Diagrams depict order and orientation of observed duplicons as defined in Figure 1. Minimal (A) and extended (B) configurations anchored in unique sequences upstream of LCR22-A. Minimal (C) and extended (D) configurations anchored in unique sequence downstream from LCR22-A. (E) Observed occurrences for upstream-anchored LCR22-A configurations A and B in different populations. (F) Observed occurrences of downstream-anchored LCR22-A configurations F and G in different populations.(G) Configurations anchored in unique sequence upstream of LCR22-D. (H) Configurations anchored in unique sequence downstream from LCR22-D. (I) Observed occurrences of downstream-anchored LCR22-D configurations D-G and D-H in different populations. For each configuration in A-D and G-H, the ID is on the left, and the number of times that configuration was observed in the data set is on the right. Duplicons for which an orientation could not be determined are represented as squares. (AFR) African; (AMR) American; (EAS) East Asian; (EUR) European; (SAS) South Asian; (*) P<0.05, (**) P<0.01; (***) P<0.001, Fisher's exact test, adjusted. Pairs of populations without asterisks in E, F, and I were not significantly different at P<0.05.*

We identified a total of 16 non-redundant partial and complete LCR22-A configurations in the set of 154 individuals (**Figure 3.3**). As seen in the fiber-FISH results, the majority of the variation involved variable copy number and orientation of SD22-4 (**Figure 3.3A-D**). One notable configuration, not observed in the fiber-FISH data set, harbored a deletion of almost the entire locus, with a minimal composition of SD22-1 to inverted SD22-6 to SD22-5 (**Figure 3.3**, configuration D).

We next evaluated the prevalence of each configuration within the data set and observed clear differences in prevalence between different configurations (**Figure 3.3**). For minimal

configurations anchored upstream of LCR22-A (**Figure 3.3A**), the most common duplicon to follow the initial cluster of SD22-1, SD22-2, and flanking SD22-6 copies was an inverted copy of SD22-4 (**Figure 3.3A**, configuration B), which accounted for 111/196 (57%) and 25/44 (57%) of the observed configurations in this group, in optical map and fiber-FISH datasets, respectively. The next most common configuration was a copy of SD22-4 in the direct orientation (**Figure 3.3A**, configuration A), which accounted for 66/196 (34%) of the observed configurations in the optical map data and 14/44 (32%) in the fiber-FISH data. Configuration A is also the structure corresponding to the beginning of both the hg19 and hg38 reference haplotypes. These results indicated that neither of the two most recent reference genomes represented the major allele at this locus.

Among the minimal configurations anchored downstream from LCR22-A (**Figure 3.3C**), a direct copy of SD22-4 preceded SD22-5 in 103/150 (69%) observed configurations (**Figure 3.3C**, configuration F), which is consistent with the reference genomes. The next most common configuration (**Figure 3.3C**, configuration G) accounted for 39/150 (26%) of observed configurations, had SD22-5 preceded by an inverted SD22-4. In the fiber-FISH dataset, 39/44 (89%) of chromosomes displayed configuration F, but only 5/44 (11%) showed SD22-5 preceded by an inverted SD22-4. The fiber-FISH samples were taken exclusively from individuals of European descent, and the distal portion of LCR22-A differed significantly among various ethnicities (**Figure 3.3F**). Among European samples in the Bionano dataset, configuration F accounted for 26/33 (79%) of observed configurations, whereas configuration G accounted for the remaining 7/33 (21%), values which were more concordant with the fiber-FISH results in individuals of European descent. For both groups of minimal configurations, anchored upstream of or downstream from LCR22-A, the shortest end-to-end haplotype containing SD22-1 to inverted SD22-6 to SD22-5 (configuration D) comprised a small minority of cases, accounting for ~3% of the observed configurations.

The extended configuration (**Figure 3.3B,D**) was observed in fewer samples because not all single molecules that matched the minimal configurations were long enough to extend into an additional duplicon. Nevertheless, this smaller dataset illustrated several distinctive patterns. The three most common upstream configurations observed, each representing ~20%-24% of the extended upstream alleles, followed the anchoring SD22-1 and SD22-2 with (1) direct SD22-4 and SD22-5, that is, the hg19 haplotype (**Figure 3.3B**, configuration A1); (2) tandem copies of indirect SD22-4 (configuration B1); and (3) an indirect and then a direct copy of SD22-4 (configuration B2). The remaining 34% of cases comprised seven configurations, each accounting for 3%-6% of observed configurations.

Among the downstream extended configurations (**Figure 3.3D**), the hg19 haplotype (configuration A1) and configuration F1, which matched the distal end of the hg38 haplotype, that is, SD22-3, SD22-4, and SD22-5, were the two predominant configurations observed – each representing 28%-30% of the observed configurations. Although the latter

configuration was also common in the fiber-FISH data (20% of alleles) (**Figure 3.2A**), the nine end-to-end haplotypes containing configuration F1 showed only one match to the exact hg38 structure, instead representing a total of six haplotypes of varying lengths and copy number of SD22-4, suggesting that this configuration class in the optical map data is likely to also represent a wide range of end-to-end haplotypes.

We next wanted to determine whether the observed configurations differed by population. Configurations A, B, F, and G (**Figure 3.3A, C**) were the only observed configurations that provided an adequate sample size for this population-based analysis. We observed substantial differences between the superpopulations for configurations F and G ($P<0.05$, Fisher's exact test), with the largest difference occurring between the African and East Asian populations (**Figure 3.3F**). Thus, at the distal end of LCR22-A, SD22-4 in the direct orientation (configuration F) was more common overall, but it accounted for only 16/41 (39%) of the observed configurations in African, compared to 24/27 (89%) of the configurations observed in East Asian (**Figure 3.3F**). At the proximal end of LCR22-A, SD22-4 in an inverted orientation (configuration B) was observed more frequently than SD22-4 in the direct orientation (configuration A) in every population (**Figure 3.3E**).

Structural variation at LCR22-D

LCR22-D was substantially less polymorphic and complex than LCR22-A, but it nonetheless harbored some large-scale structural variation. Following the same procedure as above, we compiled configurations for LCR22-D from the optical map data from 154 individuals. We observed six upstream-anchored configurations (D-A to -F), four of which involved the paralog of SD22-4 that is present in the proximal half of LCR22-D (**Figure 3.3G**). In the downstream-anchored half of LCR22-D, we only observed a 64-kb inversion (**Figure 3.3H**). Because these two regions were distant from one another, we analyzed them separately to minimize the length of the molecules required to identify each configuration. Among the upstream-anchored configurations, the most predominant was configuration D-A, which represents the configuration observed in the reference genome, accounting for 94/113 (83%) of all observed configurations (compared to 95% in the fiber-FISH data) (**Figure 3.2B**). The next most common configurations were a full inversion or partial duplication of SD22-4 (configurations D-C and D-D), accounting for 7/113 (6%) and 8/113 (7%) of the observed configurations, respectively (**Figure 3.3G**). In the fiber-FISH data, we observed a partial duplication (D-D) in 2/44 (4.5%) alleles, although we did not see any inversions of SD22-4. We detected no population-based differences among these upstream-anchored LCR22-D configurations.

Within the observed downstream-anchored configurations of LCR22-D, configuration D-H accounted for 140/221 (63%) of the observed configurations. Thus, we observed configuration D-H, with an inversion, more frequently than configuration D-G, which represents the configuration observed in the reference genome. Configuration D-H

accounted for 30/39 (77%) of the observed configurations in Europeans, which was consistent with the fiber-FISH data from European samples in which 19/25 (76%) individuals carried the inverted D-H configuration (**Figure 3.2B**). In the optical map cohort, we observed the D-H variant more frequently in all five superpopulations, with no statistically significant differences between populations (**Figure 3.3I**).

Thus, Bionano optical mapping not only confirmed the fiber-FISH assemblies, but also extended these findings demonstrating large-scale structural variation in LCR22-A and LCR22-D in a cohort of 154 individuals from five superpopulations.

3.3 Discussion

The LCR22 reference sequences have contained gaps since the first human genome assembly was released (Cole et al. 2008; Schneider et al. 2017). Although whole-genome short-read sequencing is now routine, alignment of short sequencing reads to the human reference sequence generally fails to detect and assemble large structural variants and repetitive regions like the LCR22s. Because of the length of the duplications, even assemblies using longer-range technologies like PacBio and 10X Genomics linked reads have been unable to assemble these regions (Berlin et al. 2015; Weisenfeld et al. 2017). To resolve these gaps, we combined fiber-FISH and Bionano optical mapping, and show that an astounding level of inter-individual variability of LCR22-A, and to a lesser extent LCR22-D, has likely impeded the assembly of a complete reference sequence for these LCR22s. These maps revealed at least 25 different alleles of LCR22-A and six variants of LCR22-D. LCR22-A alleles ranged in size from ~250kb to ~2000kb. Most of these alleles could be decomposed into six core duplicons (SD22-1 to SD22-6), with duplicons presenting in different orientations and at variable positions within the LCR22. The most frequent LCR22-A haplotype had the following structure: SD22-1, SD22-6 (inverted orientation), SD22-4 (direct orientation), SD22-6 (inverted orientation), SD22-5, which made up ~25% of all mapped alleles. Its structure is very similar to one of the first LCR22-A sequences proposed (Shaikh et al. 2000), a haplotype which was presented in hg19. Thus far, only one smaller haplotype was detected, in which SD22-1 was directly followed by SD22-6 (indirect orientation) and SD22-5. This might indicate the requirement of a minimal haplotype to maintain a viable gene dosage.

None of the 19 normal, diploid parents in the cohort were homozygous for LCR22-A, which suggests the existence of a high number of different haplotypes in humans. Consequently, any homologous recombination between two (different or identical) alleles of LCR22-A will likely generate a novel allele with a duplicon composition different from the parent-of-origin. Furthermore, configurations of LCR22-A and LCR22-D varied in frequency among populations. Some of these configurations might be more vulnerable to NAHR than others. Consequently, variation in the frequency of the 22q11.2DS among populations (Botto et al. 2003; McDonald-McGinn et al. 2005) may result from frequency differences of LCR22-A and

LCR22-D configurations and their respective vulnerability to NAHR. Because our sample size is relatively small, we expect that the alleles we observed are likely to be a small subset of all haplotypes that may exist in the population.

Studies on genome-wide LCR diversity have identified numerous LCR clusters, mainly in pericentromeric and subtelomeric regions (Goidts et al. 2006b). However, none of those come close to the level of complexity and the number of haplotypes found in the LCR22s. A few studies using either whole genome sequencing read depth-based predictions, digital droplet PCR, or custom BAC arrays have revealed copy number variability between individuals within regions containing LCRs (Sudmant et al. 2010, 2015; Handsaker et al. 2015; Dennis and Eichler 2016). Eight distinct haplotypes have been described for the LCR clusters on chromosome 17q21.31, ranging in size from 1.08Mb to 1.49Mb (Steinberg et al. 2012). Similarly, the 1000 Genomes Project observed copy number variation ranging from 2 to 11 copies of a ~900kb region (Chr15:20,353,991-27,802,370) in 15q11-q12 (Siva 2008; Sudmant et al. 2010). Such repeat expansions have mainly been found to be human-specific when compared to their orthologs in great ape genomes (Goidts et al. 2006a). Moreover, significant variation between different human populations suggests that these genomic rearrangements happened recently or are still ongoing (Dennis and Eichler 2016). However, a majority of these studies are based on short-read whole-genome sequencing data, which are not as reliable for determining true copy number and complex architecture of regions containing LCRs.

Although 22q11.2DS is the most frequent microdeletion syndrome, the underlying cause for the wide spectrum and variability of phenotypes observed has not been fully elucidated. Variation of genes embedded in copy number variable regions like LCR22s has been so far ignored. We suggest that copy number variable genes embedded in the LCR22s could explain some of the phenotypic variability observed in individuals with the 22q11.2DS and human in general. SD22-3 contains at least two known active genes (TMEM191B and RIMBP3) (**Figure 3.1B**). Not every allele of LCR22-A features this duplicon, but neither is it observed to be present in more than one copy. Both TMEM191B and RIMBP3 have paralogs in LCR22-D (TMEM191C, RIMBP3B, and -C). The genes PRODH, DGCR5, and DGCR6 reside in SD22-5 (Guo et al. 2018), which was retained in even the smallest mapped allele of LCR22-A. In all mapped individuals with the typical 22q11.2 deletion, this duplicon is deleted, confirming previous observations of its hemizygosity in most patients (Liu et al. 2002; Michaelovsky et al. 2012; Guo et al. 2016, 2018). Additionally, the presence of pseudogenes (PI4KAP1 and P2, GGT3P, DGCR6L, BCRP2, GGT2) and lncRNAs (FAM230A, FAM230B, and at least seven non-characterized paralogs) in different copy numbers could influence gene expression. However, in the absence of an unambiguous reference sequence, it remains challenging to investigate the gene activity of each duplicon. Moreover, the observed size differences of LCR22-A might exert a spatial effect on chromatin looping in the cell, thereby altering topologically associated domains (Bonev and Cavalli 2016). The

phenotypic effect of variable repeat architecture could be minor for intact alleles but could alter gene expression completely when the LCR22s are rearranged.

In summary, high-resolution optical mapping has allowed us to reveal an extraordinary level of variability within LCR22s. Our map of this genomic region is, to date, the most comprehensive for the LCR22s in the human genome reference sequence. Further, this map provides a framework for the alignment of both short and long read sequences which will ultimately close the remaining reference gaps and enable sequence-based analysis of the LCR22s. Understanding the LCR22 variation will shed light on the mechanisms leading to 22q11.2 rearrangements and the different frequencies of the variation among populations. This knowledge will likely guide future prenatal counseling and testing for 22q11.2-related disorders. It seems plausible that the region influences important human traits, considering the region encompasses nine known active genes and comprises 54 different RNAs. Thus, it is likely that the LCR22 variability has phenotypic consequences, which may play a role in phenotypic variability in the 22q11.2DS and affect other traits in the normal population. The ability to visualize and reconstruct complete and intact LCR22 haplotypes will greatly enhance our ability to start unraveling these important correlations.

3.4 Materials & Methods

Patients and EBV cell lines

Patients with the 22q11.2DS were diagnosed using either a FISH assay with TUPLE1/ARSA probes (Abbot Molecular), the multiplex ligation-dependent probe amplification (MLPA) Salsa P250 DiGeorge diagnostic probe kit (MRC-Holland), or with the CytoSure Constitutional v3 (4x180k) (OGT). All individuals in the study were informed of the project's outlines and gave written consent for their EBV cell lines and DNA to be used for sequencing and genotyping purposes. The study was approved by the Medical Ethics committee of the University hospital/KU Leuven (S52418), the Institutional Review Board approved research protocol (COMIRB 07-0386) at the University of Colorado Denver, School of Medicine, and the Children's Hospital of Philadelphia under Institutional Review Board (IRB) protocol 07-005352. Fiber-FISH mapping was performed on Epstein-Barr virus transformed lymphoblastoid cell lines from peripheral blood from probands and their parents. Eleven patients were recruited during routine consultations in the hospital of Leuven, one at the Children's Hospital of Philadelphia, and two at the Albert Einstein College of Medicine (IRB: 1999-201-047). HapMap control cell lines were obtained from the Coriell Cell Repository and cultured according to standard protocols.

In silico characterization of repeat subunits in LCR22s

All segmental duplication track positions were downloaded in BED format from UCSC in the region Chr22:18,000,000-25,500,000 (hg38), including paralogous LCRs located elsewhere in the genome. These were merged with BEDTools v2.17.0 (Quinlan and Hall 2010), and

sequences were retrieved with the UCSC Table Browser (Kent et al. 2002). These were then self-aligned using BLASTN v2.2.28+ (Altschul et al. 1990) and filtered for reciprocal BLAST hits, alignments <100bp, and alignments <99% identity. If multiple queries aligned to the same subject segmental duplication at different positions, the segmental duplication was split into multiple units. Unit positions were converted to BED format, self-aligned again, and similarly filtered. Clusters of units aligning to each other were each considered a subunit family.

BAC DNA, long-range PCR probe design, and labeling

Using the subunit sequences library, 14 fluorescent probes were designed (**Supplementary Table S3.1**). For each of these 14 subunits, long-range PCR primer pairs were designed, producing amplicons between 2946bp and 9794bp (**Supplementary Table S3.1**). PCR reactions were performed with the TAKARA LAv2 kit (Takara Bio) using the standard gDNA protocol. Template gDNA was extracted from the same cell line for all reactions, to reduce amplicon variation between batches.

BAC clones were obtained from BacPac Resources (CHORI) as *E. coli* stab cultures, which were grown according to recommendations. BAC DNA was extracted using the Nucleobond Xtra BAC kit (Macherey-Nagel).

Subunit PCR Amplicons and BAC DNA were purified and antibody-labeled by random prime amplification (BioPrime DNA Labeling System; Invitrogen). An indirect detection system with primary labels Biotin-dUTP, Digoxigenin-dUTP, and Fluorescein-dUTP was used. The use of three labels allowed production of six detectable probe colors: three of each label separately and three of each pairwise combination.

DNA combing, FISH, and fiber pattern assembly

DNA fibers were stretched using the Genomic Vision extraction kit and combing system for a total of 33 human cell lines using standard methodology. Cultured EBV cells were embedded in an agarose plug and subsequently lysed and washed. Next, the agarose was dissolved and long DNA fragments were resuspended. Using an automated combing system, DNA fibers were consistently stretched on a coated glass coverslip. YOYO-1 staining and scanning allowed visualization and evaluation of DNA fibers at this step. Coverslips with combed DNA were hybridized with the designed probe pattern and washed using the manufacturer's standard protocol. Probes were detected by indirect labeling with BV480 Streptavidin (pseudocolored red; BD Biosciences; 564876), Cy3 IgG Fraction Monoclonal Mouse Anti-Fluorescein (pseudocolored green; Jackson Immunoresearch; 200-162-037), and Alexa Fluor 647 IgG Fraction Monoclonal Mouse Anti-Digoxigenin (pseudocolored blue; Jackson Immunoresearch; 200-602-156). Probe mixes produced pseudocolors cyan, magenta, and yellow. Slides with labeled DNA were mounted in the provided scanner

adapters and scanned at three excitation channels on a customized automated fluorescence microscope (Genomic Vision).

Images were compiled to one complete slide recording and visualized in FiberStudio (Genomic Vision). Slides were manually screened, and fiber signals were cropped to single image files. Individual images were visually aligned based on matching colors and distances between different probes. Fibers were tiled to complete alleles for LCR22-A, -B, -C, and -D, and compared to hg38 probe positions in the UCSC Genome Browser. Chimeric fiber patterns and false positive signals caused by noise were eliminated by filtering for overlapping patterns identical in color sequence and spacing.

Assembly of artificial LCR22 reference sequences

To confirm the fiber-FISH assemblies, Bionano assays were performed on an overlapping cohort of seven individuals. To compare results from the two methods, fiber-FISH results were converted *in silico* into the optical map format. Using the hg38 reference genome sequences of SD22-1 to SD22-6 and LCR22-D, the sequence of each allele was predicted based on the orientation and copy number of subunits detected in the fiber-FISH assemblies. Those sequences were then *in silico* labeled at recognition sites of the enzyme used for Bionano optical mapping, generating CMAP data files for all LCR22 repeats.

Bionano high molecular weight DNA extraction and labeling

High-molecular weight DNA was extracted and processed for Bionano genome mapping using standard methods and protocols provided by the vendor (Bionano Genomics). Cells were embedded in thin low-melting-point agarose plugs (CHEF Genomic DNA Plug Kit, Bio-Rad). The agarose plugs were incubated with Proteinase K at 50°C overnight. The plugs were washed and then solubilized with GELase (Epicentre). The purified DNA was subjected to 45min of drop-dialysis, allowed to homogenize at room temperature overnight. DNA quality was assessed using pulsed-field gel electrophoresis.

The DNA was labeled using the Bionano Prep Early Access Direct Labeling and Staining (DLS) Kit (Bionano Genomics). DLS labels DNA using an epigenetic mark rather than by introducing single-strand nicks. 750ng of purified genomic DNA was labeled by incubating with DL-Green dye and DLE-1 Enzyme in DLE-1 Buffer for 2 hours at 37°C, followed by heat inactivation of the enzyme for 20 min at 70°C. The labeled DNA was treated with Proteinase K at 50°C for 1 hour, and excess DL-Green dye was removed by membrane adsorption. The labeled DNA was stained with an intercalating dye and left to stand at room temperature for at least 2 hours. The DNA was loaded onto the Bionano Genomics Saphyr Chip and linearized and visualized using the Saphyr system. The DNA backbone length and locations of fluorescent labels along each molecule were detected using Saphyr's image detection software. Single-molecule maps were assembled *de novo* into genome maps using Bionano Solve with the default settings.

Detection of structural variation within LCR22s

Structural variation in the LCR22s was evaluated in Bionano genome map data labeled using the Nt.BspQI nickase enzyme from 154 individuals representing 26 diverse populations from five superpopulations (Levy-Sakin et al. 2019). Assembled contigs mapping to LCR22s were realigned to chromosome 22 using RefAligner from Bionano Solve 3.1.

LCR22 haplotype identification from Bionano data

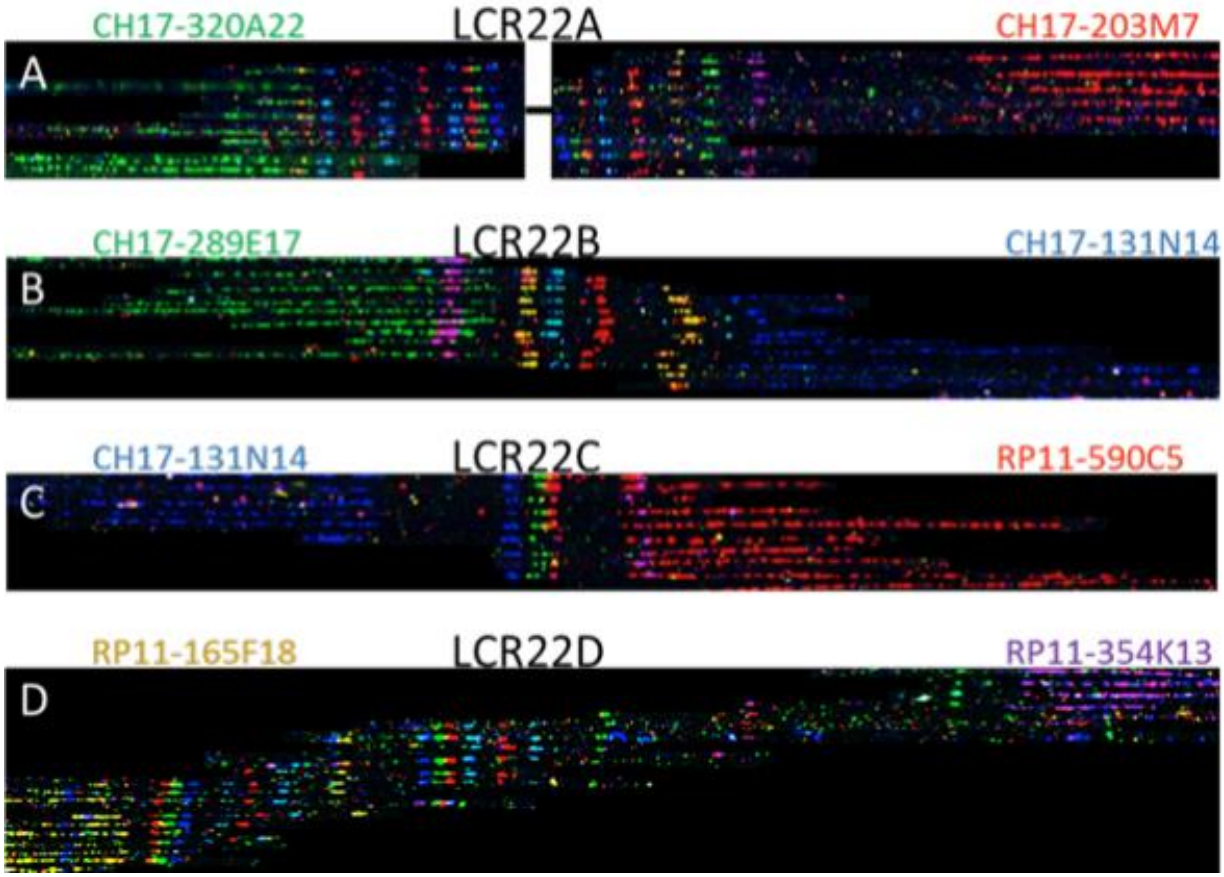
A catalog of configurations for each locus was generated by compiling the configurations observed in the assembled Bionano contigs from the full data set derived from normal individuals and verifying that each entry was supported by single molecules in at least one sample. Configurations were first grouped into categories in which the members were mutually exclusive, so that longer configurations would be analyzed separately from those that were subsets of them. Using this approach, a "minimal" set of configurations that were anchored upstream of or downstream from the repetitive region were constructed for LCR22-A (**Figure 3.3A, C**). An "extended" set was also created that expanded on the minimal configurations, where available (**Figure 3.3B, D**). LCR22-D contained a variable proximal region, as well as a distal region that contained a single structural variant. These two regions were analyzed separately (**Figure 3.3G-I**).

A package called OMGenSV was used to genotype each set of configurations in all the samples. *In silico* labeled representations of each configuration were created in Bionano CMAP format using OMGenSV's get_cmap_subsets.py and add_cmap_files.py scripts to combine 1Mb of flanking unique region from the reference chromosome with representative assembled contigs observed in the normal population-based samples. For all configurations in a given group, their CMAP representations were kept as consistent between one another as possible, that is, containing the same flanking areas. For each grouped set of configurations described above, single molecules from each sample were used to determine which configuration(s) the sample contained. Each observed configuration in a given sample was counted once, and the overall prevalence of any configuration was calculated by dividing the number of times that particular configuration was observed by the total number of all configurations observed for that locus in the relevant group.

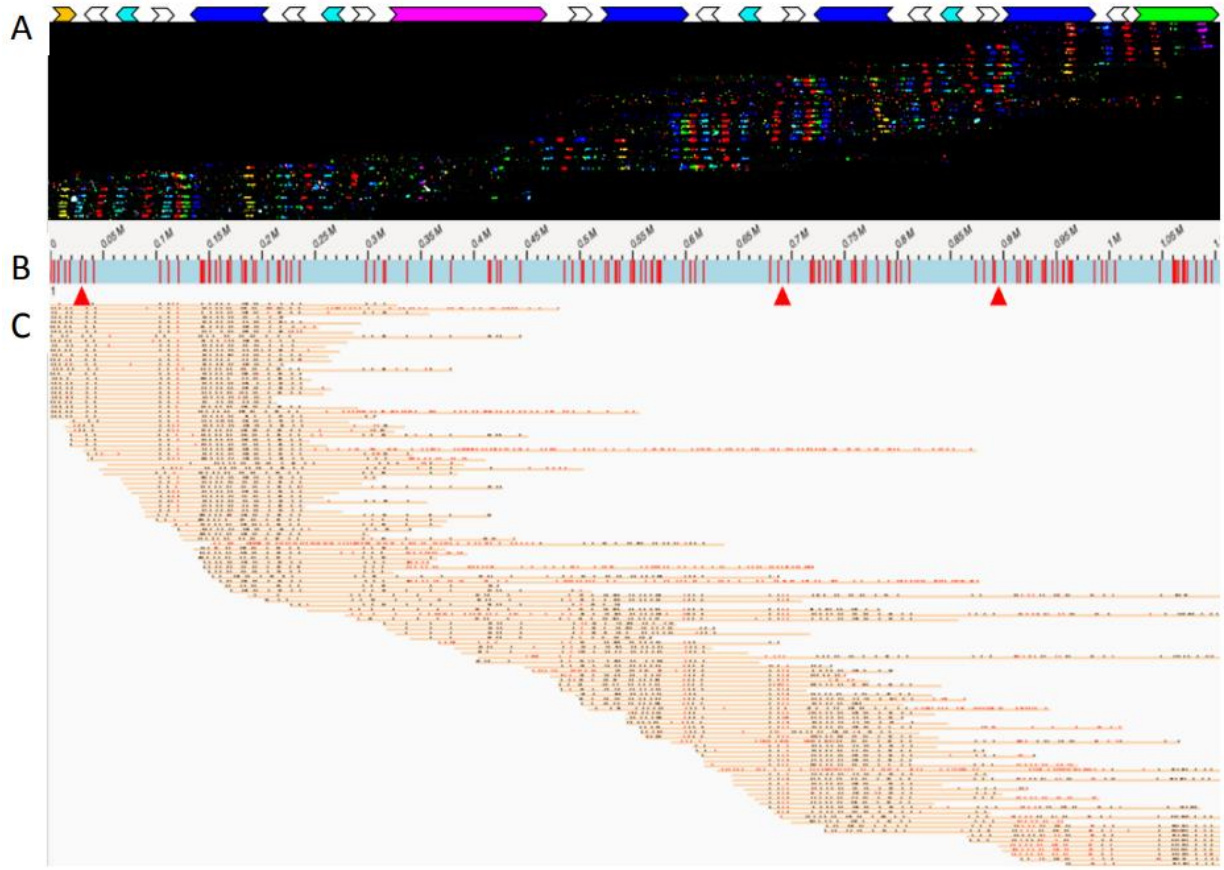
Data access

Bionano optical mapping from this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA418343. Scripts used in this study are available at https://github.com/yuliamostovoy/OMGenSV.

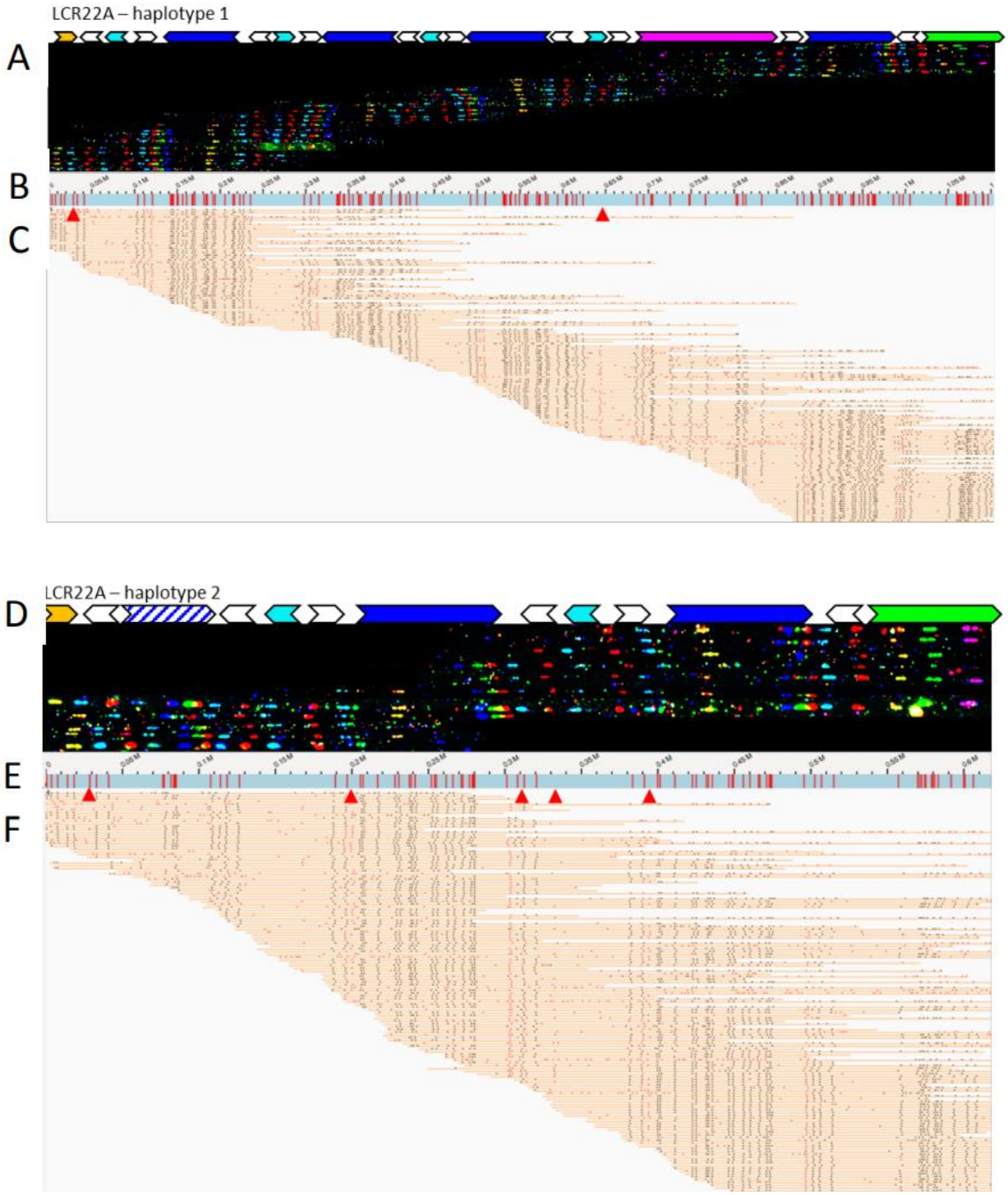
3.5 Supplementary Materials



Supplementary Figure S3.1: *Location of unique BAC probes flanking LCR22-A, -B, -C, and -D used in fiber-FISH to enable identification and directionality of each LCR22. (A) LCR22-A is flanked by CH17-320A22 (green) proximally and CH17-203M7 (red) distally. (B) LCR22-B's proximal flank is CH17-289E17 (green) and the distal flank is CH17-131N14 (blue). (C) LCR22-C is nested between CH17-131N14 (blue) and RP11-590C5 (red). (D) LCR22-D is embedded between RP11-165F18 (yellow) and RP11-354K13 (magenta).*



Supplementary Figure S3.2: *LCR22-A in haploid cell line CHM1. (A) Fiber-FISH assembly and its decomposition to duplicons. (B) Predicted optical map generated by in silico DLS labeling of hg38 duplicon sequences in the order observed by fiber-FISH depicted by blue rectangle with red lines (DLS labels). (C) Optical map aligned single molecules support the predicted haplotype structure with three discrepancies attributed to polymorphisms shown by red arrowheads.*



Supplementary Figure S3.3: *LCR22-A haplotypes in HapMap cell line GM12878. (A-C) LCR22-A allele 1 and (D-F) LCR22-A allele 2. (A,D) Fiber-FISH assembly and its decomposition to duplicons. (B,E) Predicted optical maps generated by in silico DLS labeling of hg38 duplicon sequences in the order observed by fiber-FISH depicted by blue rectangle with red lines (DLS labels). (C,F) Optical map aligned single molecules support the predicted haplotype structure and present no gross discrepancies. Red arrowheads denote mismatches between expected and observed maps.*

***Supplementary Table S3.1**: Long range PCR product coordinates and BAC probes in hg38 and their respective fluorescent labels that were designed to constitute the fiber-FISH pattern for the LCR22s.*

Probe ID	Probe coordinates (hg38)	Color	Probe size (basepairs)	Forward primer	Reverse primer
A1	Chr22:18838945-18844821	Blue	5877	CTTTCTAGATTGACCACTCAGGAGTTAC	CTTAGGAGCTTTTCTTCTAGTTGCAGT
A2	Chr22:18846775-18854471	Green	7697	GTCAAGACAACTCTAAGAAGGTTTTCC	TCTACTGAATCACTTGTCAAGAAGC
A3	Chr22:18863891-18870589	Cyan	6699	GACCCGCTAACTCATTTTATACATC	GTTTATTCCACCTGTCAGTCTCACT
A4	Chr22:18926025-18934677	Green	8653	ATAAAGGTAGTTACCTGGTTCCAAGAC	AGCCCTAAGGTTTCTTGTCTAGATTC
A5	Chr22:18952693-18960789	Magenta	8096	GGATCTACGGAGTCTTCTAAGAGATTT	CATAATAGTTAGAACTGTCTCTCTGGGCTA
B1	Chr22:20313062-20319545	Yellow	6484	GTACAAGTATAGGGCTGTAGGTGCT	AGTGTTCGGAAGAGGTCTCTAAGAT
B2	Chr22:20339592-20344598	Red	5007	CTTTGGAACAAAGCCACAGTAGTAT	AATGAACTTCCACAGTACCTTCTTG
B3	Chr22:20369408-20379200	Yellow	9793	ATACCAAGAGAATCCCCTTACCAGT	CTTTTGTTGGCAGTAGTGTTCCTAT
D1	Chr22:21057588-21062418	Blue	4831	GTGAGAGAGATTAGGATCCCTTTTCAT	TTAATACCACTAGGCTCAACCCGTAT
D2	Chr22:21115365-21122584	Red	7220	TAACTGTGAAGGATTCTTACTCTAGTGTCC	GATCATCTCTCTCTGCCAAAATAACAG
D3	Chr22:21151184-21159924	Cyan	8741	AACAAGCTCTTGACATTCTCTGAGT	AGCATTATTATCACTGCAGCTACCAC
D5	Chr22:21486822-21489767	Magenta	2946	GGTCAGGTACTTCTTATCTGAGAACAT	CAAATAGATGGGAGTGTGTTTCTTC
D6	Chr22:21557969-21561954	Green	3986	CACCATCCCAGTACATAATGACTTC	ATTGGCACCATAGAGCAGTACTAAC
D7	Chr22:21221650-21229324	Yellow	7675	GAGGCCTTAGCAGAAATAACAGAAG	CAGTGCTTTACCACAGAGTGTTTTA
CH17-320A22	Chr22:17943114-18171048	Green	227935	BAC	
CH17-203M7	Chr22:19078955-19323096	Red	244142	BAC	
CH17-289E17	Chr22:20073602-20301293	Green	227692	BAC	
CH17-131N14	Chr22:20398072-20629441	Blue	231370	BAC	
RP11-590C5	Chr22:20726003-20906345	Red	180343	BAC	
RP11-165F18	Chr22:20934046-21099834	Yellow	165789	BAC	
RP11-354K13	Chr22:21588266-21764059	Magenta	175794	BAC	

CHAPTER 4

22q11.2 low copy repeats expanded in the human lineage

Lisanne Vervoort[1], Nicolas Dierckxsens[1], Zjef Pereboom[2,3], Oronzo Capozzi[4], Mariano Rocchi[4], Tamim H. Shaikh[5], and Joris R. Vermeesch[1]

[1] Department of Human Genetics, KU Leuven, Leuven, Belgium

[2] Centre for Research and Conservation, Royal Zoological Society of Antwerp, Antwerp, Belgium

[3] Evolutionary Ecology Group, Department of Biology, Antwerp University, Antwerp, Belgium

[4] Department of Biology, University of Bari, Bari, Italy

[5] Section of Genetics and Metabolism, Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO, United States

Status: Published in *Frontiers in Genetics*

Abstract

Segmental duplications or low copy repeats constitute duplicated regions interspersed in the human genome, currently neglected in standard analyses due to their extreme complexity. Recent functional studies have indicated the potential of genes within LCRs in synaptogenesis, neuronal migration, and neocortical expansion in the human lineage. One of the regions with the highest proportion of duplicated sequence is the 22q11.2 locus, carrying eight LCRs, and rearrangements between them cause the 22q11.2 deletion syndrome. The LCR22-A block is hypervariable in the human population. It remains unknown whether this variability also exists in non-human primates, since research is strongly hampered by the presence of sequence gaps in the human and non-human primate reference genomes. To chart the LCR22 haplotypes and the associated inter- and intra-species variability, we *de novo* assembled the region in non-human primates by a combination of optical mapping techniques. A minimal and likely ancient haplotype is present in the chimpanzee, bonobo, and rhesus monkey without intra-species variation. In addition, the optical maps identified assembly errors and closed gaps in the orthologous chromosome 22 reference sequences. These findings indicate the LCR22 expansion to be unique to the human population, which might indicate involvement of the region in human evolution and adaptation. Those maps will enable LCR22-specific functional studies and investigate potential associations with the phenotypic variability in the 22q11.2 deletion syndrome.

4 22q11.2 low copy repeats expanded in the human lineage

4.1 Introduction

Segmental duplications or low copy repeats constitute over 6.6% of the genome (Nurk et al. 2022) and are complex patchworks of duplicated DNA fragments varying in length with over 90% sequence identity (Bailey et al. 2001, 2002a). The impact of these LCRs on primate and human evolution is increasingly recognized (Dennis and Eichler 2016; Dennis et al. 2017). It is estimated that the origin of the LCRs coincide with the divergence of New and Old World Monkeys, 35-40 million years ago (Bailey and Eichler 2006). However, a genomic duplication burst was observed in the great ape lineage, creating lineage-specific LCRs which are highly copy number variable (Marques-Bonet et al. 2009b). These LCR-containing regions in other great ape reference genomes are also enriched for gaps, since they are subject to similar assembly difficulties as those encountered in the assembly of these regions in the human reference genome (Mikkelsen et al. 2005; Gordon et al. 2016).

In humans, the 22q11.2 region contains a relatively higher proportion of LCRs compared with the rest of the genome. The origin of the human chromosome 22 LCRs is concordant with the evolutionary timeline of LCRs in general. No duplicated orthologous LCR22 sequences are present in the mouse (Puech et al. 1997; Shaikh et al. 2000). The segmental duplication structure is present in the non-human primates, indicating the coincidence of their origin with the New-Old World Monkey divergence (Shaikh et al. 2000). FISH mapping of functional gene loci showed lineage-specific variation in the non-human primates, underlining the instability of the locus (Bailey et al. 2002b; Babcock et al. 2007). In addition, young *Alu* SINE elements were uncovered at the boundaries of these expansions (Guo et al. 2011). However, data interpretation was performed against hg19, which harbors major differences in the LCR22 structure compared to hg38. In addition, techniques to resolve the exact structure of the LCR22s were lacking. Hence, *de novo* assembly of these haplotypes and interpretation against hg38 would provide a more accurate map of the evolutionary history of this complex locus.

We demonstrated hypervariability in the organization and the copy number of duplicons within LCR22s, especially LCR22-A (Demaerel et al. 2019). By combining fiber-FISH and Bionano optical mapping we assembled the LCR22s *de novo* and uncovered over 30 haplotypes of LCR22-A, with alleles ranging in size from 250 to 2000kb within 169 normal diploid individuals (Demaerel et al. 2019). This LCR22-A catalog was expanded by haplotyping the complete alleles of 30 22q11.2DS families (Pastor et al. 2020). To determine whether this extreme haplotype variability is human-specific, we set out to chart the inter- and intra-species variability of these LCR22s in non-human primates (**Supplementary Table S4.1**). The LCR22 structures of the great apes, including five chimpanzees (*Pan troglodytes*), one bonobo (*Pan paniscus*), two gorillas (*Gorilla gorilla* and *Gorilla berengei*

graueri), six orangutans (*Pongo pygmaeus* and *Pongo abelii*), and one rhesus monkey (Old World Monkey, Macaca Malutta) were analyzed by using the LCR22-specific fiber-FISH. To map the broader region, one representative of each species was analyzed by Bionano optical mapping. Interpreting these optical mapping data against the existing reference genomes enabled us to correct assembly errors and close gaps in the syntenic 22q11.2 loci. We demonstrate the non-human primate haplotypes to be less complex compared to humans. No intra-species variability similar to humans was observed suggesting that the hypervariability of the human LCR22-A haplotype is of recent origin.

4.2 Results

4.2.1 Assembly of chromosome 22 and the 22q11.2 locus in non-human primates

The great apes reference genomes contain several assembly gaps in chromosome 22 and especially the human 22q11.2 syntenic region. To assess the accuracy of the assemblies and close the gaps, we randomly selected one representative of the chimpanzee, bonobo, gorilla, orangutan, and rhesus monkey to be processed by Bionano optical mapping. Subsequently, the assembled chromosome 22 alleles were compared to their genomic reference sequences (**Figure 4.1**). Chromosomal locations of LCR22-A until -D in human and the other investigated species are provided in **Supplementary Table S4.2**.

The comparison of the Chimpanzee reference (chromosome 22 of panTro6/Clint_PTRv2, January 2018) with the assembled Bionano haplotype uncovered one large misassembly, visible as an inversion variant in the Bionano plot, including LCR22-A until -D (**Figure 4.1A-B**). The genes GAB4 and CCT8L2, located at the centromere in human hg38, are present at the distal end of this misassembly in the chimpanzee reference. In addition, genes RIMBP3C and LRRC74B are delineating the proximal start in the chimpanzee reference, while they are part of LCR22-D in humans. Fiber-FISH using BAC probes RP11-165F18 (proximal LCR22-D) and RP11-354K13 (distal LCR22-D) (**Supplementary Table S4.3**) validates the optical mapping assembly. No large yellow labeled regions, which indicate inconsistencies between the Bionano assembly and the reference, or differences in length are noticed for the LCR22s specifically (**Figure 4.1B**), except for LCR22-D, which is also disrupted by the misassembly.

Optical mapping assemblies based on the bonobo genomic reference of chromosome 22 resulted in two contigs (chromosome 22 of panPan3/Mhudiblu_PPA_v0, May 2020) (**Figure 4.1C**). During the analysis, complex multi-path regions splitted into two maps, since they are prone to misassembly. The distal part of the second contig (upper band) is identical to the reference except for one insertion. The first contig (lower band and zoom in **Figure 4.1D**) represents the 22q11 syntenic locus. In contrast to the bonobo reference genome where a 550kb region between LCR22-B and -C is missing, the optical map shows that this region is present and the LCR22 organization is identical to the chimp and human

organization. This missing sequence in the bonobo reference genome is actually present as a separate contig (chrUn_NW_023259866v1), but not integrated into the reference genome assembly. The BAC probe CH17-131N14 (inter LCR22-B/C) in the fiber-FISH assay validated the Bionano assembly.

The optical map of Gorilla (**Figure 4.1E**, lower band) is largely corresponding with the genomic reference of chromosome 22 (**Figure 4.1E**, upper band, chromosome 22 of gorGor6/ Kamilah_GGO_v0, August 2019), except for the human 22q11.2 syntenic region (**Figure 4.1F**). In contrast to the reference, our optical map includes a 900kb insertion, which contains the region between the LCR22-C and -D allele. Part of the sequence of this region is represented in an unassembled contig chrUn_NW_022154665v1 of the gorilla reference genome. This contig contains the genes PI4KA, SERPIND1, SNAP29, CRKL, AIFM3, and LZTR1, located in the human genome between LCR22-C and -D, has a length of ~300kb and therefore is incomplete. Based on the large LCR22-D expansions identified by our fiber-FISH assemblies, it can be postulated that the remainder of the 900kb insertion is LCR22-D sequence.

Although the distal part of chromosome 22 of the orangutan reference genome (ponAbe3/Susie_PABv2, January 2018) is generally consistent with the Bionano assembly (**Figure 4.1G**), some inconsistencies are visible in the proximal part (**Figure 4.1H**). The reference genome predicts the following LCR22 composition: A-E-D-C-B-F-G-H. However, in comparison to this reference, the orientation between LCR22-F and -H switched, visualized as an inversion variant between the reference and the assembly (**Figure 4.1H**). Individual Bionano reads validate the presence of the LCR22-F/H inversion (**Supplementary Figure S4.1**), with a coverage of 75X and 100X of the proximal and distal inversion breakpoint, respectively. The general coverage over the LCR22 blocks is between 70X and 110X. In addition, the Bionano assembly was able to differentiate between two LCR22-A alleles that were found for this orangutan (Pongo pygmaeus 8, **Supplementary Table S4.4**).

The rhesus monkey assemblies generated based on our Bionano data were largely consistent with the chromosome 10 genome reference of the rhesus monkey which contains the 22q11.2 syntenic locus (Mmul10/rheMac10, February 2019). In contrast to the great ape syntenic 22q11.2 assemblies, not only this region was correctly mapped but also the LCR22s (**Supplementary Figure S4.2**). Some rearrangements were observed in the centromeric locus, which is characterized by the presence of sequence gaps.

Considering that the 22q11.2 syntenic optical maps were validated by fiber-FISH, that the new assemblies are concordant across the great apes including human and that some unassembled contigs suggest the existence of those regions in the reference genome, we postulate the novel assemblies provide a more accurate representation of the great ape reference genomes.

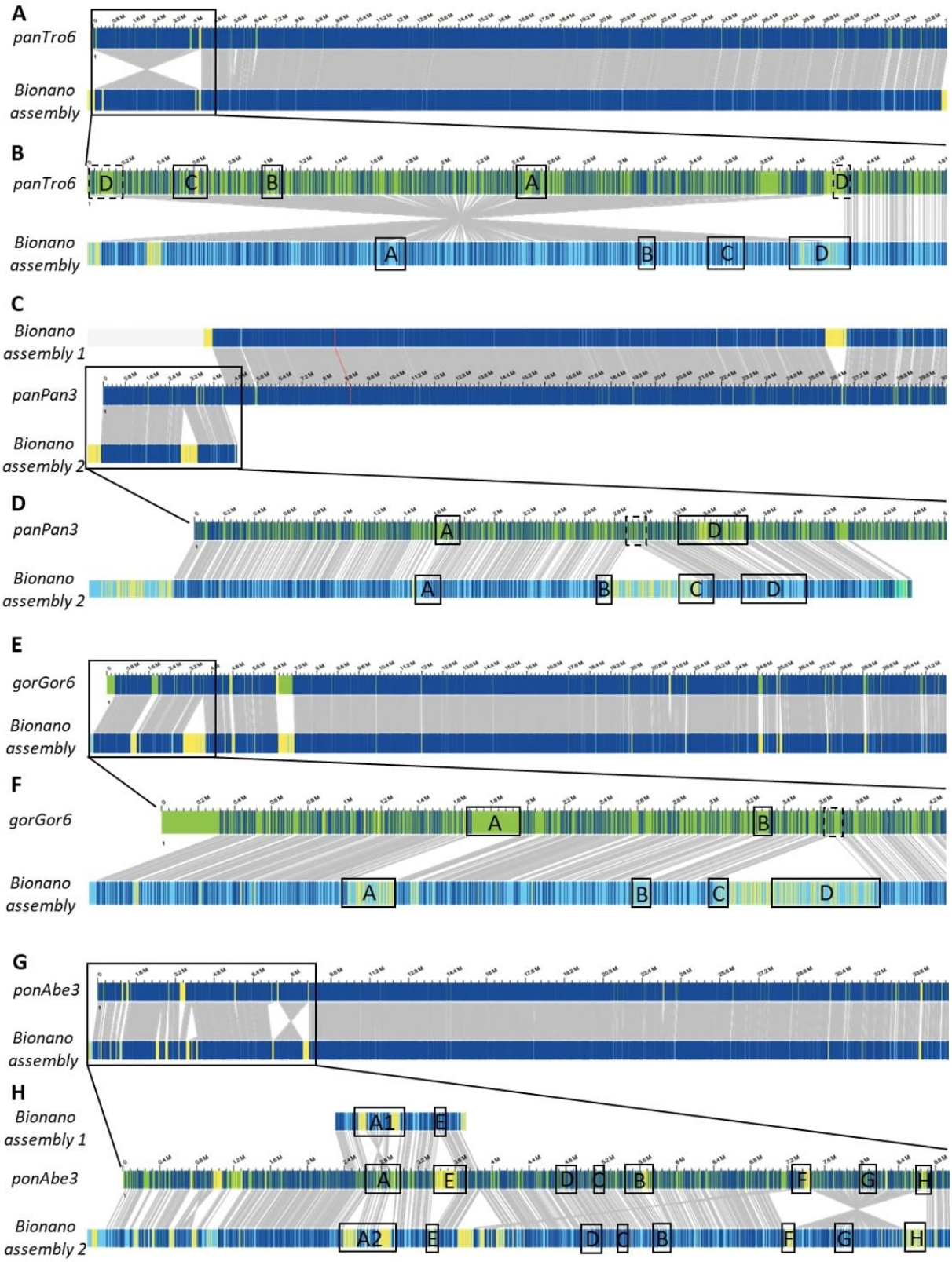


***Figure 4.1**: **Bionano optical mapping comparison to chromosome 22 reference of Great apes.** Chromosome 22-wide comparisons between the Bionano assembly and the reference of the chimpanzee (A, panTro6/Clint_PTRv2), bonobo (C, panPan3/Mhudiblu_PPA_v0), gorilla (E, gorGor6/Kamilah_GGO_v0), and orangutan (G, ponAbe3/Susie_PABv2). A zoom to the 22q11.2 locus is depicted for the chimpanzee (B, chr22:0-4,832,825), bonobo (D, chr22:0-5,007,102), gorilla (F, chr22:0-4,281,955), and orangutan (H, chr22:0-8,985,772), where the LCR22 blocks are represented. Blue labels in the Bionano optical maps represent aligned labels, and yellow labels unaligned. These unaligned labels are non-similarities between the investigated genomes. They can be present either at the nucleotide level or in very complex regions as segmental duplications. Gray lines between the maps indicated orthologous loci.*

4.2.2 Conservation of the 22q11.2 locus compared to the human reference

To investigate the level of conservation relative to humans, the generated non-human primate optical maps were screened for rearrangements against the human reference genome. The resulting 22q11.2 syntenic assemblies were validated by fiber-FISH experiments using BAC probes targeting the regions flanking the proximal LCR22s (schematic representation in **Figure 4.2A** and **Supplementary Table S4.3**). Due to the low mapping rate between the rhesus monkey sample and the human reference genome, the Bionano analysis in this non-human primate could not be performed and the composition (**Figure 4.2F**) is only based on fiber-FISH results.

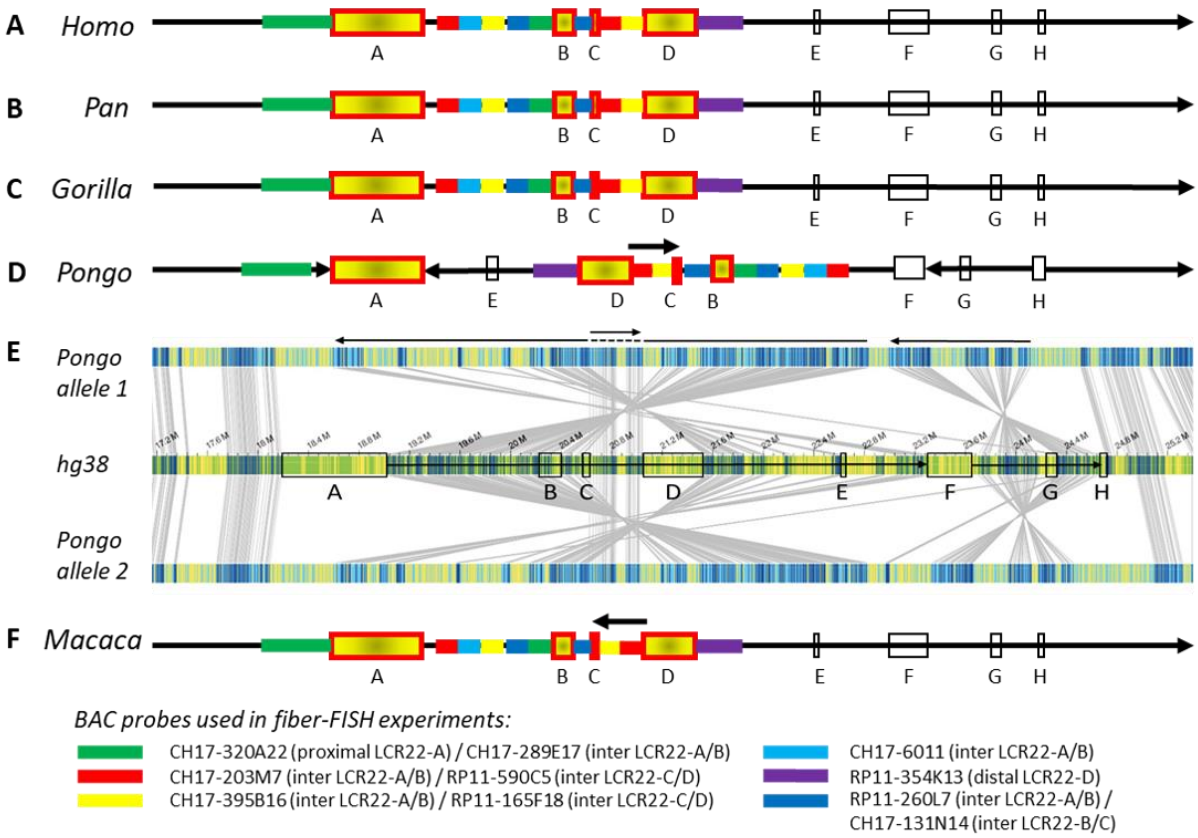


Figure 4.2: Composition of the 22q11.2 locus in human and non-human primates. *Schematic representations of the 22q11.2 region, including LCR22-A through -H, based on Bionano optical mapping and fiber-FISH. As represented in (A) the human reference genome hg38, (B) chimpanzee and bonobo, (C) gorilla, and (D) orangutan. (E) Bionano optical mapping results of orangutan compared to the human reference genome. The middle bar represents the human hg38 reference genome with blocks indicating the LCR22s (corresponding to A). The top and bottom bar represent the assembled haplotypes for this orangutan. Grey lines between the maps indicate orthologous loci. Blue labels in the maps are aligned labels, and yellow labels unaligned. These unaligned labels are non-similarities between the genomes. They can be present either at the nucleotide level or in very complex regions as segmental duplications. Arrows depict rearrangements between the human and the orangutan genomes. (F) Schematic 22q11.2 representation of the rhesus monkey, only based on fiber-FISH results. Chromosomal locations of the BAC probes can be found in Supplementary Table S3. Cartoons are not to scale.*

The order and organization of LCR22-A through -H in chimpanzee (**Figure 4.2B** and **Supplementary Figure S4.3A**), bonobo (**Figure 4.2B** and **Supplementary Figure S4.3B**), and gorilla (**Figure 4.2C** and **Supplementary Figure S4.3C**) is identical

to human. In contrast, three large rearrangements were observed in the syntenic 22q11.2 locus of the orangutan (**Figure 4.2D-E**). First, the region between LCR22-F and -H, including LCR22-G, is inverted. Second, an inversion is present between the LCR22-A and -F blocks. Third, the orientation between LCR22-C and -D is not inverted compared with the human reference. This could be interpreted as an extra inversion between LCR22-C and -D following the rearrangement between LCR22-A and -F. The composition of these LCR22 blocks in the orangutan could also be derived from the Bionano assembly against the orangutan reference genome (**Figure 4.1G-H**). However, investigating this locus in the rhesus monkey by fiber-FISH uncovered the presence of this LCR22-C/D inversion, without the larger LCR22-A/F inversion (**Figure 4.2F**). Therefore, this LCR22-C/D inversion probably represents the ancient LCR22 block organization, while an inversion in a common ancestor of the other great apes created the haplotype as present nowadays in human. Hence, despite the unstable nature of the LCR22s themselves, the structural organization between the LCR22 blocks is conserved between gorilla, bonobo, chimpanzee, and human. Inversions, typically flanked by LCRs, are present in the orangutan and rhesus monkey haplotype.

4.2.3 Evolutionary analysis of LCR22-A

The current reference genomes of great apes, except for the chimpanzee, are enriched for sequence gaps within these loci orthologous to the LCR22s. As a consequence, it was not possible to fully rely on the reference sequences and alleles had to be *de novo* assembled. For this, an LCR22-specific fiber-FISH method was applied, which has proven its value to resolve these complex structures in humans (**Figure 4.3A**; Demaerel et al. 2019).

Based on the extensive variability observed in the overall size and duplicon content of human LCR22-A (**Figure 4.3B-C**), we wanted to determine whether similar variation exists in the other great apes and rhesus monkey. Toward this end, five chimpanzees, one bonobo, two gorillas, six orangutans, and one rhesus monkey were analyzed (**Supplementary Table S4.1**). In contrast to the human variability, no structural variation was observed in any of the ten chromosomes of LCR22-A investigated in the chimpanzee samples (**Supplementary Figure S4.4**). In addition, both bonobo chromosomes had the same composition as those in the chimpanzee. This LCR22-A configuration (**Figure 4.3C**) is similar to the smallest haplotype observed in human suggesting that as the likely ancestor. However, this haplotype is rare in humans and only observed as a heterozygous allele in 5 of 169 human samples analyzed (Demaerel et al. 2019).

In the gorilla, the proximal and distal end are similar to the chimpanzee haplotype, except for a small insertion (**Figure 4.3C**). This is considered as a gorilla-specific insertion, since it is not present in the other non-human primate or human haplotypes. The same allele was observed in all four chromosomes of both gorilla cell lines. In addition to the large-scale rearrangements in the orangutan, we also observed major differences in the LCR22

compositions compared to the alleles of the other great apes (**Figure 4.3C**). First, the SD22-5 (green) duplicon, the distal delineating LCR22-A end in other great apes, is located in the middle of the allele, surrounded by SD22-6 duplicons. Second, tandem repeats of probe compositions (indicated between brackets in **Figure 4.3C**) characterize the proximal and distal end of the allele. This characteristic is different from the interspersed mosaic nature of the LCR22s in human. In addition, structural variation is observed within these repeats in the six orangutan samples (**Supplementary Table S4.4**). Thus, the haplotypes observed in the orangutan are very different from those observed in other great apes (**Figure 4.3C**). In contrast, the rhesus monkey haplotype is mostly identical to the small chimpanzee haplotype composition, except for an ~30kb insertion of unknown origin separating the SD22-5 and SD22-6 duplicons.

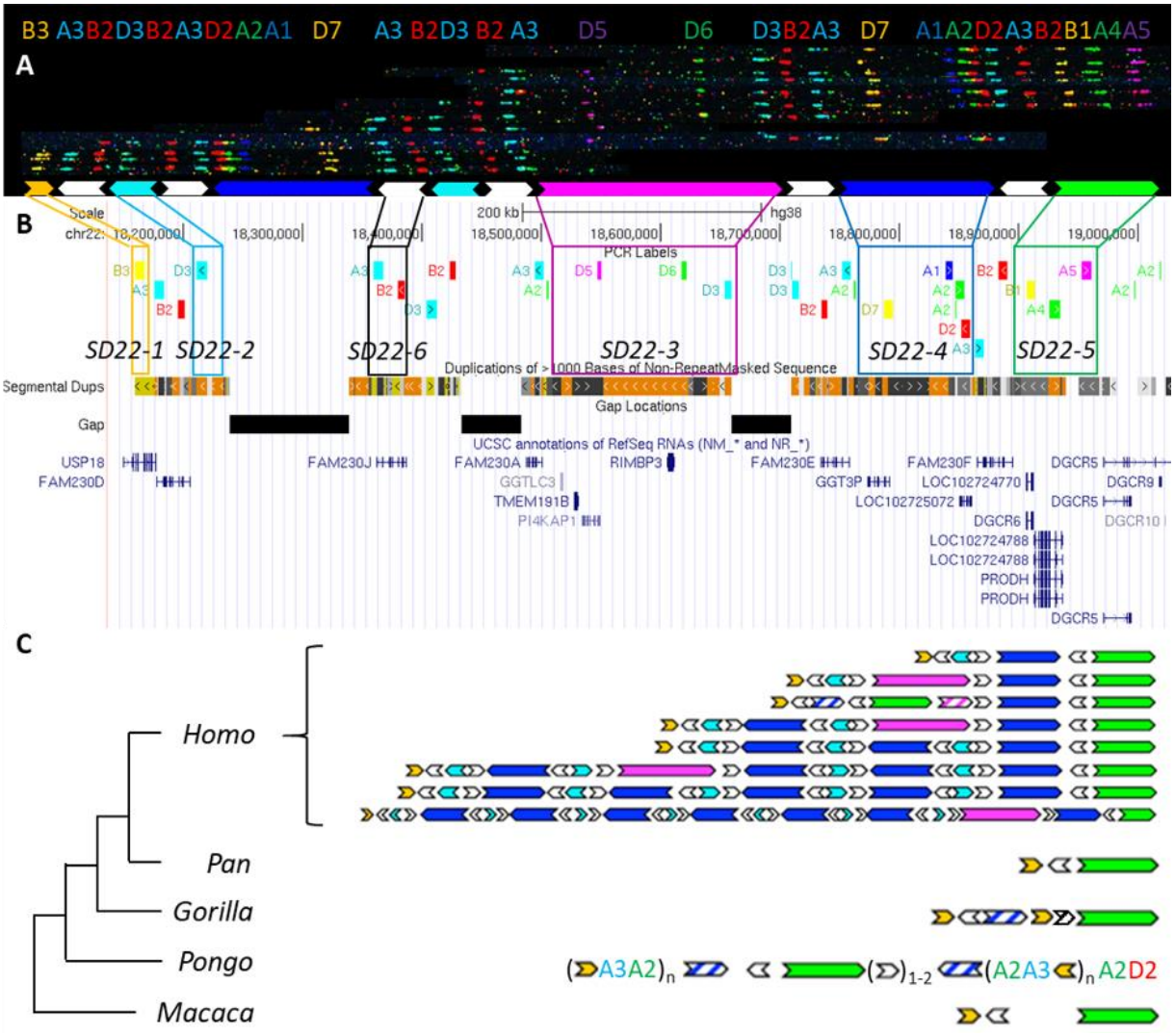


***Figure 4.3: Human duplication structure and evolutionary analysis of LCR22-A**. (A) De novo assembly of LCR22-A haplotype. Individual fibers with probe patterns are collected during the analysis of the fiber-FISH slide and later compiled based on matching colors and distances between the probes. SD22 duplicons are assigned to specific probe compositions. (B) UCSC Genome Browser hg38 reference screenshot, with tracks for fiber-FISH probe BLAT positions, segmental duplications, gaps, and RefSeq genes. Assigned duplicons in (A) are decomposed to their corresponding fiber-FISH probes in this reference screenshot. (C) Evolutionary tree representation of the observed LCR22-A haplotypes. Only a subset of assembled haplotypes are depicted for human, to emphasize the human*

hypervariability. Filled, colored arrows represent copies of duplicons, and hatched arrows represent partial copies of duplicons of the same color.

In order to validate these results, we correlated the fiber-FISH data with the corresponding chimpanzee reference genome. The human locus chr22:18,044,268-19,017,737 including the LCR22-A allele, can be traced to the chimpanzee locus chr22:2,635,159-2,386,886 in the most recent reference genome (Clint_PTRv2/panTro6, January 2018). The fiber-FISH probe order predicted from this sequence exactly matches the obtained fiber-FISH pattern. Hence, this extra independent chimpanzee allele confirms the presence of a single LCR22-A haplotype in chimpanzee.

In conclusion, structural variation of the LCR22-A haplotype is observed in orangutans and humans, although different duplicons and structures are involved.

4.2.4 Evolutionary analysis of LCR22-B/C/D

While LCR22-A is hypervariable in human genomes, LCR22-B and LCR22-C showed no variations, and only six different alleles were observed for LCR22-D (**Figure 4.4A-D**, Demaerel et al. 2019). To evaluate the evolution of these LCR22s and assess intra-species variation in non-human primates, we investigated the syntenic LCR22-B, -C, and -D haplotypes in great apes and rhesus monkey by fiber-FISH as well. Since LCR22-B and -C could be small and hard to distinguish above fiber-FISH noise, the probe set was supplemented with BAC probes flanking these LCR22s (**Supplementary Table S4.3**).

For LCR22-B, the chimpanzee and bonobo were identical to the human haplotype, while the gorilla haplotype was similar, with the deletion of one duplicon (SD22-2) (**Figure 4.4E**). In the orangutan, the distal part is substituted by two subunits (A3-D2). An extra insertion between these two subunits creates the haplotype of the rhesus monkey. LCR22-C carries lineage-specific insertions and deletions in the *Pan* and *Gorilla* genus, while in the orangutan and rhesus monkey it is reduced to only one subunit (D1) (**Figure 4.4E**). The human LCR22-D haplotype is subjected to structural variation, mainly in the SD22-3 duplicon (Demaerel et al. 2019). One variant, an internal inversion (indicated by the magenta arrow in **Figure 4.4C-D**), is present in 37% of the human haplotypes. The same variant was observed in a heterozygous state in two LCR22-D chimpanzee chromosomes (**Figure 4.4E** and **Supplementary Figure S4.4**), suggesting this variant precedes the split of the human lineage. The proximal start and distal end were conserved in *Gorilla*, with extra insertions compared to the human and *Pan* haplotype (**Figure 4.4E**). No structural variation was found at the distal end in these four investigated chromosomes. The LCR22-D haplotype in orangutan and rhesus monkey is composed of only two probes (**Figure 4.4E**).

To conclude, LCR22-B, -C, and -D haplotypes start to evolve toward their human structures in a common ancestor of *Gorilla, Pan*, and *Homo*, based on the very short haplotypes found in orangutan and rhesus monkey.

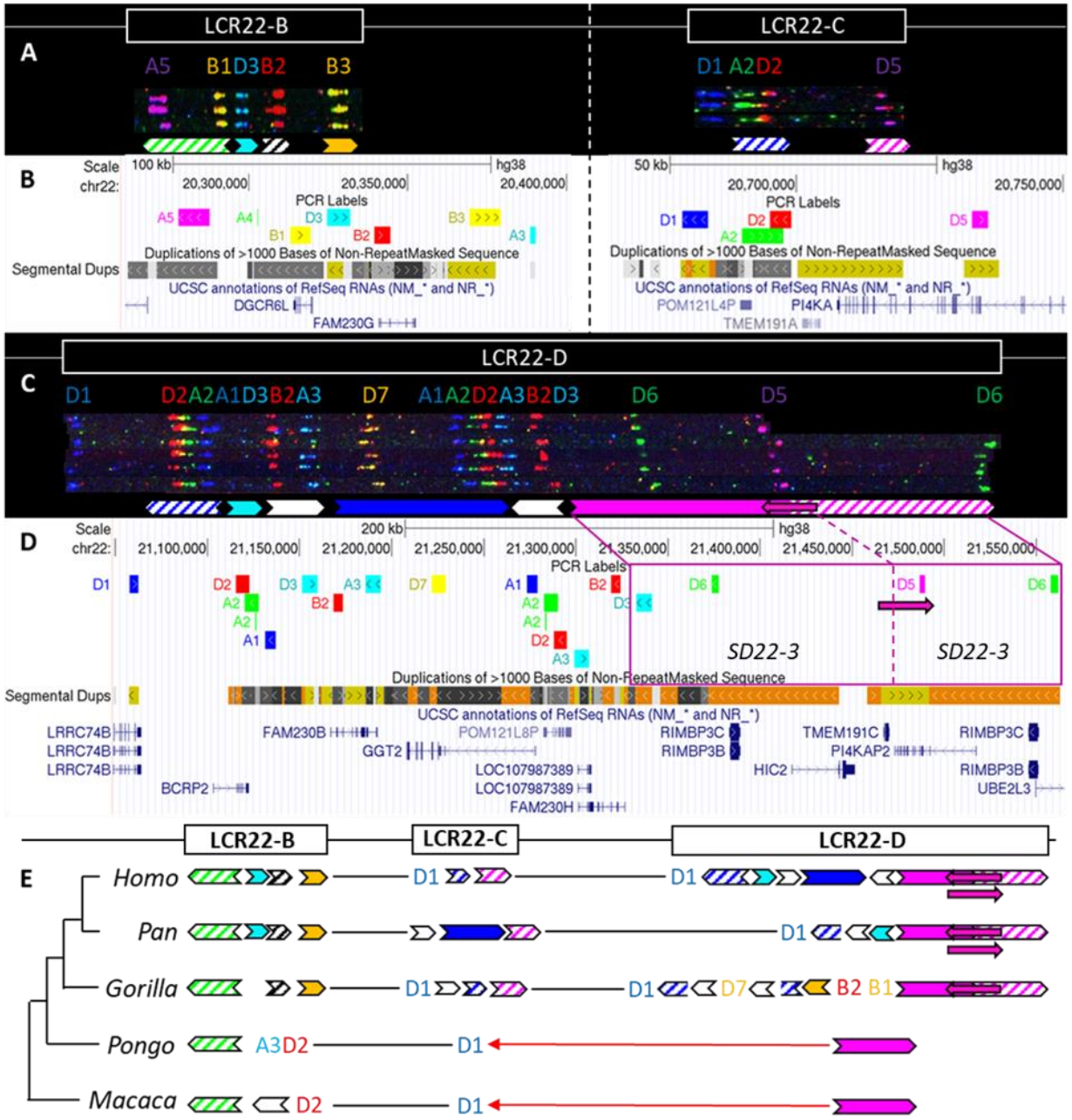


Figure 4.4: Human duplicon structure and evolutionary analysis of LCR22-B, -C, and -D. (*A) De novo assembly of the LCR22-B (left) and LCR22-C (right) haplotype. SD22 duplicons are assigned to specific probe combinations, based on the probe composition in LCR22-A (Figure 3A). (B) UCSC Genome Browser hg38 reference screenshot of LCR22-B (left) and LCR22-C (right), with tracks for fiber-FISH probe BLAT positions, segmental duplications, and RefSeq genes. (C) De novo assembly of the LCR22-D haplotype based on matching colors and distances between the probes. SD22 duplicons are assigned to specific probe combinations. (D) UCSC Genome Browser hg38 reference screenshot, with tracks for fiber-FISH probe BLAT positions, segmental duplications, and RefSeq genes. The extended SD22-3 duplicon is decomposed to the corresponding fiber-FISH probes in the reference genome. (E) Evolutionary tree representation of the observed LCR22-B, -C, and -D haplotypes. Filled, colored arrows represent copies of duplicons, and hatched arrows represent partial copies of duplicons of the same color.*

4.3 Discussion

FISH mapping studies of metaphase chromosomes from great apes using 22q11.2 BAC probes and analysis of sequencing data had demonstrated the LCR22 expansion to precede the divergence of Old and New world monkeys, and suggested species-specific LCR22 variation had occurred during primate speciation (Shaikh et al. 2000; Bailey et al. 2002b; Babcock et al. 2007; Guo et al. 2011). However, the FISH studies were mainly focusing on interrogation of the copy number of a limited number of genic segments and sequencing analysis was inevitably interpreted against human reference genome 37 (hg19), carrying important inconsistencies compared to the most recent reference genome hg38. By *de novo* assembling the LCR22s using LCR22-specific probes in the fiber-FISH assay we resolved the haplotype composition in five chimpanzees, one bonobo, two gorillas, six orangutans and a rhesus monkey. This approach provides a paradigm to map complex genomic regions and will leverage larger scale analyses of the LCR22s. The evolutionary analysis of the complex segmental duplications on chromosome 22 in different members of each species reveals a human-specific expansion of the LCR22-A haplotype, subject to structural variation in the human population.

Human-specific expansions of LCR22s possibly introduced additional substrates for LCR22-mediated rearrangements which could result in genomic disorders associated with the 22q11.2 locus. As demonstrated by Demaerel et al. (2019), the region of overlap between LCR22-A and LCR22-D is within a long stretch of homology encompassing SD22-4 flanked by SD22-6 on both sides, where recombination was shown to have taken place in case of an LCR22-A/D deletion. This locus is not present in any of the LCR22 blocks of the *Pan* genus. Pastor et al. (2020) narrowed this region to SD22-6, the duplicon encompassing the FAM230 gene member. Guo et al. (2016) predicted the rearrangement breakpoint was located in the BCR (Breakpoint Cluster Region) locus, present in the distal part of SD22-4 (end of arrow). This locus was present twice in the *Pan* haplotype, once in LCR22-C and once in LCR22-D, but in opposite orientation preventing recombination leading to deletions and duplications. In the human lineage, the prevalence of both SD22-4 and SD22-6 increases in LCR22-A and -D. Hence, human-specific expansion of the region likely increases the susceptibility of chromosome 22q11.2 to rearrangements, similar to observations made in other diseases resulting from LCR-mediated rearrangement (Sudmant et al. 2013).

Chimpanzee and Rhesus LCR22-A haplotypes were the smallest amongst the analyzed apes and monomorphic in all screened specimens. While the analyses of several independent individuals and subspecies of Pan and Gorilla suggest that intraspecies variability is absent or limited, having access to a larger population of great ape specimen would strengthen such a conclusion. Because the endangered species act does not allow exchange of great ape tissue material between the United States and other countries, such analyses have been hampered. The small LCR22-A haplotype is likely the ancestral haplotype, with lineage-

specific insertions and deletions. This ancestral haplotype is composed three core duplicons (SD22-1, SD22-6, and SD22-5). Compared with most human haplotypes, three other core duplicons are missing (SD22-2, SD22-3, and SD22-4). These elements are present in, respectively, LCR22-B/D, LCR22-D, and LCR22-C of the *Pan* genus. Babcock et al. (2003) presented a model of insertion of duplicons into LCR22-A combining homologous recombination in the absence of a crossover with non-homologous repair. The model was proposed for an interchromosomal recombination, but can be applied for intrachromosomal events as well. Following insertion in the LCR22-A block, allelic homologous recombination is a possible mechanism for the creation and expansion of new haplotypes. Since *Alu* elements are frequently delineating LCR blocks in general and on chromosome 22 specifically, they form a perfect substrate for this type of rearrangements (Babcock et al. 2003; Bailey et al. 2003; Guo et al. 2011).

Structural variation of LCR22-A was observed in the orangutan lineage as well (**Figure 4.3C**). In the investigated samples, the main variation was present in the number of the repeat units at the proximal start and the distal end of the haplotype, consisting of the SD22-1 (human) duplicon and probes A3 and A2. However, the copy number of this yellow duplicon is fixed in the human population whereas other duplicons are subjected to copy number variation, e.g., SD22-3, SD22-4, SD22-2, and SD22-6. Therefore, other genes will be subjected to changes in the copy number and the expansion in combination with the structural variation observed in humans can be considered human-specific.

This study provides the hitherto highest resolution map of the LCR22s across our closest evolutionary relatives, improving the accuracy of the most recent non-human primate reference genomes (**Figure 4.1**). Although the reference assembly was in some species based on Bionano optical mapping data as well, use of different labeling enzymes (BspQ I vs. DLE-1) could explain inconsistencies between the assemblies. Bionano optical mapping identified three LCR22-mediated inversions in the orangutan lineage, and one in the rhesus monkey. A previous study focusing on the identification of inversion variants between human and primate genomes, observed the inversion between LCR22-C/D in the rhesus monkey, but was not able to identify any in the orangutan (Catacchio et al. 2018). The heterozygous inversion within the distal end of LCR22-D, present in humans and chimpanzee (**Supplementary Figure S4.4**), was previously identified by Bionano optical mapping in the chimpanzee, supporting the presence of this polymorphism in the chimpanzee population (Kronenberg et al. 2018). The extreme LCR22 amplification in gorilla, as described by Babcock et al. (2007), was not identified in this study. It seems likely that some of the LCR22 duplicons are amplified at other regions in the gorilla genome. Since metaphase and interphase FISH studies have a lower level of resolution, the exact location of these amplifications is not known but some amplifications appear to be located at telomeric bands. Hence, they will not be identified by our LCR22 targeted fiber-FISH analysis.

It remains to be uncovered whether this LCR22 variability influences the human phenotype, which elements are under selective pressure or rather the expansion is due to genetic drift. Human-specific expansions were also observed in LCRs present on other chromosomes that are known to cause genomic disorders (Boettger et al. 2012; Antonacci et al. 2014) and are possibly associated with human adaptation and evolution (Dennis and Eichler 2016). For example, human-specific BOLA2 duplications on chromosome 16p11.2 and variation of DUF1220 domains on chromosome 1q21 are associated with iron homeostasis (Giannuzzi et al. 2019) and brain size alterations (Dumas et al. 2012), respectively. Gene duplications are a source for transcript innovation and expansion of the transcript diversity due to exon shuffling, novel splice variants, and fusion transcripts by the juxtaposition of duplicated subunits (Nuttle et al. 2016; Dougherty et al. 2017; McCartney et al. 2019). The human-specific SRGAP2C gene on chromosome 1 is an example of neofunctionalization (Dennis et al. 2012). The LCR-located gene, created by incomplete duplication, exerts an antagonistic effect on the ancestral SRGAP2A transcripts, resulting in human-specific neocortical changes (Charrier et al. 2012; Dennis et al. 2012). Another example is the partial intrachromosomal duplication of ARHGAP11A (chromosome 15) leading to ARHGAP11B, which is associated with brain adaptations during evolution (Florio et al. 2015). Hence, human-specific (incomplete) duplications of genic segments can render those genes into functional paralogs with possible innovating functions. These genes present evidence of positive selection and show a general increase in copy number in the human lineage (Marques-Bonet et al. 2009a).

The LCRs on chromosome 22 might be considered as an extreme source for expansion of the transcript catalog. Transcriptome studies may help to unravel the role of these human-specific expansions, since the presence of specific paralogs and their possible functional importance might be underestimated. Due to the duplicated nature of the LCR22s, paralogs share a high level of sequence identity. Therefore, short-read data are not always able to resolve the differences between transcripts arising from different paralogs and long-read full-length transcriptome analysis will be required. In addition, tools to obtain the full-length sequences of the LCR22s and map the paralog variability will be essential to fully comprehend the extent of sequence variation present.

In summary, optical mapping of the LCR22s unraveled lineage-specific differences between non-human primates and demonstrated the LCR22-A expansions and variability unique to the human population. It seems likely this expansion renders the region unstable and triggers NAHR resulting in the 22q11 deletions or duplications. To counter the paradox that LCR22 expansions reduce overall fitness, we hypothesize an important role for the region, previously described as the ‘core duplicon hypothesis’ (Johnson et al. 2006; Jiang et al. 2007; Marques-Bonet et al. 2009b). Further research will be needed to unravel the functional importance of LCR22 expansion, including the role of paralog-specific transcripts.

4.4 Materials & Methods

Sample collection and cell culture

Four chimpanzee samples (*Pan troglodytes* 7, 8, 15, and 17), one gorilla cell line (*Gorilla gorilla* 1), and five orangutans (*Pongo pygmaeus* 6, 7, 8, 9, and 10) were kindly provided by Professor Mariano Rocchi (University of Bari, Italy) or purchased via the Biomedical Primate Research Centre (BPRC, Rijswijk, The Netherlands). All these samples were EBV transfected cell lines and cultured according to standard protocols. One chimpanzee fibroblast cell line was purchased from the Coriell Cell Repository (AG 06939A). One gorilla fibroblast cell line (Gorilla Kaisi) was originally obtained from the Antwerp Zoo (Antwerp, Belgium). The orangutan fibroblast cell line and the rhesus monkey kidney cell line were obtained from the European Collection of Authenticated Cell Cultures (ECACC) Repository. One EBV cell line was established from bonobo Banya from the Planckendael Zoo (Mechelen, Belgium). More information on the samples is provided in **Supplementary Table S4.1**. Blood was obtained during regular health checks of the animals.

Fiber-FISH

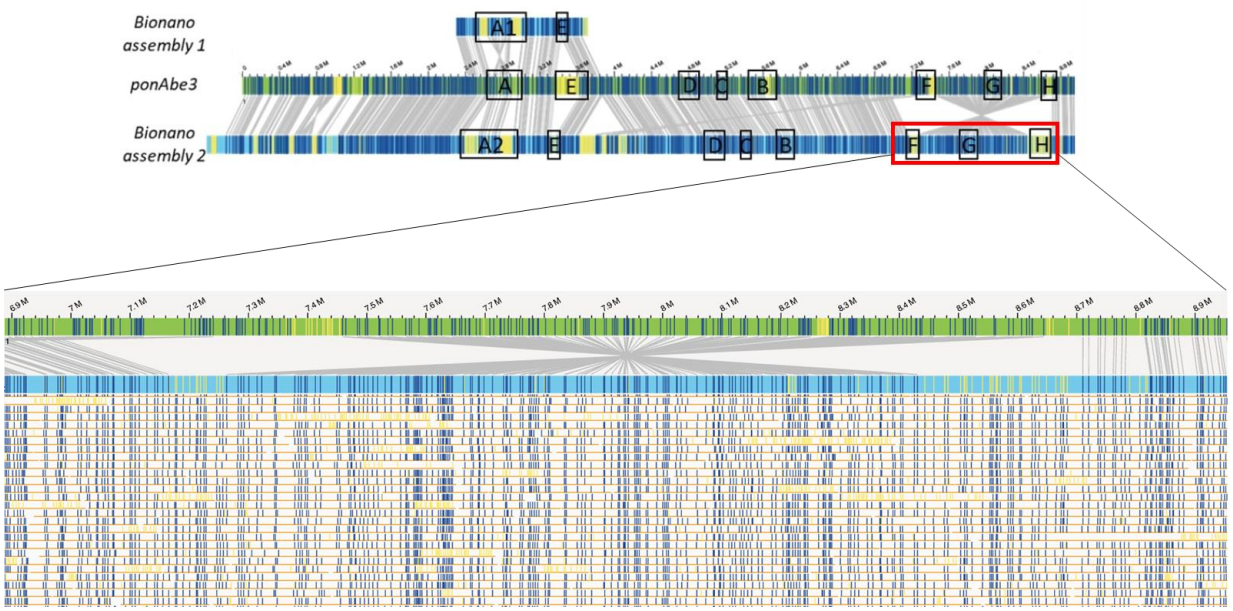
The LCR22-specific fiber-FISH is a targeted optical mapping method developed to *de novo* assemble the LCR22 haplotypes at subunit level. In contrast to techniques using shorter DNA fragments, fiber-FISH was shown to be capable of spanning the LCR22s (Demaerel et al. 2019). The procedures for slide preparation and hybridization were followed as described (Demaerel et al. 2019). The LCR22-specific customized probe set was supplemented with BAC probes targeting the unique regions between the LCR22s (**Supplementary Table S4.3**). To differentiate between probes of the same color in the non-human primate haplotypes (for example B1, B3, and D7 are all yellow), we performed color-changing experiments. Probes of the same color were re-labeled to another color (for example the yellow probes B1 to cyan and D7 to red) and the experiment was performed again using the normal probes (red, blue, cyan, magenta, green) and the re-labeled yellow probes. In that way, changes in the assembled patterns indicate the correct probe composition. These analyses were repeated for each species for the uncertain probe compositions.

Bionano optical mapping

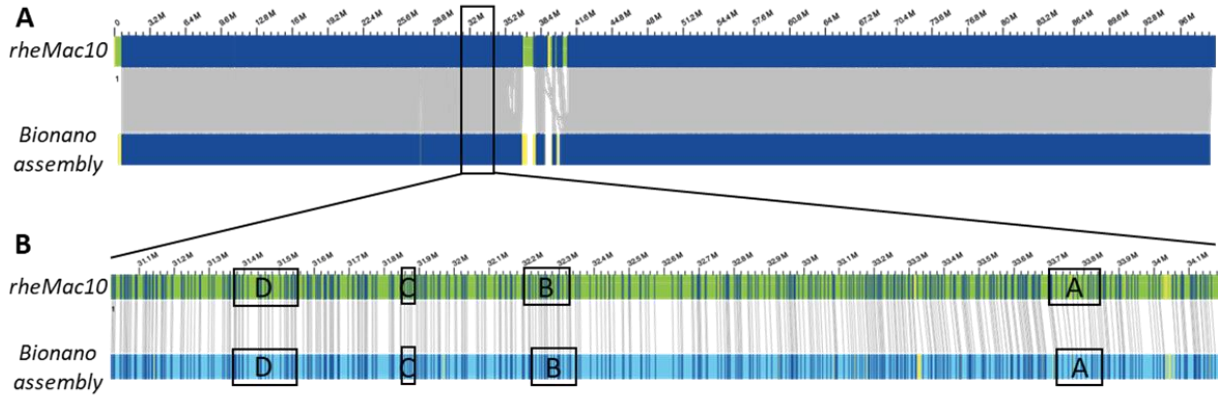
To obtain the optical mapping data, high-molecular weight DNA from one chimpanzee (*Pan troglodytes* 15), one bonobo (Bonobo Banya), one gorilla (*Gorilla gorilla* 1), one orangutan (*Pongo pygmaeus* 8), and the rhesus monkey was extracted using the SP Blood & Cell Culture DNA Isolation kit (Bionano Genomics) and labeled using the DLS DNA labeling kit (DLE-1 labeling enzyme, Bionano Genomics). Samples were loaded onto Saphyr Chips G2.3 (Bionano Genomics), linearized, and visualized using the Saphyr Instrument (Bionano Genomics), according to the Saphyr System User Guide. All analyses were performed in Bionano Access (Bionano Genomics). General quality assessment via the Molecule Quality

Report uncovered N50 values of 227kb, 215kb, 231kb, 252kb, and 177kb for the chimpanzee, bonobo, gorilla, orangutan, and rhesus monkey, respectively. First, a *de novo* assembly was performed against the most recent human reference genome hg38. The effective coverage was 191X, 182X, 185X, 71X, and 11X, for the chimpanzee, bonobo, gorilla, orangutan, and rhesus monkey, respectively. An extra *de novo* assembly was performed against chromosome 22 of the corresponding non-human primate reference genome (chromosome 10 in case of the rhesus monkey). Effective coverages reached 355X, 318X, 297X, 340X, and 302X, for chimpanzee, bonobo, gorilla, orangutan, and rhesus monkey, respectively. Structural variants could be detected at the genome-wide level in the generated circos plot. The 22q11.2 region was visually inspected for structural rearrangements by zooming in to this region and comparing the compiled haplotypes with the hg38 reference or the non-human primate reference.

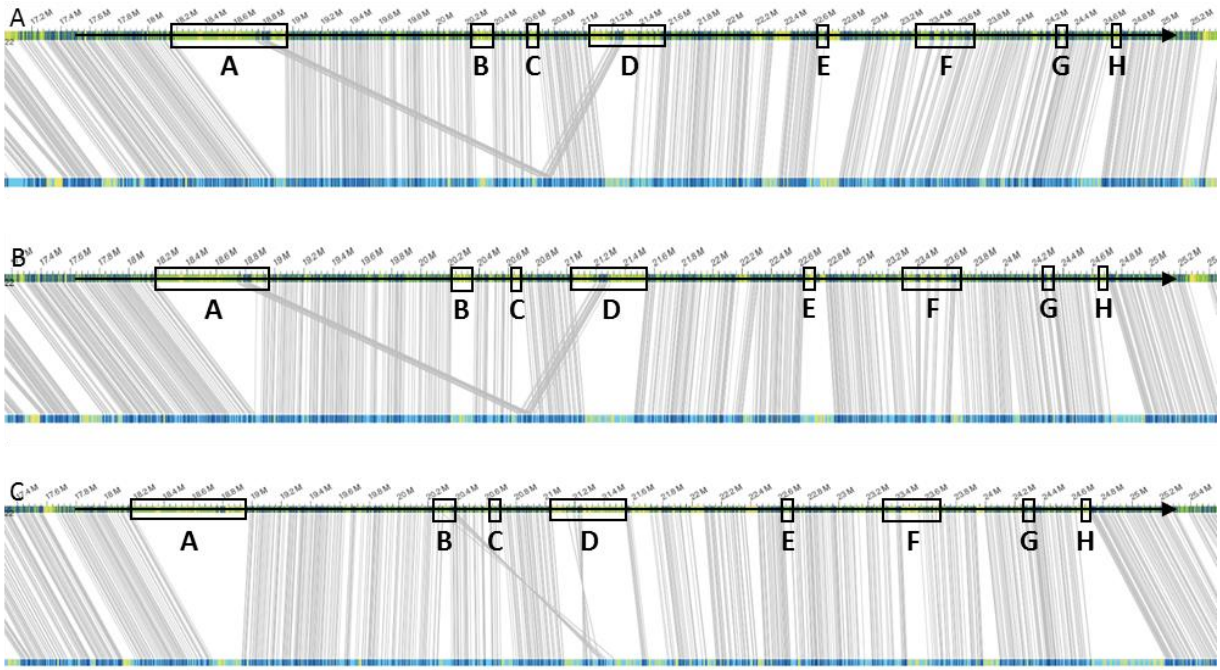
4.5 Supplementary Materials



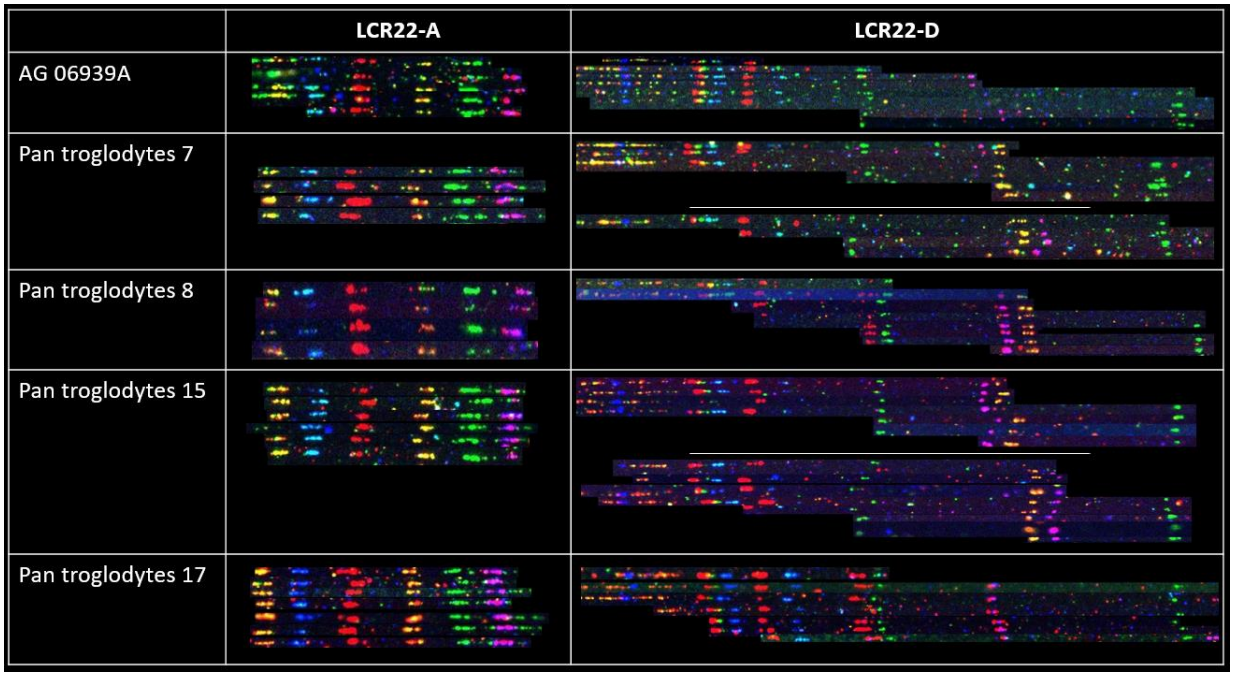
Supplementary Figure S4.1: *Bionano molecule coverage of identified misassembly in the orangutan reference genome. Individual reads shown over the 'inversion' to support the corrected Bionano assembly. Only part of the reads is visualized.*



Supplementary Figure S4.2: *Bionano optical mapping comparison to rhesus macaque reference genome (chromosome 10). (A) Chromosome 10 – wide comparison between the reference genome chromosome 10 (Mmul_10/rheMac10, February 2019) and the assembled Bionano allele. Small rearrangements are visible around the centromere locus (36-40Mb). (B) Zoom to the syntenic LCR22-containing locus, without the presence of large rearrangements between the reference and the Bionano assembly.*



Supplementary Figure S4.3: *Bionano optical mapping of the 22q11.2 region in chimpanzee, bonobo, and gorilla against the human reference genome (hg38). Regional organization of the 22q11.2 locus in (A) chimpanzee, (B) bonobo, and (C) gorilla. De novo assembled non-human primate maps are compared to the human reference genome (hg38). The top bar represents the human hg38 reference genome with blocks indicating the LCR22s. The bottom bar represents the assembled non-human primate haplotype. Grey lines between the maps indicate orthologous signals between them. Blue labels in the maps are aligned labels, and yellow labels unaligned.*



Supplementary Figure S4.4: *De novo assembled LCR22-A and -D haplotypes in the six investigated chimpanzee samples. Two chimpanzees (Pan troglodytes 7 and 15) showed structural variation distal in the LCR22-D haplotype. A white line distinguishes the two haplotypes. An extra probe was added to the probe set of Pan troglodytes 7, 8, and 15 to distinguish between the two haplotypes.*

Supplementary Table S4.1: *Overview of non-human primate samples.*

Name	Species	Cell type	Sex	Origin	Fiber-FISH	Bionano
AG06939A	*Pan troglodytes*	fibroblast	Male	N.I.A. Aging Cell Repository (Coriell Institute)	x	
Pan troglodytes 7	*Pan troglodytes*	EBV	Male	Budapest Zoo	x	
Pan troglodytes 8	*Pan troglodytes*	EBV	Male	Yale University	x	
Pan troglodytes 15	*Pan troglodytes*	EBV	Female	Biomedical Primate Research Centre Rijswijk	x	x
Pan troglodytes 17	*Pan troglodytes*	EBV	Female	Biomedical Primate Research Centre Rijswijk	x	
Bonobo Banya	*Pan paniscus*	EBV	Female	Planckendael Zoo, Mechelen, Belgium	x	x
Gorilla Kaisi	*Gorilla berengei graueri*	fibroblast	Female	Antwerp Zoo, Belgium (stock UZ Leuven)	x	
Gorilla gorilla 1	*Gorilla gorilla*	EBV	Female	Bari	x	x
Orangutan UZL	*Pongo pygmaeus*	EBV	Female	ECACC Cell Repository (stock UZ Leuven)	x	
Pongo pygmaeus 6	*Pongo pygmaeus (Borneo)*	EBV	Male	Biomedical Primate Research Centre Rijswijk	x	
Pongo pygmaeus 7	*Pongo abelii (Sumatra)*	EBV	Female	Biomedical Primate Research Centre Rijswijk	x	
Pongo pygmaeus 8	*Pongo pygmaeus (Borneo)*	EBV	Male	Biomedical Primate Research Centre Rijswijk	x	x
Pongo pygmaeus 9	*Pongo abelii (Sumatra)*	EBV	Female	Biomedical Primate Research Centre Rijswijk	x	
Pongo pygmaeus 10	*Pongo abelii (Sumatra)*	EBV	Male	Biomedical Primate Research Centre Rijswijk	x	
Rhesus Macaque	*Macaca mulatta*	kidney		ECACC Cell Repository (stock UZ Leuven)	x	x

Supplementary Table S4.2: *Genomic coordinates of LCR22 blocks in human and non-human primate reference genomes.*

Species	Reference genome	LCR22-A	LCR22-B	LCR22-C	LCR22-D
Human	GRCh38/hg38	chr22:18,156,276-19,035,473	chr22:20,141,014-20,377,631	chr22:20,667,276-20,738,272	chr22:21,009,379-21,562,091
	GRCh37/hg19	chr22:18,639,043-19,022,986	chr22:20,128,537-20,731,921	chr22:21,021,564-21,092,560	chr22:21,363,668-21,916,380
Chimpanzee	Clint_PTRv2/panTro6	chr22:2,353,333-2,542,130	chr22:997,403-1,242,036	chr22:514,361-730,768	[1] chr22:1-243,113 / chr22:4,151,191-4,297,287
Bonobo	Mhudiblu_PPA_v0/panPan3	chr22:1,587,844-1,792,189	[2] chr22:2,869,844-2,971,167	[2] chr22:2,869,844-2,971,167	chr22:3,266,103-3,684,242
Gorilla	Kamilah_GGO_v0/gorGor6	chr22:1,685,017-1,967,583 (gap)	chr22:3,253,235-3,351,078	[3] chr22:3,626,885-3,739,972	[3] chr22:3,626,885-3,739,972
Orangutan	Susie_PABv2/ponAbe3	chr22:2,653,659-2,944,486	chr22:5,446,413-5,813,781	chr22:5,132,839-5,207,783	chr22:4,659,259-4,821,562
Rhesus macaque	Mmul_10/rheMac10	chr10:33,611,508-33,853,537	chr10:32,304,036-32,490,155	chr10:31,483,289-31,521,729	chr10:31,380,657-31,856,575

Genomic coordinates of human hg19, chimpanzee, and rhesus macaque are based on the UCSC convert tool (to other genomes). Genomic coordinates of bonobo, gorilla, and orangutan are based on gene locations delineating LCR22-A (USP18 and DGCR2), LCR22-B (ZDHHC8 and ZNF74), LCR22-C (MED15 and PI4KA), and LCR22-D (LRRC74B and UBE2L3).

[1] Boundaries of misassembly (inversion compared to Bionano data) identified in reference genome (Figure 1A-B)

[2] LCR22-B and -C present as a 'merged' LCR22 block in the bonobo reference genome (Figure 1C-D)

[3] LCR22-C and -D present as a 'merged' LCR22 block in the gorilla reference genome (Figure 1E-F)

Supplementary Table S4.3: *Chromosomal locations in the human reference genome hg38 of BAC probes used in the fiber-FISH assay.*

Relative LCR22 position	BAC probe	Chromosomal location (hg38)
Proximal LCR22-A	CH17-320A22	chr22: 17,9431,14-18,171,048
Distal LCR22-A	CH17-222C16	chr22: 18,997,743-19,219,922
Distal LCR22-A / Proximal LCR22-B	CH17-203M7	chr22: 19,078,955-19,323,096
Distal LCR22-A / Proximal LCR22-B	CH17-6O11	chr22: 19,358,444-19,562,811
Distal LCR22-A / Proximal LCR22-B	CH17-395B16	chr22: 19,581,392-19,840,220
Distal LCR22-A / Proximal LCR22-B	RP11-260L7	chr22: 19,891,660-20,073,444
Proximal LCR22-B	CH17-289E17	chr22: 20,073,602-20,301,293
Distal LCR22-B / Proximal LCR22-C	CH17-131N14	chr22: 20,398,072-20,629,441
Distal LCR22-C	RP11-590C5	chr22: 20,726,003-20,906,345
Proximal LCR22-D	RP11-165F18	chr22: 20,934,046-21,099,834
Distal LCR22-D	RP11-354K13	chr22: 21,588,266-21,764,059

Supplementary Table S4.4: *LCR22-A structural variation in the orangutan samples.*

Sample	Allele 1		Allele 2	
	Proximal	Distal	Proximal	Distal
Orangutan UZL	6	None	1	None
Pongo pygmaeus 6	1	1	2	None
Pongo pygmaeus 7	4	1	2	None
Pongo pygmaeus 8	2	1	2	3
Pongo pygmaeus 9	2	None	6	None
Pongo pygmaeus 10	3	None	4	2

CHAPTER 5

Investigation of allelic homologous recombination as a mechanism to create new LCR22-A haplotypes

Lisanne Vervoort[1], Maciej Geremek[2], Beata Nowakowska[2], and Joris R. Vermeesch[1]

[1] Department of Human Genetics, KU Leuven, Leuven, Belgium

[2] Department of Medical Genetics, Institute of Mother and Child, Warsaw, Poland

Abstract

Low copy repeats are enriched in copy number variation and are hotspots for NAHR resulting in genomic disorders. Recently, the largest LCR on chromosome 22 and main driver of 22q11.2 deletions, was shown to be hypervariable with sizes ranging from 250kb up to 2Mb. The origin of this hypervariability remains vague. We hypothesized that allelic homologous recombination within the LCR22 would generate novel LCR22-A haplotypes. Hence, we screened for LCR22-A recombinations in CEPH families and subsequently performed fiber-FISH mapping of the LCR22s to identify the LCR22 substructures in both parents and offspring. We identified eight LCR22-A recombinations in the eight families, confirming that LCR22-A is a recombination hotspot. All parental and offspring LCR22-A haplotypes showed Mendelian inheritance without structural LCR22-A haplotype changes. Both the small sample size and the limited resolution of the markers used to map the cross-overs do not allow to exclude that meiotic recombination is a driver of variability. Alternatively, this result might indicate that the variability could be generated by other mechanisms.

5 Investigation of allelic homologous recombination as a mechanism to create new LCR22-A haplotypes

5.1 Introduction

Low copy repeats (LCR) are characterized by the presence of several paralogues either on the same or on other chromosomes. Those regions can misalign during meiosis, causing non-allelic homologous recombination (NAHR) resulting in genomic rearrangements (Lupski and Stankiewicz 2005). The resulting microdeletions and reciprocal microduplications are frequently causing developmental disorders (Inoue and Lupski 2002). Hence, the presence of LCRs has a negative impact on the genomic stability of specific loci.

LCRs are composed of different repeat subunits which are variable in orientation, composition and copy number in the human population. In contrast to the genome-wide single nucleotide sequence variation, interindividual LCR differences involve copy number alterations of fragments, comprising many nucleotides (Vollger et al. 2022). These alterations likely arise from allelic homologous recombination (AHR), as was demonstrated for copy number variation in the β-defensin gene (Bakar, Hollox, and Armour 2009). Bailey, Liu, and Eichler (2003) proposed an *Alu*-*Alu* mediated recombination model for the duplications of full-length mosaic LCR elements, based on the enrichment of *Alu* elements at the LCR junctions. Interestingly, the *Alu* retroposition activity burst in primates coincides with the origin of LCRs (Bailey et al. 2003; Bailey and Eichler 2006).

Recombination within LCRs is hard to study due to the high sequence identity, hampering the accurate genotyping of SNPs within LCRs (De Raedt et al. 2006). Based on the *Alu*-AHR hypothesis (Bailey et al. 2003), one would predict an overlap of the NAHR and AHR hotspots in the human genome. Indeed, in the 17q11.2 (NF type I) and the 17p12 (CMT1A/HNPP) locus NAHR and AHR hotspots coincide (De Raedt et al. 2006; Lindsay et al. 2006). However, opposite results were published for the same 17p12 locus (Inoue et al. 2001), as well as for the 17p11.2 (Smith-Magenis syndrome) locus (Bi et al. 2002). Hence, no conclusive statement can be made for the positional overlap of NAHR and AHR hotspots as this may be locus-specific.

The 22q11.2 locus is shaped by the presence of eight LCR blocks (LCR22-A until -H), increasing the complexity and instability of the region. Deletions between these blocks result in the 22q11.2 deletion syndrome (McDonald-McGinn et al. 2015). This is the most common microdeletion syndrome in humans, with an estimated incidence of 1 in 2148 live births (Blagojevic et al. 2021), indicating the high *de novo* NAHR frequency rate. In 85% of the patients, the rearrangement took place between LCR22-A and -D, resulting in a 3Mb

deletion. In addition, hypervariability of the LCR22-A structure was discovered in the human population with haplotypes ranging in size between 250kb and 2000kb (Demaerel et al. 2019). However, it is not yet known how this range of haplotypes was created and whether this process is still ongoing.

Aside from the high *de novo* NAHR rate in the 22q11.2 region, the locus shows high levels of AHR as well. Torres-Juan et al. (2007) constructed a pedigree-based linkage map based on 440 informative meiosis and 27 locus-specific recombination events typed in 34 families. Three 500kb regions displayed a high recombination rate: the first includes LCR22-A, the second LCR22-B and -C, and the third comprises part of the immunoglobulin light chain (IGL) locus. The highest recombination frequency was calculated for LCR22-A, in comparison to the other LCR22s and the whole 22q11.2 region (Torres-Juan et al. 2007). Hence, the LCR22-A repeat which is the most recurrent hotspot for 22q11.2-mediated NAHR events, shows the highest AHR level as well.

We hypothesize that the AHR mechanism could also be responsible for the expansion of the LCR22-A haplotype diversity in the human population. To test this hypothesis, we screened for individuals with a recombination within LCR22-A and subsequently mapped the LCR22-A haplotype structure at subunit level.

5.2 Results

5.2.1 Identification of LCR22-A recombination

Homologous recombination between two LCR22-A alleles might generate a novel allele with a duplicon composition different from the parent-of-origin, the parent in whom the recombination occurred. The alignment of different repeats within LCR22-A between both alleles, would result in novel duplicon compositions (**Figure 5.1**). To test this hypothesis, we set out to identify *de novo* LCR22 duplicon structures. We predict those might arise following meiotic recombination within the LCR22s. To identify recombinations within LCR22-A, chromosomes 22 have to be haplotyped in families. Haplotyping is possible if genotype data are available for all the family members and if, in addition to the parents, at least two siblings or the grandparents have been genotyped.

To identify such families, we made use of the publicly available data of eight large CEPH (Centre d'Etude du Polymorphism Humain) pedigrees (884, 1331, 1332, 1347, 1362, 1413, 1416, and 102) (Dausset et al. 1990). We inspected 82 short tandem repeat (STR) and SNP markers in the 22q11.2 locus typed by the CEPH (**Supplementary Table S5.1**).

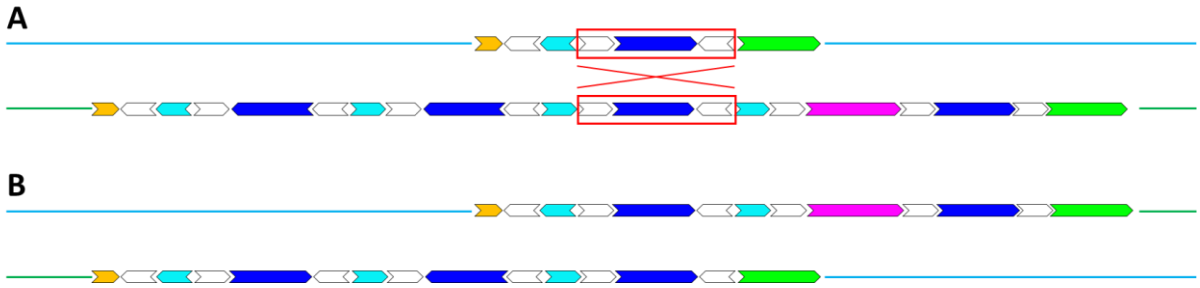


Figure 5.1: Hypothetical model of new LCR22-A haplotype creation via the allelic homologous recombination mechanism. *(A) The two different LCR22-A alleles of a heterozygous individual are displayed. Interchromosomal recombination can take place between a stretch of identical duplicons in the same orientation, indicated by the red boxes. (B) Crossover will produce two 'hybrid' LCR22-A alleles with different compositions.*

AFM288we5 (Chr22:18,108,551; 50kb proximal from LCR22-A start) and WIAF-843 (Chr22:18,994,470; in unique sequence within the distal LCR22-A end) are the STR markers flanking the proximal and distal side of LCR22-A, respectively. If cross-overs occurred in the parental haplotype between AFM288we5 and WIAF-843, the cross-over might have occurred within LCR22-A. Screening of 138 individuals in eight pedigrees identified eight individuals with recombination in this locus (**Table 5.1**). In families 884, 1332, 1413 and 102, the linkage analysis was only based on STR markers (**Figure 5.2**, example of family 884). In families 1331, 1347, 1362, and 1416, STR and SNP markers were taken into account (**Figure 5.3**, example of family 1331).

Table 5.1: Overview of CEPH families investigated for recombination over LCR22-A.

Family	Father	Mother	Child (recombination)
884	GM13113	GM13114	GM13123 (M)
1331	GM07057	GM06990	GM07023 (P) *GM06983 (P)**
1332	GM10848	GM10849	GM12095 (M)
1362	GM10860	GM10861	GM11982 (M/P) GM11985 (M) GM11987 (P)
102			No cell line available

*The recombination is of maternal origin (M) or paternal origin (P). *No fiber-FISH assay was performed on individual GM06983 since no cell line was available.*

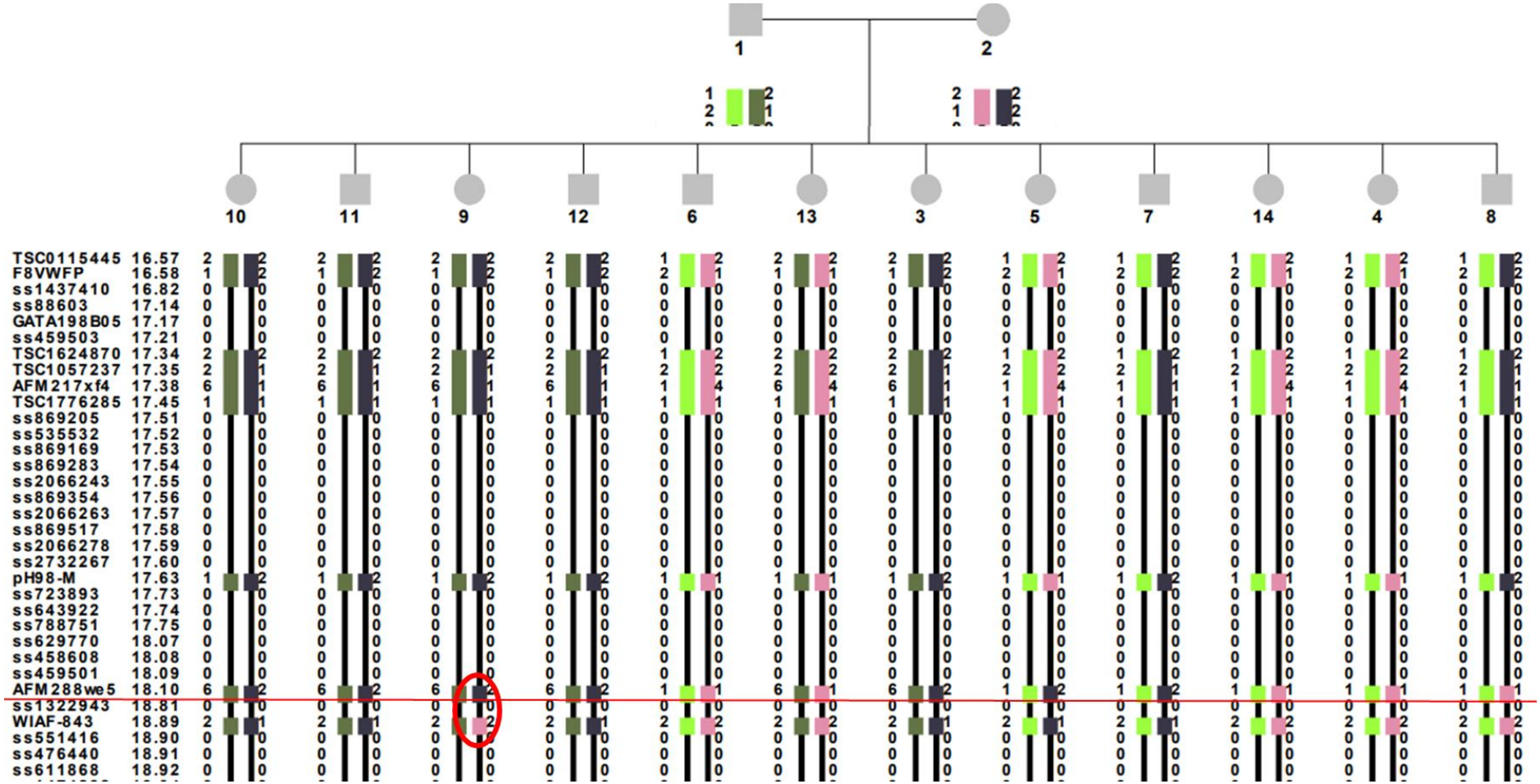


Figure 5.2: Chromosome 22 haplotypes of family 884. *Haplotypes of the parents are not fully depicted, but are indicating the two colors used for representation of maternal and paternal alleles to interpret crossovers. Cross-overs between probes AFM288we5 and WIAF-843 are indicated by the red line. The recombination within the LCR22-A locus of individual 9 is indicated by the red circle. The recombination occurred during a maternal meiosis. Not all investigated markers are shown.*

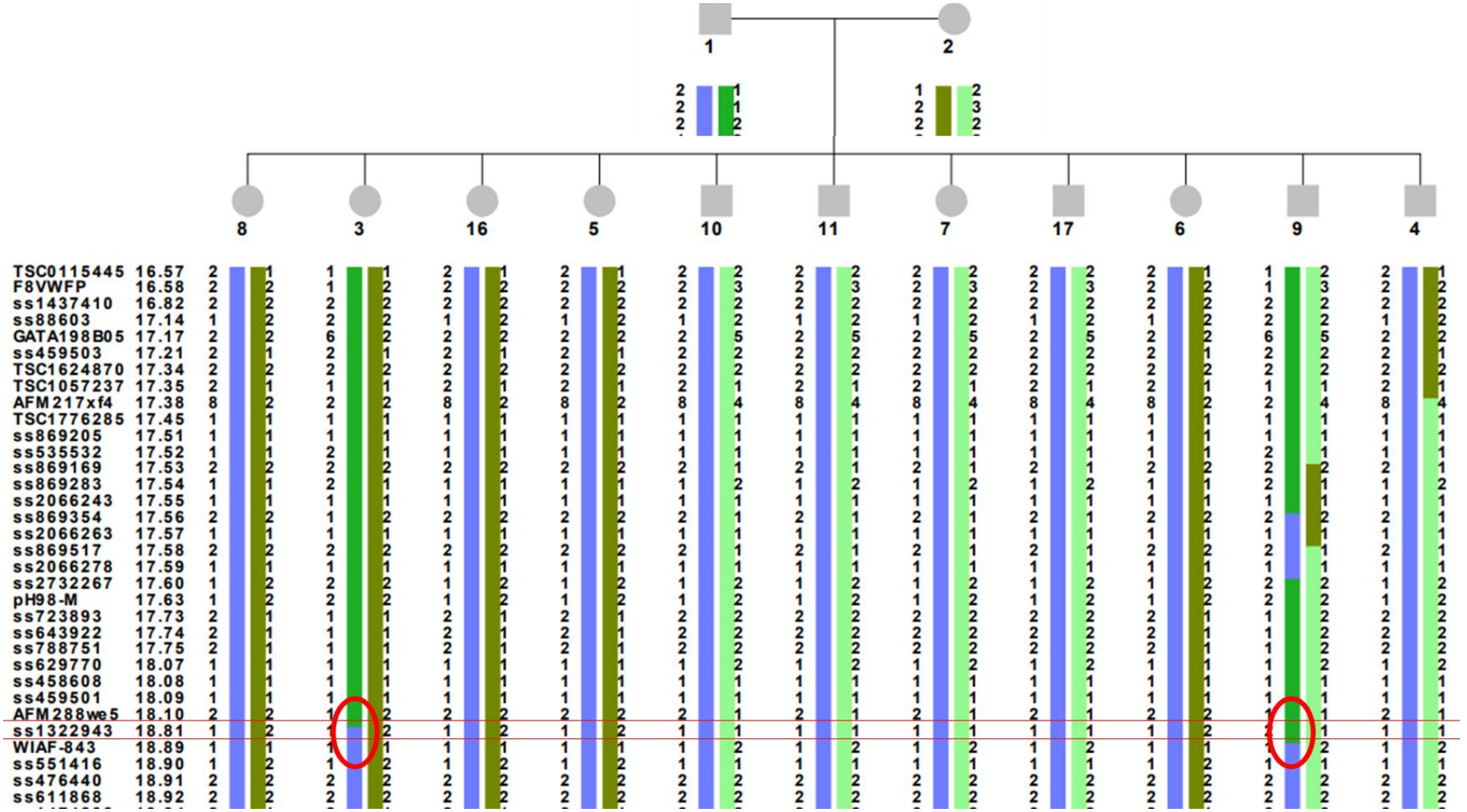


Figure 5.3: Chromosome 22 haplotypes of family 1331. *Haplotypes of the parents are not fully depicted, but are indicating the two colors used for representation of maternal and paternal alleles to interpret crossovers. Cross-overs between probes AFM288we5 and WIAF-843 are indicated by the red lines. The recombination within the LCR22-A locus is indicated by the red circles. The recombination in individuals 3 and 9 occurred during a paternal meiosis. Not all investigated markers are shown.*

5.2.2 Crossover within LCR22-A does not result in *de novo* structural variation

To screen whether recombination within the LCR22-A locus generated new LCR22-A haplotypes, the LCR22s were assembled using fiber-FISH for the individuals who had a recombination event in the LCR22-A locus, and for their parents. In family 1331, only GM07023 was tested for LCR22-A haplotype change. For family 102, one individual was identified, but no cell line was available to test the haplotype change.

In all families (884, 1331, 1332, and 1362), a Mendelian segregation of the parental haplotypes is observed. Hence, no changes in the number or orientation of the LCR22-A subunits were observed. In addition to the parents and the children where recombination was identified, the grandparents and three of the eight additional siblings (GM11984, GM11989, and GM11990) of family 1362 were haplotyped with fiber-FISH as well (**Figure 5.4A-B**) to follow the grandparental and parental allele transmission. All parental haplotypes were detected in the children, and both maternal and paternal haplotypes are present in children with and without recombination. Hence, the recombination between markers AFM288we5 and WIAF-843 did not generate a novel LCR22-A haplotype (**Figure 5.4A-B**).

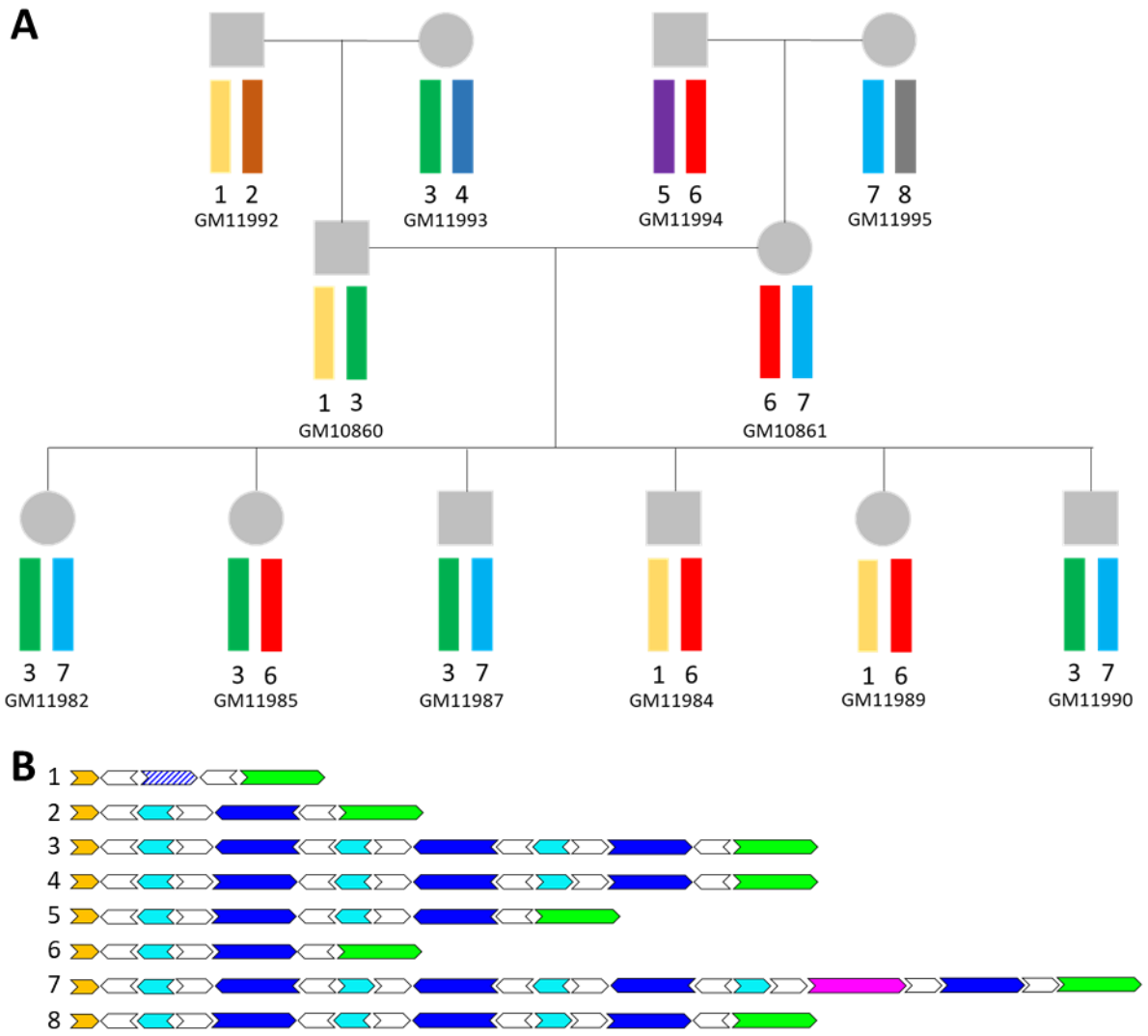


Figure 5.4: Segregation pedigree of LCR22-A haplotypes in family 1362*. (A) Pedigree structure of family 1362 including grandparents, parents, and six children: three with LCR22-A recombination (GM11982, GM11985, and GM11987) and three without recombination (GM11984, GM11989, and GM11990). Since no alterations were observed in the composed haplotypes, the colors depicting the LCR22-A haplotypes remain unchanged. (B) LCR22-A haplotypes de novo assembled in the members of family 1362.*

5.3 Discussion

Optical mapping studies uncovered human hypervariability of LCR22-A (Demaerel et al. 2019). To test whether this haplotype diversity originates via AHR, as suggested for other copy number variable loci (Bakar et al. 2009), we used publicly available STR/SNP datasets to detect recombination over LCR22-A and mapped their haplotypes via fiber-FISH. LCR22-A recombinations were identified in five of the eight investigated CEPH families. In one family, even three children carried recombinations over LCR22-A (**Table 5.1**).

No recombination-mediated alterations in haplotype structure were observed at the subunit level between the parent-of-origin and the child. We envision two potential explanations. First, recombinations could have occurred within LCR22-A but without resulting in haplotype conformation change. For example, if the crossover takes place in the proximal or distal

subunits, this will not result in a new, hybrid haplotype. To identify where the recombination occurs within the LCR22s, the cross-over locus would need to be fine-mapped at nucleotide level. Second, due to the difficulties associated with the identification of SNPs and STRs in LCR sequence in general (De Raedt et al. 2006), the closest proximal marker used in the analysis is localized in unique sequence 50kb proximal from the start of LCR22-A. As a consequence, the recombination observed between markers AFM288we5 and WIAF-843 could have taken place in this unique sequence rather than within the LCR22 block.

However, our data could also suggest that AHR is not the main driver for the creation of new haplotypes, but that other mechanisms play a role. Non-homologous end-joining will act as a repair mechanism when double-strand breaks occur. This pathway mostly invades random DNA sites resulting in loss of the overall genomic structure of the specific locus (Currall et al. 2013). Replication-based mechanisms as Fork Stalling and Template Switching and microhomology-mediated break-induced replication will typically lead to 'breakpoint signatures' (insertions, deletions, inversions) and complex rearrangements (Ottaviani et al. 2014; Carvalho et al. 2009). However, instead of randomly arranged compositions, all haplotypes are composed out of the same duplicons at subunit level and constitute a similar mosaic structure. Hence, although AHR is the most straightforward theory to explain haplotype alterations, mapping crossovers at nucleotide resolution will validate or reject this hypothesis.

Pedigree-based investigation of haplotype blocks will only detect interchromosomal meiotic recombination events. However, intrachromosomal meiotic recombination can create structural variation as well. This has been demonstrated at the Y-chromosome (Lange et al. 2013). LCR22-A intrachromosomal AHR recombinations will always result in a change of the structural organization (**Figure 5.5**). In addition, recombination between sister chromatids during mitosis in the germ cells can result in LCR22-A structural variation. Unfortunately, these recombination types are not detectable via standard genetic tests (SNP array, whole-genome sequencing data) since the haplotype proximal and distal from the LCR22 is identical compared to non-recombination individuals. Hence, occasional observation will be based on the examination of optical mapping data from control families.

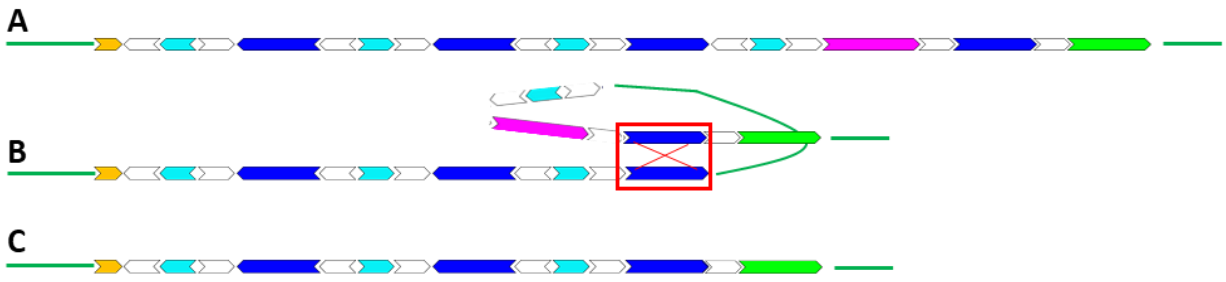


Figure 5.5: Hypothetical model of new LCR22-A haplotype creation via the intrachromosomal homologous recombination mechanism. *(A) Original LCR22-A haplotype. (B) Intrachromosomal recombination can take place between identical duplicons in the same orientation, indicated by the red boxes. (C) Crossover will produce a new LCR22-A allele with different composition.*

Limitations of the study include the small sample size and the low resolution. To explore the hypothesis further or reject the potential that novel haplotypes arise via regular recombination, it would be essential to expand the dataset. For the identification of LCR22-A recombination, we can use SNP/STR information from families with at least two siblings or the grandparents, to ensure proper phasing. In addition, cell lines have to be available to perform optical mapping. This is why high-throughput methods as single-sperm typing (Arnheim et al. 1991) are not possible. The enrichment of *Alu* repeats in the 22q11.2 LCRs (Babcock et al. 2003; Guo et al. 2011) suggests that crossovers take place via an *Alu*-mediated recombination model (Bailey et al. 2003). The resolution of the fiber-FISH method (5-10kb) is sufficient to detect alterations in the LCR22-A haplotype structure, but not to investigate repeat elements. Hence, the improvements in long-read sequencing methodologies to map LCRs at nucleotide resolution (Nurk et al. 2022; Vollger et al. 2022) will allow us to compile the internal LCR22-A SNP pattern, narrow down the AHR locus, and infer the recombination mechanism.

To conclude, no haplotype change could be detected at subunit resolution in our limited dataset (n=8). Expansion of the sample size will be essential in combination with nucleotide-level LCR22-specific recombination maps to infer the exact mechanism. In addition, we were able to retrieve eight 22q11.2 AHR events in eight families, indicative for the high recombination rate of the locus.

5.4 Materials & Methods

Pedigree linkage analysis

Genotype data were retrieved from the Centre d'Etude du Polymorphism Humain (CEPH, Fondation Jean Dausset, Paris, France) database (Dausset et al. 1990). Genotypes are available for 61 reference families of whom the EBV cell lines are available in the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository (Coriell Institute). For chromosome 22, a total of 920 markers were available, of which some are microsatellites (STRs) and other SNPs. Eight families were fully genotyped for the STR markers and therefore used in our further analysis to test recombination over LCR22-A: 102, 884, 1331, 1332, 1347, 1362, 1413, and 1416.

Merlin (version 1.1.2) (Abecasis et al. 2002) was used for pedigree linkage analysis. To run the package, a pedigree file, data file, and map file are necessary. The pedigree file can be downloaded for chromosome 22 and describes the relationships between individuals within a family, including phenotypic traits and genotypes for the markers investigated. Incomplete families were removed and genotypes extracted. The data file describes the structure of the pedigree file, for example M1 is marker one. The map file contains information about the markers to analyze and their chromosomal location, either in centimorgan or the physical location.

To test and optimize the Merlin program, analyses were performed using the Marshfield map of chromosome 22 as the map input file (Broman et al. 1998). This is a comprehensive human genetic map including 101 STR markers of the 920 genotype markers for chromosome 22 to check general recombination over the chromosome. To examine recombination over LCR22-A, a detailed map file was created including 82 STR and SNP markers surrounding LCR22-A (**Supplementary Figure S5.1**), covering the locus between 16.48Mb and 21.90Mb. Results were visualized using HaploPainter (Thiele and Nürnberg 2005), by importing the pedigree structure (adapted pedigree file without additional marker or disease information), the haplotypes (Merlin output), and the correct map file (Marshfield map or LCR22-specific map).

Sample collection

Cell lines were ordered from individuals that showed recombination over LCR22-A via the detailed recombination analysis in the LCR22 locus (**Table 5.1**). To identify haplotype pattern changes, cell lines of the parents were ordered as well (**Table 5.1**). These cell lines are available via the Coriell Institute. For family 1362, the cell lines of the grandparents and three non-recombination siblings were ordered as well to study the inheritance pattern of the LCR22-A haplotypes. No cell lines were available for individuals GM06983 of family 1331, and the recombination individual of family 102.

Haplotype composition using LCR22-specific fiber-FISH

To haplotype the LCRs on chromosome 22 and especially LCR22-A, we used the LCR22-specific fiber-FISH technique (Demaerel et al. 2019). Long DNA fibers were extracted starting from cell lines. Slides were prepared as described and hybridized using the LCR22-specific customized probe set. Following automated scanning of the slides (FiberVision, Genomic Vision), the data were analyzed by manually indicating regions of interest and afterwards, haplotypes were *de novo* assembled based on matching colors and distances between the probes. Patterns assembled for the 'recombination individuals' were compared to the parental patterns to check for haplotype alterations.

5.5 Supplementary Materials

***Supplementary table S5.2**: LCR22-A map file including 82 SNP and STR markers.*

Marker	Chr22 (Mb)	Marker	Chr22 (Mb)	Marker	Chr22 (Mb)
TSC0115445	16.57	ss1322943	18.81	ss1353140	19.99
F8VWFP	16.58	WIAF-843	18.89	ss2459898	20.10
ss1437410	16.82	ss551416	18.90	ss2451772	20.11
ss88603	17.14	ss476440	18.91	ss2459902	20.13
GATA198B05	17.17	ss611868	18.92	ss1353557	20.18
ss459503	17.21	ss1474298	19.04	ss1315906	20.19
TSC1624870	17.34	ss2401249	19.07	ss89887	20.20
TSC1057237	17.35	ss1733822	19.13	ss75778	20.21
AFM217xf4	17.38	ss2631131	19.18	ss1467075	20.23
TSC1776285	17.45	ss128211	19.19	ss2196097	20.24
ss869205	17.51	ss1459162	19.22	ss1315925	20.25
ss535532	17.52	ss1735193	19.23	ss1472458	20.26
ss869169	17.53	ss2116655	19.24	ss2670789	20.27
ss869283	17.54	ss1734714	19.25	ss1730575	20.44
ss2066243	17.55	ss1734704	19.26	ss518504	20.45
ss869354	17.56	ss546632	19.28	ss869331	20.46
ss2066263	17.57	ss1470498	19.41	ss458882	20.55
ss869517	17.58	ss2390529	19.47	ss459060	20.56
ss2066278	17.59	ss84944	19.49	ss826616	20.78
ss2732267	17.60	ss2390992	19.52	ss763863	20.86
pH98-M	17.63	ss77812	19.89	ss696363	20.88
ss723893	17.73	ss84948	19.90	ss709393	20.89
ss643922	17.74	ss1467674	19.91	ss84986	20.93
ss788751	17.75	ss128222	19.92	ss84988	20.95
ss629770	18.07	ss88859	19.96	AFM292va9	21.66
ss458608	18.08	ss481084	19.96	ss92524	21.77
ss459501	18.09	ss1315234	19.98	AFMa037zd1	21.90
AFM288we5	18.10				

CHAPTER 6

Atypical chromosome 22q11.2 deletions are complex rearrangements and have different mechanistic origins

Lisanne Vervoort[1], Wolfram Demaerel[1]*, Laura Y. Rengifo[1], Adrian Odrzywolski[1,2], Elfi Vergaelen[1], Matthew S. Hestand[1,3,4], Jeroen Breckpot[1], Koen Devriendt[1], Ann Swillen[1], Donna McDonald-McGinn[5,6], Ania M. Fiksinski[7,8], Janneke R. Zinkstok[7], Bernice E. Morrow[9], Tracy Heung[8,10], Jacob A.S. Vorstman[7,8], Anne S. Bassett[8,10], Eva W.C. Chow[10,11], Vandana Shashi[12], International 22q11 Brain and Behavior Consortium and Joris Vermeesch[1]*

[1]*Department of Human Genetics, KU Leuven, Leuven, Belgium*
[2]*Department of Biochemistry and Molecular Biology, Medical University of Lublin, Lublin, Poland*
[3]*Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA*
[4]*Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA*
[5]*Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA*
[6]*Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA*
[7]*Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands*
[8]*The Dalglish Family 22q Clinic and Center for Addiction and Mental Health, Toronto, ON, Canada*
[9]*Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA*
[10]*Department of Psychiatry, University of Toronto, Toronto, ON, Canada*
[11]*Clinical Genetics Service, Centre for Addiction and Mental Health, Toronto, ON, Canada*
[12]*Department of Pediatrics, Duke University School of Medicine, Durham, NC, USA*
**The first two authors should be regarded as joint First Authors.*

Status: Published in *Human Molecular Genetics*

Abstract

The majority (99%) of individuals with 22q11.2 deletion syndrome (22q11.2DS) have a deletion that is caused by non-allelic homologous recombination between two of four low copy repeat clusters on chromosome 22q11.2 (LCR22s). However, in a small subset of patients, atypical deletions are observed with at least one deletion breakpoint within unique sequence between the LCR22s. The position of the chromosome breakpoints and the mechanisms driving those atypical deletions remain poorly studied. Our large-scale, whole genome sequencing study of >1500 subjects with 22q11.2DS identified six unrelated individuals with atypical deletions of different types. Using a combination of whole genome sequencing data and fiber-FISH, we mapped the rearranged alleles in these subjects. In four of them, the distal breakpoints mapped within one of the LCR22s and we found that the deletions likely occurred by replication-based mechanisms. Interestingly, in two of them, an inversion probably preceded inter-chromosomal 'allelic' homologous recombination between differently oriented LCR22-D alleles. Inversion associated allelic homologous recombination (AHR) may well be a common mechanism driving (atypical) deletions on 22q11.2.

6 Atypical chromosome 22q11.2 deletions are complex rearrangements and have different mechanistic origins

6.1 Introduction

The LCR22s form an ideal substrate for NAHR events because of their large size and high sequence homology (Shaw and Lupski 2004). The non-allelic LCR22s can serve as mediators of misalignment during meiosis. Subsequent crossover between the homologous chromosomes results in deletions and reciprocal duplications, whereas intrachromosomal crossover can only cause deletions in the genome (Gu et al. 2008). In 90% of individuals with 22q11.2DS, a 3Mb deletion occurs between the two largest LCR22s, LCR22-A and LCR22-D (Burnside 2015; Guo et al. 2018). This LCR22-A/D deletion is present as a *de novo* event in 90% of the diagnosed individuals (McDonald-McGinn et al. 2015). The reciprocal 22q11.2 microduplication syndrome (MIM# 608363) was described as well (Portnoï 2009).

In addition to the most common 3Mb LCR22-A/D deletion, several other rearrangements between LCR22s exist. Approximately 9% of individuals have nested LCR22-A/B (1.5Mb) or LCR22-A/C (2Mb) deletions, with similar phenotypes as those with the common LCR22-A/D deletion (McDonald-McGinn et al. 2015; Guo et al. 2018). Deletions involving the distal LCR22s (i.e. LCR22-E, -F, -G and -H) are less frequently observed and are associated with a heterogeneous phenotype, including developmental delay, congenital heart defects, and a higher risk of preterm birth (Burnside 2015). Furthermore, a newly recognized recurrent deletion just distal to LCR22-A was described in 2.3% of 22q11.2DS subjects and it was termed LCR22-A+ (Guo et al. 2018). All the breakpoints were localized to a 12kb segmental duplication, thus acting as a hotspot for meiotic rearrangements. Recombination with LCR22-B or LCR22-D results in a 1.3Mb LCR22-A+/B or a 2.8Mb LCR22-A+/D deletion, respectively (Guo et al. 2018).

Aside from the 22q11.2 deletions with endpoints within the LCR22s, atypical 22q11.2 deletions have been described with at least one breakpoint in the unique sequence between LCR22s (Beaujard et al. 2009). In the majority of published atypical deletions, one of the breakpoints resides in an LCR22, while the second breakpoint is located in unique sequence between the LCR22s. In only two of the reported cases were both breakpoints nested in unique sequences (Amati et al. 1999; Weksberg et al. 2007). Thus far, those atypical deletions have been detected with standard techniques including STR marker analysis, SNP arrays, and FISH. Probes covering the unique 22q11.2 region were used in the FISH assays, complementary to the commercial probes (TUPLE1, ARSA), which only detect 22q11.2 deletions involving the HIRA gene. Consequently, the breakpoints were never cloned and sequenced, and the mechanisms underlying those rearrangements remained unclear (Beaujard et al. 2009; Amati et al. 1999; Weksberg et al. 2007; Carlson et al. 1997; Levy

et al. 1995; McQuade et al. 1999; Nogueira et al. 2008; O'Donnell et al. 1997; Shaikh et al. 2000; Uddin et al. 2006).

In this study, we leveraged a large-scale, whole-genome sequencing study on >1500 subjects with 22q11.2DS (Gur et al. 2017) to map the atypical deletion breakpoints. These subjects were part of a large international consortium referred to as the IBBC (Gur et al. 2017). Six individuals were found to have atypical deletions with a proximal breakpoint between LCR22-A and LCR22-B (**Figure 6.1**). To improve our understanding of the mechanisms causing those atypical deletions and possibly the driving forces of NAHR in the common deletions, we charted the rearrangements at sequence resolution by cloning and sequencing the breakpoints. In addition, fiber-FISH was applied to resolve the complex architecture of the rearranged alleles. The deletions could be divided into two groups, where the first one provides signatures of replication-based mechanisms at the breakpoints. The second group is characterized by an AHR preceded by an inversion, which is, to our knowledge, not described yet.

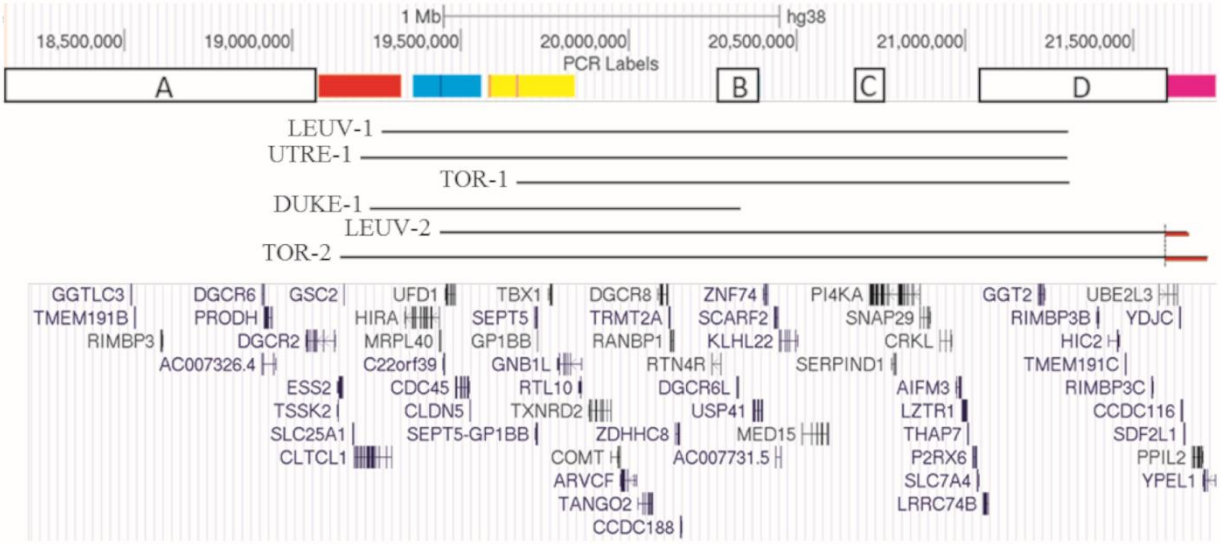


Figure 6.1: Overview of atypical deletion lengths and genes in the 22q11.2 locus. *UCSC Genome Browser screenshot with tracks for BACs and genes (GENCODE v29) in the 22q11.2 region. To visualize atypical breakpoints in the unique region between LCR22-A and LCR22-B, the locus is covered with labeled BAC probes: CH17-203M7 (red), CH17-6O11 (cyan) and CH17-395B16 (yellow). Distal from LCR22-D, the BAC probe RP11-354K13 (magenta) was added to visualize the inversion rearrangement in LEUV-2 and TOR-2. Deletion and duplication sizes of the patients are visualized using black and red lines, respectively.*

6.2 Results

6.2.1 Non-recurrent, atypical 22q11.2DS breakpoint regions detected by coverage plotting

Affymetrix 6.0 SNP microarray and whole genome sequencing analyses of the IBBC cohort were used to identify deletion sizes. Six individuals (LEUV-1, UTRE-1, TOR-1, DUKE-1, LEUV-2 and TOR-2) harbored atypical deletions with a proximal breakpoint between LCR22-A and

LCR22-B (**Figure 6.1**). Analysis of the coverage plots of these six individuals, based on a BWA-MEM alignment against hg38 (Li 2013), confirmed the presence of atypical deletions (**Figure 6.2**). Phenotypic information of the probands is provided in **Supplementary Table S6.1**.

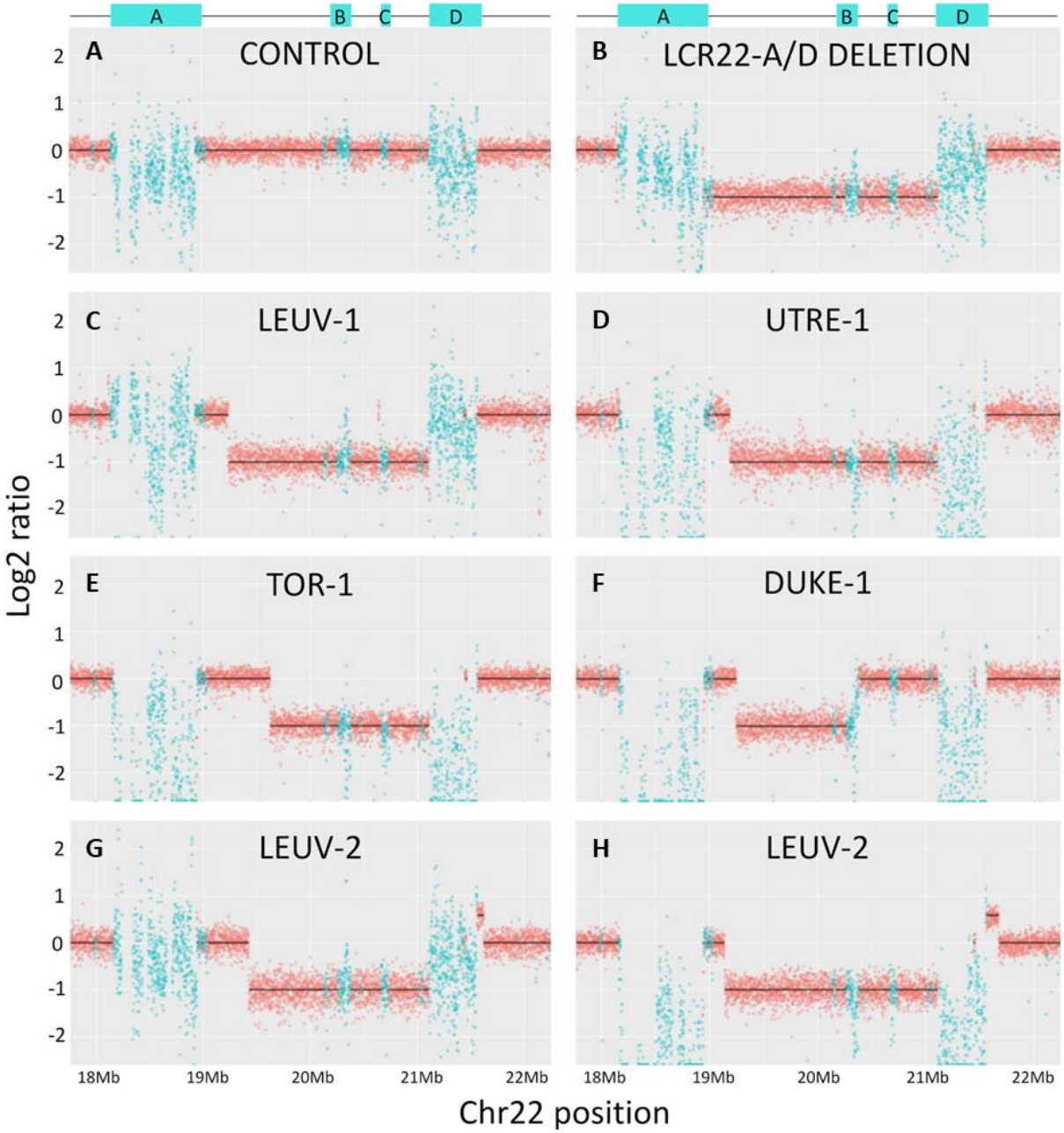


***Figure 6.2: Coverage plots of 22q11.2DS patients with an atypical deletion (hg38**). Overview of coverage plots with log2 ratio depicted in the x-axis and chr22 position showed in the y-axis. Blue regions indicate LCR22s, with LCR22-A for the largest proximal box, followed by smaller boxes LCR22-B and LCR22-C, to end with the distal LCR22-D. Coverage in the blue regions is not representative, since the high inter- and intra-chromosomal duplication events hamper correct mapping of these repeats. (A) Coverage plot of a control individual, not carrying a 22q11.2 deletion. (B) Coverage plot of an individual harboring a common 3Mb LCR22-A/D deletion. (C-H) Coverage plots of the six individuals with an atypical 22q11.2 deletion.*

Coverage plots were generated for a control individual without deletion (**Figure 6.2A**), an individual carrying the common 3Mb 22q11.2 deletion (**Figure 6.2B**) and the six individuals carrying atypical 22q11.2 deletions (**Figure 6.2C-H**). The log2 ratios of the reads within

the LCR22 blocks are not representative, due to the presence of paralogous sequences. Different deletion types were compared based on the coverage in the unique sequence surrounding the LCR22s. In a typical 3Mb 22q11.2 deletion between LCR22-A and -D, the log2 ratio drops from 0 to -1 over LCR22-A and increases over LCR22-D to the diploid state of 0 (**Figure 6.2B**). In these atypical subjects, proximal coverage drops are observed in the unique, non-LCR22 region between LCR22-A and -B, rather than being embedded in LCR22-A (**Figure 6.2C-H**). The log2 ratio of subjects LEUV-1, UTRE-1, TOR-1 and DUKE-1 increases to the chromosome's average distally from an LCR22, indicative of a deletion with a typical distal breakpoint embedded in the LCR22. In individuals LEUV-1, UTRE-1 and TOR-1 log2 ratio is -1 up to LCR22-D (**Figure 6.2C-E**). Similarly, log2 ratio is -1 up to LCR22-B for individual DUKE-1 (**Figure 6.2F**). However, in subjects LEUV-2 and TOR-2 the plots represent an increase of the log2 ratio to 0.58 in part of the unique sequence distal from LCR22-D (**Figure 6.2G-H**), before dropping to the normal diploid state. This suggests the presence of a duplication of this distal part. Hence, coverage plots of the whole-genome sequencing data uncovered two subtypes of atypical deletions and defined the global rearrangement regions.

6.2.2 Sequence resolution mapping of breakpoints

Proximal breakpoints were then refined based on whole-genome sequencing data (**Table 6.1**). The nucleotide position of the breakpoint is corresponding to the coverage drop from diploid to haploid by visual inspection of the data in Integrative Genomics Viewer (IGV) (Robinson et al. 2011). In individuals LEUV-1, UTRE-1, TOR-1 and DUKE-1, this proximal breakpoint was precisely mapped to chr22:19,251,190; chr22:19,184,629; chr22:19,627,753; and chr22:19,244,529, respectively (hg38). Hence, all proximal breakpoints differ in coordinates.

Table 6.1: Rearrangement-spanning read pair analysis of whole genome sequencing data

Patient	Coverage drop	Nucleotide position (hg38)	Gene	Fiber-FISH probe/BAC
LEUV-1	Proximal Distal	Chr22:19,251,190 LCR22-D (1)	CLTCL1 GGT2	CH17-203M7 Proximal D7
UTRE-1	Proximal Distal	Chr22:19,184,629 LCR22-D (1)	CLTCL1 GGT2	CH17-203M7 Proximal D7
TOR-1	Proximal Distal	Chr22:19,627,753 LCR22-D (2)	/ /	CH17-395B16 D3
DUKE-1	Proximal Distal	Chr22:19,244,529 LCR22-B (1)	CLTCL1 /	CH17-203M7 Proximal B2
LEUV-2	Proximal Distal	Chr22:19,427,384 Chr22:21,625,347	HIRA /	CH17-6O11 RP11-354K13
TOR-2	Proximal Distal	Chr22:19,140,370 Chr22:21,674,345	ESS2 PPIL2	CH17-203M7 RP11-354K13

Overview of the exact coverage drop positions observed in Figure 2. Exact nucleotide positions in hg38 are presented, with additional information of annotated genes, BACs and fiber-FISH probes in the

locus. For breakpoints in the LCR22s, the specific LCR22 is given, together with the number of mapping locations in that LCR22.

The genomes of the subjects were paired-end sequenced, allowing for an accurate detection of insertions, deletions and inversions based on the general fragment length of the library and orientation of the reads. Read pairs can be mapped with the BLAST-like alignment tool (BLAT) in USCS Genome Browser (Kent 2002). Primers were developed to clone the breakpoint (**Supplementary Table S6.2**, **Supplementary Figure S6.1**). Since the distal breakpoints of subjects LEUV-1, UTRE-1, TOR-1 and DUKE-1, are in LCR22 repeat sequence, we expected several BLAT results for the breakpoint-spanning read pair sequences. For subject TOR-1, these sequences match to four regions in LCR22-A and two regions in LCR22-D in hg38, all within probe D3 of the fiber-FISH pattern. Therefore, the forward primer was designed in the unique sequence proximal from the breakpoint, the reverse primer in this LCR22 sequence. The generated PCR product, specific for this atypical patient, was Sanger sequenced (**Supplementary Figure S6.1B**). The first part mapped to the unique sequence predicted by the whole genome sequencing data with an 18bp deletion compared to the reference genome, the last part of the sequence can be mapped to LCR22 subunit D3. This 18bp deletion is a known polymorphism in the population (rs530634277), inherited from the father, who is the parent-of-origin in whom the rearrangement occurred. Both sides of the breakpoint locus share a homologous region of 132bp. Therefore, we were not able to exactly pinpoint the nucleotide position of breakpoint junction. For subject UTRE-1, there was only one BLAT results in LCR22-D, proximal from the yellow D7 fiber-FISH probe. The PCR generated a patient-specific product (**Supplementary Figure S6.1C**). Consecutive Sanger sequencing unraveled the presence of unique 22q11.2 sequence, followed by a fragment mapping on the negative strand in LCR22-D (with an internal deletion of 130bp), ended by sequence mapping to LCR22-D on the positive strand. No DNA nor cell line was available for subject DUKE-1 to validate and clone the breakpoint.

BLAT mapping of the breakpoint-spanning reads of proband LEUV-1 matched one position in LCR22-D. Notably, both sequenced ends of the read mapped in the same orientation on hg38. This observation is indicative for the inversion of a DNA segment, creating a read pair with two sequences in forward orientation with respect to the reference genome. Hence, the reverse breakpoint-cloning primer had to be in the same orientation as the forward primer, allowing cloning of the patient-specific breakpoint with the inversion as a prerequisite. The first part of the Sanger sequenced fragment mapped to the unique 22q11.2 region, the last part to LCR22 sequence. In addition, breakpoint-spanning reads feature a polyA insertion. Illumina and Sanger sequencing encounter problems to sequence this long stretch of adenine nucleotides, since the polymerase is making mistakes in this repetitive nature. Therefore, PacBio SMRT sequencing was performed to exactly calculate the length of the polyA insertion. Long read sequencing of the breakpoint-spanning amplicon validated the presence of a 22bp polyA at the rearrangement breakpoint, which is neither present at the

proximal, nor the distal breakpoint coordinates in the reference genome (**Supplementary Figure S6.2A**). Sanger sequencing of the parental control products showed the absence of this polyA segment in the parental chromosomes. Hence, the insertion can be considered as a rearrangement-related event.

The coverage plots of subjects LEUV-2 and TOR-2 present an extra duplication distal from LCR22-D. Whole genome sequencing data analysis revealed proximal coverage drops uniquely mapping to chr22:19,427,384 and chr22:19,140,370, respectively. The distal trisomic to disomic coverage drop observed in the coverage plots is uniquely mapped to chr22:21,625,347 and chr22:21,674,345, respectively. In the analysis of the whole sequencing data, read pairs were observed where the first fragment mapped to this proximal coverage drop and the second fragment mapped to the distal coverage drop. In addition, both fragments of the read pair mapped in a forward orientation with respect to the reference genome. This link can be explained by the presence of an inversion between both loci. Primers were designed to validate this inversion junction in subjects TOR-2 and LEUV-2 (**Supplementary Figure S6.1D-E**). Both sequences were blunt-end ligated without the presence of insertions or deletions.

Detailed read pair analysis of the whole genome sequencing data allowed us to pinpoint exact deletion breakpoints in the first subgroup, and to identify an inversion rearrangement in the second subgroup. Nevertheless, overall architecture of the region remained unclear.

6.2.3 Fiber-FISH assemblies uncover the structural composition of the rearranged 22q11.2 allele

To overcome the biased mapping of sequencing data and resolve the structure of these atypical rearranged alleles in five subjects (LEUV-1, LEUV-2, TOR-1, TOR-2 and UTRE-1), a *de novo* assembly was performed by using fiber-FISH (Yadav and Sharma 2019). In this technique, long DNA molecules (>200kb) were extracted from cells and stretched onto coverslips. These fibers were subsequently hybridized with labeled probes targeting the LCR22 subunits (**Figure 6.3A-B**). In contrast to current sequencing technologies, fiber-FISH was shown to be capable of spanning the LCR22s (Demaerel et al. 2019). To visualize the proximal breakpoints in the region between LCR22-A and -B in the atypical subjects, labeled BAC probes were added to the standard probe composition to visualize the unique sequence in the 22q11.2 locus (**Figure 6.1**).

In the *de novo* assembled patterns, one represents that of a normal LCR22-B, -C and -D allele on the remaining, non-deleted allele. Other patterns are indicative for LCR22-A heterozygosity (data not shown). The proximal breakpoint (chr22:19,627,753) of patient TOR-1 is embedded in the yellow-labeled BAC CH17-395B16, which is directly fused to the probe pattern of LCR22-D (**Figure 6.3C**). A magenta BAC RP11-354K13 was added distally from LCR22-D. The rearranged allele features CH17-395B16, interrupted by the probe

composition D3 (cyan), B2 (red), A3 (cyan), D7 (yellow), A1 (blue), A2 (green), D2 (red), A3 (cyan), B2 (red), D3 (cyan), D6 (green), D5 (magenta), D6 (green) and magenta BAC RP11-354K13 (**Figure 6.3C**). The breakpoint locus observed within this fiber-FISH pattern is concordant with breakpoint-spanning reads mapping to LCR22-D. For individual UTRE-1, a similar pattern was assembled, for which the red BAC CH17-203M7 was directly fused to a pattern suggestive for LCR22-D (**Figure 6.3D**). The LCR22-D breakpoint observed by fiber-FISH was concordant with the breakpoint suggested by the sequencing results.

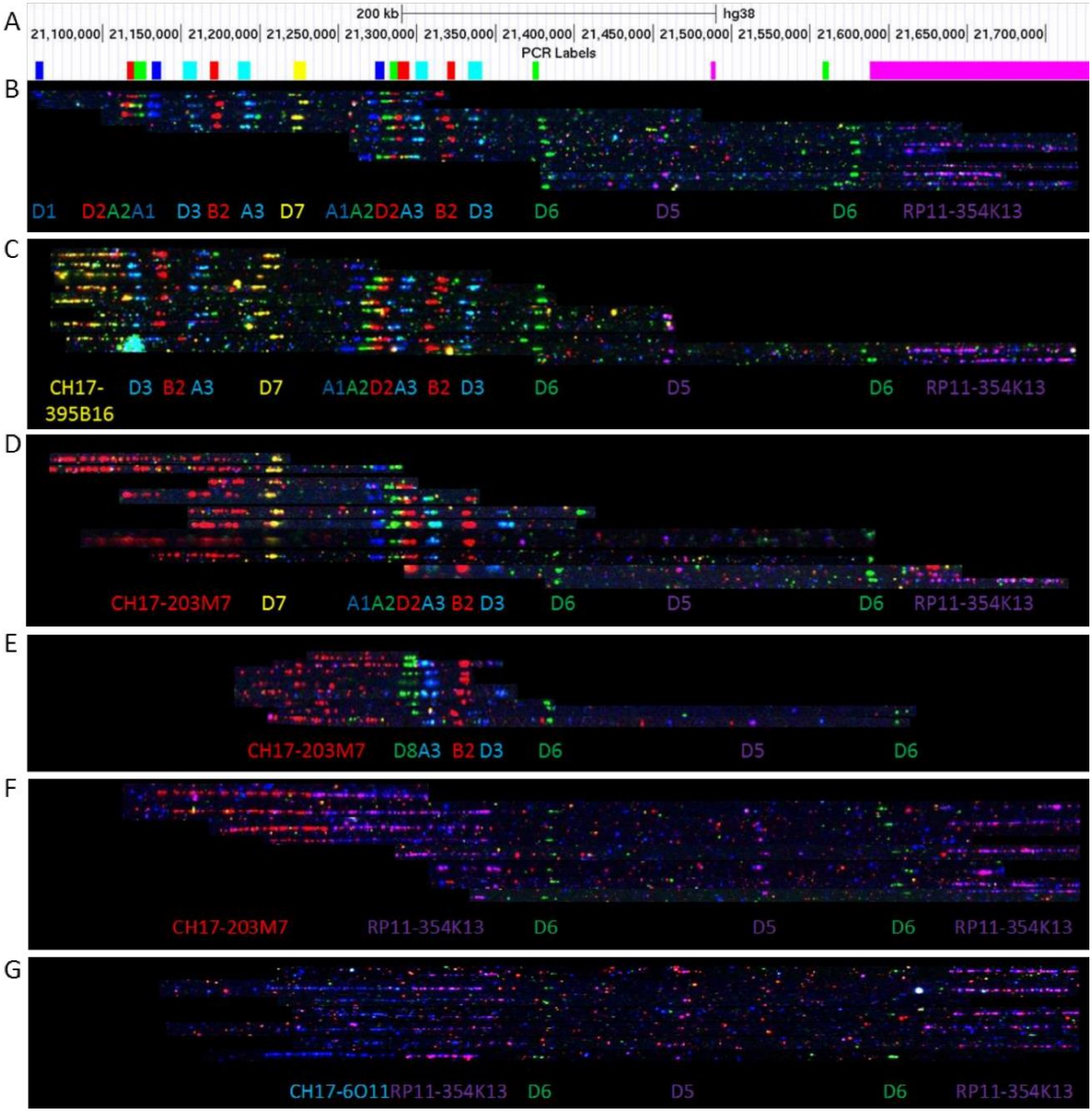


Figure 6.3: Fiber-FISH analysis of the rearranged allele in atypical 22q11.2DS patients. *(A) UCSC Genome Browser screenshot with track for fiber-FISH probe composition of LCR22-D. (B) Fiber-FISH pattern of a normal, non-rearranged LCR22-D allele. (C) Fiber-FISH results for the rearranged allele in patient TOR-1. (D) The de novo assembly of the rearranged allele in individual UTRE-1. (E) In individual LEUV-1, the probe set was supplemented with green probe D8 to visualize a supposed inversion. (F) Allele de novo assembly of individual TOR-2 uncovered the juxtaposition of BACs CH17-203M7 (red) and RP11-354K13 (magenta). (G) The same observation was made for individual LEUV-2, with a fusion of BACs CH17-6O11 (cyan) and RP11-354K13 (magenta).*

To allow the detection of an internal LCR22-D inversion in proband LEUV-1, probe D8 was designed within LCR22-D to locally increase the pattern density. In the fiber-FISH pattern of the rearranged allele, red BAC CH17-203M7 is directly proceeded with probes D8 (green), A3 (cyan), B2 (red), D3 (cyan), D6 (green), D5 (magenta) and D6 (green) (**Figure 6.3E**). This pattern suggests an indirect orientation of SD22-4 (probe order D2, A2, A1 and D8 from proximal to distal) in LCR22-D (Demaerel et al. 2019), prior to the deleterious rearrangement that fused CH17-203M7 to LCR22-D (**Supplementary Figure S6.2C-D**). The presence of this inversion embedded in the rearranged LCR22-D is validated by the orientations of the read pairs spanning the rearrangement in the whole genome sequencing data of proband LEUV-1. This SD22-4 orientation in LCR22-D has been observed in 6% of the population (Demaerel et al. 2019). However, no fused BAC signals were present, nor was SD22-4 found to be inverted in the LCR22-D alleles of the parent-of-origin, suggesting that inversion and deletion event occurred *de novo* (**Supplementary Figure S6.2A-B**).

To accommodate the inversion breakpoints in individual TOR-2, BAC probes CH17-203M7 (red) and RP11-354K13 (magenta) were added to the LCR22 probe set. The rearranged allele displays a fusion of BACs CH17-203M7 and a fragment of RP11-354K13 immediately followed by probes D6 (green), D5 (magenta), D6 (green) and a full-length magenta signal of RP11-354K13 (**Figure 6.3F**). This probe composition suggests AHR to have occurred between two LCR22-D alleles. However, prior to or concomitant with the deleterious rearrangement, an inversion between chr22:19,140,370 and chr22:21,674,354 occurred on one allele, resulting in segmental duplications oriented in the same direction. Consequently, the rearranged LCR22-D harbors RP11-354K13 on both sides. A similar pattern was observed for patient LEUV-2 with a fusion of the cyan (CH17-6O11) and magenta (RP11-354K13) BAC (**Figure 6.3G**).

We reasoned that the inversion would have preceded AHR and might be present in one of the parents, driving the 22q11.2 rearrangement. Alternatively, the inversion could have arisen *de novo*. To investigate this, PCR with primers spanning the inversion breakpoint was performed on DNA derived from peripheral blood lymphocytes from both parents of TOR-2 and LEUV-2. In none, these inversion-specific PCRs produced a positive amplicon (**Supplementary Figure S6.1D-E**). Similarly, fiber-FISH haplotypes assembled for the parental EBV cell lines using both LCR22 subunit probes and BAC probes were normal and the patterns were concordant with previous observations (Demaerel et al. 2019). No tissues other than EBV cell lines were tested.

6.3 Discussion

Six out of >1500 individuals from the IBBC cohort carried an atypical nested 22q11.2 deletion. These cases are incidental findings, since the presence of a nested and/or atypical deletion was one of the exclusion criteria of the IBBC project. However, samples were included accidentally or due to the limited resolution of the FISH assays. The overall

incidence of these atypical deletions in the general 22q11.2DS population was estimated at 2% (Campbell et al. 2018). The phenotypic spectrum of this cohort with atypical 22q11.2 deletions is not significantly different of that observed in patients with the common 3Mb 22q11.2DS.

We mapped the rearranged chromosome 22 in these six IBBC individuals to characterize those rearrangements and their causal mechanism. Analyses of read pairs spanning the rearrangements identified the unique deletion breakpoint regions in four (LEUV-1, TOR-1, UTRE-1 and DUKE-1) subjects. Unexpectedly, in two individuals (LEUV-2 and TOR-2) an inversion of a fragment including LCR22-B, -C and -D was present. Since no sequencing technology is capable of spanning the LCR22s, the overarching structure of those rearranged alleles remained elusive. Using fiber-FISH, the deletion breakpoint regions were mapped at subunit resolution within the LCR22s for three of these individuals. In the two probands with an inversion spanning LCR22-B, -C and -D, we hypothesize the final rearrangement is a consequence of two separate events: first an inversion followed by an inter-chromosomal AHR between differently oriented LCR22-D alleles (**Figure 6.4A-B**). Additionally, an internal LCR22-D inversion of SD22-4 (Demaerel et al. 2019) was present in proband LEUV-1. Surprisingly, these inversions were not observed in EBV cell lines of any of the parents. Hence, they occurred in the germline precursor or they coincide with the homologous recombination.

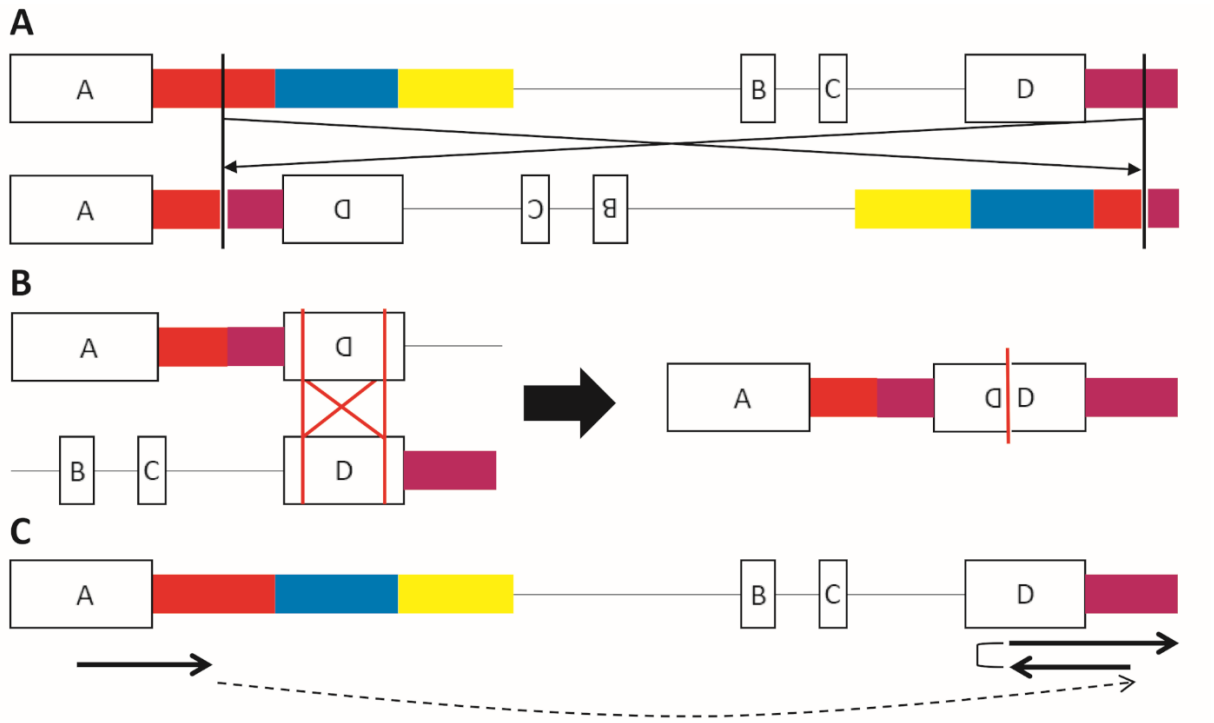


Figure 6.4: Mechanisms to create the rearranged allele in proband TOR-2. *To obtain the observed fiber-FISH pattern of TOR-2 (Figure 3F), a two-step mechanism is supposed. (A) First, an inversion of the normal allele results in the juxtaposition of the red and magenta BAC probe, distal from LCR22-A. (B) As a second event, the AHR between two LCR22-D repeats of a different allele causes a deletion. Hence, partial presence of the red BAC probe, deletion of the locus, and duplication of the magenta BAC probe is explained. (C) Alternatively, FoSTeS/MMBIR replication-dependent mechanisms generate such complex architectures.*

For AHR and NAHR to occur, at least 300bp of perfect sequence identify are required (Gu et al. 2008). Since LCR22s contain several shared subunits they are common substrates for NAHR. Absence of LCR22 sequence at one side of the atypical rearrangement suggests that other mechanisms drive these atypical deletions. Non-homologous end-joining or microhomology-mediated end-joining occasionally introduce complex rearrangements (Rodgers and Mcvey 2016). Both repair mechanisms are invoked if spontaneous double strand breaks occur in a cell. Hence, to explain the observed rearrangements, two or more double strand breaks must have occurred simultaneously prior to erroneous repair. Alternatively, replication-based mechanisms as FoSTeS or MMBIR could have led to these rearrangements (Ottaviani et al. 2014). Resulting breakpoint junctions of these events are characterized by signatures as insertions, deletions, inversion and microhomology traces (Carvalho et al. 2009). Several consecutive fork stalls do occur, and generate complex rearrangement patterns (Ottaviani et al. 2014). Microhomology of up to 132bp was detected in TOR-1 and 17 nucleotides in LEUV-1 between the two breakpoint regions. Additionally, 30 adenine base pairs were inserted at the breakpoint of LEUV-1, consistent with polymerase slippage events (Beck et al. 2019). PolyA insertions were previously linked to LINE1 endonuclease-dependent *de novo* insertions (Wimmer et al. 2011). However, there is no evidence for a *de novo* insertion of LINE1 sequence at these breakpoints. The complex architecture of the rearranged fragment of UTRE-1 (**Supplementary Figure S1C**) can be explained by the FoSTeS mechanism, where template switches occurred during DNA replication(Carvalho et al. 2009). Although the LCR22s are not directly implicated in the (proximal) breakpoints of these non-recurrent atypical deletions, Carvalho et al. (2013) suggests a mediating role for LCRs in general as a destabilizing factor making the locus sensitive to rearrangements.

In two out of six individuals, rearranged allele patterns suggest an inversion preceded the deletion (**Figure 6.4A**). Subsequent to these inversions, AHR between the inverted and a normal LCR22-D produced the observed rearranged allele (**Figure 6.4B**). Although an inversions was present, parts of the LCR22-D locus still have the same orientation and can be considered as substrates for AHR. Fiber-FISH assemblies suggest that the recombination has taken place distally in LCR22-D (**Figure 6.3F-G**). Since these breakpoints are embedded within LCR22-D, the breakpoint region could not be delineated at sequence resolution with short-read data only. Read pairs in TOR-2 and LEUV-2 only explain the observed inversion, since the mapping of the AHR reads within the LCR22s is biased. In the other atypical nested deletions, reads do map to the deleterious event, which occurred between the unique sequence and an LCR22. In LEUV-1, however, the breakpoint-spanning mates feature the same orientation compared to the reference genome hg38. This is indicative for the inversion of the SD22-4 duplicon in LCR22-D (Demaerel et al. 2019).

Since the inversions are hypothesized to have occurred prior to the deleterious rearrangements, these could be present in the parent-of-origin as well. Inversion

polymorphisms between LCRs are frequently observed in the parent-of-origin of patients with genomic disorders, predisposing these alleles to NAHR (Shaw and Lupski 2004; Osborne et al. 2001; Gimelli et al. 2003; Jafri et al. 2011). However, none of the inversion breakpoint PCRs, nor fiber-FISH did detect the inversions in the parents. This does not rule out the possibility that germline mosaicism for the inversions could be present and subsequently, NAHR between a normal and inverted chromosome22q11.2 produced these alleles. Alternatively, FoSTeS/MMBIR could have generated these complex rearrangements during replication, which could explain why the parents are not carriers of the inversions observed in the probands (**Figure 6.4C**). In this model, it would be coincidental that the template switching occurred at the homologous region within LCR22-D.

In summary, fiber-FISH allowed us to validate six atypical deletions detected by whole-genome sequencing, and map the rearrangements within the LCR22s. In two cases, the rearrangements are not merely deletions but are complex rearrangements characterized by the presence of a deletion, duplication and an inversion. Scrutinizing these breakpoint regions paves the way to enhance our understanding of the LCR22 architecture and to a better genotype-phenotype correlation.

6.4 Materials & Methods

Patient resource

The IBBC consortium performed MLPA, Affymetrix 6.0 SNP microarrays, and whole-genome sequencing on >1500 patients with a 22q11.2 deletion (Gur et al. 2017). Patients were diagnosed with 22q11.2DS using the FISH assay with TUPLE1/ARSA probes (Abbot Molecular, Abbot Park, Illinois, USA), the MLPA SALSA P250 DiGeorge diagnostic probe kit (MRC-Holland) or the CytoSure Constitutional v3 (4x180k) (OGT, Oxfordshire, UK).

Six patients (LEUV-1, LEUV-2, TOR-1, TOR-2, UTRE-1 and DUKE-1) were identified with an atypical deletion in the IBBC cohort. For this study, EBV cell lines of two proband-parent trios were recruited from Leuven (probands LEUV-1 and LEUV-2), one duo from Utrecht (proband UTRE-1 with her father) and two trios from Toronto (probands TOR-1 and TOR-2). Cell line transformation was carried out in Leuven for the families from Leuven and Utrecht, and in Toronto for the Toronto families. Genomic DNA was extracted from the cell lines with the DNeasy Blood and Tissue kit (Qiagen). For patient DUKE-1 (recruited from Duke), no cell line nor DNA was available for experiments. An informed consent was signed by all participants of the study, regarding the use of their EBV cell lines and DNA for sequencing and genotyping purposes. The study was approved by the Medical Ethics Committee of the University Hospital/KU Leuven (S52418) and of the University Medical Centre of Utrecht (08/354). The Institutional Review Board approved the research protocol for the study of the Clinical Genetics Research Program at the Centre for Addiction and Mental Health (REB# 114/2001-02).

Refined whole genome sequencing analysis

The patients included in the IBBC cohort were whole genome sequenced at the HudsonAlpha Genome Sequencing Center (Huntsville, AL) on an Illumina HiSeq2500 platform for the first 100 samples (including sample TOR-1) and on Illumina HiSeq X Ten for the remaining samples (including samples LEUV-1, LEUV-2, TOR-2, UTRE-1, DUKE-1). HiSeq2500 runs produced 100bp paired-end reads and HiSeq X Ten runs produced 151bp paired-end reads. Reads were aligned to genome build hg38 with BWA-MEM (Li 2013) to enable manual inspection of breakpoints using read pair analysis, visualized in the IGV (Robinson et al. 2011). The average coverage depth of diploid loci on chromosome 22 is 62X, 47X, 35X, 36X, 46X and 37X for patient LEUV-1, UTRE-1, TOR-1, DUKE-1, LEUV-2 and TOR-2, respectively.

Copy number variations of chromosome 22 were detected using the Control-FREEC tool (Boeva et al. 2011). Default settings were applied, except for windows size which was set to 10kb (**Figure 6.2**). Obtained copy number ratio values and called segments of CNVs were then used to generate the plots with the R package ggplot2 (Wickham 2009). Coverage plots should therefore show a 50% reduction of the unique sequence coverage depth for deletions, present as a drop from 0 to -1 on the log2 ratio scale. Reciprocally, duplicated sequence is observed as a 50% coverage depth increase, concordant with a log2 ratio increase to 0.58. Within the LCR22s, significant sequence paralogs lead to collapsing read mapping on the (incomplete) reference genome. In addition, interindividual read depth variability hampers the identification of a narrow breakpoint region within the repeats.

Fiber-FISH

DNA fibers were stretched onto coverslips using the Genomic Vision extraction kit and combing system (Genomic Vision, Paris, France). Coverslip probe hybridization and *de novo* allele assembly was performed as previously described (Demaerel et al. 2019). The standard LCR22 probe pattern consists of fourteen fluorescent probes, designed using the characterized subunit sequences library. This probe set was supplemented with BAC probes to visualize unique sequence amid the LCR22s (Figure 1). BAC DNA was extracted from BAC clones (BacPac Resources, CHORI, Oakland) using the Nucleobond Xtra BAC kit (Macherey-Nagel). Subsequent labeling of the BAC probes was performed with the Bioprime DNA labeling system (Invitrogen). Labeled BAC probes are CH17-203M7 (red), CH17-6O11 (cyan), CH17-395B16 (yellow) and RP11-354K13 (magenta). An additional probe D8 was developed to validate the presence of an internal LCR22-D inversion of SD22-4 (Demaerel et al. 2019) in LEUV-1. The forward primer (5'-GTCTTGTCAAGGTGGAATGA-3') and reverse primer (5'-TCTGTCTCTGTGCCTCAGTT-3') produced an amplicon of 7516bp, using the TAKARA LA v2 kit (Takara Bio Inc.). The probe was subsequently labeled with fluorescein-dUTP, creating a pseudocolored green signal on the slides (**Figure 6.3E**).

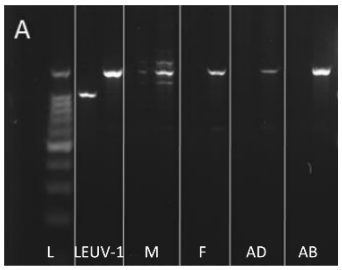
PCR validation of patient-specific (inversion) breakpoints

To validate the positions of rearrangement breakpoints, primer pairs were generated for PCR amplification based on sequencing reads spanning the breakpoint (**Supplementary Table S6.2**). To determine recurrence of breakpoints, the reaction was performed on DNA of the patient, one or two parents, and additional 22q11.2DS patients with a different breakpoint location (Supplementary Figure S1). An additional primer pair was developed to generate a control product on the non-rearranged allele of the patient or on both alleles of individuals without the specific deletion. A reaction mixture of 50µL was prepared according to the Taq DNA polymerase protocol (Invitrogen). The amplification reaction started with an initial denaturation at 94°C for 3min, followed by 25/30 cycles of 45s at 94°C (denaturation), 30s at 57°C (annealing) and 70s at 72°C (extension). A final elongation step of 10min at 72°C was included. For the LEUV-1 PCR, the extension time was 90s. Product presence or absence was then examined on a 2% agarose gel.

Long-read sequencing of breakpoint amplicon of LEUV-1

The breakpoint amplicon generated by primer pair 'Breakpoint LEUV-1' (**Supplementary Table S6.2)** was prepared for long-read sequencing according to the Template Preparation and Sequencing protocol (Template Prep kit 3.0, Pacific Biosciences, Menlo Park, CA). This library was spiked in on a single SMRT cell on a PacBio RSII using a DNA/polymerase binding kit P6 v2 (Pacific Biosciences, Menlo Park, CA) and DNA Sequencing Reagent kit 4.0 v2 (Pacific Biosciences, Menlo Park, CA). The analysis was performed using the RS_Long_Amplicon_Analysis.1 pipeline (software: Smrtanalysis_2.3.0) with the following settings: minimum sub-read length 950bp, default barcode score 22, and default amount of sub-reads 2000. On this SMRT cell 27909 reads were assigned to the barcode of this amplicon.

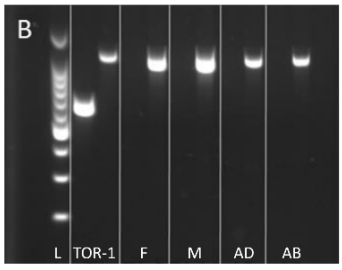
6.5 Supplementary Materials



GGGATTACAGGCATGAGTCACTGAGCCCAGCTAGTACCACATTCTGGATCACTGTAGGTTTGTGGCAAGTTTTAAAATTGGGAAGTG
TAAGTACTCCAACTTTTATTTGTTTTATGGTTGTTTTGGCTATTCTGGGTTCCTTGCAATTCCACATAAATTTTAGGATCATCTTGCCAA
CTTCTACAAAGAAGCTAGCTGGTATTCCAATAGAGATTATGTTGAATTTGTAGACCAATTTGAGGAGTACTTCCATCCTAAAAGTATTA
AATCTTCGAGTCCATGAACATGCAATATTTTTCGATTTAGATCTTCTTTAATTACTTCCAAACATGTTTTGTAGTTTTCAGAATGTAACTT
TTGCACTTTTT (Chr22:19,251,190)

TTTTTTTTTTTTTTTTTTTTT (polyA/T insertion, 22bp)

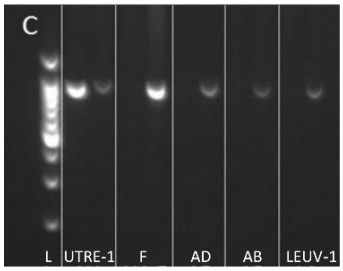
TTTGAGACGGAGTCTCGCTCTGTCGCCCAGGCTGGAGTGCAGTGGCGGGATCTCGGCTCACTGCAAGCTCCGCCTCCCGGGTTCAC
GCCATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCGCCCGCCACTACGCCCGGCTAATTTTTTGTATTTTTAGTAGAGG
CGGGGTTTCACCGTGATCGCGATCTCCTGACCTCGTGATCCGCCTGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGCATGAGCCAC
CGTGCCTGGCCTGTTTTGTTGATTTTAAAGCCCAGGGCAGTAGTCTTGGAAAATGTCCCACATCGTGGATATGATGTGCTACATCT
(LCR22-D)



TCCTGATGGAGTGCTGCCATTTGGAAGGGGTAACTTTCGTGGTCATTTTTTTAAATCATGGTAAAATATACATAACTACATAACATAAA
ATTTACCATTCTTTCTTTTCTTTTCTTTTCTTTTCAGACGGAG / (-18bp) / GCCGGACTGCAGTGGCGTTC (Chr22:19,627,753)

*TCTCGGCTCACTGCAAGCTCCGCCTCCCGTGTTCACGCCATTCTCCTGCGTCAGCCTCCCGAGCAGGGGACTATAGGCGCCCGCCACC
ACGCCTGGCTAATTTTTTGTATTTTTAGTAGAGACGGGGTTTCA (region of overlap)*

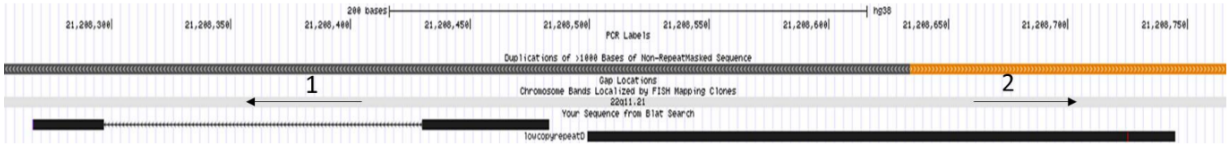
CCGTGTTAGCCAGCATGATCCCCCTCTTCTGACCTTGTGATCCACCCGCCTTGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCC
ACTGCGCCCGGCCTACCTGTGAACTTAATATCTTAATGTTTTAATGTTGCTAATCAGAATCCTTTTATTTCAACTTGAAAAACTGCCTTA
TAAGGCAGGTGCAGTGGTGATGAACTCCCTTAGAATTTTTTTGGTGGGAGTCTGGGAAAGACCTTATCATCTCTTTTTCATTTCTGAA
GGACAGCTTTACAAGTT (LCR22-D)

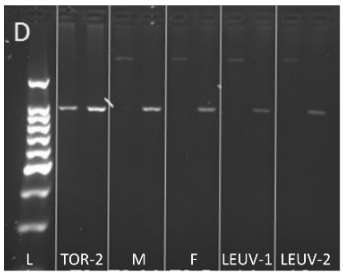


CTTGCCCAGGGCCACTCCACCTGGAGTTGGCCATGATCCA//CTGACTGTGACCTGTGCTCTCCAGGGATACCCTGTGGCGGTCCCA
GCTCTGCATGCCTGCCATCGCCTGCAGCACACACCTCTGAGTTCCGTTTGGGAGACCCCTGAGGACGCCCAGTGCCCACATCTGCAT
GGACCCGATGTGCTGAGCTGGCTCCGGAAAGAAGTCTCCCCCGCTGGCATGCTGCTCATCGATCGCTGCCTCCTGACCCACACTGTT
CCTACAAAGGCACCAGGATTCTGACCCTCAGGTGAGAGCTCCAGGTGCCTTTGGGTCTCCGCTTTGAGGACCTGTGACGCCCGCTC
ATCTGTCTC (Chr22:19,184,629)

TAGTGCCTGGGCCATCTGGAGCCCCTGTGACATGAGAGCCAAGCCCCCTGCTC / (-130bp) /
CTCCCACCCCCAGGCAGTGACTGCAGCCCTTAT (LCR22-D, negative strand, 1)

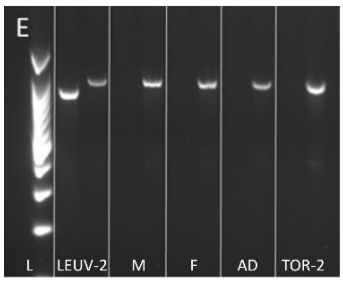
CATTGTCCACTCTGTGATGATCCAGGCCTCCAGCCCAGGATGCCCTGGGACCCCACACCGTGACTCAGTTTCTCCAACCCCTGGCCC
ACCTGGTCAATGTTTCTCTCCACTGTCGTGACGTTGGGCAGAAGCTTGTTGTGCAGCCGGGGCTCCTCCACGGCCCTCTTCACGTCA
TAGCCGAACCAGAGGTTGTAGATGATGGCCTGGGGCATGGGAGTGTGATCAGCGTGGCTTGGGGGCTGTGCA (LCR22-D, 2)





AAGAGGAGGACACCCAGGCCCAGGAATCATAKCCCCCAACATGCAGGGCTGGAATCCTCTCCCCTGGTCCTGGCTTCTGCCCATGA
AAGCCCACACTGCCAAGGACTCCCCACGGGGGCAGAATCCACCCACCCTGACTGCCCAGAGAGCCCTCAGGTGGTACCCTACAGAT
TCAGGGCCAGCTCTGCTCATGTGCTGGATGCCAAGGGAATTACGTGGTCTGATTCAGACTGGAGAAAGCAATACAGGTGTTAGAGC
AAGAGCTCGTGATCACAGGCAGGCTCCTCCCCCACCAAA (Chr22:19,140,370)

TGTTTTCAGTTTCTGTGGTGTGTCTGCAGTACAGCTCTTCCACAGCAAACTTTCCCATAAAACAAAGTCTGGACAACCGGAGGTGGG
AACTGCTGGGGGCGGGGCTGTGGGGAGACCCTGGAAGGGAGGCAGAGCTATGTGCTGCTGGGGAGCACCTCACTCAGGGCCAC
AATCCAGCAGCTGCTGCCACCTGGTGTGGGCTGGGCCCAGCTATGCCTACCCAGGCCCAGGGGAGACACAGCGGGAGCCTCCTGG
GGCTGGGCCCTAAGCTGGCAGCAGGTCCTTAAGCTGCCCTCCCCGCCCTCCGTGGGCCCTGCCTGCTGGCCAACCCAACTGAC
CGCATCTGCTCCTCTAACCACCTTCGTTCCTTCCTTTGTCATTCAACTACCCCCACCTCCCTACATCACTACAGGTGTGGAGGAGACCA
CCAGAGGGCCAGCAGGGTCCCTATTCAGACACAGGACCTGAGCAGAGAGACTCCCCAACACAACAGGCTCCTCCTGAGCCCCCAA
AAGTCACATGGGCCAGCCAGCCCTTGGCCCAGGAAC (Chr22:21,674,345)

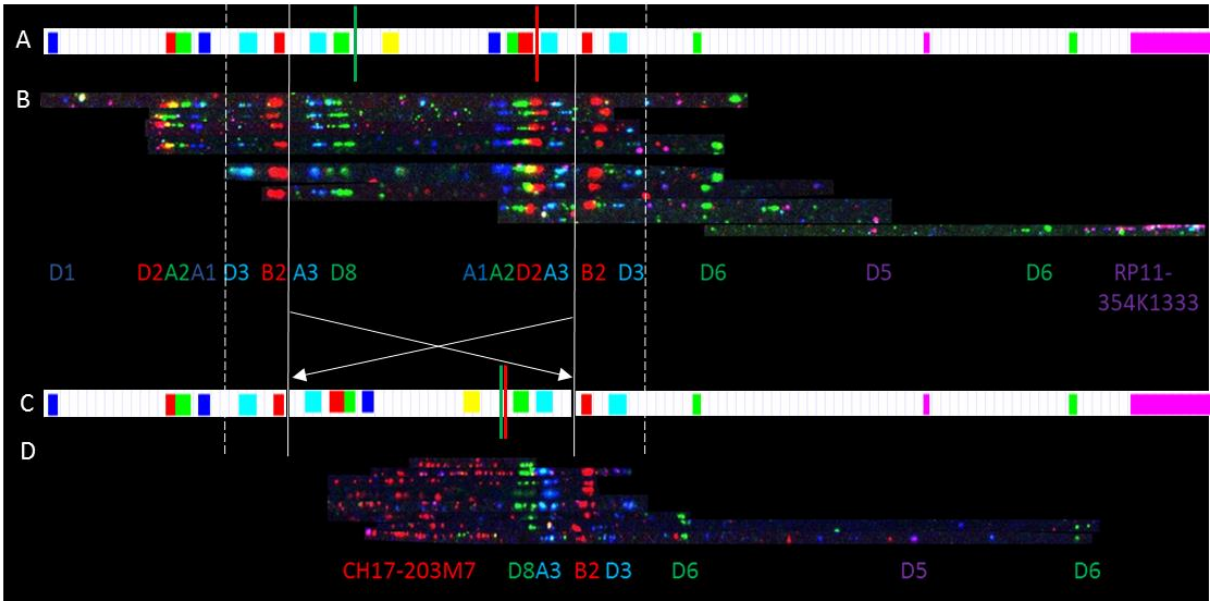


AAAAATCCACAGCAAAATCTCACCCAGTTTTCTTCTGACTCTCAGATGTACTCTGGTTATTTTCCTAAATCAACCTACTGTCTTCAGGC
ATAAGTCAGGAGAGAAATTGCTGAGAGAATTTGCATTTTCTGAGCTACACATGGAGCAGGTGAGGCAAAGAATTACTGACTTCATG
AACAGATTTAGGAAGAGATGGAAAAACTTTTTATGTCACATATTTAGTAAGCAGCAGGGCACAGAGACTCCAGAGCAGGATATCTG
ATTCCAAAAACAGAAGACAAAATTCTTGTCCTCAAGAAACTTACAGTCTAGTCGGAACTGTAAGACAAATACACCAAAAGGCAAATA
ACTTATTCAGAGCGGGGTTTCTCACCCTTAGCACTATTGATATTTGGGGCTAGATAATTCTCCATGAGTGGGCTGTCCTGCATGCTATA
CGATGTTTAACAGATTTGTCCCCTACCCACTTGATACCAGTAGCACGTACCCTGCCCAACCCTGTTGTGACAACCACAAATGTCTCTAG
ACATTGCCCTGTAGAACCACTGATCTAAAGCAACTGCCATGTTTATCTCTCAACCACTGCAATCATATAACCTCACCATTAGGATCCTTT
CCCTGCCAATTCTCAGACATAGGTCTGCAGTGCCCTCCGGCAGCA (Chr22:19,427,384)

CGGTGGTCCTTTCTACTCTGATCCTCTGAGGCCTCCAGGGTTCCTAGGGGCCCTCTATGAACAAGAACACGTTCCAATAATCTGGGA
AA (Chr22:21,625,347)

Supplementary Figure S6.1*: Rearrangement-specific PCR validation and product sequencing. Presence or absence of a breakpoint-specific PCR amplicon was evaluated in a 2% agarose gel. The BenchTop 100bp DNA ladder was used for size estimation. The first lane after the ladder is the patient lane, followed by one or two lanes for the parent(s). Additional lanes are other patients with different deletion or inversion breakpoints. For every individual a breakpoint and control reaction was performed. Breakpoint products are displayed on the left side of the individual's lane, control products on the right side. Primer pairs are listed in Supplementary Table S2. PCR products were Sanger sequenced and obtained sequences are displayed at the right side of the figure with corresponding genomic locations. (A) The deletion breakpoint-specific amplicon of individual LEUV-1 was sequenced*

with PacBio long-read sequencing technology. (B) Deletion breakpoint-specific and control PCR reactions of individual TOR-1. (C) The genomic organization of the breakpoint-specific amplicon of individual UTRE-1 is explained with a UCSC Genome Browser screenshot of the LCR22-D region. Numbers in the sequence box are corresponding to the fragments indicated in the UCSC Genome Browser screenshot. (D-E) PCR reactions over the inversion breakpoint and control region in patients TOR-2 (D) and LEUV-2 (E). L=ladder, M=mother, F=father, AD=patients with LCR22-A/D deletion, AB=patient with LCR22-A/B deletion.



Supplementary Figure S6.2: *Inversion event prior to deleterious rearrangement in LEUV-1. (A) White track displays the probe composition of a normal, non-rearranged LCR22-D allele. The red line indicates the breakpoint region deduced from the fiber-FISH map (**Figure 6.2D**). The green line represents the breakpoint location derived from the sequencing results. (B) The LCR22-D allele of the mother of LEUV-1, he parent-of-origin, displays the reference composition. (C) If the deletion is preceded by an inversion between the two white lines (white arrows), the expected and observed breakpoint coincide. White lines indicate the minimum size of the inversion, dotted white lines the maximum size, according to homology. (D) The hypothesized pattern of (C) corresponds to the mapped rearranged allele of patient LEUV-1.*

***Supplementary Table S6.1**: Overview of phenotypic features.*

	LEUV-1	UTRE-1	TOR-1	LEUV-2	TOR-2	DUKE-1
Pharyngeal	/	Cleft palate	Velopharyngeal insufficiency	Velopharyngeal insufficiency	Velopharyngeal insufficiency	/
Cardiac	Interrupted aortic arch type B, bicuspid aortic valve, ventricular septum defect	Atrium septum defect	Ventricular septum defect, patent ductus arteriosus, pulmonary valve stenosis	Perimembraneous ventricular septum defect, cervical aortic arch	/	Truncus arteriosus
Facial	Small tubular nose, smooth philtrum, thin upper lip, dysplastic ears	Thin upper lip, mild hypertelorism, bulbous nose tip, low placed ears	Retruded jaw, short, narrow and upslanting palpebral fissures	Prominent nose, short upslanting palpebral fissures, thin upper lip, broad nasal bridge	Narrow face, large nose with broad nasal bridge and bulbous tip, retrognathia	Hooded palpebral fissures, hypertelorism, small overfolded ears, pinched nasal tip
Cervical	Hypoplasia left hemi-arch of C1	/	Scoliosis	C2-C3 fusion	Moderate thoracic scoliosis	Unknown
Development	Mild developmental delay, poor visual perceptual skills (WISC-III testing @10y2m, FSIQ 76, VIQ 86, PIQ 71)	Mild-moderate intellectual disability (WISC-III testing @14y, FSIQ 48)	Mild intellectual disability (WAIS-R testing @24y, FSIQ 61)	Mild intellectual disability (WISC-III testing @13y, FSIQ 67, VIQ 75, PIQ 64)	Normal intellect (WAIS-III testing @19y, FSIQ 86, VIQ 91, PIQ 80)	Normal intellect (WISC-IV testing @13y, FSIQ 88)
Behavior	Socially shy, otherwise normal behavior, no psychiatric diagnosis	No formal diagnosis, autistic features (mostly rigidity), acoustic hallucinations @18-19y	Schizoaffective disorder (onset, 22y), social anxiety disorder	Attention deficit disorder, major depressive disorder, generalized anxiety disorder	Attention deficit disorder, autism spectrum disorder	Anxiety disorder
Other	Stenosis of right external auditory canal	Frequent otitis media and hypothyroidism	Obesity	/	Bilateral hearing loss	/

Supplementary Table S6.2: *Primer pairs.*

Target locus	Forward primer	Reverse primer
LEUV-1 BP	GTCATCTATTCTCAAGTTAGTACCACA	ATCCACGATGTGGGACATTT
LEUV-1 control		CCCCCAGAAGATATGAAGCA
TOR-1 BP	CAAGCTGGGTGGTTTGATGC	ATCAGGAAGGCCACAACTTGT
TOR-1 control		ATGGGTGAAGCCAATGTGGT
UTRE-1 BP	ACCCCATTCAGAGACCAGGA	GTCCATGTGCCTGACGATCA
UTRE-1 control		AAGCCTCCTGGAGTCTTCCT
LEUV-2 invBP	TCTGGTCCCCACAGAACTC	GTACCATTGCTCCCAGTGCA (fw)
LEUV-2 control		TGGTTCACACTTTGATGGCA
TOR-2 invBP	TTCTCCTCCTCCTCTCCAGC	TTGCTTGGCCAGACCTATGG (fw)
TOR-2 control	TTGCTTGGCCAGACCTATGG	TCAGCTAGAGTGGTGGGACA

Overview of the primers used to validate the patient-specific (inversion) breakpoints and control reactions. BP=breakpoint, invBP=inversion breakpoint, fw=forward orientation in the reference genome hg38.

CHAPTER 7

Different loci for NAHR and PATRR-mediated recombination drive the high incidence of 22q11.2 deletion syndrome

Lisanne Vervoort[1], Nicolas Dierckxsens[1], Bo Zhou[2], Ruben Cools[1], Tracy Heung[3], Ann Swillen[1], Donna McDonald-McGinn[4], Beverly S. Emanuel[4], Hilde Van Esch[1], Anne Bassett[3], Alexander E. Urban[2], and Joris R. Vermeesch[1]

[1] Department of Human Genetics, KU Leuven, Leuven, Belgium

[2] Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA

[3] Department of Psychiatry, University of Toronto, Toronto, ON, Canada

[4] Division of Human Genetics, Children's Hospital of Philadelphia and Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Status: Will be prepared for submission

Abstract

Non-allelic homologous recombination between low copy repeats on chromosome 22 results in the 22q11.2 deletion syndrome, the most common microdeletion disorder in humans. Due to the complexity of the LCR22s, the exact crossover sites had never been mapped at nucleotide level. To chart these rearrangements at subunit level, optical mapping was performed on 20 patients and 17 parents-of-origin with different 22q11.2 deletion sizes. The application of whole-genome and targeted ultra-long read sequencing approaches in combination with an LCR22-specific haplotype-resolved *de novo* assembler allowed us to refine the rearrangement locus in five patients. Specific recombination loci were identified for deletions involving LCR22-B and -C. However, a subset of LCR22-A/B deletions showed a rearrangement pattern which was not predicted by NAHR. Nucleotide level resolution of this deletion type showed the rearrangement took place in a palindromic AT-rich repeat, which may be mediated by non-homologous end-joining. Although the LCR22-B and -C recombination hotspots are present in LCR22-A and -D as well, the crossover site of one standard LCR22-A/D deletion was located in a different subunit. Hence, several recombination loci are responsible for the 22q11.2 deletions. This may explain why 22q11.2 deletion syndrome is the most common microdeletion disorder.

7 Different hotspots for NAHR and PATRR-mediated recombination drive the high incidence of 22q11.2 deletion syndrome

7.1 Introduction

The 22q11.2 locus is one of the most complex and genomic unstable loci of the human genome, due to the presence of eight LCR22s (LCR22-A until -H). These LCR22s are mosaic patchworks of DNA subunits (>1kb) that share a high sequence identity (>95%) (Bailey et al. 2002b). Because of the nearly identical sequence, non-allelic homologous pairing of the paralogues can occur resulting in NAHR. The resulting deletion syndrome is known as 22q11.2DS, the most frequent human genomic disorder (Blagojevic et al. 2021).

The genes within the LCR22s, the mechanism(s) causing the rearrangement, and the consequences of the rearrangement for the 22q11.2DS phenotype remain largely unknown. This is because (I) an accurate reference genome is missing and (II) current sequencing and assembly technologies do not enable haplotype resolved assemblies of the LCR22s. The presence of gaps in LCR22-A of hg38 is caused by (I) the difficulty to assemble large repeat loci in general (Vollger et al. 2019) and (II) a high variability in the size and structural organization of the LCR22-A haplotype in the human population, with copy number and orientation variations of six duplicons (SD22-1 until SD22-6, **Figure 7.1**) (Demaerel et al. 2019).

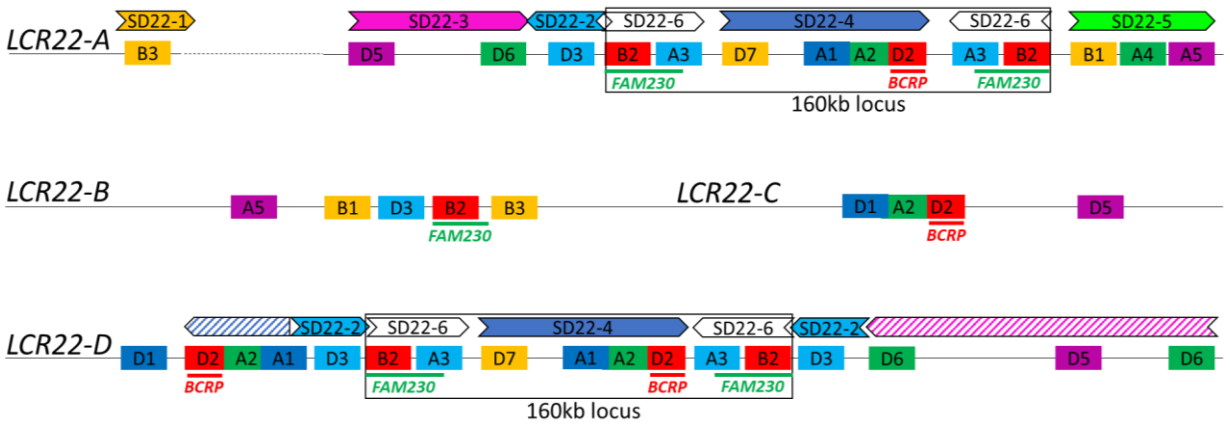


Figure 7.1: Recombination loci on the LCR22 map. *The four proximal LCR22s are depicted schematically based on the fiber-FISH probe composition from Chapter 3. LCR22-A is represented as the smallest haplotype including all possible SD22s. The dotted line reflects the variability in size, copy number, and composition. The SD22 duplicons are illustrated above the fiber-FISH probe compositions of LCR22-A and -D based on Demaerel et al. (2019). The BCR module* (Shaikh et al. 2007; Guo et al. 2016)*, FAM230* (Pastor et al. 2020)*, and 160kb locus are indicated in the LCR22s. Drawings are not to scale.*

Demaerel et al. (2019) and Pastor et al. (2020) used optical mapping techniques to map the rearrangements to the LCR22 subunits. The seven LCR22-A/D recombinations mapped in Demaerel et al. (2019) all occurred in a 160kb module, composed of SD22-4 flanked by SD22-6 on each site (**Figure 7.1**). Interestingly, this region does contain the *BCR* locus, previously indicated as the predicted 22q11.2 recombination locus (Shaikh et al. 2007; Guo

et al. 2016). Pastor et al. (2020) mapped 30 LCR22-A/D deletion trios using Bionano optical mapping. In six of these trios, the *BCR* module was not included in the rearrangement locus. In contrast, all 30 family maps demonstrate the association of the *FAM230* locus, located in SD22-6 (**Figure 7.1**), with the LCR22-A/D crossovers. These results counter the hypothesis of the BCR module to be the universal 22q11.2 recombination site (Pastor et al. 2020). To map the landscape of rearrangements we performed optical mapping on 20 more patients and parents covering different 22q11.2 (nested) deletions.

With the advent of long read sequencing technologies, we reasoned it might become possible to sequence through the LCR22s and identify the recombination regions. Because of the size and complexity, standard long read assemblers do not yet allow this. Here, we optimized whole-genome ultra-long read sequencing and CTLR-Seq (CRISPR-Cas9 targeted ultra-long read sequencing) (Jiang et al. 2015; Zhou et al.) in combination with a novel *de novo* assembler algorithm allowing for haplotype-aware assembly of the LCR22s. Those strategies allowed us to sequence the cross-over sites in five.

7.2 Results

7.2.1 LCR22-C rearrangements involve the A2-D2 module

LCR22-C is with four subunits, D1, A2, D2 and D5, the smallest proximal LCR22, covering a length of ~70kb (hg38) (**Figure 7.2A**). Fiber-FISH was performed on four 22q11.2DS patients with an LCR22-A/C rearrangement (**Supplementary Table S7.1**). For three also the parental LCR22s could be analyzed. The LCR22s were mapped using fiber-FISH and parents-of-origin were determined based on LCR22-A transmission. In all, the parent presented two different LCR22-A alleles and a single LCR22-C allele. This confirms that LCR22-C is conserved whereas LCR22-A is extremely variable. All were wild type indicating the deletions occurred *de novo* in the patients. For all four patients, the LCR22-A/C fusion haplotype was characterized by the subunits A1-A2-D2-D5 at the distal end, in which A1 is LCR22-A specific, D5 is LCR22-C specific, and A2-D2 are present on both LCR22-A and -C. Hence, recombination occurred within the A2-D2 module (**Figure 7.2B**).

In a fifth family, the patient carries a deletion between LCR22-C and -D (**Supplementary Table S7.1**). The deletion occurred on the paternal allele and fiber-FISH of this individual showed the wild type LCR22-C and -D haplotype. The fiber-FISH pattern of the recombined LCR22-C/D contains the proximal start of LCR22-C (D1), followed by the A2-D2 module shared by both LCR22-C and D, and continuing with the probe pattern of LCR22-D (A3-B2-…) (**Figure 7.2C**).

Thus, based on the fiber-FISH data of five recombinations involving LCR22-C, all rearrangements occurred within the A2-D2 module.

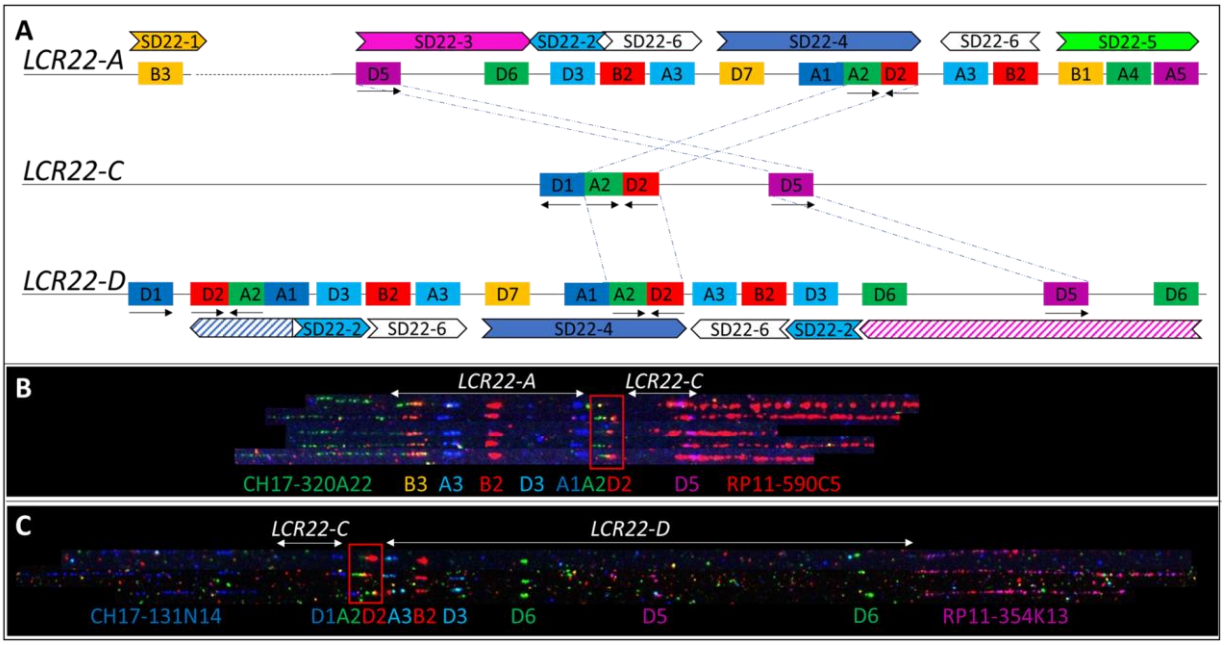


Figure 7.2: 22q11.2 rearrangements involving LCR22-C. *(A) Fiber-FISH probe composition of LCR22-A, -C, and -D. The SD22 duplicons are illustrated above LCR22-A and below LCR22-D. A 'hypothetical' haplotype is representing all possible SD22 duplicons from LCR22-A, since SD22-2, -3, and -4 are copy number and orientation variable. The arrows show the orientation of subunits that are shared between LCR22-C and LCR22-A or -D to interpret NAHR possibilities. Based on the orientations, NAHR is possible between A2-D2 modules or D5 (dotted blue line). Drawings are not to scale. (B) All LCR22-A/C deletions (n=4) show an identical A1-A2-D2-D5 pattern at the distal end of the crossover in which A2-D2 is the shared module (red box). (C) Rearranged haplotype of one LCR22-C/D deletion with A2-D2 as the identified crossover site (red box).*

In one family, the LCR22-A was small enough to sequence and partially assemble both parental and patient LCR22 haplotypes. Sequencing generated 104Gb (N50~100kb) and 60Gb (N50~100kb) of output for the patient and parent, respectively (**Supplementary Table S7.2**). The different LCR22 alleles were *de novo* assembled using NOVOLoci, a haplotype-aware assembler able to read through (part of) the LCR22s (unpublished). A multiple alignment was performed between the shared modules of the three haplotypes and SNPs unique to LCR22-A or -C were identified in the rearranged allele, based on the parental SNPs. The region of recombination was narrowed to 800bp (**Figure 7.3B**). This locus is located proximal of the A2 subunit which harbors an *Alu* SINE element and proximal of the first exon of the *POM121L1* pseudogene (**Figure 7.3C**).

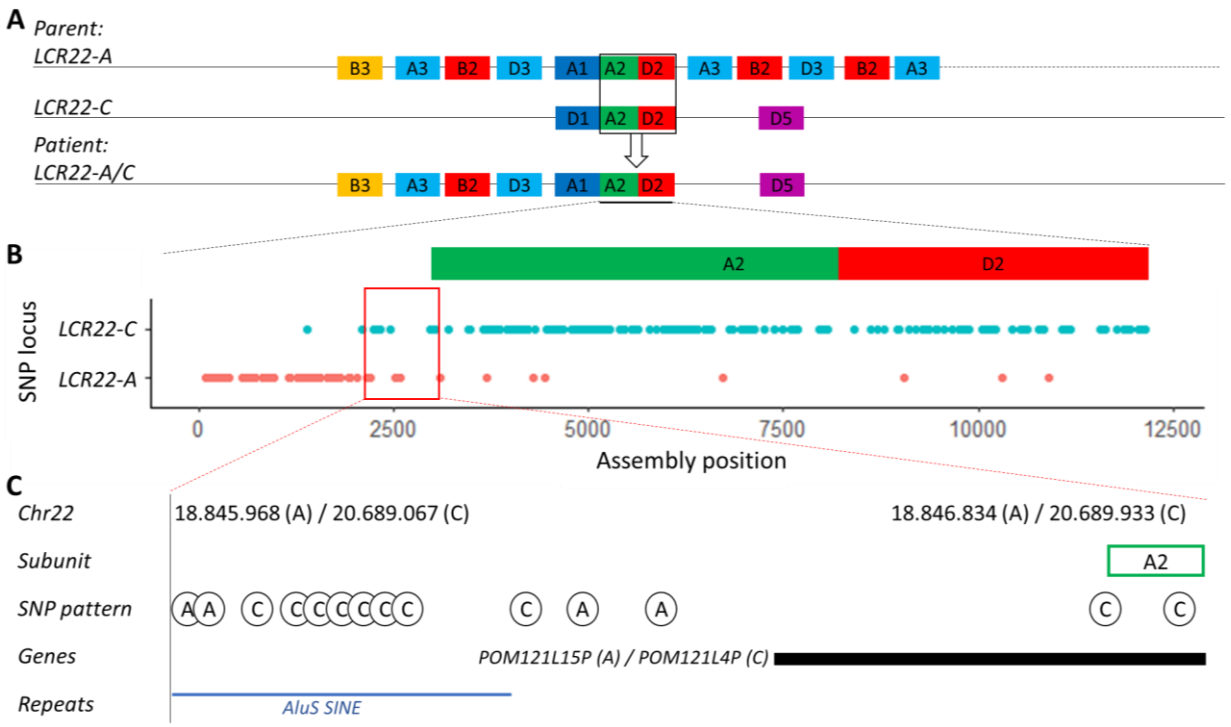


Figure 7.3: Crossover site in LCR22-A/C deletion of family AB3002. *(A) Schematic representation of the Fiber-FISH haplotypes of family AB3002. Only the proximal part of LCR22-A is represented, indicated by the dotted line at the distal end. The black box indicates the shared subunits between the two parental alleles where crossover had taken place. Drawings are not to scale. (B) Zoom of the shared LCR22-A and -C locus and representation of the LCR22-specific SNPs based on alignment of the LCR22-A (parent), LCR22-C (parent), and LCR22-A/C (patient) assemblies. SNPs present in the parental LCR22-A and patient LCR22-A/C, but not in the parental LCR22-C, are considered as LCR22-A specific SNPs (red, lower band). SNPs present in the parental LCR22-C and patient LCR22-A/C, but not in the parental LCR22-A, are considered as LCR22-C specific SNPs (blue, upper band). The LCR22-specific SNPs are plotted along the patient assembly position. A SNP-specific density switch is observed from LCR22-A to LCR22-C in a locus proximal from subunit A2 (red box), considered as the crossover site. (C) This crossover site is present at both LCR22-A (Chr22:18,845,968-18,846,834) and LCR22-C (Chr22:20,689,067-20,689,933) in hg38 and harbors an AluS SINE element and part of the first exon of the POM121L1 pseudogene.*

7.2.2 LCR22-B deletions can be mediated by palindromic AT repeat instability

LCR22-B contains subunits A5-B1-D3-B2-B3 and is approximately 120kb in size. Based on subunit sequence and orientation similarities between LCR22-B and LCR22-A or -D, the crossovers are expected to occur in the D3-B2 module (**Figure 7.4A**). Fiber-FISH analysis of both the parent and the patient carrying an LCR22-B/D deletion confirms that the chromosomes rearranged within D3-B2 (**Figure 7.4D**). Depending on the LCR22-A structure, several rearrangements are possible in an LCR22-A/B recombination. Each recombined LCR22 is predicted to carry the D3-B2-B3 subunits derived from distal LCR22-B (**Figure 7.4C**). Surprisingly, in three out of five LCR22-A/B deletions, two B2 probes juxtaposed in the fiber-FISH pattern (**Figure 7.4B**). In this pattern, the proximal B2 is part of LCR22-A and the distal one is part of LCR22-B, pinpointing the breakpoint locus in the sequence between these two probes. This B2-B2 juxtaposition (JV2001, JV2003, and AB2001) is not predicted by NAHR.

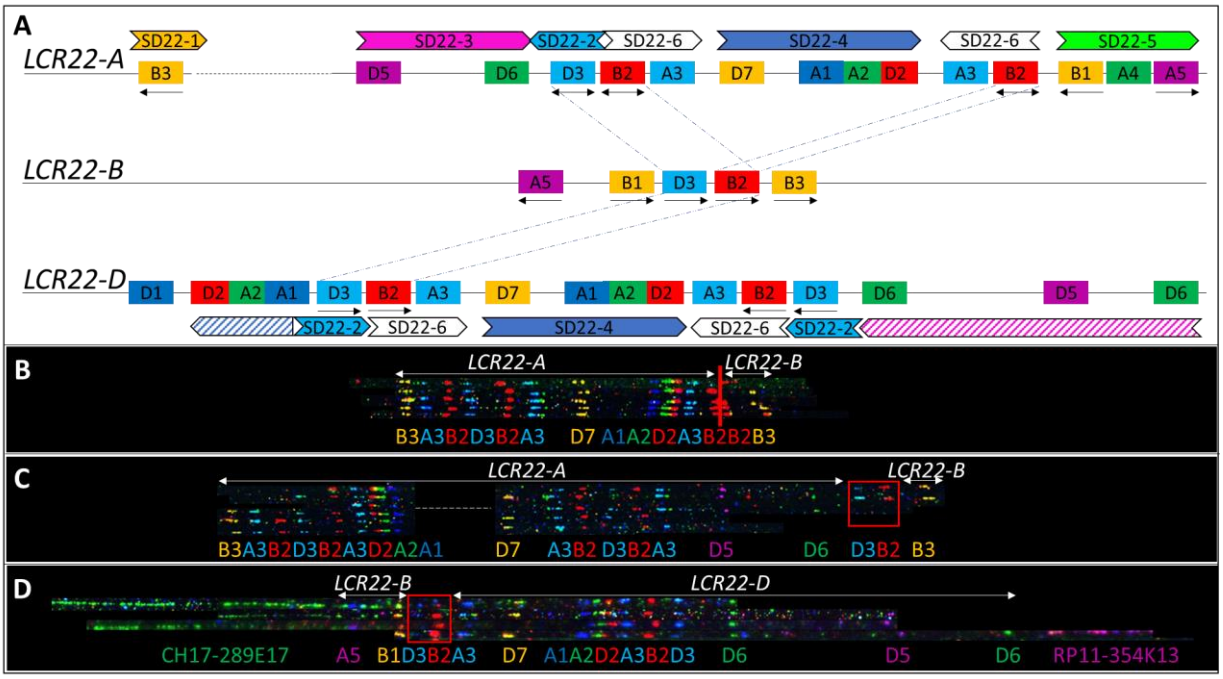


Figure 7.4: 22q11.2 rearrangements involving LCR22-B. *(A) Fiber-FISH probe composition of LCR22-A, -B, and -D. The SD22 duplicons are illustrated above LCR22-A and below LCR22-D. A 'hypothetical' haplotype is representing all possible SD22 duplicons from LCR22-A, since SD22-2, -3, and -4 are copy number and orientation variable. The arrows show the orientation of subunits that are shared between LCR22-B and LCR22-A or -D to interpret NAHR possibilities. Based on the orientations, NAHR is possible between D3 and B2 subunits (dotted blue line). Drawings are not to scale. (B) First group of LCR22-A/B deletions characterized by the juxtaposition of two red B2 subunits in the rearranged allele, creating a non-standard haplotype composition. (C) 'Standard' LCR22-A/B deletion with an LCR22-A to -B pattern change observed in subunit cluster D3-B2. (D) Haplotype of an LCR22-B/D deletion, consistent with the composition predicted by subunit overlap.*

To determine the LCR22-A/B sequence and identify the mechanism causing this rearrangement, the genomes of patient and parent of family AB2001 were sequenced using ultra-long whole-genome sequencing. For each individual, over 100Gb of data were retrieved with N50 values over 70kb (**Supplementary Table S7.2**). The data were scrutinized for reads covering the rearranging LCR22-A or -B alleles in the parent and the LCR22-A/B hybrid haplotype in the patient. By comparison of the parental and patient repeat composition, the rearrangement occurred within the palindromic AT-rich repeat (**Figure 7.5**). Surprisingly, a 50 bp LINE element was inserted while this element was not present in the parental LCR22-A nor LCR22-B allele (**Figure 7.5**).

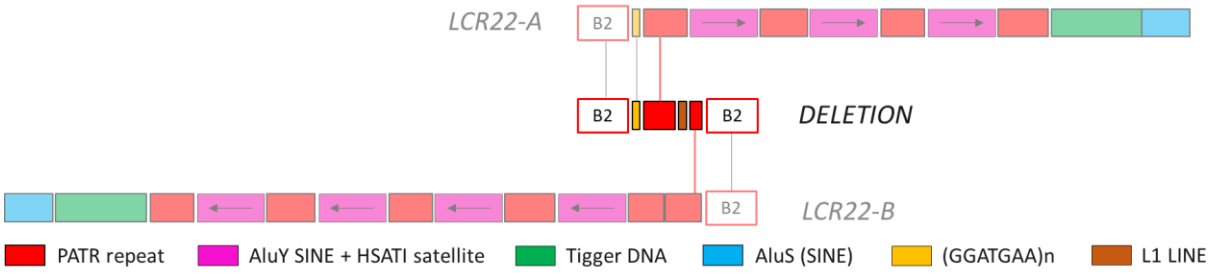


Figure 7.5: Repeat composition of the LCR22-A/B rearrangement site in family AB2001. *Based on whole-genome ultra-long read sequencing data, the repeats were determined in proximity of probes B2 in the patient, and the non-rearranged LCR22-A and -B allele of the parent. Arrows in the pink boxes show the difference in orientation between LCR22-A and -B. PATR = palindromic AT-*

rich repeat, SINE = short interspersed element, LINE = long interspersed element, HSATI = human satellite I.

In a second family (JV2001), a mosaic LCR22-A/B deletion was identified in the patient via low-pass sequencing. The mosaicism was validated by arrayCGH which showed a logR value of -0.24 in the chr22:18877787-20311613 locus, indicating that around 30% of blood cells were mosaic for the deletion in the region corresponding to LCR22-A until -B. To differentiate between mosaicism and chimerism, a SNP array was performed. B-allele frequency values of 0, 0.37, 0.62, and 1 were observed in the locus between LCR22-A and -B, indicating the presence of mosaic aberration of around 40% which falls in line with the result of arrayCGH. Dual color interphase-FISH (TUPLE1/ARSA probe set) showed the deletion to be present in 51%, 89%, and 56% of blood, urine, and buccal mucosa cells. Fiber-FISH uncovered the presence of three alleles: (I) a normal LCR22-A and -B haplotype, (II) the LCR22-A/B deletion haplotype, and (III) an allele carrying an inversion between LCR22-A and -B (**Supplementary Figure S7.1**). To confirm the presence of an inversion, interphase-FISH was performed using two differentially labeled BAC probes (CH17-222C16 and CH17-389E17) within the inversion/deletion region and two BAC probes (CH17-320A22 and RP11-354K13) flanking the region. A total of 71 cells were screened by two independent investigators. The internal BAC probes were deleted on one allele in 45% of cells and the orientation was switched, indicative of an inversion on a single allele in 44%. Hence, the interphase-FISH confirmed the presence of the three haplotypes and uncovered that each cell carried a wild type 22q11.2 and a deletion or an inversion (**Supplementary Figure S7.2**).

We hypothesized that the deletion was created from the inversion allele in an early stage during embryogenesis. Ultra-long whole-genome sequencing data from the patient (160Gb in total, N50>66kb, **Supplementary Table S7.2**) were scrutinized for deletion and inversion reads. Within these reads, the repeat sequence of the crossover site in the deletion (B2-B2) and the inversion (B2-D3 and D3-B2 in proximal and distal inversion locus, respectively) were determined using RepeatMasker (Smit et al.). Alignment of the repeat modules showed transition from proximal to distal inversion repeat composition in the PATRR (**Figure 7.6).** The inversion loci could be determined in the same way by examining the ultra-long Nanopore sequencing data from the parent-of-origin (**Supplementary Table S7.2**). The repeat compositions of the original LCR22-A (B2-D3) and -B (D3-B2) crossover loci pinpoint the transition point in a PATRR as well (**Figure 7.6**). The most parsimonious explanation is that two consecutive PATRR-mediated events created an LCR22-A/B inversion and deletion allele in patient JV2001 (**Figure 7.6**).

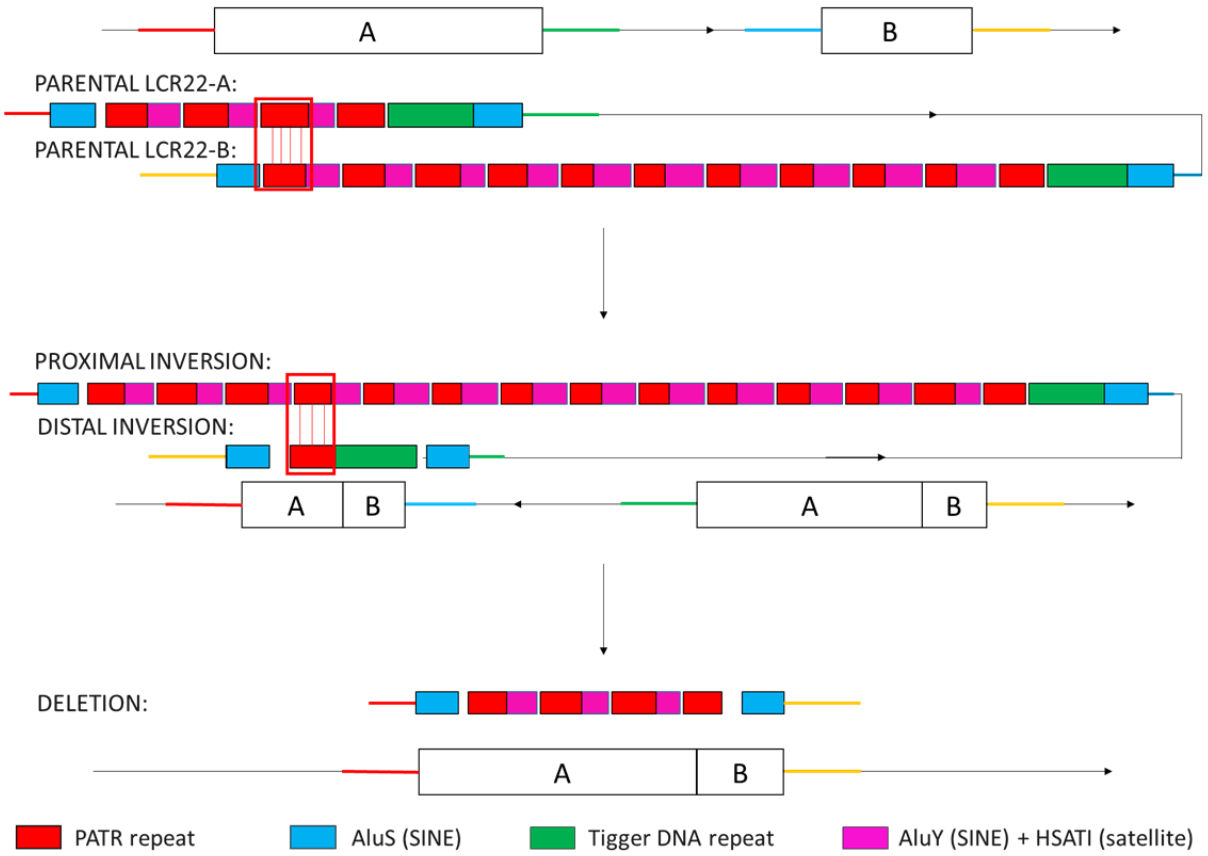


Figure 7.6: PATRR-mediated recombination creates an inversion and deletion allele in family JV2001. *Repeat transition pattern of the original 'breakpoint-creating' alleles in the parent-of-origin (parental LCR22-A and LCR22-B) and the resulting alleles (proximal and distal inversion) in the patient. Second event between PATRR of proximal and distal inversion in the patient leads to the deletion allele. Only the repeat structures of interest are showed and not the whole LCR22s.*

In a third patient (JV2003), the B2-B2 composition was targeted via long-range PCR using a single primer and subsequent PacBio sequencing (**Supplementary Figure S7.3**). A 1.8kb PCR product, from the end of the proximal B2 until the start of the distal B2, was generated in the patient and his two children with 22q11.2DS, but not in three control individuals, nor in other 22q11.2DS patients (two LCR22-A/D, one 'standard' LCR22-A/B, and one atypical 22q11.2 deletion). Single-molecule real-time sequencing showed two PATRRs flanking an Alu SINE element and HSATI satellite (**Supplementary Figure S7.3**). Since no parents were available for this patient, the composition was compared against reference genome hg38, localizing the rearrangement in a PATRR (**Supplementary Figure S7.3**).

In conclusion, within LCR22-B rearrangements at least two different rearrangement groups can be identified: the first characterized by crossover within the D3-B2 module (**Figure 7.4C-D**) and the second by the B2-B2 juxtaposition (**Figure 7.4B**). In this last subgroup, PATRRs seem to play a role in the rearrangement mechanism.

7.2.3 LCR22-A/D crossover site identified in *GGT* gene sequence

Due to the size and complexity of the subunits, rearrangements involving both LCR22-A and -D are the most complex to analyze. In addition, LCR22-A is not accurately represented in

hg38 nor in the T2T-CHM13 reference genome (Nurk et al. 2022). Based on the fiber-FISH probe composition, a 160kb locus is shared between LCR22-A and -D, containing probes B2-A3-D7-A1-A2-D2-A3-B2 (**Figure 7.7A**). Depending on the LCR22-A haplotype structure, this 160kb locus can be present multiple times on both alleles (copy number variation between 0 and 8 detected), increasing the complexity. Hence, it is difficult to pinpoint the position of the crossover.

Fiber-FISH mapping of parents-of-origin and patients with an LCR22-A/D deletion uncovers two subgroups of LCR22-A/D deletions: (I) in seven out of nine duos, the shared 160kb locus was the overlapping locus (**Figure 7.7B**)(Demaerel et al. 2019), (II) in two out of nine duos, a smaller recombination region of 20kb could be delineated, consisting of probes D2-A2-A1, containing the BCR locus (**Figure 7.7C**). Unfortunately, this small region was located in the middle of the rearranged and/or parental alleles, making it complex to cover via sequencing and subsequent assembly.

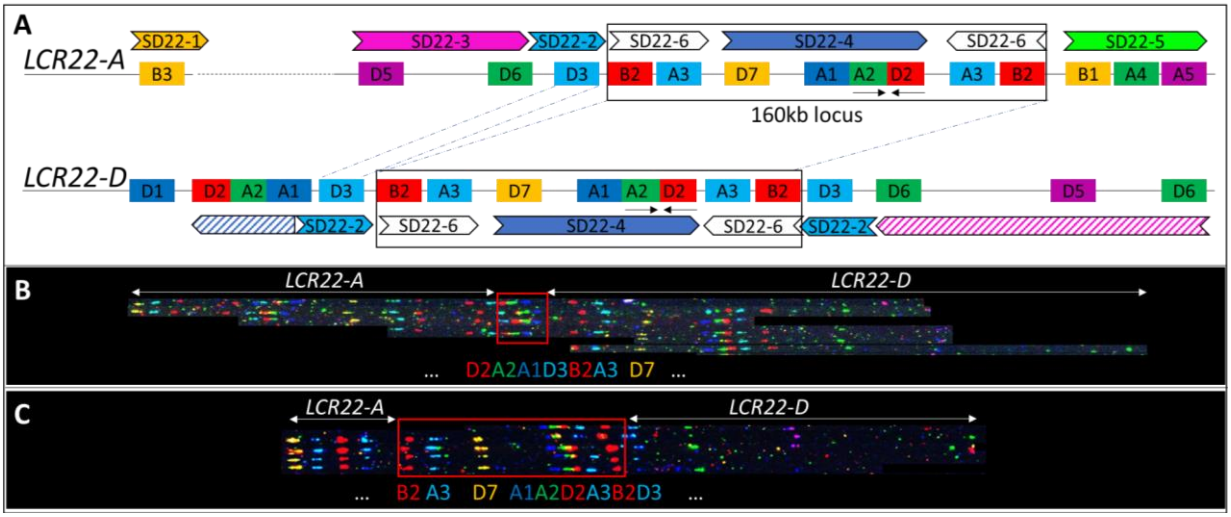


Figure 7.7: 22q11.2 rearrangements between LCR22-A and -D. *(A) Fiber-FISH probe composition of LCR22-A, and -D. The SD22 duplicons are illustrated above LCR22-A and below LCR22-D. A 'hypothetical' haplotype is representing all possible SD22 duplicons from LCR22-A, since SD22-2, -3, and -4 are copy number and orientation variable. The arrows show the orientation of subunits that are shared between LCR22-A and -D to interpret NAHR possibilities. Based on the orientations, NAHR is possible between D3 subunits and the 160kb shared module (dotted blue line). (B) In two families, the 20kb locus D2-A2-A1 was identified as the crossover site. (C) In seven families, the 160kb locus is delineating the rearrangement locus. Drawings are not to scale.*

Since most LCR22-A haplotypes are too large to assemble from whole-genome sequencing data, even from the ultra-long whole-genome sequencing data, we resorted to CRISPR-targeted ultra-long read sequencing (CTLR-Seq). CTRL-Seq combines the advantages of ultra-long read and targeted sequencing (Jiang et al. 2015). To map the position of the cross-overs in the long LCR22-A/D rearrangement at nucleotide resolution, CRISPR-Cas9 guide RNAs were designed to target the parental LCR22-A and -D alleles and the rearranged LCR22-A/D allele in the patient (**Table 7.1**). The parental LCR22-A (partly, ~280kb), LCR22-D (~350kb), and the rearranged LCR22-A/D (partly, ~350kb) were targeted (**Figure 7.8A**, **Supplementary Figure S7.4**) by a combination of locus-specific guide

RNAs. The fragments were separated in and isolated from the agarose gel and long-read sequenced. The rearranged LCR22-A/D and parental LCR22-A and -D segments were sequenced at a depth of 100x, 30x, and 70x, respectively. The single haplotype of LCR22-A/D and parental LCR22-A, and two parental LCR22-D alleles were *de novo* assembled using NOVOLoci. By comparing SNPs located at the distal end of the rearranged LCR22-A/D allele and the two LCR22-D alleles, the parental LCR22-D haplotype involved in the rearrangement could be identified. An alignment was performed between the shared loci (160kb module) of the patient LCR22-A/D, and parental LCR22-A and LCR22-D allele. The SNP pattern changes from LCR22-A specific SNPs observed on the rearranged allele, to LCR22-D specific SNPs in a locus encompassing the D7 subunit (**Figure 7.8B**). This 300bp transition region occurs within an a L2 LINE element intronic in the *GGT* gene (**Figure 7.8C**).

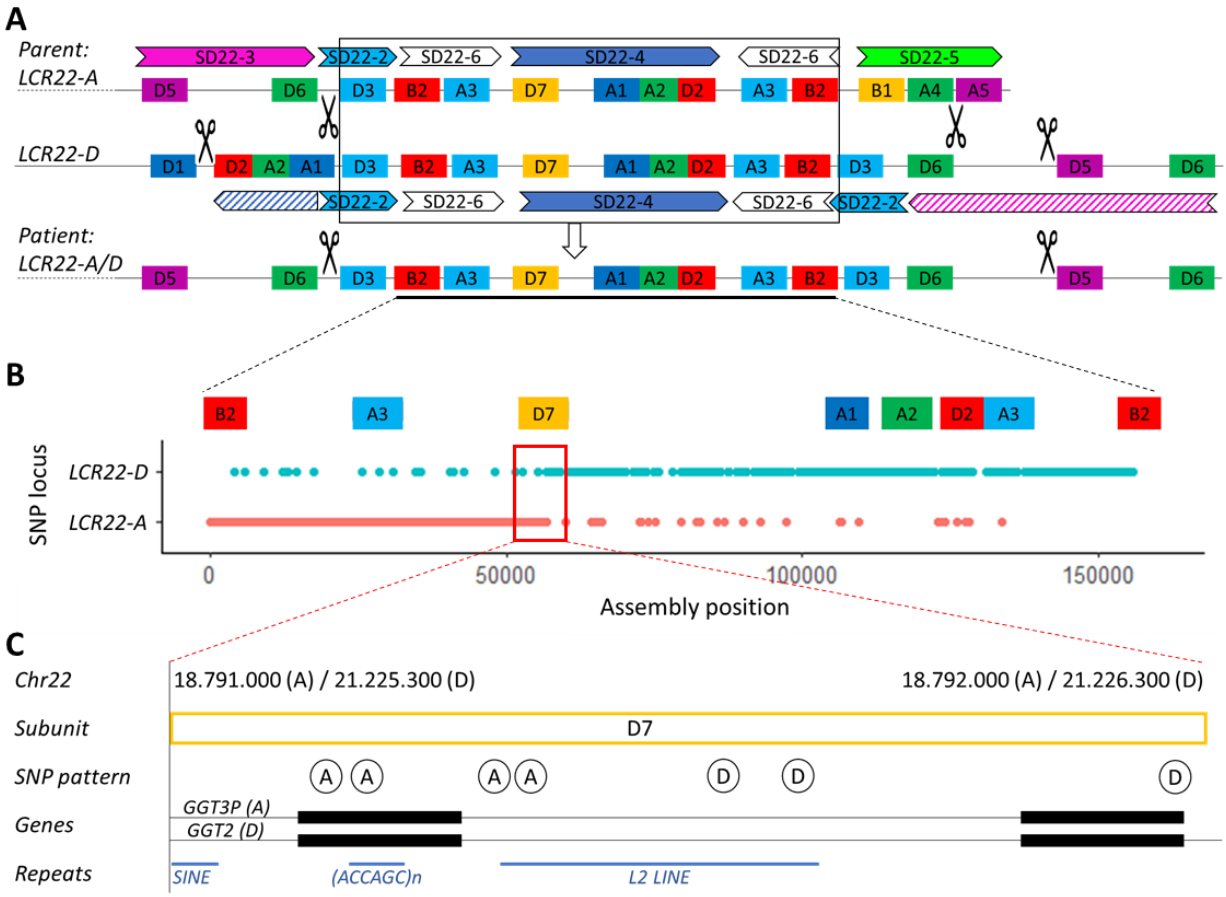


Figure 7.8: Crossover site in LCR22-A/D deletion of family JV1004 via CTLR-Seq. *(A) Schematic representation of the Fiber-FISH haplotypes of family JV1004: the distal part of the parental LCR22-A including both CRISPR-Cas9 target sites, the parental LCR22-D (both alleles will be sequenced), and the distal part of the rearranged LCR22-A/D allele in the patient. The black box indicates the shared subunits between the two parental alleles where crossover had taken place. The scissors indicate the CRISPR target sites. Drawings are not to scale. (B) Zoom of the shared LCR22-A and -D locus and representation of the LCR22-specific SNPs based on alignment of the LCR22-A (parent), LCR22-D (parent), and LCR22-A/D (patient) assemblies. SNPs present in the parental LCR22-A and patient LCR22-A/D, but not in the parental LCR22-D, are considered as LCR22-A specific SNPs (red, lower band). SNPs present in the parental LCR22-D and patient LCR22-A/D, but not in the parental LCR22-A, are considered as LCR22-D specific SNPs (blue, upper band). The LCR22-specific SNPs are plotted along the patient assembly position. A SNP-specific density switch is observed from LCR22-A to LCR22-D in a locus encompassing subunit D7 (red box), considered as the crossover site. (C) This crossover site is present at both LCR22-A (Chr22:18,791,000-18,792,000) and LCR22-D*

(Chr22:21,225,300-21,226,300) in hg38. It harbors repeat elements (SINE, (ACCAGC)n, and L2 LINE) and genic fragments of GGT3P (LCR22-A) or GGT2 (LCR22-D). The LCR22-A to -D switch (SNP pattern track) is observed in intronic sequence of the GGT gene and in a L2 LINE element.

7.3 Discussion

Due to the limitations of short- and standard long-read sequencing methods in LCR22 research and the absence of an accurate reference genome, the recombination loci of the 22q11.2DS within the LCR22s are still uncharted. As a consequence, the mechanisms and genes involved in the recombination remain unknown. In this study, we mapped the recombination sites in 20 families and pinpointed the locus at nucleotide level using long-read sequencing approaches in five. In these five cases, we identified four loci in LCR22-A where the recombination had taken place. In addition, the involvement of PATRRs suggest not only NAHR but also breakage-mediated repair mechanisms contribute to the high incidence of the syndrome.

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 for the generation of double-strand breaks, which are repaired via non-homologous end-joining (Kato et al. 2012). The LCR22-B PATRR is known to drive recurrent 22q11.2 translocations (Kurahashi et al. 2006), but has not been reported to be involved in 22q11.2 deletions. Here, we show involvement of the PATRR sequences at LCR22-A and -B to create LCR22-A/B deletions. In addition, in patient JV2001, PATRRs mediated the meiotic 1.5Mb LCR22-A/B inversion and the subsequent mitotic 1.5Mb LCR22-A/B deletion. It is known that PATRR size polymorphisms influence the rearrangement frequency of the de novo t(11;22) translocations (Kato et al. 2006; Tong et al. 2010). For example, larger and symmetric PATRRs on chromosome 11 and 22 are more prone to t(11;22) translocations (Kato et al. 2006; Tong et al. 2010). Interestingly, we uncovered structural polymorphisms in the PATRR-AluY-HSATI triplets, with copy number ranging between 0 and 13 in the four investigated individuals (patient and parent of JV2001 and AB2001, **Figure 7.5**, **Figure 7.6**). It seems likely that this CNV will affect rearrangement frequency as well. In addition, Correll-Tash et al. (2021) showed that the secondary structure formation and double-strand breaks occur both during meiosis and mitosis. Here, we identified an individual with a sequel of a likely meiotic followed by a mitotic PATTR-driven rearrangement (JV2001, **Supplementary Table S7.1**).

Mosaicism of the 22q11.2 deletion is rare, and a few cases were described with levels ranging from 11% until 85% of deletion cells observed in the index case (Consevage et al. 1996; Halder et al. 2008; Chen et al. 2019; Patel et al. 2006; Chen et al. 2004). All reports were based on standard interphase or metaphase FISH testing. Therefore, it was not possible to delineate the exact deletion region within the 22q11.2 locus. 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 (Patel et al. 2006). It would be of interest to map the LCR22 haplotypes of more mosaic cases to investigate whether inversions or other specific LCR22 structures are involved as well.

Previous studies have been performed to identify the regions involved in 22q11.2 NAHR. Shaikh et al. (2007) used long-range PCR with paralogous-specific primers to amplify and subsequently sequence the breakpoints of two distal 22q11.2 deletions. The Breakpoint Cluster Region (*BCR*) module was identified as the recombination locus in both an LCR22-D/E and an LCR22-E/F deletion (Shaikh et al. 2007). Guo et al. (2016) charted shared and paralogous sequence polymorphisms between LCR22-A and –D by sequencing BACs containing (parts of) the LCR22s. Based on this variation map, whole-genome sequencing data of two LCR22-A/D patients and their parents were screened for uniquely mapping read pairs within these LCR22s. The results suggested that this same BCR module was involved in the recombination mechanism of two recurrent LCR22-A/D deletions (Guo et al. 2016). Hence, these studies propose the BCR module to be the hotspot for 22q11.2 chromosomal rearrangements (**Figure 7.1**). Our data suggest that at least three additional loci exist where recombination can take place. Although the recombination locus of AB3002 contains the BCR module at subunit level, the exact crossover site is located 5kb upstream from the BCR gene (**Figure 7.3**). For the three PATRR-mediated LCR22-A/B deletions, the recombination is located in the *FAM230* sequence, concordant to the results of Pastor et al. (2020). The 300bp crossover locus of family JV1004 (**Figure 7.8**) is located in subunit D7, which is not even present in the smaller LCR22-B and -C. Hence, different recombination loci are present in the shared modules between the two involved LCR22s and therefore create variability in the crossover locus.

The identification of multiple subunits driving NAHR is also observed in other genomic disorders. For example, in neurofibromatosis (NF1, 17q11.2) type I deletions, 92.3% of the rearrangements cluster in the PRS1 and PRS2 subunits, with a length of 5.2kb and 4.8kb, respectively (Summerer et al. 2018). A 3kb hotspot was identified as the rearrangement hotspot in the majority (78.7%) of patients with Sotos syndrome (5q35.3) with differences observed at the nucleotide level (Visser et al. 2005a). The crossover sites of additional patients were located at different loci within the involved 5q35.5 LCRs (Visser et al. 2005b). Hence, large LCRs at specific genomic loci harbor several crossover sites causing the same genomic disorder.

The relatively small size of our sample collection have to be taken into account to interpret the results correctly. Fiber-FISH rearrangement patterns of extra samples can be predicted based on the shared probe modules between the two involved LCR22s, except for the PATRR-mediated crossovers. Here, extra samples would allow us to calculate the percentage of PATRR-mediated LCR22-A/B deletions and whether this type of rearrangement can cause LCR22-B/D or the larger LCR22-A/D deletion as well. Sequencing of extra duos will give a

broader view on the crossover variety at nucleotide level, since only one LCR22-A/C (AB3001), one LCR22-A/D (JV1004) and three PATRR involved LCR22-A/B (JV2001, JV2003, AB2001) deletions were resolved at this resolution. Nanopore sequencing introduces random errors during sequencing. As a consequence, the LCR22-specific SNP plots (**Figure 7.3B**, **Figure 7.8B**) show presence of occasional SNPs belonging to the distal LCR22 in the proximal SNP dense locus and vice versa. These are probably sequencing errors and could be corrected by increasing the coverage or additional 10X Genomics linked-read sequencing. Especially for the larger LCR22-A and -D alleles, high coverage is necessary to distinguish large shared segments between or within LCR22s.

In conclusion, we mapped the crossover sites of 20 families with 22q11.2DS at subunit level and five of them at nucleotide level. We uncovered that there was variability of the crossover site within the LCR22s and different mechanisms may be involved in different deletion types. It will be important to further investigate what elements, beside the repeat sequences, contribute to the 'recombination treshold'. In addition, it will be important to unravel the effect of the specific crossovers on the expression of LCR22 genes and 3D chromosomal structure, which might be associated to the phenotypic level.

7.4 Materials & Methods

Sample collection

A total of 37 Epstein-Barr virus transformed (EBV) cell lines, of which 20 index patients and 17 parents-of-origin, were collected for the study. Four samples were collected from Albert Einstein College of Medicine (New York, Bernice Morrow), 12 from University of Toronto (Anne Bassett), and 21 from University Hospital Leuven (Joris Vermeesch). Seven of the nine selected LCR22-A/D deletion duos were previously used in the study of Demaerel et al. (2019) and their patterns were re-analyzed for crossover site delineation. All 22q11.2DS patients and parents had given written consent to participate in the study. The EBV cell lines are the start materials for the fiber-FISH and sequencing experiments. Study approval was obtained from the Medical Ethics Committee of the University Hospital/KU Leuven (S52418), the Institutional Review Board of the Clinical Genetics Research Program at the Centre for Addiction and Mental Health (REB# 114/2001-02), and at the Albert Einstein College of Medicine (IRB# 1999-201-047). Extra information regarding the samples is available in **Supplementary Table S7.1**.

Fiber-FISH

To haplotype the LCRs on chromosome 22, we used LCR22-specific fiber-FISH (Demaerel et al. 2019). Long DNA fibers were extracted from EBV cell lines from probands and their parents using the Genomic Vision extraction kit (Genomic Vision). Slides were prepared as described and hybridized using the LCR22-specific customized probe set (Demaerel et al. 2019). Following automated 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 using matching colors and distances between the probes as anchors and haplotype coverage of at least 5X was aimed. Patterns of recombined LCR22s were compared to the parental patterns to identify the haplotype alteration position.

Ultra-long read sequencing of Oxford Nanopore Technologies

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, 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 **Supplementary Table S7.2**. Reads from the fastq files were mapped against hg38 using Minimap2 (Li 2018) and visualized in IGV (Robinson et al. 2011).

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

In patients and parents of families AB2001 and JV2001, 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 (**Table 7.1**). 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.).

Table 7.1: SNP identification for haplotyping of ultra-long whole-genome data.

	Family JV2001	**Family AB2001**
Patient LCR22-A/B	Manual read identification	Proximal: Chr22: 18,157,142 Distal: Chr22: 20,364,402
Parental LCR22-A	Chr22: 18,176,068	Chr22: 18,152,009
Parental LCR22-B	Chr22:20,341,319	Chr22:20,352,666

The SNPs were used to group the reads in two haplotypes. The SNP initially chosen is displayed in the table and was linked to other SNPs proximal and distal.

De novo assembly and sequence alignment

Ultra-long Nanopore reads were aligned to the human reference genome (hg38) with Minimap2 (Li 2018). To facilitate visualization of the alignment with IGV (Robinson et al. 2011), the 22q11 region was isolated with samtools (Li et al. 2009). *De novo* assemblies were performed with NOVOLoci, a targeted haplotype-aware assembler (unpublished). NOVOLoci needs a seed sequence to initiate the assembly and outputs separate assemblies for each haplotype. For the CTLR-Seq libraries, the target sequences were used as seed

sequence for the assemblies, while for the whole-genome libraries, non-duplicated sequences downstream from the target regions were selected. To identify the rearrangement region, a multiple alignment between shared subunits among the two parental alleles was performed with mafft (Katoh et al. 2019). To zoom in on the crossover locus, a customized script was used to identify unique SNPs between the shared subunits to reveal the transition between the two LCR22s.

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

Long-range PCR was performed using the TaKaRa LA PCR kit (TaKaRa Bio). PCR conditions were optimized, taking into account the presence of AT-repeats (Inagaki et al. 2005), by testing several times and temperatures for the extension-annealing phase: 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, Menlo Park, CA). Four libraries (22q11.2DS patient JV2002, his two children with a 22q11.2 deletion and the mother of his children) were pooled and loaded onto a single SMRT cell on a PacBio RSII using a DNA/polymerase binding kit P6 v2 (loading concentration 25pM) and DNA Sequencing Reagent kit 4.0 v2 (Pacific Biosciences, Menlo Park, CA). The RS_Long_Amplicon_Analysis.1 pipeline was used for analysis.

Low-pass sequencing in patient JV2001

The mosaic LCR22-A/B 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) in patient JV2001

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 in patient JV2001

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 in mosaic patient JV2001

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.) in blue (CH17-320A22, Aqua 431dUTP), green (CH17-222C16, Spectrum Green), red (CH17-389E17, Spectrum Orange), or orange (RP11-354K13, combination of Spectrum Green and Spectrum Orange). EBV cells of the mosaic patient, a normal control, and a non-mosaic heterozygous LCR22-A/B deletion patient were fixed and slides for FISH prepared. Two investigators (one without prior knowledge of the inversion) screened the slides independent from each other. Aside from the blue (CH17-320A22) and orange (RP11-354K13) BAC probe that were used as control, the presence and orientation of the green (CH17-222C16) and red (CH17-389E17) BAC probe were essential to score a chromosome as normal, inversion, or deletion. In the control and deletion individual, both screeners found over 85% of cells carrying two normal or one normal and one deleted allele, respectively. In the mosaic patient, one investigator found 50% normal-deletion cells and 38% normal-inversion cells, and the other found 33% normal-deletion cells, and 57% normal-inversion cells.

Design and efficiency testing of crRNAs

CrRNAs were designed via https://chopchop.cbu.uib.no (**Table 7.1).** Selection criteria were GC content (40-60%), self-complementarity (0), mismatches and efficiency (>50%). Off-target cleavage sites and efficiency were predicted *in silico* via the CRISPR-Cas9 guide RNA design checker (Integrated DNA Technologies) as well. The crRNA magenta A4 has two additional cleavage sites in LCR22-D, which are inevitable due to the repeat nature of the locus.

Table 7.1: Designed crRNAs for targeted CTLR-Seq.

crRNA	Target sequence	Chromosomal location
A4 magenta	TACGGCACCGCCAACACCTC	Chr22: 18624055 / 21371214 / 24184201
A2 distal	GGTGACCGGCCCAACCTCGG	Chr22: 18948145
D3 proximal	GATTTCGTATCTTTACCCAC	Chr22: 21100459
D1 distal (HIC2)	GTCATCCAAGCTCGGTATCA	Chr22: 21445403

Efficiency of the crRNAs was tested by *in vitro* cutting of a DNA fragment spanning the target locus. Primers were designed to create a fragment (170-300bp) over this target locus via standard PCR. A 10µL solution with 30nM of the substrate DNA (PCR fragment) was prepared. The concentrations of the locus-specific crRNAs and universal tracrRNA (Integrated DNA Technologies) were quantified using the Qubit Broad Range RNA kit (ThermoFisher). During a complexation reaction (5min at 95°C), the tracrRNA was saturated with a 1.5-fold molar excess of total crRNA. This cr-tracrRNA complex stock was diluted to a 300nM solution. The DNA was *in vitro* digested by incubating the cr-tracrRNA complex (30nM final concentration) and the Cas9 nuclease (30nM final concentration) for 10 minutes at 25°C and adding the substrate DNA (3nM final concentration, incubate at 37°C for 15 minutes). To stop the reaction, 1µL of proteinase K was added to each sample (room temperature, 10 minutes). The fragments were purified using 2X Ampure beads (to retain very small fragments). The length of the fragments was assessed on the BioAnalyzer DNA1000 chip.

CRISPR-targeted ultra-long read sequencing (CTLR-Seq)

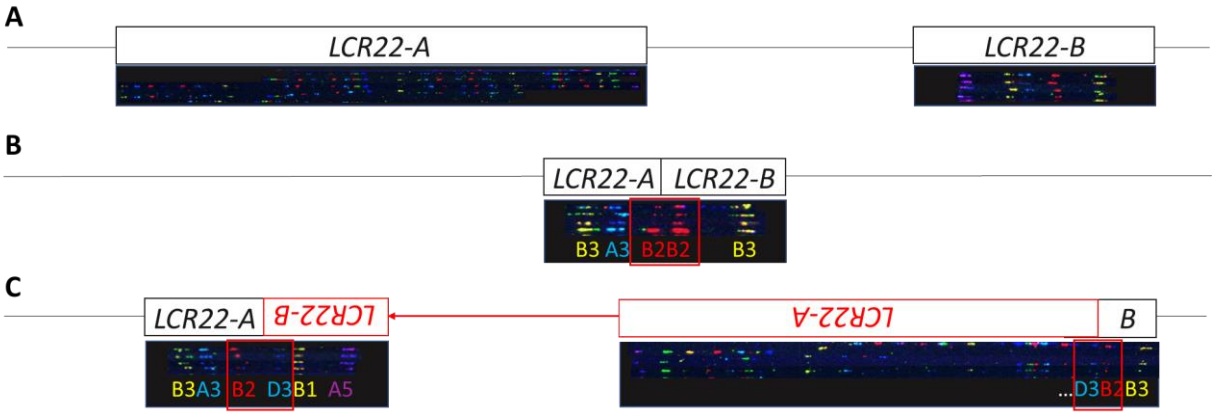
The CATCH protocol (Jiang et al. 2015) aims to enrich for targeted fragments with a large size (100kb-1Mb). For CTLR-Seq (Zhou et al.), two million cells were collected and washed two or three times in phosphate buffered saline. The cells were resuspended in the M2 suspension buffer (SageScience) and quantified using the Qubit High Sensitivity kit (ThermoFisher). The gel cassette was prepared by removing bubbles, replacing the running buffer (SageScience), and running a test to check the current. Following current quality approval, 70µL of the cell suspension was loaded in the sample well, based on the quantification and corresponding to 6-7.5µg of DNA (diluted with M2 buffer, SageScience). The reagent well was filled with 180µL of HLS Lysis Buffer A1 (SageScience). After sealing the cassette, the extraction program was started for three hours (100-300kb or 500kb depending on the target size). Afterwards, the contents from the sample and reagent wells are replaced by 80µL of the CRISPR-Cas9 mix and 220µL of HLS enzyme buffer C (SageScience), respectively. The 80µL CRISPR-Cas9 mix is composed of 16µL of Cas9 nuclease (100µM, New England Biolabs, 20µM), 20µL of Enzyme Buffer F (4X, SageScience), and 44µL of guideRNA complex (single guideRNA complex or combination, concentration of 11.8µM). This mix was injected to the sample well during an electrophoresis step, followed by an enzyme treatment step of 30 minutes. The content of the reagent well was replaced by 180µL lysis reagent A1 (SageScience) and the cassette resealed before the separation and collection step of four hours. After this phase, the liquid from the elution wells was collected using wide-bore tips to prevent shearing of the DNA.

To check whether the targeted fragment is effectively in one of the elution modules, a qPCR reaction was performed using probes targeting the HIC2 locus (ThermoFisher, catalogue number Hs_04400291), PRODH (ThermoFisher, catalogue number Hs_04502371), and

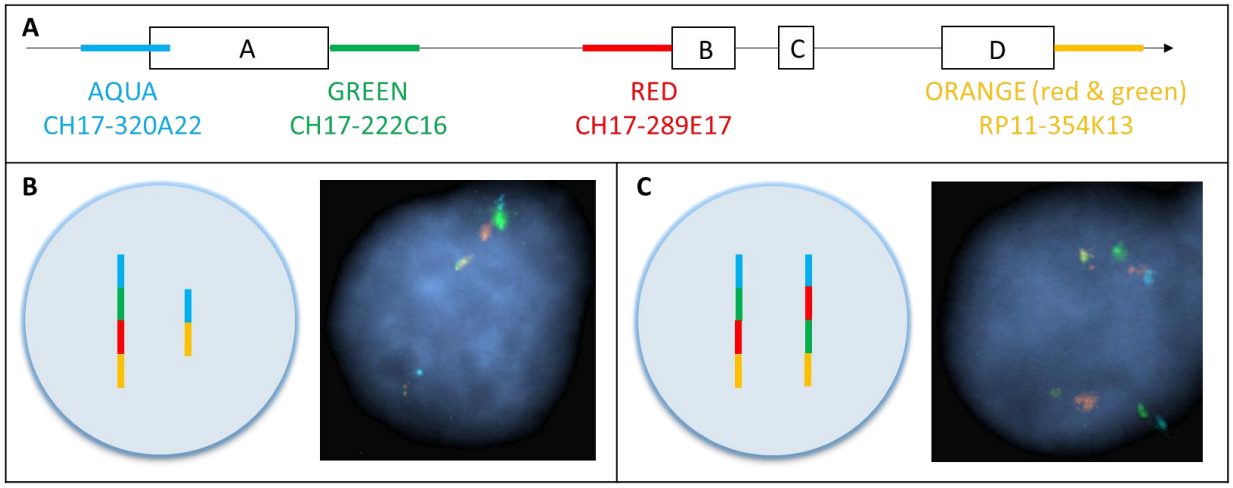
SD22-3 (ThermoFisher, custom oligo: 5'-GAGGGTCTGGATGCTCTCCT-3'). A qPCR master mix was prepared containing 5µL of TaqMan qPCR master mix (ThermoFisher), 0.5µL of RNase P control probe (ThermoFisher), 0.5µL of the target probe (ThermoFisher), and 2µL of cyclodextrin (SageScience) per sample. In the qPCR plate, 8µL of the qPCR master mix was combined with 2µL the diluted sample (5X dilution with cyclodextrin). The qPCR reaction was as followed: 10 minutes at 95°C, followed by 50 cycli of 15 seconds at 95°C – 1 minute at 60°C. The number of copies of RNase P and the target were calculated using the ΔΔCt method and plotted to check enrichment.

A nanopore library was created of the DNA in the elution modules that contained the target, based on the qPCR results. First, the buffer was exchanged using the Sage Hi-Bead workflow (SageScience) using 0.6X Hi-Bead suspension and DNA was resuspended in TE+ Hi-Bead solution (SageScience). Second, a nanopore adapter (AMX-F) was ligated via the ligation sequencing protocol (ONT, SQK-LSK110), followed by a AMPure (Beckman Coulter) bead clean-up using 0.45X beads and overnight elution in 24µL of EB buffer (ONT). The next day, the solution containing 24µL of the sample, 51µL of loading beads (LBII), and 75µL of Sequencing Buffer (SBII) was loaded onto a R9 Promethion flowcell and sequenced during 24-72 hours. Run statistics are provided in **Supplementary Table S7.2**.

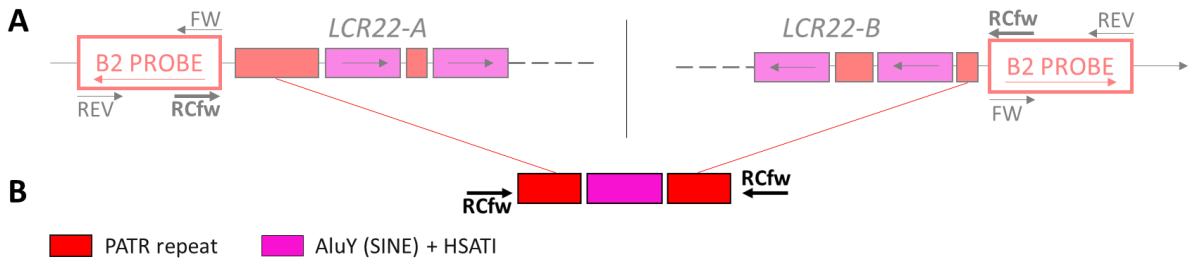
7.5 Supplementary Materials



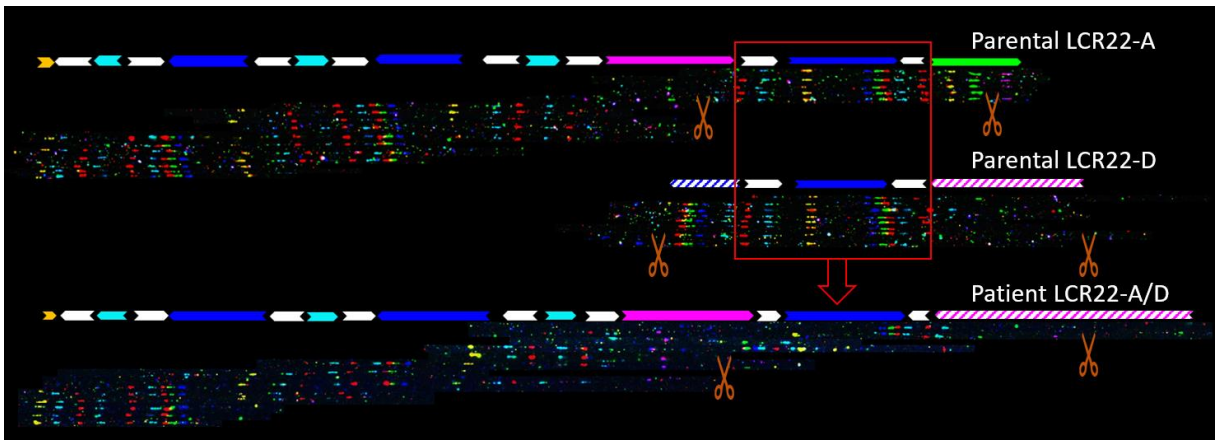
Supplementary Figure S7.1: *De novo assembly of LCR22 haplotypes of patient JV2001 using fiber-FISH. (A) Patterns showing the normal composition of LCR22-A and LCR22-B. (B) The deletion haplotype, visualized as a merged LCR22 block between LCR22-A and -B. (C) The composition of these patterns are indicating the presence of an inversion between LCR22 blocks A and B. The red box indicates the rearrangement locus in the deletion allele, based on the composition of the inversion alleles (as hypothesized that the deletion is caused by rearrangement of the inverted blocks).*



Supplementary Figure S7.2: *Targeted LCR22-A/B inversion/deletion interphase-FISH in patient JV2001. (A) Schematic representation of the 22q11.2 locus with the proximal LCR22s A, B, C, and D. BAC probes used in the targeted interphase-FISH assay and their corresponding labeling color are displayed at their chromosomal location. The aqua and orange probe are used as control probes, the green and red to validate the presence of an inversion/deletion. (B) Cell composition 1: normal – deletion. Schematic overview and corresponding targeted interphase-FISH result. (C) Cell composition 2: normal – inversion. Schematic overview and corresponding targeted interphase-FISH result.*



Supplementary Figure S7.3: *Zoom of the breakpoint locus in JV2003. (A) Hg38 repeat element structure distal and proximal from fiber-FISH probe B2 in LCR22-A and -B, respectively. Primers (FW = forward and REV = reverse) to generate the standard B2 fiber-FISH probe B2 are indicated. The reverse complement of the forward primer (RCfw) is used in the long-range PCR reaction. (B) Structural organization of the fragment obtained by long-range PCR using RCfw in patient JV2003. Since the forward and reverse primer in this reaction is identical, we were not able to determine the orientation of the fragment. PATR = palindromic AT-rich repeat, SINE = short interspersed element, HSATI = human satellite I.*



Supplementary Figure S7.4*: Fiber-FISH patterns of parent-of-origin and patient of family JV1004. The fiber-FISH probe compositions and SD22 compositions of parental LCR22-A, LCR22-D, and rearranged LCR22-A/D haplotypes of family JV1004 are displayed. The red box indicates the shared probe content between parental LCR22-A and -D. Crossover within this locus generated the rearranged LCR22-A/D allele in the patient (red arrow). Scissor icons indicate the loci targeted by CTLR-Seq by using CRISPR-Cas9 guideRNAs.*

Supplementary Table S7.1: *Overview of the samples with additional family, deletion type, sequencing, and breakpoint information.*

Site	Family	Individual	Deletion type	Sequencing	Breakpoint identification
Toronto	AB1002	Patient	LCR22-A/D	/	Fiber-FISH: 20kb shared locus (A1/A2/D2)
		Parent-of-origin (F)		/	
Toronto	AB1009	Patient	LCR22-A/D	/	Fiber-FISH: 20kb shared locus (A1/A2/D2)
		Parent-of-origin (F)		/	
Leuven	JV1001	Patient	LCR22-A/D	/	Fiber-FISH: 160kb shared locus
		Parent-of-origin (F)		/	
Leuven	JV1002	Patient	LCR22-A/D	/	Fiber-FISH: 160kb shared locus
		Parent-of-origin (F)		/	
Leuven	JV1003	Patient	LCR22-A/D	/	Fiber-FISH: 160kb shared locus
		Parent-of-origin (M)		/	
Leuven	JV1004	Patient	LCR22-A/D	CTLR-Sequencing	LINE element in D7 subunit (~200bp)
		Parent-of-origin (M)		CTLR-Sequencing	
Leuven	JV1005	Patient	LCR22-A/D	/	Fiber-FISH: 160kb shared locus
		Parent-of-origin (M)		/	
New York	BM1452	Patient	LCR22-A/D	/	Fiber-FISH: 160kb shared locus
		Parent-of-origin (F)		/	
New York	BM1453	Patient	LCR22-A/D	/	Fiber-FISH: 160kb shared locus
		Parent-of-origin (M)		/	
Toronto	AB2001	Patient	LCR22-A/B	Ultra-long read sequencing (ONT)	PATRR
		Parent-of-origin (F)		Ultra-long read sequencing (ONT)	
Leuven	JV2001	Patient	LCR22-A/B (mosaic)	Ultra-long read sequencing (ONT)	PATRR
		Parent-of-origin (F)		Ultra-long read sequencing (ONT)	

Leuven	JV2002	Patient	LCR22-A/B	/	Fiber-FISH: 21kb shared locus (D3/B2)
		Parent-of-origin (F)		/	
Leuven	JV2003	Patient	LCR22-A/B	Long-range PCR + PacBio sequencing PCR fragment	PATRR (no parent-of-origin available)
Leuven	JV2004	Patient	LCR22-A/B	/	Fiber-FISH: 21kb shared locus (D3/B2)
		Parent-of-origin (F)		/	
Toronto	AB3001	Patient	LCR22-A/C	/	Fiber-FISH: 15kb shared locus (A2/D2)
		Parent-of-origin (F)		/	
Toronto	AB3002	Patient	LCR22-A/C	Ultra-long read sequencing (ONT)	AluS SINE element proximal from A2 (~800bp)
		Parent-of-origin (F)		Ultra-long read sequencing (ONT)	
Leuven	JV3001	Patient	LCR22-A/C	/	Fiber-FISH: 15kb shared locus (A2/D2)
		Parent-of-origin (M)		/	
Leuven	JV3002	Patient	LCR22-A/C	/	Fiber-FISH: 15kb shared locus (A2/D2) (no parent-of-origin available)
Leuven	JV4001	Patient	LCR22-B/D	/	Fiber-FISH: 21kb shared locus (D3/B2)
Toronto	AB5001	Patient	LCR22-C/D	/	Fiber-FISH: 15kb shared locus (A2/D2)
		Parent-of-origin (F)		/	

Abbreviations: F = father / M = mother / ONT = Oxford Nanopore Technologies / CTLR-Sequencing = CRISPR-targeted ultra-long read sequencing

***Supplementary Table S7.2**: Statistics of sequencing runs.*

Sample	Sequencing device	Approach	Sequencing output	N50
JV1004 patient	Promethion (R9) – Stanford	CTLR-Seq LCR22-A/D	8.5Gb	20.85kb
JV1004 parent	Promethion (R9) – Stanford	CTLR-Seq LCR22-A	3.4Gb	45kb
	Promethion (R9) – Stanford	CTLR-Seq LCR22-D	5.7Gb	60kb
AB2001 Patient	Promethion (R9) – Leuven	Whole-genome Ultra-long	42.9Gb	61kb
	Promethion (R9) – Leuven	Whole-genome Ultra-long	75Gb	81kb
AB2001 Parent	Promethion (R9) – Leuven	Whole-genome Ultra-long	29.9Gb	103kb
	Promethion (R9) – Leuven	Whole-genome Ultra-long	75Gb	99kb
JV2001 Patient	Promethion (R9) – Leuven	Whole-genome Ultra-long	52.6Gb	75kb
	Promethion (R9) – Leuven	Whole-genome Ultra-long	111.4Gb	66kb
JV2001 Parent	Promethion (R9) – Leuven	Whole-genome Ultra-long	26.6Gb	126kb
	Promethion (R9) – Leuven	Whole-genome Ultra-long	14.8Gb	132kb
JV2003 Patient	PacBio RSII – Leuven	Long-range PCR	(482 barcode reads)	1.8kb
AB3002 Patient	Promethion (R9) – Leuven	Whole-genome ultra-long	6Gb	114kb
	Promethion (R9) – Leuven	Whole-genome ultra-long	12Gb	132kb
	Promethion (R9) – Leuven	Whole-genome ultra-long	86Gb	104kb
AB3002 Parent	Promethion (R9) – Leuven	Whole-genome ultra-long	4.7Gb	32kb
	Promethion (R9) – Leuven	Whole-genome ultra-long	55.7Gb	107kb

CHAPTER 8
GENERAL DISCUSSION

8 GENERAL DISCUSSION

The study of the 22q11.2DS is hampered by the absence of a proper genomic reference sequence of the 22q11.2 LCRs. The lack of a reference genome is due to the complexity of those LCR22s that are characterized by a complex patchwork of subunits, present in multiple and variable copy numbers and sharing a high percentage of sequence identity. As a consequence, genome sequences cannot be assembled properly and the reference genomes have an inaccurate representation of the LCR22s, containing gaps. Hence, to uncover LCR22 structure and eventual variability, development of *de novo* assembly strategies was crucial.

During this thesis, we uncovered unprecedented variability in both the LCR22 subunit structure and the non-allelic recombination hotspots. By using optical mapping techniques, we were for the first time able to map the LCR22s at subunit level. Mapping of controls and 22q11.2DS patients showed hypervariability of the LCR22-A haplotype structure (Demaerel et al. 2019). By mapping the LCR22s in the great apes, we demonstrated that, at least for LCR22-A, the regional expansion and variability is human-specific (Vervoort et al. 2021). By using specialized (ultra) long-read sequencing approaches, we were able to generate a sequence map of both the wild-type and rearranged alleles, which allowed us to map the non-allelic homologous recombination sites resulting in the 22q11.2DS. We demonstrate that not only different loci but also different mechanisms cause both the recurrent and the atypical deletions (Vervoort et al. 2019; unpublished). The discovered variability of the LCR22-A haplotype, recombination loci and mechanisms driving NAHR can be considered as fundamental new insights for the 22q11.2DS community.

8.1 Impact of human LCR22 variability

8.1.1 The reference genome

The LCRs on chromosome 22q11.2 have been one of the most complex loci of the human genome with three remaining sequence gaps in the hg38 reference genome. Those gaps complicate the 22q11.2 assembly. If the gaps would contain unknown sequence information, the reference remains incomplete and reads cannot be assigned to those loci. If the gaps would reflect different copies of subunits present elsewhere in the reference genome, short reads will be assigned to other loci. In addition, it remains unclear how many paralogous genes and pseudogenes are located in the LCR22s and, as a consequence, they cannot be studied.

Using a combination of optical mapping techniques, we were able to unravel the LCR22 composition at subunit level and uncovered an enormous variability in the human population. The presence and extent of this variability was not known and is a reasonable explanation why it has thus far been impossible to assemble the region. All different duplicons were represented as separate contigs in hg38 (**Figure 8.1A**). However, in

retrospect is has become clear that they could not be linked in the correct order, orientation, and copy number, because of the extensive variability. In this study of 33 individuals none was homozygous for LCR22-A, with size variations between 250kb and 2Mb.

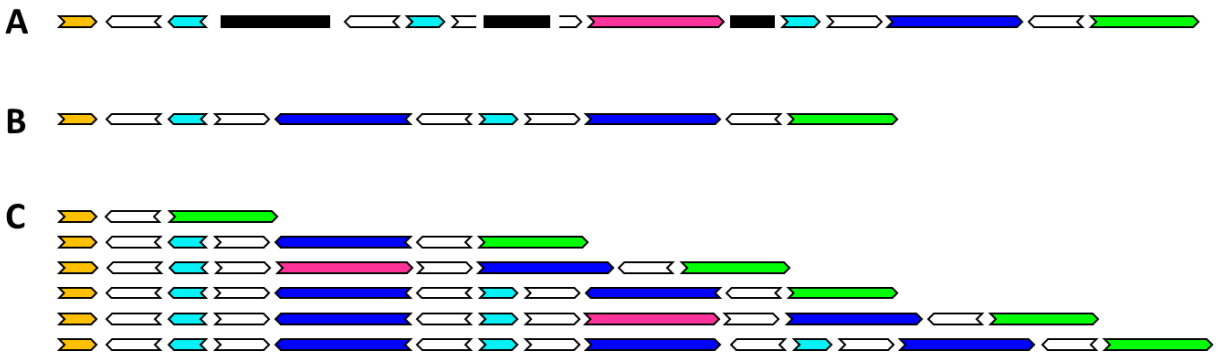


Figure 8.1: LCR22-A haplotypes of the current reference genomes and population. *(A) Duplicon structure of the LCR22-A haplotype depicted in hg38. (B) Duplicon structure of the LCR22-A allele assembled in the CHM13 sample from the telomere-to-telomere consortium. (C) Limited overview of the LCR22-A structural variation level present in the human population. A more comprehensive overview of assembled LCR22-A haplotypes can be found in Chapter 3.*

Recently, the telomere-to-telomere consortium generated the first complete, gapless genome (Nurk et al. 2022). To generate this genome, the haploid CHM13 cell line was assembled using different sequencing and mapping methods. Using this haploid cell line with only a single LCR22-A allele, they were able to sequence all LCR22s. Projecting this sequence to the duplicon structures defined by our optical mapping patterns (**Figure 8.1B**), this assembly represents a common LCR22-A haplotype (**Figure 3.5**). Unfortunately, the T2T-CHM13 LCR22-A lacks the SD22-3 duplicon. The absence of this duplicon may hamper future analyses using the T2T-CHM13 as a reference genome. Nevertheless, the release of this new reference genome makes it now possible to compare novel sequences against a more accurate, reliable reference genome and is a major step forward. It will be essential to construct new resources and databases to collect and visualize the variation of this locus (**Figure 8.1C**). Mapping and visualizing large scale structural variation is one of the main aims of the Human Pangenome Reference Consortium (Wang et al. 2022).

It will be essential to construct new resources and databases to collect and visualize the human structural variation (**Figure 8.1C**), as aimed by the Human Pangenome Reference Consortium (Wang et al. 2022). They will establish a database of 350 phased genomes, generating a total of 700 haplotypes (Wang et al. 2022). For the 22q11.2 locus, duplicon and nucleotide variation will be essential to include. First, different haplotypes need to be depicted at duplicon level including incidences and population distribution, as well as the classification of copy number variable duplicons. Second, since the individuals will be sequenced, single nucleotide variation information will become available, which can be grouped per haplotype and per duplicon. The development of bioinformatic tools to solve and construct complex structural variants in a high-throughput way will enhance future 22q11.2 research including crossover site identification on a larger scale. Hence, this

initiative will expand the catalogue of gross 22q11.2 structural variation at duplicon and sequence level generating a valuable control population dataset.

8.1.2 Consequences for the transcriptome

LCR22-specific structural variation can cause (I) gene-dosage effects in the LCR22-specific transcripts based on copy number variation or new fusion transcripts associated with specific rearrangements, (II) effects on gene expression in LCR22 flanking genes, and (III) exert a genome-wide impact. Of particular interest are the genes located within the LCR22-A sequence. Since an accurate gapless reference genome is only recently available, genes and transcripts within the LCR22s are poorly characterized (**Table 8.1**).

Table 8.1: LCR22-A specific genes.

SD	Probe	Gene	CNV	Expression	Function
SD22-1	B3	*USP18*	1	EBV cells	Protein de-ubiquitination (Kang and Jeon 2020)
SD22-6	A3	*FAM230*	1-*18*	Testis	Long intergenic non-coding RNA (Delihas 2018, 2020)
SD22-3		*TMEM191*	0-*1*	Testis	Transmembrane protein
SD22-3	D6	*RIMBP3*	0-*1*	Outside nervous system	RIM binding protein: important in synaptic vesicle fusion and Ca^{2+} channel function (Mittelstaedt and Schoch 2007)
SD22-3	D5	*PI4KAP*	0-*1*	Placenta, brain, testis	Phosphatidylinositol kinase
SD22-4	D7	*GGT3*	0-*8*	Kidney, intestine, duodenum	Gamma glutamyltransferase: glutathione metabolism (Figlewicz et al. 1993)
SD22-5	A4	*PRODH*	1-*2*	Brain, nerve, skin	Proline dehydrogenase enzyme in proline degradation pathway (Bender et al. 2005)
SD22-5	B1	*DGCR6*	1-*2*	Heart, brain, testis	Neural crest cell migration, pharyngeal arch development (Edelmann et al. 2001)

The SD22 and probe column show the corresponding SD22 and subunit where the gene is located. Copy number variation is based on the number observed in the fiber-FISH patterns from chapter 3. Expression is derived from the GTEx expression profiles (Lonsdale et al. 2013)*. The general function is described including references if available.*

Transcriptome studies may help to unravel the functional importance of these human-specific expansions and the role of the LCR22-A-specific genes (**Table 8.1**), mainly focusing on genes that are (I) copy number variable, and (II) absent or fixed in the chimpanzee lineage, suggesting an evolutionary role. For example, mutations and deletions of the *PI4KA* gene (LCR22-C) lead to aberrant myelination, polymicrogyria, cerebellar hypoplasia, and arthrogryposis (Alvarez-Prats et al. 2018; Pagnamenta et al. 2015). *PI4KAP2*, located distal in LCR22-D (SD22-3) and therefore present in all humans, was previously identified as a

deregulated pseudogene in Huntington's disease (Costa et al. 2012). Interestingly, the *PI4KAP1* pseudogene is not present in every human, since it is located in the SD22-3 duplicon of LCR22-A. One study found that the gene was consistently upregulated across four immune cell subpopulations in a posttraumatic stress disorder test setting (Kuan et al. 2019). Another interesting transcript is *RIMBP3*, located in the SD22-3 duplicon as well, of which two copies (*RIMBP3B* and *RIMBP3C*) are present in LCR22-D, and one (*RIMBP3*) in LCR22-A. *RIMBP3* is exclusively expressed in mammals and is, in contrary to *RIMBP1* and *RIMBP2*, ubiquitously expressed except in the nervous system (Mittelstaedt and Schoch 2007). An additional family worthwhile to investigate would be the *FAM230* transcripts, a family of long intergenic non-coding RNAs which are specifically formed by sequence duplications (Delihas 2018). They are located in the SD22-6 duplicon, which is highly copy number variable depending on the haplotype length. This long non-coding RNA is evolutionary grouped with a gamma-glutamyltransferase (*GGT*) protein gene family in the repeat sequence of SD22-4 (Delihas 2020). Hence, further characterization of the LCR22-specific genes will be essential to infer function and importance of the LCR22s themselves.

A large subset of the LCR22-specific genes is transcribed in the testis (**Table 8.1**). This is in line with the 'out of the testes' hypothesis described by Kaessmann (2010), stating that the testis is a catalyst for the transcription of otherwise silenced genes. An interplay of factors (promotor demethylation, increase in components involved in transcription pathway) will lead to an open chromatin state in spermatocytes and spermatids. As a consequence, natural selection will favor the expression of advantageous 'new genes', eventually resulting in the expression of the gene in new tissues. It will be interesting to explore whether these paralogous copies might have contributed to new or altered gene functions.

LCR22-focused transcriptome analyses have to deal with similar problems as described at the DNA level. Standard RNA-Seq protocols using Illumina short-reads are too short to map paralogous variants and as a consequence, no differentiation between the transcripts of the different LCR22s is possible. To reduce the complexity of the analysis, these 'multi-mapping' transcripts are removed from standard pipelines. Hence, the development of a qualitative method to study these specific transcripts will be necessary to measure the impact of the discovered LCR22 variability on gene expression.

To qualitatively map all the LCR22-specific genes and paralogues, a targeted long-read RNA sequencing method was developed in our laboratory (unpublished, preliminary results). IsoSeq, PacBio long-read sequencing of full transcripts, was used in combination with BAC-based capture enrichment of LCR22-specific genes. The advantage of the BAC probe-based enrichment strategy is that no prior knowledge about the transcript composition is necessary, as opposed to oligonucleotide probe strategies (Dougherty et al. 2018). The protocol was tested on cDNA extracted from the EBV cell lines from one 22q11.2DS patient and the two parents. Biotin-streptavidin capture using the LCR22-spanning BAC probes was

followed by amplification, PacBio library preparation, and sequencing on one Sequel SMRT cell. Analysis showed that 13-20% of the clusters were on target (LCR22 transcripts or overlapping BAC transcripts), demonstrating an enrichment ranging from 124-386X (unpublished results). Despite this high enrichment, only transcripts from two LCR22-specific genes (*USP18* and *UBE2L3*) were retrieved, confirming the GTEx expression profiles (Lonsdale et al. 2013). Hence, tissue-specificity will be important to take into account in further studies.

8.1.3 Consequences at the 3D organizational level

Both regular and haplotype-resolved Hi-C analysis mapping long-range chromosomal interactions on 22q11.2DS patients demonstrated that LCR22-A and -D act as topologically associated domains (TAD) (Zhang et al. 2018). In addition, by performing chromatin immunoprecipitation sequencing analysis of regulatory histone marks and RNA-Seq analysis of gene expression patterns, expression of genes within and flanking the 22q11.2 region were altered: within the deletion boundaries, in the deletion flanking regions, along chromosome 22q, and genome-wide (Zhang et al. 2018). Hence, LCR22s may play an important role in the 3D organizational chromatin structure, containing sequence features that constitute strong topological boundaries and likely control long-range regulation of transcription. Thus, a nucleotide difference of over 1.75Mb between the smallest and largest LCR22-A haplotype identified, will likely affect TAD boundaries and can exert an effect at the transcriptional level on genes flanking the deletion region. The generation of haplotype-specific chromosomal contact maps will allow the identification of TADs. To determine the 3D consequences of haplotype differences, chromosome conformation capture protocols are available for both short-read (Hi-C) (Lieberman-aiden et al. 2009) and long-read (Pore-C) (Ulahannan et al. 2019) sequencing.

8.2 Predisposition for 22q11.2 rearrangements

8.2.1 Factors that may increase recombination frequency

The prevalence of 22q11.2DS is estimated at 1 in 2148 live births, based on newborn screening data in Ontario (Blagojevic et al. 2021). This prevalence is significantly higher compared to other recurrent genomic disorders such as Williams-Beuren syndrome (1 in 7500) (Cuscó et al. 2008), Smith-Magenis and Potocki-Lupski syndrome (17p11.2 rearrangements, both 1 in 25000) (Neira-Fresneda and Potocki 2015). Therefore, 22q11.2DS is the most common microdeletion syndrome in humans. However, it remained unclear what are the exact genetic drivers of the deletion leading to this high frequency in the human population. Our data pinpoint several potential causes for this high recombination frequency.

First, intra-individual LCR22-A structural heterozygosity can facilitate misalignment and NAHR during meiosis. Structural variation heterozygosity was previously reported as a susceptibility factor for genomic rearrangements in the 7q11.23 locus, resulting in Williams-Beuren syndrome (Cuscó et al. 2008). Not only the general presence of structural differences between the two LCR22-A haplotypes can be important, but the exact structural compositions as well. Pastor et al. (2020) examined the LCR22-A and -D haplotype frequencies between parents-of-origin and non-transmitting parents, but no significant differences were observed. In addition, the rearranged allele structure of some 22q11.2DS patients (TOR-2 and LEUV-2 in Chapter 6 and JV2001 in Chapter 7) may have occurred via two recombination events: first, an inversion between or mediated by LCR22s, and second, the deletion-causing (N)AHR. Inversions are known to play an important role in the origin of some genomic disorders (Puig et al. 2015). However, despite several efforts, we were not able to detect inversions in the parents of 22q11.2DS children so far (Gebhardt et al. 2003; unpublished results). Hence, more information on LCR22-mediated inversion prevalence in the human population is essential to elucidate the mechanism of the observed 'two-step' rearrangements and the impact at the genomic instability level.

Second, the involvement of a variety of rearrangement mechanisms other than NAHR may increase the frequency of 22q11.2 rearrangement events. Sequence mapping of atypical deletions showed involvement of replication-based mechanisms as fork stalling and template switching or microhomology-mediated break-induced repair. These events were characterized by the presence of indels and microhomology traces at the breakpoints. In addition, the crossover site in a subset of the LCR22-A/B deletions are within the palindromic AT-rich repeat. Via the formation of hairpin and cruciform structures, PATRRs are known to be involved in rearrangements via the non-homologous end-joining pathway. Mapping more families will identify the ratio of rearrangements that is caused by PATRRs and whether they play a role in the more common LCR22-A/D recombination as well.

Third, the high AHR rate in the LCR22-A locus could act as a mediator for NAHR. The locus appears to be a hotspot for recombination. Analyses of polymorphic markers in families showed frequent recombination of the locus (Torres-Juan et al. 2007). In addition, marker analysis showed that certain families appear to have a higher recombination tendency with clustering of events in the 22q11.2 locus. Potentially LCR22 structure might affect recombination frequency. Mechanistically, AHR–NAHR are probably linked and predisposition for 22q11.2 rearrangements in certain families might be correlated with AHR frequency.

Probably an interplay of these factors will lead to the 22q11.2 rearrangement predisposition in certain individuals. As a consequence, it seems likely that longer haplotypes are more susceptible for rearrangements, compared to shorter ones. To some extent this is supported by the high prevalence of the LCR22-A/D deletion, the two largest LCR22s each containing

several paralogous subunits. This specific deletion type constitutes 85% of all *de novo* 22q11.2DS cases, significantly higher than the frequencies of the nested, distal, and atypical deletions, involving smaller LCR22s or unique sequence. Larger LCR22-A haplotypes contain extra paralogous subunits which can trigger non-allelic homologous recombination. We demonstrate recombination in at least four different subunits, each of which is present in multiple copies in LCR22-A and D. Hence, it seems likely that a higher number of paralogous units will lead to a higher recombination rate.

8.2.2 LCR22 structure as predisposing factor?

To discover whether LCR22 haplotype length and/or structure are susceptibility factors for the 22q11.2DS, the haplotypes of parents-of-origin and controls have to be mapped and compared. For this purpose, 22q11.2DS trios are the ideal start material, since the parent-of-origin and the non-transmitting parent (control) are both present. Comparisons of the two groups can be made for the length of the alleles and presence or absence of certain SD22 duplicons (for example, SD22-3 in LCR22-A). If a trend is observed between the two groups, a power analysis can be performed to check how many individuals are necessary to obtain a significant result. Some questions have to be taken into account in the experimental set-up. For example, in the case of intrachromosomal NAHR, only the recombining allele is considered to have higher recombination tendency. The question remains whether structure of both LCR22-A alleles is involved in the predisposition to the process of interchromosomal NAHR between LCR22-A and -D?

If a rearrangement predisposition relationship can be established, the results should be interpreted in terms of genetic counseling and risk prediction. As a consequence, specific LCR22-A structures involved in the recombination can be over- or underrepresented in certain populations (Demaerel et al. 2019), causing their over- or underrepresentation in specific study designs (McDonald-McGinn et al. 2005; Kruszka et al. 2018). In the long-term future, when whole-genome long-read sequencing will be used in a high-throughput setting, individualized risk prediction for 22q11.2DS can be implemented in the clinic. The question remains whether this prediction will add high benefit compared to the associated cost, since recurrence risk for 22q11.2DS is low in families of *de novo* cases (McDonald-McGinn et al. 2015), suggesting that the NAHR recombination frequency is low even in parents with LCR22-A structures that predispose to 22q11.2DS.

8.3 Multi-omics to understand phenotypic variability in 22q11.2DS

Despite extensive research efforts (Cleynen et al. 2021; Breetvelt et al. 2022; Davies et al. 2020), the genetic basis for neuropsychiatric disorders in the 22q11.2DS remain largely unsolved (Vermeesch 2022). It will be essential to perform genome-wide association studies at different levels including large cohort groups to unravel an association. In addition, since LCR genes in the genome are known to have played a role in human neuronal development

(Dennis and Eichler 2016), LCR22 genomic structure and transcripts have to be taken into account.

8.3.1 Genotype-phenotype association studies

To compare a group of 22q11.2DS patients with and without a neuropsychiatric disease, sufficiently large samples of both subgroups have to be collected. For this, the IBBC consortium offers a good starting point. Phenotypic information, with a focus on neuropsychiatric elements via the Diagnostic and Statistical Manual of Mental Disorders diagnostic criteria, was collected for the genome-wide association study of genetic contributors for schizophrenia risk (Cleynen et al. 2021). In the study, the phenotypic and short-read sequencing data of 519 patients with 22q11.2DS was used. Since that study, the cohort has been extended and a total of over 1900 patients are available via this consortium (Gur et al. 2017).

The LCR22 structures of the two groups of patients (with and without schizophrenia) have to be mapped using fiber-FISH or Nanopore ultra-long read sequencing and compared to each other. In addition, Nanopore sequencing data offer the possibility to extract and explore methylation data, adding an important layer of information. To reach significance, it will be essential for the 22q11.2 research community to further expand the cohorts and extend both phenotyping and biobank efforts.

8.3.2 induced pluripotent stem cell transcriptomes to unravel phenotypes

To associate transcriptomic alterations with the neuropsychiatric phenotype, tissue-specificity has to be taken into account. Hence, transcriptomic differential expression studies need to be carried out in human induced pluripotent stem cells (iPSCs) differentiated to neurons.

IPSCs are a powerful *in vivo* tool to interpret *in vitro* results at the cellular level, to identify new modifiers or markers, and to recapitulate disease phenotypes in tissues otherwise difficult to access. Murine iPSCs were used to unravel the function and mutation of the 22q11.2 *TBX1* gene in corticogenesis (Flore et al. 2017), in endo- and mesodermal differentiation (Yan et al. 2014), and in the cardiopharyngeal mesoderm (Nomaru et al. 2021). However, to study the human 22q11.2DS related brain pathology, human iPSCs is the preferential model. In addition, effects of LCR22 structural variation cannot be studied in murine iPSC, due to absence of the LCR22 blocks (Puech et al. 1997).

Patient-derived human iPSCs carrying a 22q11.2 deletion (typically deleted LCR22-A/D locus) were previously generated, differentiated towards different neuronal cell types, and compared with control individuals at the cellular level. The 22q11.2 deletion iPSCs are characterized by reduced neuronal differentiation efficiency and alterations in neuronal characteristics compared to control cell lines (Toyoshima et al. 2016). Khan et al. (2020)

differentiated 15 controls and 15 22q11.2 deletion iPSCs towards 2D cortical neurons and 3D cerebral cortical organoids. Gene expression and neurophysiological experiments revealed alterations in expression of excitability genes and defects at the neuronal activity level, respectively (Khan et al. 2020). In addition, neuronal differentiation allowed to link miRNA interactions (Zhao et al. 2015), expression profiles (Lin et al. 2016; Nehme et al. 2021), and mitochondrial compensation (Li et al. 2021a) of 22q11.2DS patient-derived iPSCs with schizophrenia penetrance. The effects of the 22q11.2 deletion were also studied in blood-brain barrier models: differentiated 22q11.2 deletion iPSCs, generated from patients with schizophrenia, showed immune imbalances promoting neuroinflammation (Crockett et al. 2021) and disruption of the barrier integrity (Li et al. 2021b), suggesting an increased risk for neuropsychiatric disorders.

Although several human iPSC studies did unveil pathophysiologic mechanisms of the 22q11.2DS to the expression of neuropsychiatric disease, human neuronal 22q11.2DS phenotypes remain poorly understood. It will be an important step to discriminate the neuropsychiatric pathophysiology between 22q11.2DS patients with and without schizophrenia at the molecular and cellular level. In addition, studies should be interpreted taking into account the unique genomic structure of the locus and the rearrangement. Afterwards, these neuronal 22q11.2DS models will provide the basis for the screening of new drug discovery and testing cell-therapy based strategies.

8.3.3 Long-term: individualized risk assessment and personalized medicine

This future research will lead to the development of preventive strategies and targeted early interventions, decreasing the associated socio-economic burden experienced by 22q11.2DS patients (Angelis et al. 2015; Benn et al. 2017). Associating structural variation with neuropsychiatric disease will have a major clinical impact in genetic diagnosis and counselling, since parents could receive more clear information about the consequence of the specific deletion in their child. Schizophrenia is diagnosed in about 25% of adults with 22q11.2DS (McDonald-McGinn et al. 2015). Personalized schizophrenia prediction is also an important step towards early diagnosis, prevention, and treatment. As an additional application, fundamental insights in the structure, role, and rearrangement mechanisms of the LCR22s pave the way towards gene editing applications. For example, organoids could be created in which the deletion is rescued using CRISPR-Cas9 and the resulting organoids could eventually replace non-functional tissue (kidney, heart, brain) or absent organs (thymus).

8.4 Conclusions

The 22q11.2DS is the most common genomic disorder in the human population. However, genetic research to the 22q11.2 region and the associated deletion syndrome was hampered due to the presence of complex LCR22s in the locus. These LCR22s are present with remaining gaps in reference genome hg38 and embed the crossover sites for the deletions. In this thesis, we developed new methods and optimized protocols to map the LCR22s and the recombination sites of the 22q11.2 rearrangements. In summary, we uncovered LCR22 variability at several levels. First, the LCR22-A locus is hypervariable in the human population, with haplotypes ranging in size between 250kb and 2Mb. This structural variation is human-specific. Second, the recombinations generating the 22q11.2 rearrangements occur at different sites in the LCR22s. A subgroup of deletions was identified where recombination took place over a palindromic AT-rich repeat, suggesting the involvement of the non-homologous end-joining pathway. In addition, LCR22s may act as mediators of atypical rearrangements as well. Future studies have to unravel how this can be linked to the transcriptomic and 3D organizational level, followed by molecular and cellular studies to unravel the link with neuropsychiatric diseases. Hence, this research will provide a paradigm for the study of other rare genetic disorders with incomplete penetrance and will advance the study of neuropsychiatric disorders in general.

BIBLIOGRAPHY

Abdullaev ET, Umarova IR, Arndt PF. 2021. Modelling segmental duplications in the human genome. *BMC Genomics* **22**.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin — Rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**: 97–101.

Aigner J, Villatoro S, Rabionet R, Roquer J, Jiménez-Conde J, Martí E, Estivill X. 2013. A common 56-kilobase deletion in a primate-specific segmental duplication creates a novel butyrophilin-like protein. *BMC Genet* **14**.

Algady W, Louzada S, Carpenter D, Brajer P, Färnert A, Rooth I, Ngasala B, Yang F, Shaw MA, Hollox EJ. 2018. The Malaria-Protective Human Glycophorin Structural Variant DUP4 Shows Somatic Mosaicism and Association with Hemoglobin Levels. *Am J Hum Genet* **103**: 769–776.

Allderdice PW, Eales B, Onyett H, Sprague W, Henderson K, Lefeuvre PA, Pal4 AG. 1983. Duplication 9q34 Syndrome. *Am J Hum Genet* **35**: 1005–1019.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Alvarez-Prats A, Bjelobaba I, Aldworth Z, Baba T, Abebe D, Kim YJ, Stojilkovic SS, Stopfer M, Balla T. 2018. Schwann-Cell-Specific Deletion of Phosphatidylinositol 4-Kinase Alpha Causes Aberrant Myelination. *Cell Rep* **23**: 2881–2890.

Amati F, Conti E, Novelli A, Bengala M, Digilio MC, Marino B, Giannotti A, Gabrielli O, Novelli G, Dallapiccola B. 1999. Atypical deletions suggest five 22q11.2 critical regions related to the DiGeorge/velo-cardio-facial syndrome. *Eur J Hum Genet* **7**: 903–909.

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**: 1293–302.

Ardui S, Race V, Zablotskaya A, Hestand MS, Van Esch H, Devriendt K, Matthijs G, Vermeesch JR. 2017. Detecting AGG Interruptions in Male and Female FMR1 Premutation Carriers by Single-Molecule Sequencing. *Hum Mutat* **38**: 324–331.

Arnheim N, Li H, Cui X. 1991. Genetic mapping by single sperm typing. *Anim Genet* **22**: 105–115.

Babcock M, Pavlicek A, Spiteri E, Kashork CD, Ioshikhes I, Shaffer LG, Jurka J, Morrow BE. 2003. Shuffling of Genes Within Low-Copy Repeats on 22q11 (LCR22) by Alu-Mediated Recombination Events During Evolution. *Genome Res* **13**: 2519–2532.

Babcock M, Yatsenko S, Hopkins J, Brenton M, Cao Q, De Jong P, Stankiewicz P, Lupski JR, Sikela JM, Morrow BE. 2007. Hominoid lineage specific amplification of low-copy repeats on 22q11.2 (LCR22s) associated with velo-cardio-facial/digeorge syndrome. *Hum Mol Genet* **16**: 2560–2571.

Bacchelli E, Cameli C, Viggiano M, Igliozzi R, Mancini A, Tancredi R, Battaglia A, Maestrini E. 2020. An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray. *Sci Reports 2020 101* **10**: 1–13.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**: 552–564.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte R V., Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002a. Recent Segmental Duplications in the Human Genome. *Science (80-)* **297**: 1003–1007.

Bailey JA, Liu G, Eichler EE. 2003. An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications. *Am J Hum Genet* **73**: 823–834.

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.

Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, Archidiacono N, Schwartz S, Rocchi M, Eichler EE. 2002b. Human-Specific Duplication and Mosaic Transcripts: The Recent Paralogous Structure of Chromosome 22. *Am J Hum Genet* **70**: 83–100.

Bakar SA, Hollox EJ, Armour JAL. 2009. Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins. *Proc Natl Acad Sci U S A* **106**: 853–858.

Barber JCK, Rosenfeld JA, Foulds N, Laird S, Bateman MS, Thomas NS, Baker S, Maloney VK, Anilkumar A, Smith WE, et al. 2013. 8p23.1 duplication syndrome; common, confirmed, and novel features in six further patients. *Am J Med Genet A* **161A**: 487–500.

Bassett AS, Chow EWC. 2008. Schizophrenia and 22q11.2 deletion syndrome. *Curr Psychiatry Rep* **10**: 148–157.

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. *Eur J Hum Genet* **23**: 1286–1293.

Beaujard MP, Chantot S, Dubois M, Keren B, Carpentier W, Mabboux P, Whalen S, Vodovar M, Siffroi JP, Portnoï MF. 2009. Atypical deletion of 22q11.2: Detection using the FISH TBX1 probe and molecular characterization with high-density SNP arrays. *Eur J Med Genet* **52**: 321–327.

Beck CR, Carvalho CMB, Akdemir ZC, Sedlazeck FJ, Song X, Meng Q, Hu J, Doddapaneni H, Chong Z, Chen ES, et al. 2019. Megabase Length Hypermutation Accompanies Human Structural Variation at 17p11.2. *Cell* **176**: 1310-1324.e10.

Bedeschi MF, Colombo L, Mari F, Hofmann K, Rauch A, Gentilin B, Renieri A, Clerici D. 2010. Unmasking of a Recessive SCARF2 Mutation by a 22q11.12 de novo Deletion in a Patient with Van den Ende-Gupta Syndrome. *Mol Syndromol* **1**: 239–245.

Ben-Shachar S, Ou Z, Shaw CA, Belmont JW, Patel MS, Hummel M, Amato S, Tartaglia N, Berg J, Sutton VR, et al. 2008. 22q11.2 distal deletion: a recurrent genomic disorder distinct from DiGeorge syndrome and velocardiofacial syndrome. *Am J Hum Genet* **82**: 214–221.

Bender H-U, Almashanu S, Steel G, Hu C-A, Lin W-W, Willis A, Pulver A, Valle D. 2005. Functional Consequences of PRODH Missense Mutations. *Am J Hum Genet* **76**: 409–420.

Benn P, Iyengar S, Crowley TB, Zackai EH, Burrows EK, Moshkevich S, McDonald-McGinn DM, Sullivan KE, Demko Z. 2017. Pediatric healthcare costs for patients with 22q11.2 deletion syndrome. *Mol Genet Genomic Med* 631–638.

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**: 623–630.

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet 2021 536* **53**: 779–786.

Bi W, Yan J, Stankiewicz P, Park SS, Walz K, Boerkoel CF, Potocki L, Shaffer LG, Devriendt K, Nowaczyk MJM, et al. 2002. Genes in a refined Smith-Magenis syndrome critical deletion interval on chromosome 17p11.2 and the syntenic region of the mouse. *Genome Res* **12**: 713–728.

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. *C open* **9**: E802–E809.

Boerma EG, Siebert R, Kluin PM, Baudis M. 2008. Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: a historical review of cytogenetics in the light of todays knowledge. *Leuk 2009 232* **23**: 225–234.

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.

Boeva V, Zinovyev A, Bleakley K, Vert J-P, Janoueix-Lerosey I, Delattre O, Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27**: 268–269.

Bondeson ML, Dahl N, Malmgren H, Kleijer WJ, Tönnesen T, Carlberg BM, Pettersson U. 1995. Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum Mol Genet* **4**: 615–621.

Bonev B, Cavalli G. 2016. Organization and function of the 3D genome. *Nat Rev Genet* **17**: 772–772.

Botto LD, May K, Fernhoff PM, Correa A, Coleman K, Rasmussen SA, Merritt RK, O'Leary LA, Wong LY, Elixson EM, et al. 2003. A population-based study of the 22q11.2 deletion: phenotype, incidence, and contribution to major birth defects in the population. *Pediatrics* **112**: 101–107.

Bovee D, Zhou Y, Haugen E, Wu Z, Hayden HS, Gillett W, Tuzun E, Cooper GM, Sampas N, Phelps K, et al. 2008. Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat Genet* **40**: 96–101.

Boyd JL, Skove SL, Rouanet JP, Pilaz LJ, Bepler T, Gordân R, Wray GA, Silver DL. 2015. Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. *Curr Biol* **25**: 772–779.

Breetvelt EJ, Smit KC, van Setten J, Merico D, Wang X, Vaartjes I, Bassett AS, Boks MPM, Szatmari P, Scherer SW, et al. 2022. A Regional Burden of Sequence-Level Variation in the 22q11.2 Region Influences Schizophrenia Risk and Educational Attainment. *Biol Psychiatry* **91**: 718–726.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* **63**: 861–869.

Budarf ML, Konkle BA, Ludlow LB, Michaud D, Li M, Yamashiro DJ, Mcdonald-mcginn D, Zackai EH, Driscoll DA. 1995. Identification of a patient with Bernard-Soulier syndrome and a deletion in the DiGeorge/Velo-cardio-facial chromosomal region in 22q11.2. *Hum Mol Genet* **4**: 763–766.

Burnside RD. 2015. 22q11 . 21 Deletion Syndromes : A Review of Proximal , Central , and Distal Deletions and Their Associated Features. *Cytogenet Genome Res* **27709**: 89–99.

Butcher NJ, Kiehl TR, Hazrati LN, Chow EWC, Rogaeva E, Lang AE, Bassett AS. 2013. Association between early-onset Parkinson disease and 22q11.2 deletion syndrome: identification of a novel genetic form of Parkinson disease and its clinical implications. *JAMA Neurol* **70**: 1359–1366.

Calderón JF, Puga AR, Guzmán ML, Astete CP, Arriaza M, Aracena M, Aravena T, Sanz P, Repetto GM. 2009. VEGFA polymorphisms and cardiovascular anomalies in 22q11 microdeletion syndrome: A case-control and family-based study. *Biol Res* **42**: 461–468.

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 Part A* **176**: 2058–2069.

Carlson C, Sirotkin H, Pandita R, Goldberg R, McKie J, Wadey R, Patanjali SR, Weissman SM, Anyane-Yeboa K, Warburton D, et al. 1997. Molecular definition of 22q11 deletions in 151 velo-cardio-facial syndrome patients. *Am J Hum Genet* **61**: 620–9.

Carter MT, St. Pierre SA, Zackai EH, Emanuel BS, Boycott KM. 2009. Phenotypic delineation of Emanuel syndrome (supernumerary derivative 22 syndrome): Clinical features of 63 individuals. *Am J Med Genet A* **149A**: 1712–1721.

Carvalho CMB, Lupski JR. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet 2016 174* **17**: 224–238.

Carvalho CMB, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, Shaw C, Peacock S, Pursley A, Tavyev YJ, et al. 2009. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum Mol Genet* **18**: 2188–2203.

Catacchio CR, Angela F, Maggiolini M, Addabbo PD, Bitonto M, Capozzi O, Signorile ML, Miroballo M, Archidiacono N, Eichler EE, et al. 2018. Inversion variants in human and primate genomes. *Genome Res* **28**: 1–11.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**: 608–611.

Chan S, Lam E, Saghbini M, Bocklandt S, Hastie A, Cao H, Holmlin E, Borodkin M. 2018. Structural variation detection and analysis using bionano optical mapping. *Methods Mol Biol* **1833**: 193-203.

Charrier C, Joshi K, Coutinho-Budd J, Kim JE, Lambert N, De Marchena J, Jin WL, Vanderhaeghen P, Ghosh A, Sassa T, et al. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**: 923–935.

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**.

Cleynen I, Engchuan W, Hestand MS, Heung T, Holleman AM, Johnston HR, Monfeuga T, McDonald-McGinn DM, Gur RE, Morrow BE, et al. 2021. Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion. *Mol Psychiatry* **26**: 4496–4510.

Cole CG, McCann OT, Collins JE, Oliver K, Willey D, Gribble SM, Yang F, McLaren K, Rogers J, Ning Z, et al. 2008. Finishing the finished human chromosome 22 sequence. *Genome Biol* **9**.

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. *Am J Cardiol* **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.

Costa V, Esposito R, Aprile M, Ciccodicola A. 2012. Non-coding rnaand pseudogenes in neurodegenerative diseases: "the (un)usual suspects." *Front Genet* **3**: 1–7.

Costain G, Chow EWC, Silversides CK, Bassett AS. 2011. Sex differences in reproductive fitness contribute to preferential maternal transmission of 22q11.2 deletions. *J Med Genet* **48**: 819–824.

Crockett AM, Ryan SK, Vásquez AH, Canning C, Kanyuch N, Kebir H, Ceja G, Gesualdi J, Zackai E, Mcdonald-Mcginn D, et al. 2021. Disruption of the blood-brain barrier in 22q11.2 deletion syndrome. *Brain* **144**: 1351–1360.

Currall BB, Chiangmai C, Talkowski ME, Morton CC. 2013. Mechanisms for Structural Variation in the Human Genome. *Curr Genet Med Rep* **1**: 81.

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

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. 1990. Centre d'Etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. *Genomics* **6**: 575–577.

Davies RW, Fiksinski AM, Breetvelt EJ, Williams NM, Hooper SR, Monfeuga T, Bassett AS, Owen MJ, Gur RE, Morrow BE, et al. 2020. Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nat Med* **26**: 1912–1918.

de la Chapelle A, Herva R, Koivisto M, Aula P. 1981. A deletion in chromosome 22 can cause DiGeorge syndrome. *Hum Genet* **57**: 253–256.

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

Delihas N. 2018. A family of long intergenic non-coding RNA genes in human chromosomal region 22q11.2 carry a DNA translocation breakpoint/AT-rich sequence. *PLoS One* **13**: 1–19.

Delihas N. 2020. Formation of human long intergenic noncoding RNA genes, pseudogenes, and protein genes: Ancestral sequences are key players. *PLoS One* **15**: 1–19.

Delio M, Guo T, McDonald-Mcginn DM, Zackai E, Herman S, Kaminetzky M, Higgins AM, Coleman K, Chow C, Jarlbrzkowski M, et al. 2013. Enhanced maternal origin of the 22q11.2 deletion in velocardiofacial and digeorge syndromes. *Am J Hum Genet* **92**: 439–447.

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.

Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, Antonacci F, Penewit K, Denman L, Raja A, et al. 2017. The evolution and population diversity of human-specific segmental duplications. *Nat Ecol Evol* **1**: 1–23.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912–922.

DiGeorge A. 1965. Discussion on a new concept of the cellular immunology. *J Pediatr* **67**: 907–908.

Dikow N, Maas B, Gaspar H, Kreiss-Nachtsheim M, Engels H, Kuechler A, Garbes L, Netzer C, Neuhann TM, Koehler U, et al. 2013. The phenotypic spectrum of duplication 5q35.2-q35.3 encompassing NSD1: is it really a reversed Sotos syndrome? *Am J Med Genet A* **161A**: 2158–2166.

Dori N, Green T, Weizman A, Gothelf D. 2017. The Effectiveness and Safety of Antipsychotic and Antidepressant Medications in Individuals with 22q11.2 Deletion Syndrome. *J Child Adolesc Psychopharmacol* **27**: 83–90.

Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, Harshman L, Duyzend MH, Ventura M, Antonacci F, et al. 2017. The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biol* **18**: 1–16.

Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE. 2018. Transcriptional fates of human-specific segmental duplications in brain. *Genome Res* **28**: 1566–1576.

Dumas LJ, O'bleness MS, Davis JM, Dickens CM, Anderson N, Keeney JG, Jackson J, Sikela M, Raznahan A, Giedd J, et al. 2012. DUF1220-domain copy number implicated in human brain-size pathology and evolution. *Am J Hum Genet* **91**: 444–454.

Dunham I, Shimizu N, Roe BA, Chissoe S, Dunham I, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402**: 489–495.

Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, Iqbal Z, Chuang HY, Humphray SJ, Halpern AL, et al. 2017. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* **27**: 157–164.

Edelmann L, Pandita RK, Spiteri E, Funke B, Goldberg R, Palanisamy N, Chaganti RSK, Magenis E, Shprintzen RJ, Morrow BE. 1999. A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum Mol Genet* **8**: 1157–1167.

Edelmann L, Stankiewicz P, Spiteri E, Pandita RK, Shaffer L, Lupski J, Morrow BE. 2001. Two functional copies of the DGCR6 gene are present on human chromosome 22q11 due to a duplication of an ancestral locus. *Genome Res* **11**: 208–217.

Emanuel BS. 2008. Molecular mechanisms and diagnosis of chromosome 22q11.2 rearrangements. *Dev Disabil Res Rev* **14**: 11–18.

Emanuel BS, Shaikh TH. 2001. Segmental duplications: an "expanding" role in genomic instability and disease. *Nat Rev Genet* **2**: 791–800.

Evans DG, Messiaen LM, Foulkes WD, Irving REA, Murray AJ, Perez-Becerril C, Rivera B, McDonald-McGinn DM, Stevenson DA, Smith MJ. 2021. Typical 22q11.2 deletion syndrome appears to confer a reduced risk of schwannoma. *Genet Med* **23**: 1779–1782.

Ewart AK, Morris CA, Atkinson D, Jin W, Sternes K, Spallone P, Stock AD, Leppert M, Keating MT. 1993. Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. *Nat Genet* **5**: 11–16.

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, et al. 2018. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* **173**: 1356-1369.e22.

Figlewicz DA, Delattre O, Guellaen G, Krizus A, Thomas G, Zucman J, Rouleau GA. 1993. Mapping of Human γ-Glutamyl Transpeptidase Genes on Chromosome 22 and Other Human Autosomes. *Genomics* **17**: 299–305.

Flore G, Cioffi S, Bilio M, Illingworth E. 2017. Cortical Development Requires Mesodermal Expression of Tbx1, a Gene Haploinsufficient in 22q11.2 Deletion Syndrome. *Cereb Cortex* **27**: 2210–2225.

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, et al. 2015. Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science (80-)* **347**: 1465–1470.

Forstner AJ, Degenhardt F, Schratt G, Nöthen MM. 2013. MicroRNAs as the cause of schizophrenia in 22q11.2 deletion carriers, and possible implications for idiopathic disease: a mini-review. *Front Mol Neurosci* **6**: 47.

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. *Eur J Hum Genet* **11**: 109–111.

Genovese G, Handsaker RE, Li H, Altemose N, Lindgren AM, Chambert K, Pasaniuc B, Price AL, Reich D, Morton CC, et al. 2013. Using population admixture to help complete maps of the human genome. *Nat Genet* **45**: 406–414.

Gheldof N, Witwicki RM, Migliavacca E, Leleu M, Didelot G, Harewood L, Rougemont J, Reymond A. 2013. Structural variation-associated expression changes are paralleled by chromatin architecture modifications. *PLoS One* **8**.

Giannuzzi G, Schmidt PJ, Porcu E, Willemin G, Munson KM, Nuttle X, Earl R, Chrast J, Hoekzema K, Risso D, et al. 2019. The Human-Specific BOLA2 Duplication Modifies Iron Homeostasis and Anemia Predisposition in Chromosome 16p11.2 Autism Individuals. *Am J Hum Genet* **105**: 947–958.

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

Goidts V, Armengol L, Schempp W, Conroy J, Nowak N, Müller S, Cooper DN, Estivill X, Enard W, Szamalek JM, et al. 2006a. Identification of large-scale human-specific copy number differences by inter-species array comparative genomic hybridization. *Hum Genet* **119**: 185–198.

Goidts V, Cooper DN, Armengol L, Schempp W, Conroy J, Estivill X, Nowak N, Hameister H, Kehrer-Sawatzki H. 2006b. Complex patterns of copy number variation at sites of segmental duplications: an important category of structural variation in the human genome. *Hum Genet* **120**: 270–284.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LDW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science (80-)* **352**.

Grati FR, Molina Gomes D, Ferreira JCPB, Dupont C, Alesi V, Gouas L, Horelli-Kuitunen N, Choy KW, García-Herrero S, de la Vega AG, et al. 2015. Prevalence of recurrent pathogenic microdeletions and microduplications in over 9500 pregnancies. *Prenat Diagn* **35**: 801–809.

Groffen J, Stephenson JR, Heisterkamp N, de Klein A, Bartram CR, Grosveld G. 1984. Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell* **36**: 93–99.

Gu W, Zhang F, Lupski JR. 2008. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**: 4.

Guo T, Diacou A, Nomaru H, Mcdonald-mcginn DM, Hestand M, Demaerel W, Zhang L, Zhao Y, Ujueta F, Shan J, et al. 2018. Deletion size analysis of 1680 22q11.2DS subjects identifies a new recombination hotspot on chromosome 22q11.2. *Hum Mol Genet* **27**: 1150–1163.

Guo T, Repetto GM, McDonald McGinn DM, Chung JH, Nomaru H, Campbell CL, Blonska A, Bassett AS, Chow EWC, Mlynarski EE, et al. 2017. Genome-Wide Association Study to Find Modifiers for Tetralogy of Fallot in the 22q11.2 Deletion Syndrome Identifies Variants in the GPR98 Locus on 5q14.3. *Circ Cardiovasc Genet* **10**: 1–10.

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.

Guo X, Delio M, Haque N, Castellanos R, Hestand MS, Vermeesch JR, Morrow BE, Zheng D. 2015. 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.

Guo X, Freyer L, Morrow B, Zheng D. 2011. Characterization of the past and current duplication activities in the human 22q11.2 region. *BMC Genomics* **12**: 71.

Gur RE, Bassett AS, McDonald-McGinn DM, Bearden CE, Chow E, Emanuel BS, Owen M, Swillen A, Van den Bree M, Vermeesch J, et al. 2017. A neurogenetic model for the study of schizophrenia spectrum disorders: the International 22q11.2 Deletion Syndrome Brain Behavior Consortium. *Mol Psychiatry* **22**: 1664–1672.

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.

Haller M, Mo Q, Imamoto A, Lamb DJ. 2017. Murine model indicates 22q11.2 signaling adaptor CRKL is a dosage-sensitive regulator of genitourinary development. *Proc Natl Acad Sci U S A* **114**: 4981–4986.

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, Mccarroll SA. 2015. Large multiallelic copy number variations in humans. *Nat Genet* **47**: 296–303.

Harel T, Lupski JR. 2018. Genomic disorders 20 years on—mechanisms for clinical manifestations. *Clin Genet* **93**: 439–449.

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

Hestand MS, Nowakowska BA, Vergaelen E, Van Houdt J, Dehaspe L, Suhl JA, Del-Favero J, Mortier G, Zackai E, Swillen A, et al. 2016. A catalog of hemizygous variation in 127 22q11 deletion patients. *Hum genome Var* **3**.

Hochstenbach R, Poot M, Nijman IJ, Renkens I, Duran KJ, Van'T Slot R, Van Binsbergen E, Van Der Zwaag B, Vogel MJ, Terhal PA, et al. 2012. Discovery of variants unmasked by hemizygous deletions. *Eur J Hum Genet* **20**: 748–753.

Hollox EJ, Zuccherato LW, Tucci S. 2022. Genome structural variation in human evolution. *Trends Genet* **38**: 45–58.

Hsiao MC, Piotrowski A, Alexander J, Callens T, Fu C, Mikhail FM, Claes KBM, Messiaen L. 2014. Palindrome-Mediated and Replication-Dependent Pathogenic Structural Rearrangements within the NF1 Gene. *Hum Mutat* **35**: 891–898.

Hsieh PH, Dang V, Vollger MR, Mao Y, Huang TH, Dishuck PC, Baker C, Cantsilieris S, Lewis AP, Munson KM, et al. 2021. Evidence for opposing selective forces operating on human-specific duplicated TCAF genes in Neanderthals and humans. *Nat Commun 2021 121* **12**: 1–14.

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, Dewar K, Katsanis N, Reiter LT, Lander ES, Devon KL, Wyman DW, Lupski JR, Birren B. 2001. The 1.4-Mb CMT1A duplication/HNPP deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes. *Genome Res* **11**: 1018–1033.

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

Ishiguro H, Koga M, Horiuchi Y, Noguchi E, Morikawa M, Suzuki Y, Arai M, Niizato K, Iritani S, Itokawa M, et al. 2010. Supportive evidence for reduced expression of GNB1L in schizophrenia. *Schizophr Bull* **36**: 756–765.

Jafri F, Fink J, Higgins RR, Tervo R. 2011. 22q13.32 Deletion and Duplication and Inversion in the Same Family: A Rare Occurrence. *ISRN Pediatr* **2011**: 829825.

Jeanne M, Vuillaume ML, Ung DC, Vancollie VE, Wagner C, Collins SC, Vonwill S, Haye D, Chelloug N, Pfundt R, et al. 2021. Haploinsufficiency of the HIRA gene located in the 22q11 deletion syndrome region is associated with abnormal neurodevelopment and impaired dendritic outgrowth. *Hum Genet* **140**: 885–896.

Jerome LA, Papaioannou VE. 2001. DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. *Nat Genet* **27**: 286–291.

Jiang W, Zhao X, Gabrieli T, Lou C, Ebenstein Y, Zhu TF. 2015. Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nat Commun 2015 61* **6**: 1–8.

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner P a, Eichler EE. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39**: 1361–1368.

Johnson ME, Cheng Z, Morrison VA, Scherer S, Ventura M, Gibbs RA, Green ED, Eichler EE. 2006. Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc Natl Acad Sci U S A* **103**: 17626–17631.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–1326.

Kang JA, Jeon YJ. 2020. Emerging roles of usp18: From biology to pathophysiology. *Int J Mol Sci* **21**: 1–18.

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 (80-)* **311**: 971.

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

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

Kelley RI, Zackai EH, Emanuel BS, Kistenmacher M, Greenberg F, Punnett HH. 1982. The association of the DiGeorge anomalad with partial monosomy of chromosome 22. *J Pediatr* **101**: 197–200.

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

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

Khan TA, Revah O, Gordon A, Yoon SJ, Krawisz AK, Goold C, Sun Y, Kim CH, Tian Y, Li MY, et al. 2020. Neuronal defects in a human cellular model of 22q11.2 deletion syndrome. *Nat Med* **26**: 1888–1898.

Kremer LS, Distelmaier F, Alhaddad B, Hempel M, Iuso A, Küpper C, Mühlhausen C, Kovacs-Nagy R, Satanovskij R, Graf E, et al. 2016. Bi-allelic Truncating Mutations in TANGO2 Cause Infancy-Onset Recurrent Metabolic Crises with Encephalocardiomyopathy. *Am J Hum Genet* **98**: 358–362.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science (80-)* **360**.

Kruszka P, Addissie YA, Mcginn DE, Porras AR, Share M, Crowley TB, Chung BHY, Mok GTK, Mak CCY, Muthukumarasamy P, et al. 2018. 22q11.2 deletion syndrome in diverse populations. *Am J Med Genet A* **173**: 879–888.

Kuan PF, Yang X, Clouston S, Ren X, Kotov R, Waszczuk M, Singh PK, Glenn ST, Gomez EC, Wang J, et al. 2019. Cell type-specific gene expression patterns associated with posttraumatic stress disorder in World Trade Center responders. *Transl Psychiatry* **9**: 1–11.

Kunishima S, Imai T, Kobayashi R, Kato M, Ogawa S, Saito H. 2013. Bernard-Soulier syndrome caused by a hemizygous GPIbβ mutation and 22q11.2 deletion. *Pediatr Int* **55**: 434–437.

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

Lakich D, Kazazian HH, Antonarakis SE, Gitschier J. 1993. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet* **5**: 236–241.

Lange J, Noordam MJ, Van Daalen SKM, Skaletsky H, Clark BA, Macville M V., Page DC, Repping S. 2013. Intrachromosomal homologous recombination between inverted amplicons on opposing Y-chromosome arms. *Genomics* **102**: 257–264.

Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY, et al. 2019. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun* **10**: 1–14.

Levy A, Demczuk S, Aurias A, Depétris D, Mattei M, Philip N. 1995. Interstitial 22q11 microdeletion excluding the ADU breakpoint in a patient with DiGeorge syndrome. *Hum Mol Genet* **12**: 2417–9.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303.3997v**: 1–3.

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.

Li J, Tran OT, Crowley TB, Moore TM, Zackai EH, Emanuel BS, McDonald-Mcginn DM, Gur RE, Wallace DC, Anderson SA. 2021a. Association of Mitochondrial Biogenesis With Variable Penetrance of Schizophrenia. *JAMA psychiatry* **78**: 911–921.

Li Y, Xia Y, Zhu H, Luu E, Huang G, Sun Y, Sun K, Markx S, Leong KW, Xu B, et al. 2021b. Investigation of Neurodevelopmental Deficits of 22 q11.2 Deletion Syndrome with a Patient-iPSC-Derived Blood-Brain Barrier Model. *Cells* **10**.

Lieberman-aiden E, Berkum NL Van, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-)* **326**: 289–293.

Lin M, Pedrosa E, Hrabovsky A, Chen J, Puliafito BR, Gilbert SR, Zheng D, Lachman HM. 2016. Integrative transcriptome network analysis of iPSC-derived neurons from schizophrenia and schizoaffective disorder patients with 22q11.2 deletion. *BMC Syst Biol* **10**.

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

Liu APY, Chow PC, Lee PPW, Mok GTK, Tang WF, Lau ET, Lam STS, Chan KY, Kan ASY, Chau AKT, et al. 2014. Under-recognition of 22q11.2 deletion in adult Chinese patients with conotruncal anomalies: implications in transitional care. *Eur J Med Genet* **57**: 306–311.

Liu H, Gogos JA, Galke BL, Lenane M, Blundell ML, Sobin C, Heath SC, Roos JL, Robertson B, Wijsman EM, et al. 2002. Genetic variation at the 22q11 PRODH2/DGCR6 locus presents an unusual pattern and increases susceptibility to schizophrenia. *Proc Natl Acad Sci USA* **99**: 3717–3722.

Logsdon GA, Vollger MR, Hsieh PH, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, et al. 2021. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**: 101–107.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580–585.

Louzada S, Komatsu J, Yang F. 2017. Fluorescence In Situ Hybridization onto DNA Fibres Generated Using Molecular Combing. 275–293.

Ludlow LB, Schick BP, Budarf ML, Driscoll DA, Zackai EH, Cohen A, Konkle BA. 1996. Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome. *J Biol Chem* **271**: 22076–22080.

Lupski JR. 2009. Genomic disorders ten years on. *Genome Med* **1**.

Lupski JR, Stankiewicz P. 2005. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* **1**: 0627–0633.

Mao Y, Catacchio CR, Hillier LDW, Porubsky D, Li R, Sulovari A, Fernandes JD, Montinaro F, Gordon DS, Storer JM, et al. 2021. A high-quality bonobo genome refines the analysis of hominid evolution. *Nat 2021 5947861* **594**: 77–81.

Marques-Bonet T, Girirajan S, Eichler EE. 2009a. The origins and impact of primate segmental duplications. *Trends Genet* **25**: 443–454.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009b. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**: 877–881.

McCartney AM, Hyland EM, Cormican P, Moran RJ, Webb AE, Lee KD, Hernandez-Rodriguez J, Prado-Martinez J, Creevey CJ, Aspden JL, et al. 2019. Gene Fusions Derived by Transcriptional Readthrough are Driven by Segmental Duplication in Human. *Genome Biol Evol* **11**: 2676–2690.

McDermid HE, Morrow BE. 2002. Genomic Disorders on 22q11. *Am J Hum Genet* **70**: 1077–1088.

McDonald-McGinn D, Fahiminiya S, Revil T, Nowakowska B, Suhl J, Bailey A, Mlynarski E, Lynch D, Yan A, Bilaniuk L, et al. 2013. Hemizygous mutations in SNAP29 unmask autosomal recessive conditions and contribute to atypical findings in patients with 22q11.2DS. *J Med Genet* **50**: 80–90.

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 Prim* **1**.

McDonald-McGinn DM, Minugh-Purvis N, Kirschner RE, Jawad A, Tonnesen MK, Catanzaro JR, Goldmuntz E, Driscoll D, LaRossa D, Emanuel BS, et al. 2005. The 22q11.2 deletion in African-American patients: an underdiagnosed population? *Am J Med Genet A* **134**: 242–246.

McQuade L, Christodoulou J, Budarf ML, Sachdev R, Wilson M, Emanuel B, Colley A. 1999. Patient with a 22q11.2 deletion with no overlap of the minimal DiGeorge syndrome critical region (MDGCR). *Am J Hum Genet* **3**: 27–33.

Michaelovsky E, Frisch A, Carmel M, Patya M, Zarchi O, Green T, Basel-Vanagaite L, Weizman A, Gothelf D. 2012. Genotype-phenotype correlation in 22q11.2 deletion syndrome. *BMC Med Genet* **13**: 122.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nat 2020 5857823* **585**: 79–84.

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang SP, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK, et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Mittelstaedt T, Schoch S. 2007. Structure and evolution of RIM-BP genes: Identification of a novel family member. *Gene* **403**: 70–79.

Mohammadi MM, Bavi O. 2021. DNA sequencing: an overview of solid-state and biological nanopore-based methods. *Biophys Rev* **14**: 99–110.

Montavon T, Thevenet L, Duboule D. 2012. Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus. *Proc Natl Acad Sci U S A* **109**: 20204–20211.

Morsheimer M, Brown Whitehorn TF, Heimall J, Sullivan KE. 2017. The immune deficiency of chromosome 22q11.2 deletion syndrome. *Am J Med Genet Part A* **173**: 2366–2372.

Mostovoy Y, Yilmaz F, Chow SK, Chu C, Lin C, Geiger EA, Meeks NJL, Chatfield KC, Coughlin CR, Surti U, et al. 2021. Genomic regions associated with microdeletion/microduplication syndromes exhibit extreme diversity of structural variation. *Genetics* **217**.

Mukai J, Liu H, Burt RA, Swor DE, Lai WS, Karayiorgou M, Gogos JA. 2004. Evidence that the gene encoding ZDHHC8 contributes to the risk of schizophrenia. *Nat Genet* **36**: 725–731.

Namjou B, Ni Y, Harley ITW, Chepelev I, Cobb B, Kottyan LC, Gaffney PM, Guthridge JM, Kaufman K, Harley JB. 2014. The effect of inversion at 8p23 on BLK association with lupus in Caucasian population. *PLoS One* **9**.

Nehme R, Pietiläinen O, Artomov M, Tegtmeyer M, Bell C, Ganna A, Singh T, Trehan A, Valakh V, Sherwood J, et al. 2021. The 22q11.2 region regulates presynaptic gene-products linked to schizophrenia. *bioRxiv* 1–74.

Neira-Fresneda J, Potocki L. 2015. Neurodevelopmental Disorders Associated with Abnormal Gene Dosage: Smith-Magenis and Potocki-Lupski Syndromes. *J Pediatr Genet* **4**: 159–167.

Nogueira SI, Hacker AM, Bellucco FTS, Christofolini DM, Kulikowski LD, Cernach MCSP, Emanuel BS, Melaragno MI. 2008. Atypical 22q11.2 deletion in a patient with DGS/VCFS spectrum. *Eur J Med Genet* **51**: 226–230.

Nomaru H, Liu Y, De Bono C, Righelli D, Cirino A, Wang W, Song H, Racedo SE, Dantas AG, Zhang L, et al. 2021. Single cell multi-omic analysis identifies a Tbx1-dependent multilineage primed population in murine cardiopharyngeal mesoderm. *Nat Commun* **12**.

Nota B, Struys EA, Pop A, Jansen EE, Fernandez Ojeda MR, Kanhai WA, Kranendijk M, Van Dooren SJM, Bevova MR, Sistermans EA, et al. 2013. Deficiency in SLC25A1, encoding the mitochondrial citrate carrier, causes combined D-2- and L-2-hydroxyglutaric aciduria. *Am J Hum Genet* **92**: 627–631.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science (80-)* **376**: 44–53.

Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Sudmant PH, Penn O, Chiatante G, Malig M, Benner C, Camponeschi F, et al. 2016. Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**: 205–209.

O'Donnell H, McKeown C, Gould C, Morrow B, Scambler P. 1997. Detection of an Atypical 22q11 Deletion That Has No Overlap with the DiGeorge Syndrome Critical Region. *Am J Hum Genet* **60**: 1544–1548.

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.

Ottaviani D, Lecain M, Sheer D. 2014. The role of microhomology in genomic structural variation. *Trends Genet* **30**: 85–94.

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

Pagnamenta AT, Howard MF, Wisniewski E, Popitsch N, Knight SJL, Keays DA, Quaghebeur G, Cox H, Cox P, Balla T, et al. 2015. Germline recessive mutations in PI4KA are associated with perisylvian polymicrogyria, cerebellar hypoplasia and arthrogryposis. *Hum Mol Genet* **24**: 3732–3741.

Papangeli I, Scambler P. 2013. The 22q11 deletion: DiGeorge and velocardiofacial syndromes and the role of TBX1. *Wiley Interdiscip Rev Dev Biol* **2**: 393–403.

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.

Payne A, Holmes N, Rakyan V, Loose M. 2019. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**: 2193–2198.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**: 1256–1260.

Portnoï MF. 2009. Microduplication 22q11.2: A new chromosomal syndrome. *Eur J Med Genet* **52**: 88–93.

Prasad SE, Howley S, Murphy KC. 2008. Candidate genes and the behavioral phenotype in 22q11.2 deletion syndrome. *Dev Disabil Res Rev* **14**: 26–34.

Puech A, Saint-Joke B, Funke B, Gilbert DJ, Sirotkin H, Copeland NG, Jenkins NA, Kucherlapati R, Morrow B, Skoultchi AI. 1997. Comparative mapping of the human 22q11 chromosomal region and the orthologous region in mice reveals complex changes in gene organization. *Proc Natl Acad Sci U S A* **94**: 14608–14613.

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

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Racedo SE, McDonald-Mcginn DM, Chung JH, Goldmuntz E, Zackai E, Emanuel BS, Zhou B, Funke B, Morrow BE. 2015. Mouse and human CRKL is dosage sensitive for cardiac outflow tract formation. *Am J Hum Genet* **96**: 235–244.

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**: 278–289.

Riggi N, Suva ML, Stamenkovic I. 2021. Ewing's sarcoma. *N Engl J Med* **384**: 154–164.

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

Rodgers K, Mcvey M. 2016. Error-prone repair of DNA double-strand breaks. *J Cell Physiol* **231**: 15–24.

Scambler PJ, Carey AH, Wyse RKH, Roach S, Dumanski JP, Nordenskjold M, Williamson R. 1991. Microdeletions within 22q11 associated with sporadic and familial DiGeorge syndrome. *Genomics* **10**: 201–206.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**: 849–864.

Shaikh TH, Budarf ML, Celle L, Zackai EH, Emanuel BS. 1999. Clustered 11q23 and 22q11 Breakpoints and 3:1 Meiotic Malsegregation in Multiple Unrelated t(11;22) Families. *Am J Hum Genet* **65**: 1595.

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.

Shi W, Massaia A, Louzada S, Handsaker J, Chow W, McCarthy S, Collins J, Hallast P, Howe K, Church DM, et al. 2019. Birth, expansion, and death of VCY-containing palindromes on the human Y chromosome. *Genome Biol* **20**: 1–12.

Sinkus ML, Graw S, Freedman R, Ross RG, Lester HA, Leonard S. 2015. The human CHRNA7 and CHRFAM7A genes: A review of the genetics, regulation, and function. *Neuropharmacology* **96**: 274–288.

Siva N. 2008. 1000 Genomes project. *Nat Biotechnol* **26**: 256.

Smit A, Hubley R, Green P. RepeatMasker.

Stalmans I, Lambrechts D, De Smet F, Jansen S, Wang J, Maity S, Kneer P, Der Ohe M Von, Swillen A, Maes C, et al. 2003. VEGF: A modifier of the de122q11 (DiGeorge) syndrome? *Nat Med* **9**: 173–182.

Stark KL, Xu B, Bagchi A, Lai WS, Liu H, Hsu R, Wan X, Pavlidis P, Mills AA, Karayiorgou M, et al. 2008. Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. *Nat Genet* **40**: 751–760.

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.

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, et al. 2013. Evolution and diversity of copy number variation in the great ape lineage. *Genome Res* **23**: 1373–1382.

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Eichler EE, et al. 2010. Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–646.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al. 2015. Global diversity, population stratification, and selection of human copy number variation. *Science (80-)* science.aab3761-.

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.

Swillen A, McDonald-Mcginn D. 2015. Developmental trajectories in 22q11.2 deletion syndrome. *Am J Med Genet Part C Semin Med Genet* **169**: 172–181.

Swillen A, Moss E, Duijff S. 2018. Neurodevelopmental outcome in 22q11.2 deletion syndrome and management. *Am J Med Genet Part A* **176**: 2160–2166.

Thiele H, Nürnberg P. 2005. HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics* **21**: 1730–1732.

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.

Torres-Juan L, Rosell J, Sánchez-de-la-Torre M, Fibla J, Heine-Suñer D. 2007. Analysis of meiotic recombination in 22q11.2, a region that frequently undergoes deletions and duplications. *BMC Med Genet* **8**: 14.

Toyoshima M, Akamatsu W, Okada Y, Ohnishi T, Balan S, Hisano Y, Iwayama Y, Toyota T, Matsumoto T, Itasaka N, et al. 2016. Analysis of induced pluripotent stem cells carrying 22q11.2 deletion. *Transl Psychiatry* **6**.

Uddin RK, Zhang Y, Siu VM, Fan YS, O'Reilly RL, Rao J, Singh SM. 2006. Breakpoint Associated with a novel 2.3 Mb deletion in the VCFS region of 22q11 and the role of Alu (SINE) in recurring microdeletions. *BMC Med Genet* **7**: 1–10.

Ulahannan N, Pendleton M, Deshpande A, Schwenk S, Behr JM, Dai X, Tyer C, Rughani P, Kudman S, Adney E, et al. 2019. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. *bioRxiv*.

Unolt M, Kammoun M, Nowakowska B, Graham GE, Crowley TB, Hestand MS, Demaerel W, Geremek M, Emanuel BS, Zackai EH, et al. 2020. Pathogenic variants in CDC45 on the remaining allele in patients with a chromosome 22q11.2 deletion result in a novel autosomal recessive condition. *Genet Med* **22**: 326–335.

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. 2018. The Third Revolution in Sequencing Technology. *Trends Genet* **34**: 666–681.

Verdura E, Rodríguez-Palmero A, Vélez-Santamaria V, Planas-Serra L, De La Calle I, Raspall-Chaure M, Roubertie A, Benkirane M, Saettini F, Pavinato L, et al. 2021. Biallelic PI4KA variants cause a novel neurodevelopmental syndrome with hypomyelinating leukodystrophy. *Brain* **144**: 2659–2669.

Vermeesch JR. 2022. The Hunt for the Chromosome 22q11.2 Deletion Syndrome Schizophrenia Genes. *Biol Psychiatry* **91**: 692–693.

Vervoort L, Demaerel W, Rengifo LY, Odrzywolski A, Vergaelen E, Hestand MS, Breckpot J, Devriendt K, Swillen A, McDonald-McGinn DM, et al. 2019. Atypical chromosome 22q11.2 deletions are complex rearrangements and have different mechanistic origins. *Hum Mol Genet* **28**: 3724–3733.

Vervoort L, Dierckxsens N, Pereboom Z, Capozzi O, Rocchi M, Shaikh TH, Vermeesch JR. 2021. 22q11.2 Low Copy Repeats Expanded in the Human Lineage. *Front Genet* **12**.

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. *Am J Hum Genet* **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.

Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP, et al. 2022. Segmental duplications and their variation in a complete human genome. *Science (80-)* **376**.

Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, Popejoy AB, Asri M, Carson C, Chaisson MJP, et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**: 437–446.

Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. 2021. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol 2021 3911* **39**: 1348–1365.

Wat MJ, Shchelochkov OA, Holder AM, Breman AM, Dagli A, Bacino C, Scaglia F, Zori RT, Cheung SW, Scott DA, et al. 2009. Chromosome 8p23.1 Deletions as a Cause of Complex Congenital Heart Defects and Diaphragmatic Hernia. *Am J Med Genet A* **149A**: 1661.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res* **27**: 757–767.

Weksberg R, Stachon AC, Squire JA, Moldovan L, Bayani J, Meyn S, Chow E, Bassett AS. 2007. Molecular characterization of deletion breakpoints in adults with 22q11 deletion syndrome. *Hum Genet* **120**: 837–845.

Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. 1st ed. Springer-Verlag New York, New York.

Wimmer K, Callens T, Wernstedt A, Messiaen L. 2011. The NF1 gene contains hotspots for L1 endonuclease-dependent De Novo insertion. *PLoS Genet* **7**.

Yadav H, Sharma P. 2019. A simple and novel DNA combing methodology for Fiber-FISH and optical mapping. *Genomics* **111**: 567–578.

Yan Y, Su M, Song Y, Tang Y, Tian X, Rood D, Lai L. 2014. Tbx1 Modulates Endodermal and Mesodermal Differentiation from Mouse Induced Pluripotent Stem Cells. *Stem Cells Dev* **23**: 1491–1500.

Zenagui R, Bernicot I, Ranisavljevic N, Haquet E, Ferrieres-Hoa A, Pellestor F, Anahory T. 2019. Inheritance of imbalances in recurrent chromosomal translocation t(11;22): clarification by PGT-SR and sperm-FISH analysis. *Reprod Biomed Online* **39**: 40–48.

Zhang X, Zhang Y, Zhu X, Purmann C, Haney MS, Ward T, Khechaduri A, Yao J, Weissman SM, Urban AE. 2018. Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. *Nat Commun* **9**.

Zhao D, Lin M, Chen J, Pedrosa E, Hrabovsky A, Fourcade HM, Zheng D, Lachman HM. 2015. MicroRNA Profiling of Neurons Generated Using Induced Pluripotent Stem Cells Derived from Patients with Schizophrenia and Schizoaffective Disorder, and 22q11.2 Del. *PLoS One* **10**.

Zhao Y, Guo T, Fiksinski A, Breetvelt E, McDonald-McGinn DM, Crowley TB, Diacou A, Schneider M, Eliez S, Swillen A, et al. 2018. Variance of IQ is partially dependent on deletion type among 1,427 22q11.2 deletion syndrome subjects. *Am J Med Genet Part A* **176**: 2172–2181.

Zhou B, Shin G, Greer SU, Vervoort L, Huang Y, Pattni R, Ho M, Wong WH, Vermeesch JR, Ji HP, et al. Complete and haplotype-specific sequence assembly of segmental duplication-mediated genome rearrangements using CRISPR-targeted ultra-long read sequencing (CTLR-Seq). *bioRxiv*.

Zinkstok JR, Boot E, Bassett AS, Hiroi N, Butcher NJ, Vingerhoets C, Vorstman JAS, van Amelsvoort TAMJ. 2019. Neurobiological perspective of 22q11.2 deletion syndrome. *The lancet Psychiatry* **6**: 951–960.

SCIENTIFIC ACKNOWLEDGEMENT

First, we would like to thank the patients and their families for participation in the projects of this study. We also acknowledge the clinicians at the Centre For Human Genetics (UZ Leuven), especially Prof. Dr. Koen Devriendt, Prof. Ann Swillen, Prof. Dr. Jeroen Breckpot, and Prof. Dr. Hilde Van Esch for the consultation, recruitment, and follow-up of the patients included in this study. In addition, we thank the Genomics Core at the Centre of Human Genetics for sequencing experiments and access to sequencing platforms and Greet Peeters for help with the laboratory work. We are grateful for the collaboration with the laboratory from Professor Alexander E. Urban (Stanford University), for contributing to this research by sharing reagents and temporarily hosting the candidate in his laboratory. Finally, we would also like to thank all contributing authors for reading and correcting the manuscripts.

This work was made possible by grants to Joris R. Vermeesch from Katholieke Universiteit Leuven (Programma financiering Vlaanderen SymBioSys PFV/10/016 and C14/18/092), Fondation Jérôme-Lejeune (project 1665), and Fonds Wetenschappelijk Onderzoek (G0E1117N and G0A2622N). Other funding include Geconcentreerde Onderzoeksacties (GOA/12/015 to Joris R. Vermeesch and Koen Devriendt), National Institute of Mental Health (5U01MH101723-02), Belgian Science Policy Office Interuniversity Attraction Poles (P7/43-BeMGI), and Stichting Marguerite-Marie Delacroix (GV/B-453 to Lisanne Vervoort). In addition, the candidate received a grant from Academische Stichting Leuven (2018/129) for participation in an international conference and a grant from Fonds Wetenschappelijk Onderzoek (V400821N) for a long research stay abroad (Stanford).

PERSONAL CONTRIBUTION

Chapter 3 – The 22q11.2 low copy repeats are characterized by unprecedented size and structural variability

The candidate (Lisanne Vervoort, LV) contributed to the experimental design of the fiber-FISH assay as well as data collection, analysis, and interpretation. The manuscript was drafted in collaboration with the contributing authors.

Chapter 4 – 22q11.2 low copy repeats expanded in the human lineage

All experiments and strategies were designed by LV. Fiber-FISH assays and Bionano optical mapping was performed by LV, as well as data analysis and interpretation. The candidate drafted the manuscript.

Chapter 5 – Investigation of allelic homologous recombination as a mechanism to create new LCR22-A haplotypes

LV was responsible for conceptualization and design of the study. Merlin analyses were performed on the Vlaamse Super Computer. LV performed fiber-FISH assays on the ordered

Coriell cell lines and subsequent data analysis and *de novo* LCR22 assembly. The candidate drafted the chapter.

Chapter 6 – Atypical chromosome 22q11.2 deletions are complex rearrangements and have different mechanistic origins

The study was designed by LV. The candidate was responsible for the experimental work (fiber-FISH assay, PCR design and optimalization, Sanger sequencing, PacBio library preparation) and data analysis (short-read whole-genome sequencing data, fiber-FISH, Sanger, and PacBio sequencing results). The candidate drafted the manuscript.

Chapter 7 – Different loci for NAHR and PATRR-mediated recombination drive the high incidence of 22q11.2 deletion syndrome

LV designed the study and performed all fiber-FISH and Nanopore sequencing experiments. The CTLR-Seq experiments were performed in the laboratory of Alexander E. Urban (Stanford University) under supervision of Dr. Bo Zhou. LV was responsible for the analysis of the fiber-FISH data, part of the sequencing data analysis (PATRR-mediated recombinations), and data interpretation, including fiber-FISH recombination identification and crossover determination at sequence level. The candidate drafted the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare there is no conflict of interest.

CURRICULUM VITAE

Personalia

Name	Lisanne Vervoort
Date of birth	April 13, 1994
Nationality	Belgian
Home address	Zonnebloemstraat 15, box 001
	2600 Berchem (Antwerp)
Mobile phone	+32 499 60 30 31
E-mail	lisannevervoort@hotmail.com

Education

2017 – 2022	Doctoral training in Biomedical Sciences PhD researcher at the Laboratory for Cytogenetics and Genome Research, Department of Human Genetics, University of Leuven, Belgium *'Low copy repeats flanking chromosome 22q11.2 deletion syndrome'* *(Co-)promotors: Joris Vermeesch, Jeroen Breckpot*
2022, January – March	Visiting Student Researcher at the Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Palo Alto (CA), USA *Supervisor: Alexander Urban*
2015 – 2017	Master in Drug Development, specialization pharmacy University of Leuven, Belgium *Magna cum laude (July 6th, 2017)*
2012 – 2015	Bachelor in Pharmaceutical Sciences University of Leuven, Belgium *Cum laude (July 1st, 2015)*
2006 – 2012	Latin – Sciences Kardinaal Van Roey Instituut, Vorselaar, Belgium

Scientific Awards

European Cytogenetics Association Conference (Virtual Conference),
3 July 2021 – 5 July 2021: *Best Poster Award*

European Society of Human Genetics Conference (Gothenburg, Sweden),
15 June 2019 – 18 June 2019: *Isabelle Oberlé Award*
(outstanding presentation in the field of intellectual disability genetics)

29th Genetics Retreat Graduate Meeting (Kerkrade, The Netherlands),
28 March 2019 – 29 March 2019: *First Prize Oral Presentation*

11th Biennial International 22q11.2 Conference (Whistler, Canada),
11 July 2018 – 13 July 2018: *Basic Science Junior Investigator Award*

Grants

2021 – 2022	Research Grant Delacroix Fund
2020	Grant for long research stay abroad *(postponed due to COVID)*
	Awarded by Fonds Wetenschappelijk Onderzoek
2018	Grant for participation in conference abroad,
	Awarded by Academische Stichting Leuven

Certificates, courses, and other research activities

2021 – 2022	Supervision of master student in Biomedical Sciences
	Thesis: 'Fiber-FISH mapping of 22q11.2 rearrangements show locus heterogeneity'
December 2019	Saphyr System Training on Bionano Technology (Bionano Genomics)
February 2019	Essential Tools for R (LStat training course)
April 2018	RNA-Seq analysis for differential expression in GenePattern (VIB)
January 2018	Introduction to the analysis of NGS data (VIB)
2017 – 2018	Permanent Education Course in Human Genetics (BeSHG)

LIST OF PUBLICATIONS

Published articles

Vervoort, L., Dierckxsens, N., Pereboom, Z., Capozzi, O., Rocchi, M., Shaikh, T.H., Vermeesch, J.R. (2021). 22q11.2 Low Copy Repeats Expanded in the Human Lineage. *Frontiers in Genetics*, 12 (7), 12:706641.

Vervoort, L., Demaerel, W., Rengifo, L.Y., Odrzywolski, A., Vergaelen, E., Hestand, M.S., Breckpot, J., Devriendt, K., Swillen, A., McDonald-McGinn, D.M., Fiksinski, A.M., Zinkstok, J.R., Morrow, B.E., Heung, T., Vorstman, J.A.S., Bassett, A.S., Chow, E.W.C., Shashi, V., International 22q11.2 Brain and Behavior Consortium, Vermeesch, J.R. (2019). Atypical chromosome 22q11.2 deletions are complex rearrangements and have different mechanistic origins. *Human molecular genetics*, 28 (22), 3724-3733.

Demaerel, W., Mostovoy, Y., Yilmaz, F., Vervoort, L., Pastor, S., Hestand, M., Swillen, A., Vergaelen, E., Geiger, E.A., Coughlin, C.R., Chow, S.K., McDonald-McGinn, D., Morrow, B.E., Kwok, P-Y., Xiao, M., Emmanuel, B.S., Shaik, T.H., Vermeesch, J. (2019). The 22q11 low copy repeats are characterized by unprecedented size and structural variability. *Genome Research*, 29 (9), 1389-1401.

Preprint articles

Zhou, B., Shin, G., Greer, S.U., Vervoort, L., Huang, Y., Pattni, R., Ho, M., Wong, W.H., Vermeesch, J.R., Ji, H.P., Urban, A.E. (2020). Complete and haplotype-specific sequence assembly of segmental duplication-mediated genome rearrangements using CRISPR-targeted ultra-long read sequencing (CTLR-Seq). *BioRxiv*, 2020.10.23.349621.

Oral presentations

Vervoort, L., Dierckxsens, N., Zhou, B., Cools, R., Heung, T., Peeters, G., Swillen, A., Breckpot, J., Pastor, S., McDonald-McGinn, D., Emanuel, B., Van Esch, H., Bassett, A., Urban, A., Vermeesch, J.R. 22q11.2 rearrangements caused by non-allelic homologous recombination and palindromic AT-rich repeat-mediated pathways. Presented at the 12th Biennial International 22q11.2 Conference, Split, Croatia, 26 Jun 2022 – 28 Jun 2022.

Vervoort, L., Dierckxsens, N., Zhou, B., Cools, R., Heung, T., Peeters, G., Swillen, A., Breckpot, J., Pastor, S., McDonald-McGinn, D., Emanuel, B., Van Esch, H., Bassett, A., Urban, A., Vermeesch, J.R. 22q11.2 rearrangements caused by NAHR and PATRR-mediated pathways. Presented at the European Human Genetics Conference, Vienna, Austria, 11 Jun 2022 – 14 Jun 2022.

Vervoort, L., Peeters, G., Dierckxsens, N., Dehaspe, L., Vancoillie, L., Van Den Bogaert, K., Melotte, C., Van Esch, H., Vermeesch J.R. 22q11.2 inversion in a mosaic 22q11.2 deletion

patient provides insights in LCR22-mediated rearrangements. Presented at the European Human Genetics Conference, Virtual Conference, 28 Aug 2021 – 31 Aug 2021.

Vervoort, L., Demaerel, W., Pereboom, Z., Rocchi, M., Vermeesch, J. (2020). LCR22q11.2 hypervariability is human specific. Presented at the 20th annual BeSHG meeting : Genome for all?, Brussels, 06 Mar 2020 - 06 Mar 2020.

Vervoort, L., Demaerel, W., Mostovoy, Y., Yilmaz, F., Pastor, S., Hestand, M., Swillen, A., Vergaelen, E., Geiger, A., Coughlin, C.R., Chow, S.K., McDonald-McGinn, D., Morrow, B.E., Kwok, P., Xiao, M., Emmanuel, B.S., Shaikh, T.H., Vermeesch, J. (2019). Optical mapping of 22q11.2 low copy repeats reveals structural hypervariability. In: *Online abstracts*, (Abstract No. C19.3). Presented at the European Human Genetics Conference, Gothenburg, Sweden, 15 Jun 2019 - 18 Jun 2019.

Vervoort, L., Demaerel, W., Hestand, M., Swillen, A., Vergaelen, E., Breckpot, J., Devriendt, K., Morrow, B.E., Emmanuel, B., Vermeesch, J. (2019). Optical mapping of 22q11.2 Low Copy Repeats reveals structural interindividual hypervariability. Presented at the 29th Genetics Retreat NVHG Graduate Meeting, Kerkrade, The Netherlands, 28 Mar 2019 – 29 Mar 2019.

Vervoort, L., Demaerel, W., Hestand, M., Swillen, A., Vergaelen, E., Breckpot, J., Devriendt, K., Morrow, B.E., Emmanuel, B., Vermeesch, J. (2019). Optical mapping of 22q11.2 Low Copy Repeats reveals structural interindividual hypervariability. In: *Abstract book*, (Abstract No. O12), (33-34). Presented at the 19th annual BeSHG meeting: Precision Medicine: Application of Genetics in Prevention and Treatment, Liège, Belgium, 15 Mar 2019 - 15 Mar 2019.

Vervoort, L., Demaerel, W., Hestand, M., Swillen, A., Vergaelen, E., Breckpot, J., Devriendt, K., Morrow, B.E., Emanuel, B., Vermeesch, J. (2018). Optical mapping of 22q11.2 low copy repeats reveals structural hypervariability. In: *Program Guide*, (Abstract No. 98), (133-133). Presented at the 11th Biennial International 22q11.2 Conference, Whistler, British Columbia, Canada, 11 July 2018 – 13 July 2018.

Invited talks

Vervoort, L., Demaerel, W., Mostovoy, Y., Yilmaz, F., Pastor, S., Hestand, M., Swillen, A., Vergaelen, E., Geiger, E.A., Coughlin, C.R., McDonald-McGinn, D., Morrow, B.E., Kwok, P-Y., Xiao, M., Emmanuel, B.S., Shaikt, T.H., Vermeesch, J. (2019). Optical Mapping of 22q11.2 Low Copy Repeats reveals structural hypervariability. In: *Abstract book*, (Abstract No. G 08), (25-25). Presented at the NVHG and BeSHG joint annual meeting, Veldhoven, the Netherlands, 19 Sep 2019 - 20 Sep 2019.

Poster Presentations

Vervoort, L., Dierckxsens, N., Pereboom, Z., Capozzi, O., Rocchi, M., Shaikh, T.H., Vermeesch, J.R. (2021). Optical mapping uncovers human-specific expansion of 22q11.2 low copy repeats. Presented at the 13th European Cytogenomics Conference, Virtual Meeting, 3 July 2021 – 5 July 2021.

Vervoort, L., Dierckxsens, N., Pereboom, Z., Capozzi, O., Rocchi, M., Shaikh, T.H., Vermeesch, J.R. (2020). Structural hypervariability of low copy repeats on chromosome 22 is human specific. (Abstract No. 3023). Presented at the American Society of Human Genetics, Virtual Meeting, 27 Oct 2020 - 30 Oct 2020.

Zhou, B., Shin, G., Vervoort, L., Greer, S., Huang, Y., Roychowdhury, T., Pattni, R., Abyzov, A., Vermeesch, J.R., Ji, H.P., Urban, A.E. (2020). Resolving the exact sequence rearrangements of large neuropsychiatric copy number variations at single base-pair resolution using CRISPR-Catch Long-Read Sequencing (CCLR-Seq). (Abstract No. 3561). Presented at the American Society of Human Genetics, Virtual Meeting, 27 Oct 2020 - 30 Oct 2020.

Vervoort, L., Demaerel, W., Pereboom, Z., Rocchi, M., Vermeesch, J. (2020). LCR22q11.2 hypervariability is human specific. In: *Abstract book*, (Abstract No. P13.04.C). Presented at the European Society of Human Genetics, Virtual Conference, 06 Jun 2020 - 09 Jun 2020.