Multi-modal investigation of the schizophrenia-associated 3q29 genomic interval reveals global genetic diversity with unique haplotypes and segments that increase the risk for non-allelic homologous recombination

Feyza Yilmaz[1,9], Umamaheswaran Gurusamy[2,9], Trenell J. Mosley[3], Yulia Mostovoy[2], Tamim H. Shaikh[4], Michael E. Zwick[5], Pui-Yan Kwok[2,6], Charles Lee[1,7], Jennifer G. Mulle[8,*]

1 - The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
2 - Cardiovascular Research Institute and Institute for Human Genetics, UCSF School of Medicine, San Francisco, CA, USA
3 - Graduate Program in Genetics and Molecular Biology, Laney Graduate School, Emory University, 201 Dowman Drive, Atlanta, GA, 30322, USA
4 - Department of Pediatrics, Section of Genetics and Metabolism, University of Colorado School of Medicine, Aurora, CO, USA
5 - Department of Genetics, School of Arts and Sciences, Rutgers University, New Brunswick, NJ, USA
6 - Department of Dermatology, UCSF School of Medicine, San Francisco, CA, USA
7 - Precision Medicine Center, The First Affiliated Hospital of Xi'an Jiaotong University, 277 West Yanta Road, Xi'an 710061, Shaanxi, People's Republic of China
8 - Department of Psychiatry, Robert Wood Johnson Medical School, Rutgers Biomedical and Health Sciences, Rutgers University, New Brunswick, NJ, USA

[9]These authors contributed equally to this work
*Corresponding author: jm2618@cabm.rutgers.edu

Abstract

Chromosomal rearrangements that alter the copy number of dosage-sensitive genes can result in genomic disorders, such as the 3q29 deletion syndrome. At the 3q29 region, non-allelic homologous recombination (NAHR) between paralogous copies of segmental duplications (SDs) leads to a recurrent ~1.6 Mbp deletion or duplication, causing neurodevelopmental and psychiatric phenotypes. However, risk factors contributing to NAHR at this locus are not well understood. In this study, we used an optical mapping approach to identify structural variations within the 3q29 interval. We identified 18 novel haplotypes among 161 unaffected individuals and used this information to characterize this region in 18 probands with either the 3q29 deletion or 3q29 duplication syndrome. A significant amount of variation in haplotype prevalence was observed between populations. Within probands, we narrowed down the breakpoints to a ~5 kbp segment within the SD blocks in 89% of the 3q29 deletion and duplication cases studied. Furthermore, all 3q29 deletion and duplication cases could be categorized into one of five distinct classes based on their breakpoints. Contrary to previous findings for other recurrent deletion and duplication loci, there was no evidence for inversions in either parent of the probands mediating the deletion or duplication seen in this syndrome.

Introduction

Genomic disorders account for a substantial fraction of physical, neurodevelopmental, and psychiatric morbidity [1–3]. Examples include deletions at the 7q11.23 region known as Williams–Beuren syndrome (WBS, OMIM 194050), deletions at the 22q11.2 locus that give rise to 22q11.2 deletion syndrome, (OMIM 611867), and reciprocal pathogenic deletions, duplications at the proximal 16p11.2 locus (OMIM 611913 and OMIM 614671), and the 3q29 deletion syndrome. [4–9]. The 3q29 deletion syndrome (OMIM 609425) was first identified by Rossi and colleagues in 2001 as a cryptic subtelomeric deletion, and later Willatt and colleagues described an additional six individuals with similar 3q29 deletions [7,10]. In 2008, reciprocal 3q29 duplications were identified in 19 individuals when analyzing the genomes of 14,698 individuals with idiopathic mental retardation using array comparative genomic hybridization (aCGH) [11]. The 3q29 deletion syndrome is associated with a >40-fold increased risk for schizophrenia, and patients are also predisposed to developmental delay, intellectual disability, autism spectrum disorder (ASD), anxiety disorders, attention-deficit/hyperactivity disorder (ADHD), congenital heart defects, and additional somatic, neurodevelopmental, and psychiatric phenotypes [12–14]. Similarly, phenotypes observed in individuals with the reciprocal 3q29 duplication syndrome include developmental delay, speech delay, intellectual disability, ocular and cardiac anomalies, microcephaly, dental anomalies, obesity, seizures, and behavioral similarities to ASD [11,14–23]. The 3q29 deletion syndrome has an estimated population prevalence of 1:30,000 [24], and 3q29 duplication syndrome has a population prevalence of 1:8,000-75,000 [14]. Although most deletions are *de novo*, a small number of inherited deletions (7%) have been reported [25].

The 3q29 deletion is typically *de novo*, yet most patients have the same ~1.6 Mbp region of chromosome 3 deleted. These observations suggest there are key risk structures flanking the 3q29 interval that predispose to non-allelic homologous recombination (NAHR) and consequently, formation of the deletions and duplications. It has previously been noted that

complex regions of the human genome, such as the 3q29 region, can predispose individuals to recurrent deletions or duplications associated with genomic disorders [26]. Therefore, characterizing the fine structure of these regions among probands, their parents, and unaffected individuals can help us understand the etiology of the genomic disorder as well as whether certain haplotype configurations are at risk for increased frequency of the genomic disorder. In this study, our aim was to determine the fine structure of the 3q29 region in probands with 3q29 deletion or 3q29 duplication syndromes and their parents. We used an optical mapping technique [27,28] to identify the haplotype structures and breakpoints in 16 probands with the 3q29 deletion and two probands with the 3q29 duplication syndrome. Additionally, we applied the same approach to identify the haplotype structures in the 3q29 region among 161 unaffected individuals from the 1000 Genomes Project and the California Initiative to Advance Precision Medicine. The results from our study have important implications for the understanding of the molecular etiology of the 3q29 deletion and duplication syndromes.

Materials and Methods

Sample collection - 1000 Genomes Project and California Initiative to Advance Precision Medicine cohorts

In this study, we analyzed a ~2 Mbp region on chromosome 3 (chr3:195,428,934-197,230,596; GRCh38), including three segmental duplication (SD) blocks, denoted as SDA, SDB, and SDC. The two SDs closest to the telomere (SDB and SDC) flank the canonical ~1.6 Mbp interval deleted in 3q29 deletion syndrome. We investigated this region in 161 unaffected individuals from 26 diverse populations consisting of five super populations, Africans (AFR) (n=37), Americans (AMR) (n=52), East Asians (EAS) (n=22), Europeans (EUR) (n=27), and South Asians (SAS) (n=23), as part of the 1000 Genomes Project (1000GP) and California Initiative to Advance Precision Medicine (CIAPM) (Table S1). CIAPM samples were collected as described previously [29]. CIAPM samples and 114 unaffected individuals from 1000GP were designated as the University of California San Francisco (UCSF) dataset. Samples from two publicly available datasets, the Human Genome Structural Variation Consortium (HGSVC) and the Human Pangenome Reference Consortium (HPRC), were included in the analyses. HGSVC included Bionano Genomics optical mapping data of three 1000GP samples (HG01573, HG02018, and GM19036), which were previously not studied (TableS1, Web Sources). 44 unaffected individuals from 1000GP had optical mapping data and phased assemblies available through HPRC, two of which (HG00733 and NA19240) were shared between HGSVC and HPRC (Table S1, Web Sources). Additionally, our dataset included optical mapping data of 114 samples that were part of the 1000GP and the CIAPM, three of which (HG00513, HG00732, and NA19239) were shared with HGSVC and not included (Table S1). Cell lines of these additional 1000GP samples were obtained from Coriell and grown in RPMI 1640 media with 15% FBS, supplemented with L-glutamine and penicillin/streptomycin, at 37°C and 5% CO_2.

Sample collection - The 3q29 Project samples

111 Subjects were recruited from the 3q29 Project registry [12] (3q29deletion.org) as
112 previously described [30]. Inclusion criteria were a validated clinical diagnosis of 3q29 deletion
113 syndrome where the subject's deletion overlapped the canonical region (chr3:195,725,000–
114 197,350,000; GRCH37) by ≥ 80%, and willingness and ability to travel to Atlanta, Georgia.
115 Exclusion criteria were any 3q29 deletion with less than 80% overlap with the canonical region
116 and nonfluency in English. 3q29 study subjects underwent deep phenotyping according to an
117 established protocol [30]. Whole blood drawn at the study visit was used for downstream optical
118 mapping analysis. Optical mapping analysis was completed for 16 probands and parental
119 genomes were analyzed when available. For 10 probands, both biological parents were
120 available for analysis (deletions were confirmed to be *de novo*). Four probands had one parent
121 available for analysis; two probands were singletons (Table S2). Two additional probands with
122 the 3q29 duplication syndrome were also included in the present study; the same protocol was
123 followed for optical mapping analysis. All subjects had genotyping arrays completed for parent-
124 of-origin analysis [31].
125
126 ***High-molecular-weight DNA extraction***
127 Ultra-high-molecular-weight DNA was extracted from cell lines of 1000GP and CIAPM
128 samples (n=114), and from whole blood samples from the 3q29 Project (n=46) according to the
129 Bionano Prep SP Fresh Cells DNA Isolation protocol, revision C (Document #30257), using a
130 Bionano SP Blood & Cell DNA Isolation Kit (catalog #80030). In short, 1.5 million cells were
131 centrifuged and resuspended in a solution containing detergents, proteinase K, and RNase A.
132 DNA was bound to a silica disk, washed, eluted, and homogenized via 1- hour end-over-end
133 rotation at 15 rpm, followed by an overnight rest at room temperature. Isolated DNA was
134 fluorescently tagged at motif CTTAAG by the enzyme DLE-1 and counter-stained using a
135 Bionano PrepTM DNA Labeling Kit – Direct Label and Stain (catalog #8005) according to the
136 Bionano Prep DLS Protocol, revision F (Document #30206). A total of 750 ng of purified

137 genomic DNA was labeled by incubating with DL-Green dye and DLE-1 Enzyme in DLE-1
138 Buffer for 2 hours at 37°C, followed by heat inactivation of the enzyme for 20 minutes at 70°C.
139 The labeled DNA was treated with Proteinase K at 50°C for 1 hour, and excess DL-Green dye
140 was removed by membrane adsorption. The DNA was stored at 4°C overnight to facilitate DNA
141 homogenization and then quantified using a Qubit dsDNA HS Assay Kit (Molecular Probes/Life
142 Technologies). The labeled DNA was stained with an intercalating dye and left to stand at room
143 temperature for at least 2 hours before loading onto a Bionano Chip. The DNA was loaded onto
144 the Bionano Genomics Saphyr Chip and linearized and visualized by the Saphyr systems. Data
145 collection was performed using Saphyr 2nd generation instruments (Part #60325) and
146 Instrument Control Software (ICS) version 4.9.19316.1. The DNA backbone length and
147 locations of fluorescent labels along each molecule were detected using the Saphyr system
148 software.
149
150 ***De novo assembly of optical genome maps***
151 Single-molecule genome maps of 1000GP, CIAPM, and The 3q29 Project samples were
152 assembled *de novo* into genome maps using the assembly pipeline Bionano Solve v3.5,
153 developed by Bionano Genomics with default settings [32] as described previously [33]. In short, a
154 pairwise comparison of DNA molecules (min 250 kbp) was generated to produce the initial
155 consensus genome maps. During an extension step, molecules were aligned to genome maps,
156 and maps were extended based on the molecules aligning past the map ends. Overlapping
157 genome maps were then merged. Extension and merge steps were repeated five times before a
158 final refinement of the genome maps. Clusters of molecules aligned to genome maps with
159 unaligned ends >30 kbp in the extension step were re-assembled to identify all alleles. To
160 identify alternate alleles with smaller size differences from the assembled allele, clusters of
161 molecules aligned to genome maps were detected with internal alignment gaps of size <50 kbp,

162 in which case the genome maps were converted into two haplotype maps. The final genome
163 maps were aligned to the reference genome GRCh38.
164
165 ***Single-molecule analysis pipeline for optical genome maps***
166 In the 3q29 SD blocks, SDA, SDB, and SDC, we defined segments according to
167 characteristic banding patterns of Bionano Genomics optical mapping labels. We color-coded
168 these segments and displayed them according to assigned colors. The order of the segments
169 are: SDA-26 kbp (magenta), SDA-5kbp (blue), SDA-11 kbp (yellow), SDA-4 kbp (red), ~33 kbp
170 (maroon, SDA and unique region), ~15 kbp (orange, unique region), ~124 kbp (green, unique
171 region), SDB-5kbp 1st copy (blue), SDB-15kbp (magenta partial copy), SDB-4 kbp (red), SDB-
172 11kbp (yellow), SDB-5kbp 2nd copy (blue), SDC-11 kbp (yellow), SDC-5 kbp (blue), and SDC-
173 15 kbp (magenta partial copy) (Table S3). Structural variations and haplotypes at the 3q29 locus
174 were analyzed and single molecules were evaluated in 161 unaffected individuals from 26
175 diverse populations from the 1000GP and CIAPM, and 46 samples from the 3q29 Project using
176 the Optical Maps to Genotype Structural Variation (OMGenSV) package as described
177 previously [34,35]. To identify 3q29 haplotypes in samples from 114 unaffected individuals which
178 were part of 1000GP and CIAPM, the assembled contigs were visualized using the “anchor”
179 mode in OMView from the OMTools package [36], and Bionano Access was used for HGSVC
180 (n=3), HPRC (n=44), and the 3q29 Project (n=46) samples. Haplotypes were manually identified
181 from these visualizations, and corresponding consensus map (cmap) files were constructed for
182 each haplotype if the contig in the 3q29 region was not contiguous, including at least 500 kbp of
183 the unique flanking region where applicable. When 3q29 SDs contained multiple SVs and were
184 too long to analyze from end to end with single molecules (>350 kbp), molecules were
185 subdivided into groups that were anchored in the proximal or distal unique regions of 3q29 SDA
186 and SDB. For each haplotype, the corresponding cmaps were compiled into a single file and
187 used as a reference input file for the OMGenSV pipeline, along with local molecules from each

188 sample and a set of “critical regions" (SDA:195,578,485-195,817,578 Mbp; SDB:195,804,017-
189 196,073,500 Mbp; SDC:197,557,633-197,743,251 Mbp; GRCh38) defining the areas on each
190 cmap that molecules need to span to support the presence of that haplotype in the sample. The
191 3q29 Project sample cohort haplotypes were identified by using 1000GP and CIAPM individuals’
192 3q29 haplotype maps as a reference, and molecule support was confirmed by following the
193 steps described above. Images of molecule support were obtained by using OMTools and
194 Bionano Access (Data and Code Availability). Fisher’s exact test was used to test the
195 significance of the 3q29 haplotypes in R Studio (version 1.4.1717). 100,000 replicates were
196 used for the Monte Carlo test. The expected value for each haplotype was calculated by the
197 following command:

198 ```
tab.exp <- round(chisq.test(data)$expected)
```

199 Then we used expected values for Fisher’s exact test:

200 ```
fisher.test(data, simul=T, B =100000)$p.value
```

201 Association plots were generated with the following command:

202 ```
tab.sel <- as.matrix(data[freq$Total >5,])
```

203 ```
assocplot(tab.sel, col=c("red","green"))
```

204 Finally, we annotated all haplotypes detected in our study with GENCODE v37 to identify
205 overlapping genomic elements, including protein-coding genes and noncoding RNAs.
206

207 ***Breakpoint mapping and trio analysis of the 3q29 Project samples***

208 For each proband (n=18, 16 probands with the 3q29 deletion and two probands with the
209 3q29 duplication syndrome), *de novo* assembled contigs were aligned to the GRCh38 human
210 genome reference assembly, and the 3q29 region was manually inspected to identify deletion
211 and duplication breakpoints. Once identified, the underlying single molecules were examined to
212 verify that deletion and duplication breakpoints were well supported by single molecules (n ≥ 3
213 molecules). Then, the sequence identity of approximate breakpoints and the corresponding

214 sequence in GRCh38 was determined by the NCBI *blastn* tool (BLASTN 2.9.0+) with the
215 following command:

```
blastn -query $file1 -subject $file2 -num_alignments 5 -
num_descriptions 5 -out $alignmentoutputfile
```

218 Only "the best alignment" was included in the final output file. Haplotypes of the parents were
219 identified as described above. Finally, the parents transmitting the intact chromosome and
220 deletion parent of origin were identified by visual pairwise comparison of the haplotypes using
221 Bionano Access.
222

223 ***Validation using orthogonal data***

224 Publicly available optical mapping data and phased assemblies of 44 unaffected
225 individuals as part of the 1000GP from the HPRC were analyzed to confirm the 3q29 haplotypes
226 which were identified by using Bionano Genomics optical mapping data. First, HPRC phased
227 assembly fasta files were converted to *in silico* maps and haplotypes were detected as
228 described above. Then a visual pairwise comparison of optical mapping and phased assembly
229 3q29 haplotypes for each sample was performed by using Bionano Access (Figure S1-S4). All
230 HPRC phased assemblies (n=44) except the paternal haplotype of HG01928 supported the
231 haplotypes identified in optical mapping data (Figure S5).
232

233 **Results**

234 The 3q29 region contains three segmental duplication (SD) blocks within a ~2 Mbp
235 region, SDA (~73 kbp), SDB (~64 kbp), and SDC (~41 kbp) (GRCh38) (Figure 1, Table S3).
236 Colored arrows depicting the 3q29 segments and their orientations were determined based on
237 the labeling pattern from the optical mapping data and the sequence identity between 3q29 SDs
238 (Table S4) as described in the Materials and Methods. In the GRCh38 human genome
239 reference assembly we identified three copies of ~26 kbp segment (magenta), four copies of ~5
240 kbp (blue), three copies of ~11 kbp (yellow), two copies of ~4 kbp (red), one copy of ~33 kbp
241 (maroon), one copy of ~15 kbp (orange), and one copy of ~124 kbp (green) segments (Figure
242 1B).

243

244 ***3q29 haplotypes in unaffected individuals from 1000GP and CIAPM***

245 In this study, we used *de novo* assembled optical genome maps having an average
246 effective coverage of 102.4 X and an average N50 of 301.96 kbp. We identified a total of 36
247 different haplotypes for the 3q29 region (195,605,705 - 197,706,750; GRCh38) among 161
248 individuals from the 1000GP and CIAPM cohort. Within the SDC region, no variation was
249 observed across samples. Therefore, all haplotypes were based on the variation observed
250 between SDA and SDB. These 161 individuals were from 26 diverse populations representing
251 five super populations: AFR, AMR, EAS, EUR, and SAS (Table S2). The haplotype size of the
252 3q29 region ranged from ~287 kbp (H28, HG03863-SAS) to 859 kbp (H24, BC02901-AMR, and
253 BC03702-EUR) (Figure 1, Table S5). The smallest haplotype, H28, is missing the proximal
254 ~440 kbp and the distal ~132.5 kbp of the largest haplotype H24 and lacks five copies of the
255 following segments: ~26 kbp, ~5 kbp, ~11 kb, and ~33 kbp (Figure S6). Of 18 previously
256 described haplotypes, we identified 13 haplotypes in our samples; H10, H11, H12, H14, and
257 H18 from Ebert and colleagues [33] were not observed. We detected 18 novel haplotypes, H19-
258 H36. The most common three haplotypes among all 161 unaffected individuals were H3

259 (14.91%), H1 (14.29%), and H2 (12.42%). The H3 haplotype (14.91%) represents the GRCh38
260 human genome reference assembly haplotype for the 3q29 region, and the H2 (12.42%)
261 haplotype was described in the most recent human genome assembly as part of the Telomere-
262 to-Telomere (T2T) Consortium [37]. We also detected 14 haplotypes that were each observed
263 once among all 161 genomes (i.e., 322 chromosomes) examined and therefore referred to as
264 "singletons". To assess the accuracy of the haplotypes identified by optical mapping data, we
265 overlaid long-read sequencing-based phased assemblies (HPRC dataset, n=44 (Table S5)) to
266 the optical mapping-based haplotypes. 20 haplotypes (11 previously identified and 9 novel)
267 assessed in this manner were confirmed by long-read phased assemblies. One discrepancy
268 was detected between the optical mapping and phased assembly haplotype structures, for one
269 chromosome from sample HG01928. The phased assembly suggested that haplotypes were
270 H1/H5, and the optical mapping data suggested that haplotypes were H4/H5. Single DNA
271 molecules (>10) from the optical mapping data were anchored to the proximal and distal unique
272 region and span at least 90% of the 3q29 region and confirmed the presence of the H4
273 haplotype in HG01928 (Figure S5). Our data suggest that the phased assembly for this
274 individual may have been incorrectly assembled.
275

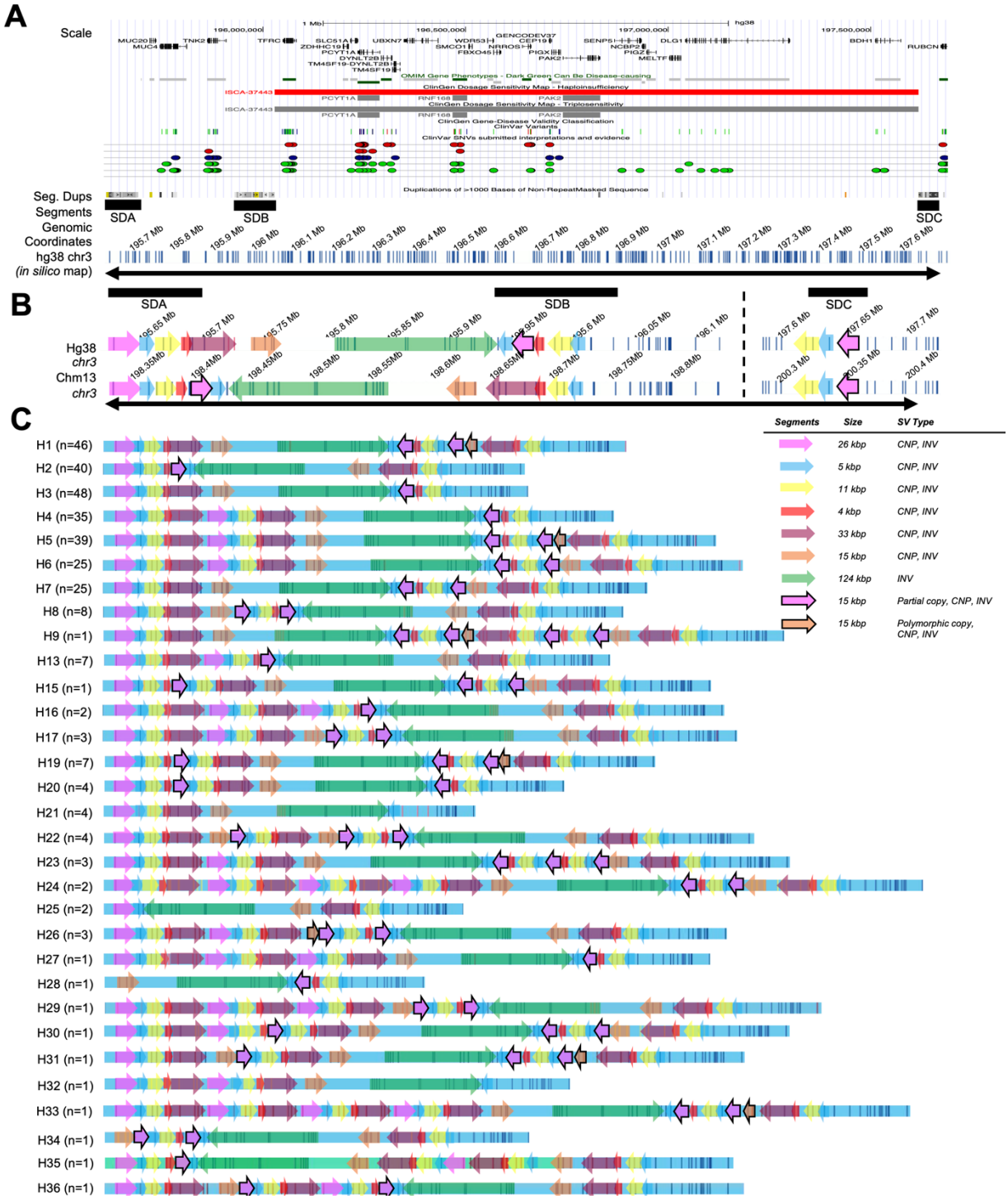

Figure 1: The segment structure of GRCh38 3q29 region and 3q29 haplotypes identified in this study.
(A) 3q29 genomic locus with SDA, SDB, SDC, OMIM Gene Phenotypes, ClinGen Dosage Sensitivity Map - Haploinsufficiency, ClinGen Dosage Sensitivity Map - Triplosensitivity, ClinGen Gene-Disease Validity Classification, ClinVar Variants, ClinVar SNVs, and Duplications of > 1000 Bases of Non-RepeatMasked Sequence represented as Tracks. The 3q29 GRCh38 in silico map is represented as the last track. (B) 3q29 segments of GRCh38 and T2T/chm13 with SDA, SDB, and SDC represented as black boxes on top and genomic segments as colored arrows overlaid on the in silico maps (white background with vertical blue lines). Dotted line: the unique region between SDB and SDC, 196.1-197.6Mbp. (C) The structure and prevalence of 13 previously reported and identified among our samples (H1-H9, H13, H15-H17) and 18 novel 3q29 haplotypes (H19-H36) of 3q29 SDA and SDB. Black arrows in panels A and B represent the region included in the analyses. Novel haplotypes were ordered by prevalence in the population. Each colored arrow, magenta, blue, yellow, maroon, orange, green, represents 3q29 segments. CNP: copy number polymorphism, INV: inversion.

We found significant differences in haplotype frequency between populations (Figure 2, *p-value = 0.000001 Fisher's exact test*). One of the most common haplotypes, H3, was observed at ≥ 7% frequency (≥ 14% in four populations) in all five super populations (Figure 2A). In the AFR population, three haplotypes account for 48% of the haplotype pool: H2 (22%), which was significantly enriched compared to other haplotypes and all other populations (*p-value = 0.009, Fisher's exact test*), H3 (15%), and H8 (11%) (Figure 2B). Interestingly, the H8 haplotype, which contains the second-largest inversion observed in the 3q29 region (~289 kbp), was observed only in the AFR population (Figure 2B, *p-value = 5.73972e-06, Fisher's exact test*). Similarly, H19 haplotype was enriched in the AFR population (p-value = *5.73972e-06*, Fisher's exact test) Furthermore, the H5 haplotype was enriched in the EAS population (41%, p-value = 7.731026e-08, Fisher's exact test compared to other haplotypes and all other populations). The H6 haplotype was enriched in the EUR population (19%, p-value = 0.0002, Fisher's exact test compared to all other populations and haplotypes). The H1 haplotype was enriched in the SAS population (26%, p-value = 0.002, compared to other haplotypes and all other populations based on the Fisher's exact test). The SVs differ from one haplotype to the other, often overlapping protein-coding genes (e.g., *MUC4* and *MUC20*), pseudogenes (e.g., *SDHAP1* and *SDHAP2*), and lincRNAs (e.g., lncRNA MUC20-OT1).

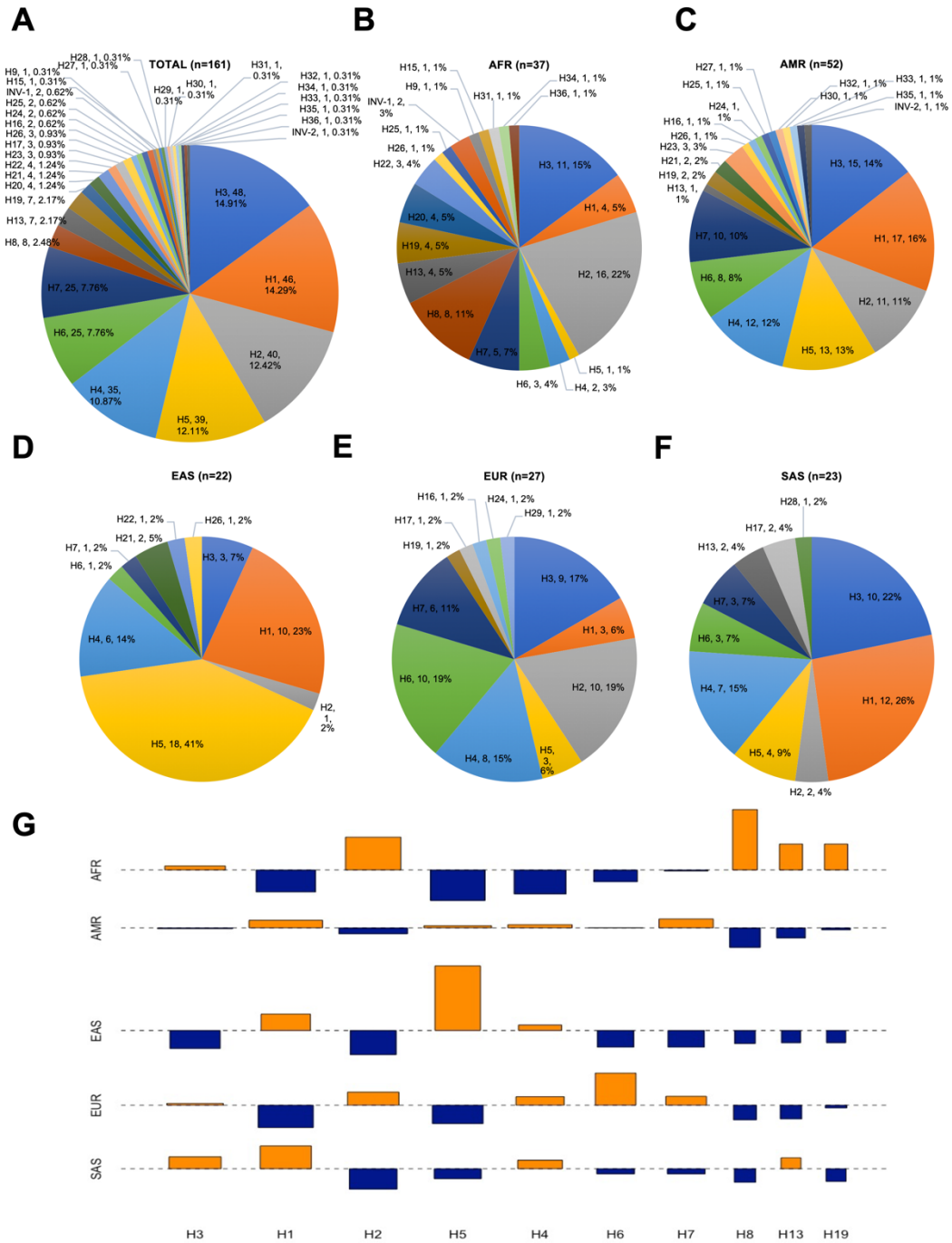

Figure 2: Prevalence of the 3q29 haplotypes represented in five major populations*. We identified 3q29 haplotypes in each sample and then calculated the prevalence of 3q29 haplotypes in five major populations. (A) The distribution of haplotypes identified in total. (B) the distribution of haplotypes in the AFR population, (C) the distribution of haplotypes in the AMR population. (D) the distribution of haplotypes in the EAS population. (E) the distribution of haplotypes in the EUR population, (F) The distribution of haplotypes in the SAS population. AFR - African, AMR - American, EAS - East Asian, EUR - European, SAS - South Asian. Each color in the pie charts represents a distinct 3q29 haplotype. The order in the labels in pie charts represents haplotype ID, count, prevalence. Singleton haplotypes are H9, H15, H27-H36, and INV-2. n represents the number of samples. (G) - Cohen-Friendly association plot depicting the relationship between haplotypes and populations. Each cell has a rectangle with height proportional to the difference between observed and expected (normalized to expected) and width proportional to the square root of expected counts so that the area of the rectangle is proportional to the difference in observed and expected counts. Expected values were calculated as described in the Materials and Methods. If the observed count is greater than expected, the rectangle rises above the baseline and is colored in red.*

321 ***Inversions among the 3q29 haplotypes***

322 Among the 36 haplotypes identified from unaffected samples, we detected three different
323 types of inversions that were > 100 kbp in size: two large inversions (~ 2.03 Mbp and ~2.13
324 Mbp) between SDA and SDC, and a smaller inversion (~289 kbp) between SDA and SDB which
325 was observed on twelve distinct haplotypes (Figure 3, Figure S3). Additionally, we observed
326 inversions of each 3q29 segment in distinct haplotypes. The most prevalent inversion was a
327 ~289 kbp inversion which was observed at a frequency > 10% in all five super populations on
328 the following haplotypes: H2, H8, H13, H16, H17, H22, H25, H26, H29, H34, H35, and H36
329 (Figure 3A, Figure S7). This previously reported inversion lies in a unique region between the
330 3q29 SDA and SDB blocks and was observed in seven novel and five previously identified
331 haplotypes in 69 unaffected individuals (76 chromosomes, Figure S3, Table S6) [38,39]. The two
332 larger >2 Mbp inversions between 3q29 SDA and SDC (Figure 3, Figure S8) were identified
333 among three unaffected individuals. The ~2.13 Mbp inversion (INV-1) (Figure 3B) was observed
334 in HG03470-AFR and potentially arose on haplotype H8, and the ~2.03 Mbp inversion (INV-2)
335 (Figure 3C) was observed in two individuals, GM19984-AFR and BC04902-AMR, and
336 potentially arose on H20 haplotype. The ~2.03 Mbp and ~2.13 Mbp inversions are similar to the
337 ~2 Mbp inversion described previously by Antonacci and colleagues (Figure 3D) [40]. However,
338 the breakpoints of the ~2 Mbp inversion were localized to a 15.5 kbp region (chr3:196,868,578 –
339 196,884,133 and chr3:198,832,975 – 198,848,521; hg17/UCSC 2004), and a more precise
340 definition was not provided because the proximal breakpoint overlapped with an assembly gap
341 in the hg17/UCSC 2004 human genome reference assembly. In this study, we have found that
342 breakpoints for all inversions clustered to a ~5 kbp (blue) paralogous SD copies within SDA and
343 SDC. These paralogous copies share ~98% sequence identity which suggests that NAHR is the
344 mechanism causing these inversions.

345

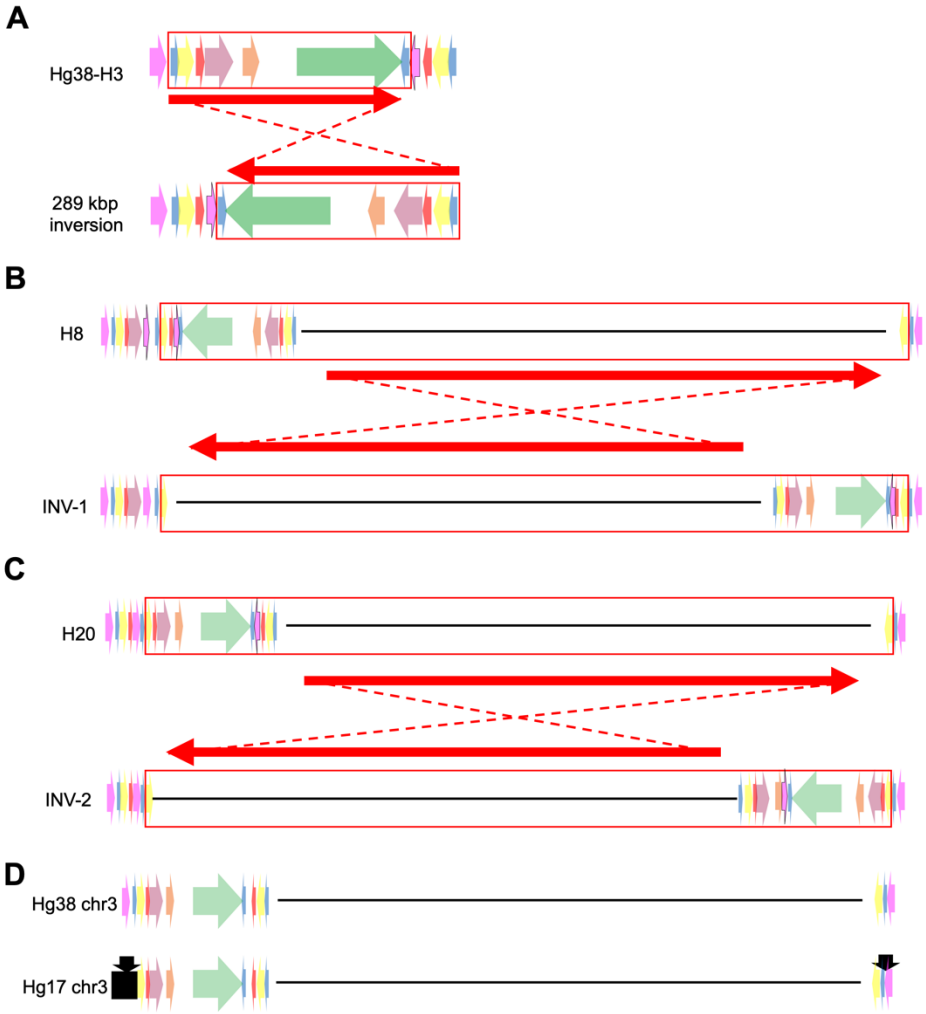

Figure 3: Inversions > 100 kbp identified from the 1000GP and CIAPM samples.
(A) Hg38/GRCh38 human genome reference assembly presented at the top and the 289 kbp inversion, including the 5, 11, 4, 33, 15, and 124 kbp segments, represented at the bottom. (B) - Upper rectangle represents the H8 haplotype of the 3q29 region. The rectangle at the bottom represents the structure of INV-1. (C) - Upper rectangle represents the H20 haplotype of the 3q29 region. The rectangle at the bottom represents the structure of INV-2. (D) - Representation of Hg38/GRCh38 and Hg17/NCBI35 reference assemblies. Black arrows in Hg17/NCBI35 represent inversion breakpoints identified by Antonacci and colleagues. Colored arrows in each panel represent 3q29 segments, red rectangles indicate the region involved in the inversion, red arrows represent the orientation of the region before and after the inversion and dashed lines show the inversion.

3q29 probands and parental haplotypes

3q29 deletion syndrome is associated with a >40-fold increased risk for schizophrenia and can also be associated with neurodevelopmental disorders. Therefore, characterizing the fine structure of the 3q29 region in probands and their parents is critical to gain a better understanding of disease etiology. In this study, we analyzed 16 probands with the 3q29 deletion and two probands with the 3q29 duplication syndrome from the 3q29 Project using

362 Bionano Genomics optical mapping technology (Figure 4A). The average effective coverage of
363 assembly was 102.4 X and the average molecule N50 was 301.96 kbp (Table S2). The
364 recurrent ~1.6 Mbp deletion in the 3q29 deletion syndrome probands results from one break in
365 SDB and one break in SDC [7,10,11]. The unique genomic sequence residing between SDA and
366 SDB is typically intact, not deleted, in the 3q29 deletion syndrome probands. We detected the
367 recurrent ~1.6 Mbp deletion in 15 probands. In one proband we detected a smaller ~1.2 Mbp
368 deletion lying completely within the ~1.6 Mbp 3q29 deletion region (Figure 4A, Family 18). This
369 smaller deletion contains 19 of the 21 genes completely within its breakpoints. The centromeric
370 breakpoint falls within the *TFRC* gene, deleting the promoter and the first four exons. On the
371 telomeric side, the *BDH1* gene is intact. There is no discernible difference in phenotypic severity
372 in this proband, despite their smaller deletion. None of the 15 deletions (~1.6 Mbp (n=14) and
373 ~1.2 Mbp (n=1)) were seen in either of the parents and were therefore considered to be *de novo*
374 events. One proband inherited the deletion (~1.6 Mbp) from his father (Family 20).

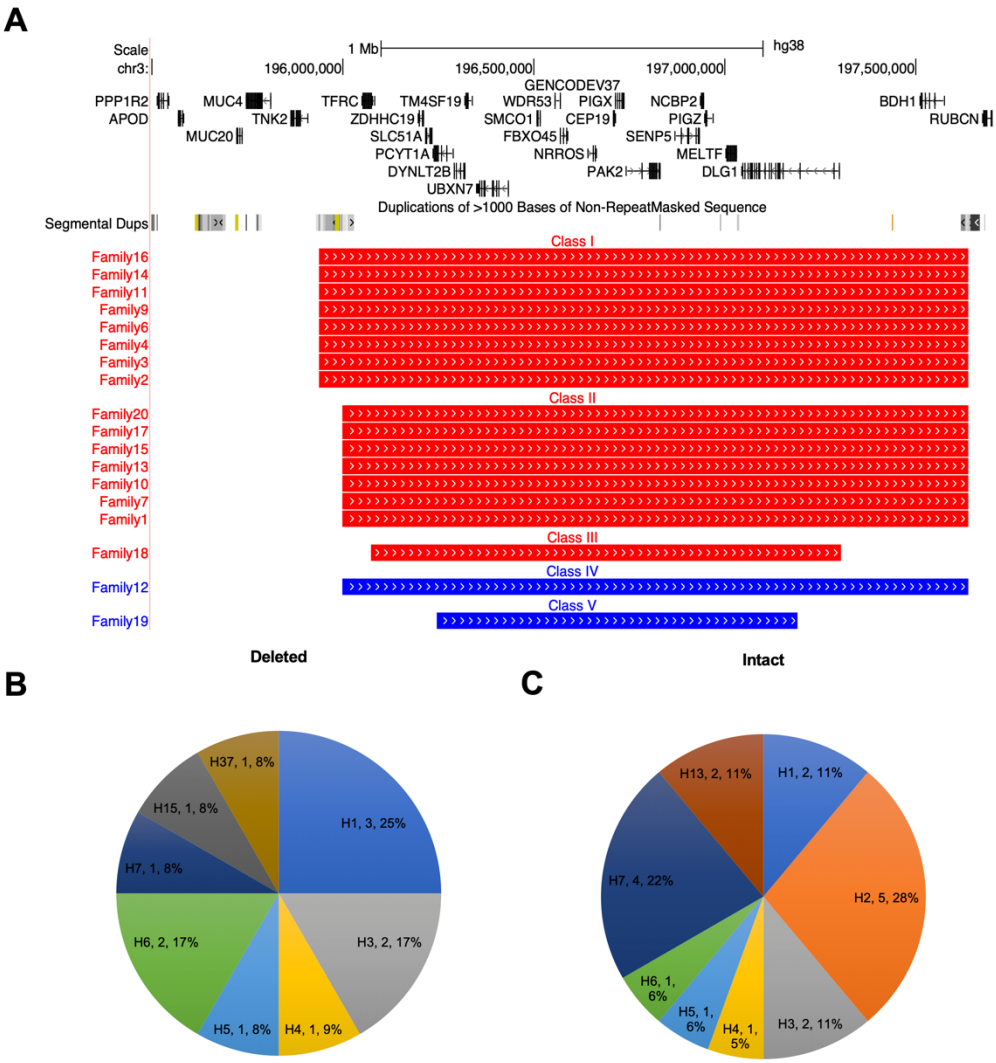

Figure 4: 3q29 Deletions and duplications reported among the studied probands. *(A) Genes and duplications of > 1000 Bases of Non-Repeat Masked Sequence represented as Tracks overlapping the 3q29 region (top). Red lines represent deletion, and blue lines duplication cases analyzed in this study (bottom). (B) Distribution of the 3q29 haplotypes observed in deleted chromosomes in probands. (C) Distribution of the 3q29 haplotypes observed in intact chromosomes in probands. The order in the labels in pie charts represents haplotype IDs, count, and prevalence.*

To investigate whether certain haplotypes predispose to the 3q29 deletion or duplication syndrome, we characterized the 3q29 region, determined the haplotype in probands (Figure 4B-C) and available parents, and finally determined which chromosomes were inherited by the probands. For ten probands, DNA from both biological parents was available (Table S7). For two probands, DNA from only one parent was available (Table S7). For four probands, DNA was not available from either of the parents (Table S7). When possible, we determined on which inherited haplotype the deletion occurred, and which chromosome was inherited in an intact

389 state for each proband. Interestingly, we identified a novel haplotype, H37, in one father (Family
390 18), who was the parent of origin of the deletion observed in the proband (Figure S9). However,
391 we did not observe an enrichment of 3q29 deletions or duplications with any particular
392 haplotype.
393
394 ***Defining five breakpoint classes for the 3q29 deletion and duplication syndromes***
395 While it has been known that breakpoints for 3q29 deletions and duplications generally
396 occur within two SD blocks (SDB and SDC), the breakpoints have not been precisely mapped.
397 Among the 16 probands with the 3q29 deletions, we identified three deletion classes (Table S8).
398 The Class I deletion breakpoints (~1.69 Mbp, 195.94-197.64 Mbp; GRCh38) were observed in
399 eight probands, with the proximal breakpoint localized to the first ~5 kbp segment (blue, Figure
400 5A) in SDB, and the distal to a ~5 kbp paralogous copy in SDC (blue, Figure 5A). Although
401 optical mapping data indicated that the deletion breakpoints localized to the ~5 kbp segments,
402 the possibility of breakpoint localization to the ~15 kbp segments (magenta, partial copies of
403 SDA-26 kbp segment) in SDB and SDC cannot be excluded. The optical mapping resolution in
404 this ~15 kbp was insufficient (only one label resides within ~26 kbp segment and zero labels in
405 the ~15 kbp segment, which is a partial copy of ~26 kbp segment), therefore sequence-level
406 resolution is required to determine whether paralogous copies of ~5 kbp or ~15 kbp cause
407 NAHR in deletions classified as Class I. The Class II deletions (~1.63 Mbp, 195.99-197.64 Mbp;
408 GRCh38) were observed in seven probands and had the proximal breakpoint localized to the
409 second ~5 kbp segment in SDB and the distal breakpoint localized to the same ~5 kbp segment
410 in SDC (Figure 5B). The Class III deletion (~1.23 Mbp, 196.01-197.3 Mbp; GRCh38) was
411 observed in only one proband, and the proximal and distal breakpoints did not overlap with any
412 SDs (Figure 5C). Sequence analysis of the ~5 kbp segments in SDB and SDC in the GRCh38
413 reference genome showed a high sequence identity to each other (≥ 98%) suggesting that
414 these paralogous copies can facilitate NAHR-mediated 3q29 deletions and duplications (Figure

5D-G, Table S7). On the other hand, the proximal and distal deletion breakpoints in Class III were not localized to SDs, sequence similarity was not detected between breakpoints which suggests that a mechanism other than NAHR has caused this deletion.

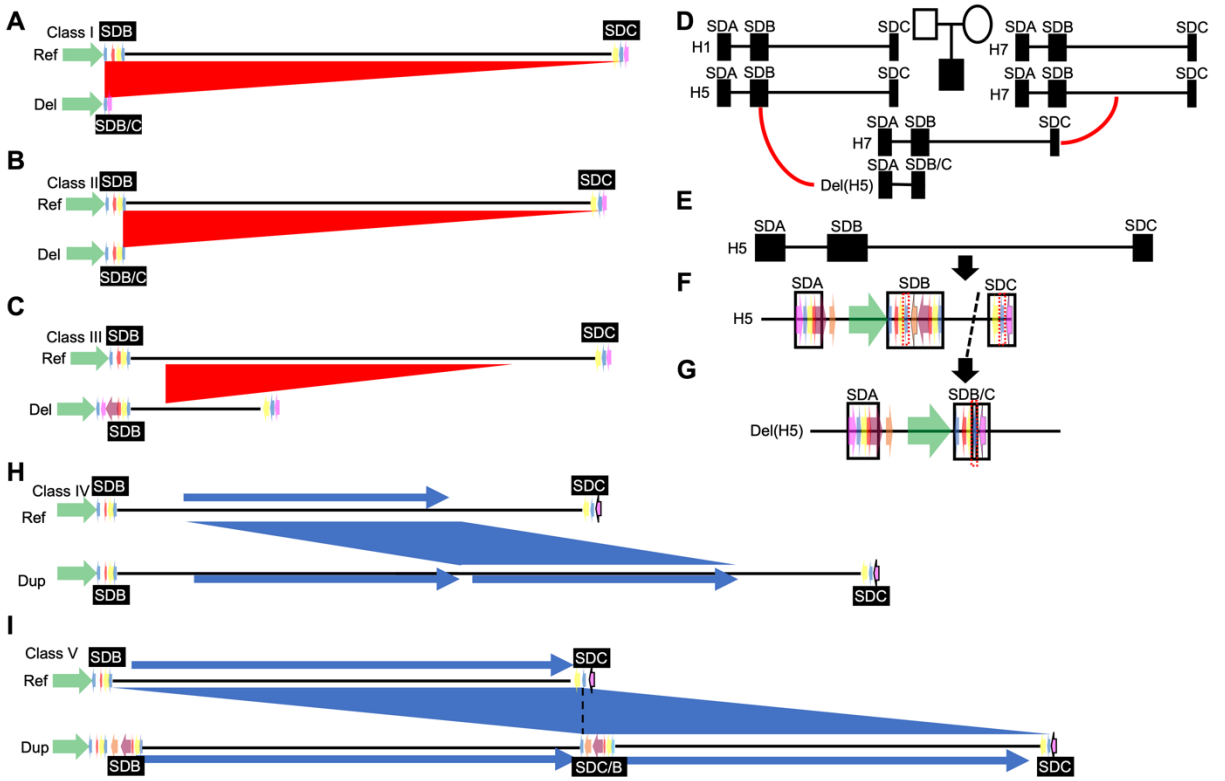

Figure 5: Deletion and duplication breakpoint classes identified in our study and depiction of NAHR-mediated deletion in Family 7 trio from the 3q29 Project. *(A-C) Schematic representation of 3q29 deletion breakpoint classes. The red triangles represent the deletion in each class.* ***(****D) Pedigree of Family 7, black rectangles with vertical dark lines represent genomes of the father, mother, and proband, red curved lines show the parent of origin and the transmitting parent, and black rectangles represent the 3q29 SDs.* ***(****E) Father's parent of origin chromosome, haplotype H5. (F) The 3q29 SDs, SDA-B-C, and the proposed NAHR mechanism.* ***(****G) Deleted chromosome occurring as a result of NAHR. Red dashed boxes indicate the proposed segments to mediate NAHR and result in 3q29 deletion in the proband.* ***(****H-I) Duplication breakpoint classes from Families 12 and 19, respectively. Black dotted line represents the duplication breakpoint. Blue rectangle: duplicated 3q29 region with the blue shaded area representing the duplication.*

Additionally, we analyzed two probands with duplications, ranging in size from ~942 kbp (~196.25-197.2 Mbp; GRCh38) to ~1.6 Mbp (~196-197.6 Mbp; GRCh38). We categorized these duplications as Class IV and Class V, respectively (Figure 5H-I). In Class IV, the proximal and distal breakpoints localized to a unique region between SDB and SDC (Figure 5H) suggesting that a mechanism other than NAHR has caused the duplication. The Class V duplication appears to be a reciprocal duplication of the recurrent ~1.6 Mbp 3q29 Class II deletion (Figure

436 5I). The proximal breakpoint in Class V localized to the second ~5 kbp segment (blue) in SDB
437 and the distal breakpoint localized to the paralogous ~5 kbp segment (blue) in SDC (Figure 5I).
438 Therefore, this duplication was likely caused by NAHR.
439 Finally, we investigated 3q29 haplotypes and phenotypes to evaluate whether there was
440 a correlation between haplotypes and phenotypes of 3q29 deletion and duplication probands,
441 however, no significant associations were identified (Table S9).
442

443 **Discussion**

444 The 3q29 deletion and duplication syndromes are rare genomic disorders. In the
445 recurrent 3q29 deletion and duplication cases, breakpoints occur within two SD blocks which
446 themselves are comprised of different SDs with > 98% sequence identity prone to NAHR [11,41,42].
447 Since these SD blocks have a complex structure containing paralogous copies of SDs with high
448 sequence identity, it is often difficult to accurately identify where the breakpoints occurred.
449 Techniques such as short-read sequencing have limited ability to resolve the breakpoint
450 locations within repetitive genomic segments such as SDs, due to the high sequence identity
451 and lengths of the paralogous copies [43]. With the availability of complementary genomic
452 technologies, including long-read sequencing and optical mapping, we are now able to
453 interrogate these regions and to identify breakpoint locations more precisely [34,35].

454 In this study, we first *de novo* assembled the 3q29 region in 161 unaffected individuals
455 from the 1000GP and CIAPM cohorts to identify distinct haplotypes. Ebert and colleagues [33]
456 previously reported 18 haplotypes in 30 unaffected individuals from the 1000GP. We have now
457 identified an additional 18 novel haplotypes in unaffected individuals using optical mapping
458 technology bringing the total number of known haplotypes for this region to 36. Six haplotypes
459 are either population-specific or have low frequency in other populations. For instance, the H8
460 haplotype was enriched in 11% of the AFR population, the H5 haplotype was enriched in 41% of
461 the EAS population, the H6 haplotype was enriched in 19% of the EUR population, and the H1
462 haplotype was enriched in 26% of the SAS population. In summary, we have conducted the
463 largest and most detailed investigation of 3q29 haplotypes in 161 unaffected individuals from 26
464 worldwide populations to date and report significant differences in haplotype frequencies
465 between the studied populations. Furthermore, the 3q29 haplotypes contain inversions and
466 copy number changes of 3q29 segments which overlap with protein-coding genes,
467 pseudogenes, and non-coding RNAs, such as *MUC20* and *MUC4*, lncRNA MUC20-OT1,
468 *SDHAP1*, *SDHAP2*, *SMBD1P*, and miRNAs. A recent study showed that lncRNA *SDHAP1*

469 upregulated the expression of *EIF4G2* by reducing miR-4465 levels in ovarian cancer cells [44]
470 which suggests that the pseudogenes may regulate gene expression through microRNAs [45].
471 Therefore, additional copies of the pseudogenes may impact patients' phenotypes by regulating
472 the protein-coding genes in the 3q29 region. However, further studies need to be conducted to
473 investigate the function of the pseudogenes in the etiology of 3q29 deletion and duplication
474 syndromes.

475 We analyzed the 3q29 haplotype structures in 16 probands with 3q29 deletion syndrome
476 to determine the breakpoint locations more precisely in probands and available parents, and to
477 identify the molecular mechanisms responsible for the deletions. In 15 of 16 probands, the
478 deletion breakpoints overlapped paralogous copies of ~5 kbp SD segments (blue**)** within SDB
479 and SDC, located in the same orientation. High sequence identity (> 98%) between the ~5 kbp
480 segments (blue) suggests that an NAHR mechanism caused the deletions in these probands,
481 consistent with previous findings [11]. However, in Class I, the possibility of breakpoint localization
482 to the ~15 kbp segments (magenta, partial copies of SDA-26 kbp segment) in SDB and SDC
483 cannot be excluded since these segments are adjacent (in the distal region) to the ~5 kbp
484 segment (blue). The sequence-level resolution is required to overcome the optical mapping
485 resolution insufficiency at the ~15 kbp segment and to determine whether paralogous copies of
486 ~5 kbp or ~15 kbp cause NAHR in deletions classified as Class I. The deletion breakpoints in
487 Class II were localized to the paralogous copies of ~5 kbp segments and the deletion
488 breakpoints for one proband, Class III, did not occur within any SD blocks in the region
489 suggesting that non-homologous end joining (NHEJ) or replication-based mechanism might be
490 responsible [46]. We also analyzed two individuals with the 3q29 duplication syndrome. One
491 individual carried the common recurrent ~1.6 Mbp duplication that appears to be the reciprocal
492 duplication of the recurrent ~1.6 Mbp 3q29 deletion, whereas the other ~1.2 Mbp 3q29 deletion
493 was not recurrent. Based on the deletion and duplication breakpoints, we have defined three
494 deletion and two duplication breakpoint classes for the 3q29 Project patient cohort.

495 Inversions have previously been hypothesized to be a risk factor for some *de novo*
496 deletions and duplications [47]. It is thought that inversions can interfere with synapsis during
497 meiosis, potentially causing DNA loops that are susceptible to misalignment and/or breakage [2].
498 For example, at the Williams-Beuren syndrome locus, there is an enrichment of inversions in the
499 parents where the *de novo* deletion arises [48,49]. However, in 2003, a group studying the 22q11.2
500 deletion syndrome showed that none of the parents of 18 probands with the 22q11.2 deletion
501 syndrome carried an inversion in this region [50]. Two inversions, ~289 kbp, and ~2 Mbp, have
502 previously been reported at the 3q29 locus [38–40]. The breakpoints in these inversions were
503 localized to inversely oriented paralogous copies of 3q29 segments between SDA and SDB,
504 and SDA and SDC, which could mediate NAHR. Among the 22 parents analyzed in the current
505 study, six were found to carry the ~289 kbp inversion, and three were inherited (H2: n=2, H13:
506 n=1) by the proband in an intact state. None of the parents carried the larger ~2 Mbp inversion
507 occurring between SDA and SDC. While the sample size is relatively small, our data provide no
508 evidence for an inversion predisposing to deletion or duplication at the 3q29 locus.
509 One limitation of our study is the lack of resolution of deletion and duplication
510 breakpoints at the sequence level. Nevertheless, with optical mapping, we refined the deletion
511 and duplication breakpoints to ~5 kbp segments. Other technologies, such as long reads from
512 Oxford Nanopore Technology (ONT) or Pacific Biosciences (PacBio) may be able to provide
513 nucleotide resolution breakpoints for these deletions and duplications. Unfortunately, the
514 available DNA samples from the 3q29 Project are not appropriate for ONT or PacBio High
515 Fidelity (HiFi) long-read sequencing since these require large amounts of high molecular weight
516 DNA.
517 In summary, we have identified a total of 19 novel haplotypes in the 3q29 region of the human
518 genome, some of which were shown to have significant enrichment in certain populations. We
519 have further shown that the majority of patients with the 3q29 deletion or duplication syndromes
520 have breakpoints within a ~5 kbp (blue) SD segment but find no evidence of inversions or other

521 SVs to act as predisposing factors. Haplotypes identified in both 3q29 patients and unaffected
522 individuals will be a valuable resource for future 3q29 deletion and duplication studies both to
523 help to identify the breakpoints more precisely and whether certain 3q29 haplotypes predispose
524 or protect from 3q29 deletion or duplication syndrome.

525

526 **Supplemental Data**

527 Figure S1: HPRC optical mapping and phased assembly 3q29 haplotype comparison – Part 1.

528 Figure S2: HPRC optical mapping and phased assembly 3q29 haplotype comparison – Part 2.

529 Figure S3: HPRC optical mapping and phased assembly 3q29 haplotype comparison – Part 3.

530 Figure S4: HPRC optical mapping and phased assembly 3q29 haplotype comparison – Part 4.

531 Figure S5: HPRC HG01928 bionano optical mapping and phased assembly 3q29 haplotype

532 comparison.

533 Figure S6: The smallest 3q29 haplotype, H28 (~287 kbp), and the largest 3q29 haplotype, H24

534 (~859 kbp), comparison.

535 Figure S7: Inversion Frequencies: Inversions identified in this study; INV-1 and INV-2, and 289

536 kbp inversion.

537 Figure S8: Large Inversions identified in this study; INV-1 and INV-2.

538 Figure S9: Novel haplotype H38 identified in one of the parents of origins (Father, Family 18)

539 from the 3q29 Project.

540 Table S1: 1000 Genomes Project and California Initiative to Advance Precision Medicine

541 Sample Cohort Optical Mapping Statistics

542 Table S2: The Emory 3q29 Project Sample Cohort Optical Mapping Statistics

543 Table S3: The 3q29 Segments

544 Table S4: Segment Sequence Identity

545 Table S5: 1000 Genomes Project and California Initiative to Advance Precision Medicine

546 Sample Cohort Haplotypes

547 Table S6: DGV Overlap

548 Table S7: The Emory 3q29 Project Sample Cohort Haplotypes

549 Table S8: Breakpoint classes identified in the 3q29 Project probands with the 3q29 deletion or

550 duplication syndrome

551 Table S9: The Emory 3q29 Project Proband Breakpoint Classes and Phenotype Comparison

552

553 **Declaration of Interests**

554 The authors declare no competing interests.

555

556 **Acknowledgments**

557 The authors are grateful to the participants and their families, without whose support this work
558 would not have been possible. In addition, we thank the Human Genome Structural Variation
559 Consortium and the Human Pangenome Reference Consortium for making their datasets
560 available. We thank Dr. Pille Hallast, Dr. Jee Young Kwon, and Dr. Kwondo Kim for providing
561 valuable feedback in editing this manuscript and Dr. Joshy George for assistance in statistical
562 analyses, and The Jackson Laboratory High-Performance Computing admins for providing an
563 environment to perform all computational analyses. J.G.M. conceived the study and analysis
564 using Bionano optical mapping. F.Y. and U.G. performed the analysis and interpretation. J.G.M.
565 obtained informed consent from parents and patients during hospital follow-up and facilitated
566 collection of blood samples. J.G.M. coordinated study and subject enrollment. J.G.M. and T.M.
567 carried out sample processing and experimentation. F.Y., and U.G drafted, and F.Y., C.L. and
568 J.G.M. critically revised the article. Final approval of the version to be published was given by
569 F.Y., U.G., T.M, Y.M., T.H.S, M.E.Z., P-Y.K., C.L. and J.G.M.

570

571 **Funding**

572 C.L. was supported by the National Institute of Health (NIH) U24HG007497 and The First
573 Affiliated Hospital of Xi'an Jiaotong University. T.H.S. and P-Y.K. were supported by grant #
574 GM120772, from the National Institute of General Medical Sciences of the National Institutes of
575 Health (NIH). Human Pangenome Reference Consortium support comes from NHGRI: A
576 Human Genome Reference Center (HGRC; RFA-HG-19-004), High-Quality Human Reference
577 Genomes (HGRQ; RFA-HG-19-002), Genome Reference Representations (GRR; RFA-HG-19-

003), and Technology development for complete sequencing of genomes (NOT-HG-19-011). JGM was supported by R01 MH110701.

Web resources

SV Genotyping: https://github.com/yuliamostovoy/OMGenSV

Bionano Solve 3.5.1: https://bionanogenomics.com/support/software-downloads/

OMTools: https://github.com/TF-Chan-Lab/OMTools

The Human Genome Structural Variation Consortium dataset: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20200117_Bionano_optical_maps/

University of California San Francisco dataset: Bianano bnx files of 114 samples (including 52 1000GP samples and 65 CIAPM samples) were available under the following accession number PRJNA588278 in the NCBI BioProject database.

The Human Pangenome Reference Consortium dataset: https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0

Data and code availability

The 3q29 Project dataset: The data supporting the findings of this study are available upon request from the authors.

HGSVC samples molecule support: 10.6084/m9.figshare.16899313

UCSF samples molecule support: 10.6084/m9.figshare.16886500

600 **References**

601 1. Finucane, B.M., Ledbetter, D.H., and Vorstman, J.A. (2021). Diagnostic genetic testing for
602 neurodevelopmental psychiatric disorders: closing the gap between recommendation and
603 clinical implementation. Curr. Opin. Genet. Dev. *68*, 1–8.

604 2. Lee, J.A., and Lupski, J.R. (2006). Genomic rearrangements and gene copy-number
605 alterations as a cause of nervous system disorders. Neuron *52*, 103–121.

606 3. Harel, T., and Lupski, J.R. (2018). Genomic disorders 20 years on-mechanisms for clinical
607 manifestations. Clin. Genet. *93*, 439–449.

608 4. Osborne, L.R., Martindale, D., Scherer, S.W., Shi, X.-M., Huizenga, J., Heng, H.H.Q., Costa,
609 T., Pober, B., Lew, L., Brinkman, J., et al. (1996). Identification of Genes from a 500-kb Region
610 at 7q11.23 That Is Commonly Deleted in Williams Syndrome Patients. Genomics *36*, 328–336.

611 5. Pober, B.R. (2010). Williams–Beuren Syndrome. N. Engl. J. Med. *362*, 239–252.

612 6. McDonald-McGinn, D.M., Sullivan, K.E., Marino, B., Philip, N., Swillen, A., Vorstman, J.A.S.,
613 Zackai, E.H., Emanuel, B.S., Vermeesch, J.R., Morrow, B.E., et al. (2015). 22q11.2 deletion
614 syndrome. Nat Rev Dis Primers *1*, 15071.

615 7. Willatt, L., Cox, J., Barber, J., Cabanas, E.D., Collins, A., Donnai, D., FitzPatrick, D.R.,
616 Maher, E., Martin, H., Parnau, J., et al. (2005). 3q29 microdeletion syndrome: clinical and
617 molecular characterization of a new syndrome. Am. J. Hum. Genet. *77*, 154–160.

618 8. Ballif, B.C., Hornor, S.A., Jenkins, E., Madan-Khetarpal, S., Surti, U., Jackson, K.E.,
619 Asamoah, A., Brock, P.L., Gowans, G.C., Conway, R.L., et al. (2007). Discovery of a previously
620 unrecognized microdeletion syndrome of 16p11.2–p12.2. Nat. Genet. *39*, 1071–1073.

621 9. Shinawi, M., Liu, P., Kang, S.-H.L., Shen, J., Belmont, J.W., Scott, D.A., Probst, F.J.,
622 Craigen, W.J., Graham, B.H., Pursley, A., et al. (2010). Recurrent reciprocal 16p11.2
623 rearrangements associated with global developmental delay, behavioural problems,
624 dysmorphism, epilepsy, and abnormal head size. J. Med. Genet. *47*, 332–341.

625 10. Rossi, E., Piccini, F., Zollino, M., Neri, G., Caselli, D., Tenconi, R., Castellan, C., Carrozzo,
626 R., Danesino, C., Zuffardi, O., et al. (2001). Cryptic telomeric rearrangements in subjects with
627 mental retardation associated with dysmorphism and congenital malformations. J. Med. Genet.
628 *38*, 417–420.

629 11. Ballif, B.C., Theisen, A., Coppinger, J., Gowans, G.C., Hersh, J.H., Madan-Khetarpal, S.,
630 Schmidt, K.R., Tervo, R., Escobar, L.F., Friedrich, C.A., et al. (2008). Expanding the clinical
631 phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal
632 microduplication. Mol. Cytogenet. *1*, 8.

633 12. Glassford, M.R., Rosenfeld, J.A., Freedman, A.A., Zwick, M.E., Mulle, J.G., and Unique
634 Rare Chromosome Disorder Support Group (2016). Novel features of 3q29 deletion syndrome:
635 Results from the 3q29 registry. American Journal of Medical Genetics Part A *170*, 999–1006.

636 13. Russo, R.S., Gambello, M.J., Murphy, M.M., Aberizk, K., Black, E., Lindsey Burrell, T.,
637 Carlock, G., Cubells, J.F., Epstein, M.T., Espana, R., et al. (2021). Deep phenotyping in 3q29
638 deletion syndrome: recommendations for clinical care. Genet. Med. 1–9.

639 14. Pollak, R.M., Zinsmeister, M.C., Murphy, M.M., Zwick, M.E., Emory 3q29 Project, and Mulle,
640 J.G. (2020). New phenotypes associated with 3q29 duplication syndrome: Results from the
641 3q29 registry. Am. J. Med. Genet. A *182*, 1152–1166.

642 15. Blanquer, F.A., Aleixandre Blanquer, F., Manchón Trives, I., Forniés Arnau, M.J., Alcaraz
643 Mas, L.A., Picó Alfonso, N., and Galán Sánchez, F. (2011). Síndrome de microduplicación
644 3q29. Anales de Pediatría *75*, 409–412.

645 16. Fernández-Jaén, A., Castellanos, M. del C., Fernández-Perrone, A.L., Fernández-
646 Mayoralas, D.M., de la Vega, A.G., Calleja-Pérez, B., Fernández, E.C., Albert, J., and Hombre,
647 M.C.S. (2014). Cerebral palsy, epilepsy, and severe intellectual disability in a patient with 3q29
648 microduplication syndrome. Am. J. Med. Genet. A *164A*, 2043–2047.

649 17. Goobie, S., Knijnenburg, J., Fitzpatrick, D., Sharkey, F.H., Lionel, A.C., Marshall, C.R.,
650 Azam, T., Shago, M., Chong, K., Mendoza-Londono, R., et al. (2008). Molecular and clinical
651 characterization of de novo and familial cases with microduplication 3q29: guidelines for copy
652 number variation case reporting. Cytogenet. Genome Res. *123*, 65–78.

653 18. Kessi, M., Peng, J., Yang, L., Duan, H., Tang, Y., and Yin, F. (2018). A Case With 4 de
654 Novo Copy Number Variations With Clinical Features That Overlap 1q43q44 Microdeletion and
655 3q29 Microduplication Syndromes. Child Neurol Open *5*, 2329048X18798200.

656 19. Lisi, E.C., Hamosh, A., Doheny, K.F., Squibb, E., Jackson, B., Galczynski, R., Thomas,
657 G.H., and Batista, D.A.S. (2008). 3q29 interstitial microduplication: A new syndrome in a three-
658 generation family. American Journal of Medical Genetics Part A *146A*, 601–609.

659 20. Lesca, G., Rudolf, G., Labalme, A., Hirsch, E., Arzimanoglou, A., Genton, P., Motte, J., de
660 Saint Martin, A., Valenti, M.-P., Boulay, C., et al. (2012). Epileptic encephalopathies of the
661 Landau-Kleffner and continuous spike and waves during slow-wave sleep types: genomic
662 dissection makes the link with autism. Epilepsia *53*, 1526–1538.

663 21. Schilter, K.F., Reis, L.M., Schneider, A., Bardakjian, T.M., Abdul-Rahman, O., Kozel, B.A.,
664 Zimmerman, H.H., Broeckel, U., and Semina, E.V. (2013). Whole-genome copy number
665 variation analysis in anophthalmia and microphthalmia. Clinical Genetics *84*, 473–481.

666 22. Tassano, E., Uccella, S., Giacomini, T., Severino, M., Siri, L., Gherzi, M., Celle, M.E., Porta,
667 S., Gimelli, G., and Ronchetto, P. (2018). 3q29 microduplication syndrome: Description of two
668 new cases and delineation of the minimal critical region. Eur. J. Med. Genet. *61*, 428–433.

669 23. Vitale, A., Labruna, G., Mancini, A., Alfieri, A., Iaffaldano, L., Nardelli, C., Pasanisi, F.,
670 Pastore, L., Buono, P., and Lombardo, B. (2018). 3q29 microduplication in a small family with
671 complex metabolic phenotype from Southern Italy. Clin. Chem. Lab. Med. *56*, e167–e170.

672 24. Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K.,
673 Arnarsdottir, S., Bjornsdottir, G., Walters, G.B., Jonsdottir, G.A., Doyle, O.M., et al. (2014).
674 CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature *505*, 361–
675 366.

676 25. Cox, D.M., and Butler, M.G. (2015). A clinical case report and literature review of the 3q29
677 microdeletion syndrome. Clin. Dysmorphol. *24*, 89–94.

678 26. Uddin, M., Sturge, M., Peddle, L., O'Rielly, D.D., and Rahman, P. (2011). Genome-wide

679 signatures of “rearrangement hotspots” within segmental duplications in humans. PLoS One *6*,
680 e28853.

681 27. Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H.,
682 Nagarajan, N., Xiao, M., et al. (2012). Genome mapping on nanochannel arrays for structural
683 variation analysis and sequence assembly. Nat. Biotechnol. *30*, 771–776.

684 28. Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A.K.Y., McCaffrey, J., Young, E.,
685 Lam, E.T., Hastie, A.R., Wong, K.H.Y., et al. (2019). Genome maps across 26 human
686 populations reveal population-specific patterns of structural variation. Nat. Commun. *10*, 1025.

687 29. Wong, K.H.Y., Ma, W., Wei, C.-Y., Yeh, E.-C., Lin, W.-J., Wang, E.H.F., Su, J.-P., Hsieh, F.-
688 J., Kao, H.-J., Chen, H.-H., et al. (2020). Towards a reference genome that captures global
689 genetic diversity. Nat. Commun. *11*, 5482.

690 30. Murphy, M.M., Lindsey Burrell, T., Cubells, J.F., España, R.A., Gambello, M.J., Goines,
691 K.C.B., Klaiman, C., Li, L., Novacek, D.M., Papetti, A., et al. (2018). Study protocol for The
692 Emory 3q29 Project: evaluation of neurodevelopmental, psychiatric, and medical symptoms in
693 3q29 deletion syndrome. BMC Psychiatry *18*, 183.

694 31. Mosley, T.J., Johnston, H.R., Cutler, D.J., Zwick, M.E., and Mulle, J.G. (2021). Sex-specific
695 recombination patterns predict parent of origin for recurrent genomic disorders. BMC Med.
696 Genomics *14*, 154.

697 32. Cao, H., Hastie, A.R., Cao, D., Lam, E.T., Sun, Y., Huang, H., Liu, X., Lin, L., Andrews, W.,
698 Chan, S., et al. (2014). Rapid detection of structural variation in a human genome using
699 nanochannel-based genome mapping technology. Gigascience *3*, 34.

700 33. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J.,
701 Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse
702 human genomes and integrated analysis of structural variation. Science *372*,.

703 34. Demaerel, W., Mostovoy, Y., Yilmaz, F., Vervoort, L., Pastor, S., Hestand, M.S., Swillen, A.,
704 Vergaelen, E., Geiger, E.A., Coughlin, C.R., et al. (2019). The 22q11 low copy repeats are
705 characterized by unprecedented size and structural variability. Genome Res. *29*, 1389–1401.

706 35. Mostovoy, Y., Yilmaz, F., Chow, S.K., Chu, C., Lin, C., Geiger, E.A., Meeks, N.J.L.,
707 Chatfield, K.C., Coughlin, C.R., Surti, U., et al. (2021). Genomic regions associated with
708 microdeletion/microduplication syndromes exhibit extreme diversity of structural variation.
709 Genetics *217*,.

710 36. Leung, A.K.-Y., Jin, N., Yip, K.Y., and Chan, T.-F. (2017). OMTools: a software package for
711 visualizing and processing optical mapping data. Bioinformatics *33*, 2933–2935.

712 37. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R.,
713 Altemose, N., Uralsky, L., Gershman, A., et al. (2021). The complete sequence of a human
714 genome.

715 38. Alsmadi, O., John, S.E., Thareja, G., Hebbar, P., Antony, D., Behbehani, K., and Thanaraj,
716 T.A. (2014). Genome at Juncture of Early Human Migration: A Systematic Analysis of Two
717 Whole Genomes and Thirteen Exomes from Kuwaiti Population Subgroup of Inferred Saudi
718 Arabian Tribe Ancestry. PLoS ONE *9*, e99069.

719 39. Thareja, G., John, S.E., Hebbar, P., Behbehani, K., Thanaraj, T.A., and Alsmadi, O. (2015).
720 Sequence and analysis of a whole genome from Kuwaiti population subgroup of Persian
721 ancestry. BMC Genomics *16*, 92.

722 40. Antonacci, F., Kidd, J.M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z., and
723 Eichler, E.E. (2009). Characterization of six human disease-associated inversion
724 polymorphisms. Hum. Mol. Genet. *18*, 2555–2566.

725 41. Rudd, M.K., Keene, J., Bunke, B., Kaminsky, E.B., Adam, M.P., Mulle, J.G., Ledbetter, D.H.,
726 and Martin, C.L. (2009). Segmental duplications mediate novel, clinically relevant chromosome
727 rearrangements. Hum. Mol. Genet. *18*, 2957–2962.

728 42. Mulle, J.G., Dodd, A.F., McGrath, J.A., Wolyniec, P.S., Mitchell, A.A., Shetty, A.C., Sobreira,
729 N.L., Valle, D., Rudd, M.K., Satten, G., et al. (2010). Microdeletions of 3q29 confer high risk for
730 schizophrenia. Am. J. Hum. Genet. *87*, 229–236.

731 43. Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing:
732 computational challenges and solutions. Nat. Rev. Genet. *13*, 36–46.

733 44. Zhao, H., Wang, A., and Zhang, Z. (2020). LncRNA SDHAP1 confers paclitaxel resistance
734 of ovarian cancer by regulating EIF4G2 expression via miR-4465. J. Biochem. *168*, 171–181.

735 45. Hu, X., Yang, L., and Mo, Y.-Y. (2018). Role of Pseudogenes in Tumorigenesis. Cancers
736 *10*,.

737 46. Carvalho, C.M.B., and Lupski, J.R. (2016). Mechanisms underlying structural variant
738 formation in genomic disorders. Nat. Rev. Genet. *17*, 224–238.

739 47. Feuk, L. (2010). Inversion variants in the human genome: role in disease and genome
740 architecture. Genome Med. *2*, 11.

741 48. Bayés, M., Magano, L.F., Rivera, N., Flores, R., and A. Pérez Jurado, L. (2003). Mutational
742 Mechanisms of Williams-Beuren Syndrome Deletions. Am. J. Hum. Genet. *73*, 131–151.

743 49. Osborne, L.R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T.,
744 Cox, S., Tsui, L.-C., et al. (2001). A 1.5 million–base pair inversion polymorphism in families
745 with Williams-Beuren syndrome. Nat. Genet. *29*, 321–325.
746
747 50. Gebhardt, G.S., Devriendt, K., Thoelen, R., Swillen, A., Pijkels, E., Gewillig, M., Fryns, J.-P.,
748 and Vermeesch, J.R. (2003). No evidence for a parental inversion polymorphism predisposing
749 to rearrangements at 22q11.2 in the DiGeorge/Velocardiofacial syndrome. European Journal of
750 Human Genetics *11*, 109–111.

751

752

Supplemental Data

Figure S1: HPRC optical mapping and phased assembly 3q29 haplotype comparison – Part 1. *EUR: European ancestry, EAS: East Asian ancestry, AMR: American ancestry, BNG: Bionano Genomics optical mapping data, PA: Phased assembly in silico mapping data, colored arrows depicting the 3q29 segments.*

Figure S2: HPRC optical mapping and phased assembly 3q29 haplotype comparison – Part 2. *AMR: American ancestry, EAS: East Asian ancestry, AFR: African ancestry, BNG: Bionano Genomics optical mapping data, PA: Phased assembly in silico mapping data, colored arrows depicting the 3q29 segments.*

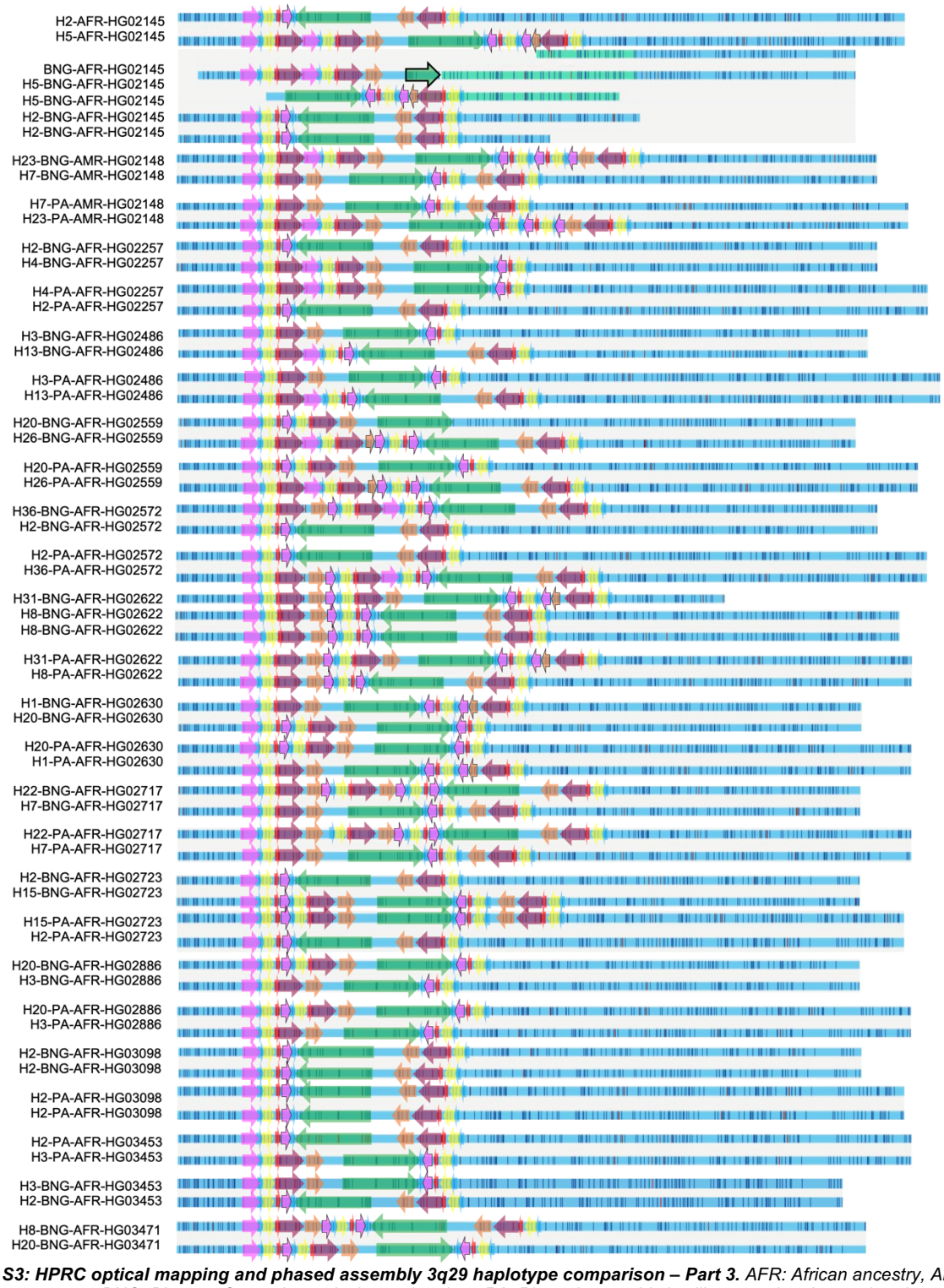

Figure S3: HPRC optical mapping and phased assembly 3q29 haplotype comparison – Part 3. *AFR: African ancestry, AMR: American ancestry, BNG: Bionano Genomics optical mapping data, PA: Phased assembly in silico mapping data, colored arrows depicting the 3q29 segments.*

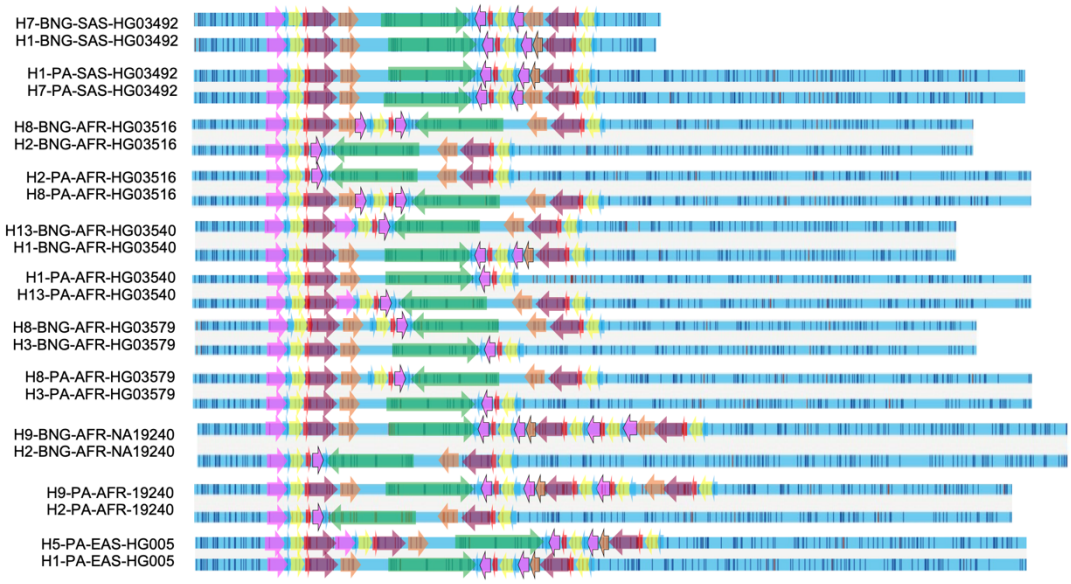

Figure S4: HPRC optical mapping and phased assembly 3q29 haplotype comparison – Part 3. *AFR: African ancestry, AMR: American ancestry, BNG: Bionano Genomics optical mapping data, PA: Phased assembly in silico mapping data, colored arrows depicting the 3q29 segments.*

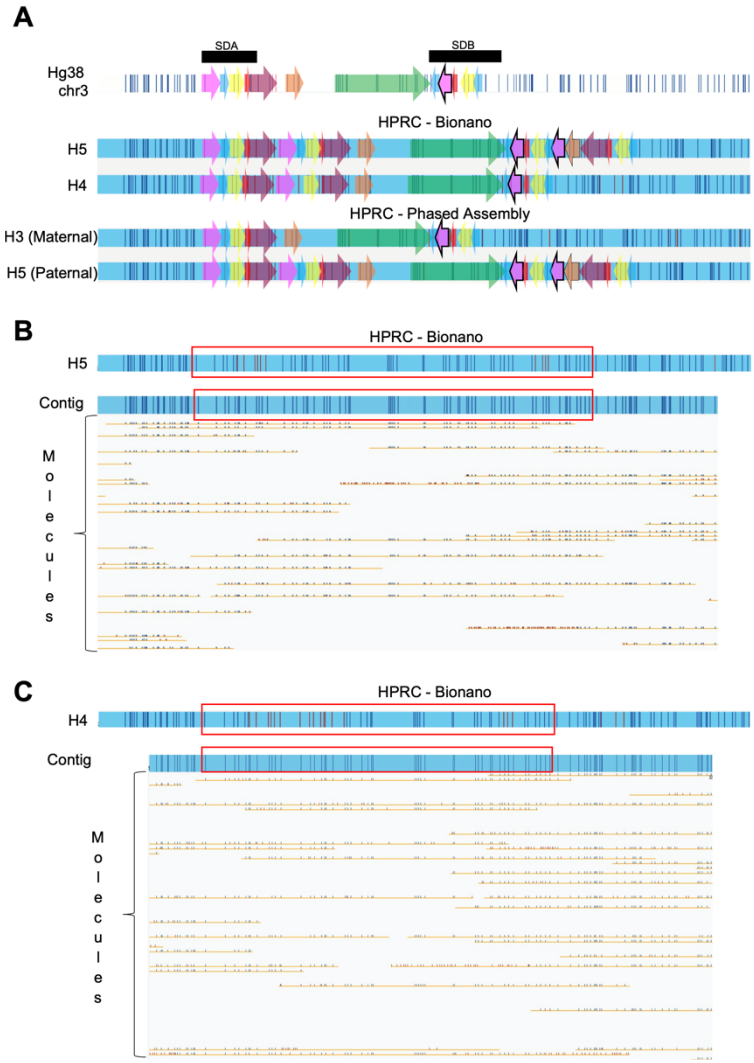

Figure S5: HPRC HG01928 bionano optical mapping and phased assembly 3q29 haplotype comparison*. A – Hg38/GRCh38 in silico map represented at the top, optical mapping haplotypes, H5 and H4 respectively, represented in the middle, and phased assembly maternal and paternal haplotypes are represented at the bottom. B - Optical mapping of the H5 haplotype contigs represented in blue and optical mapping molecules represented as yellow lines showing single molecule support for the H5 haplotype. C - Optical mapping of the H4 haplotype contigs represented in blue and optical mapping molecules represented as yellow lines showing single molecule support for the H4 haplotype. The red boxes in B and C shows the 3q29 region in each haplotype.*

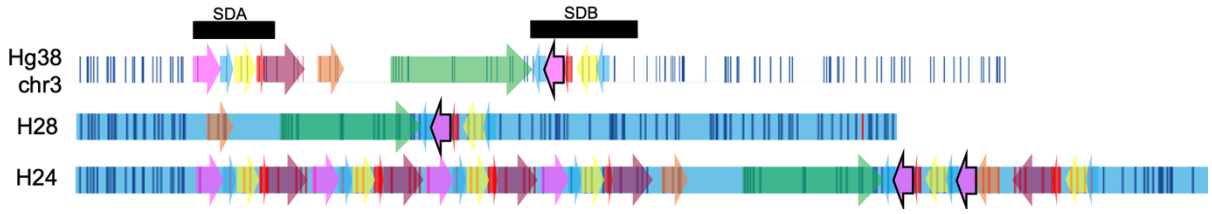

Figure S6: Comparison of the smallest, H28 (~287 kbp), and the largest, H24 (~859 kbp), 3q29 haplotypes. *Colored blocks represent 3q29 segments, white rectangles with vertical lines represent Hg38/GRCh38 in silico map, and blue rectangles with vertical lines represent H28 and H24 haplotypes.*

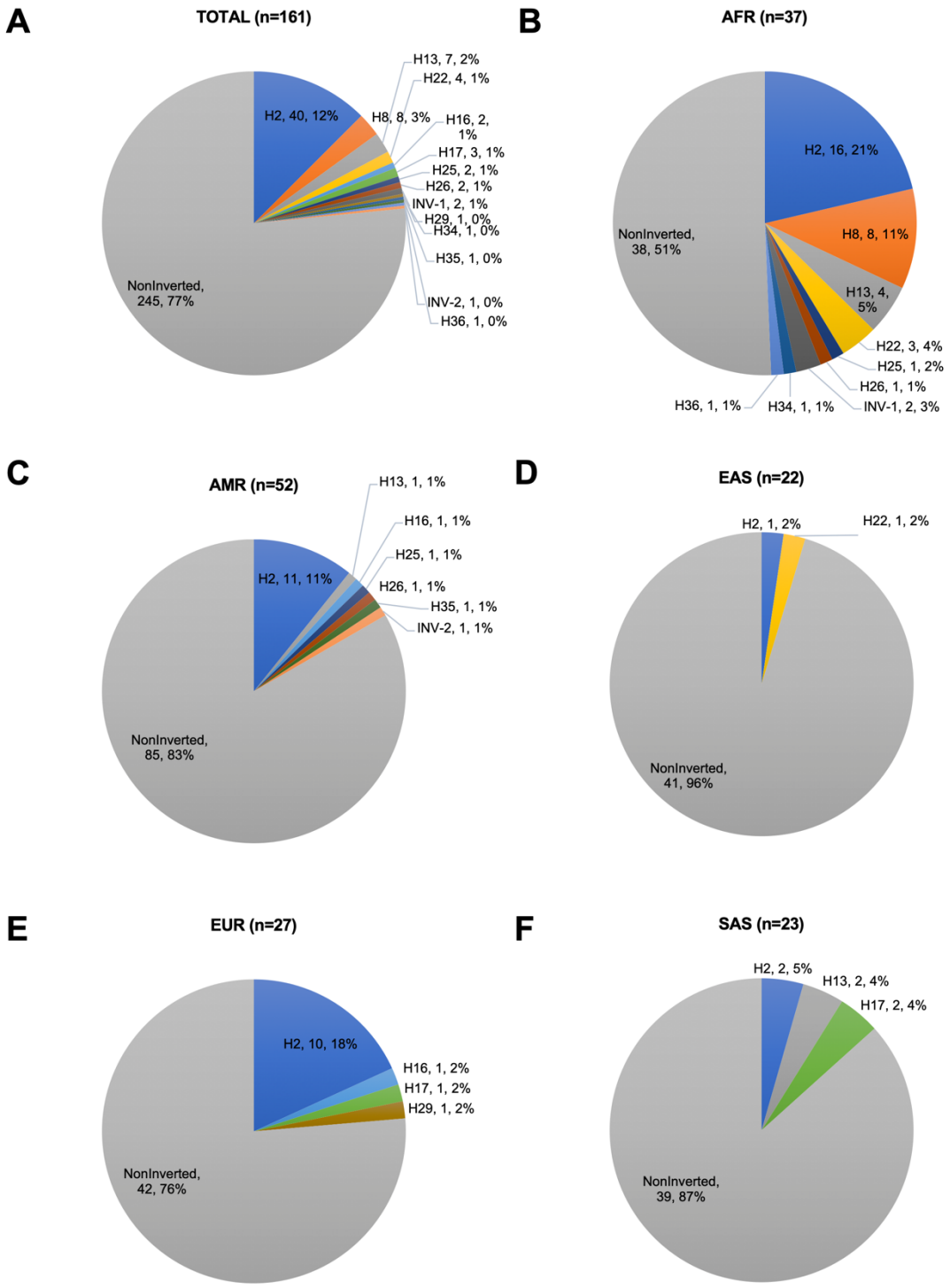

Figure S7: Inversion Frequencies: Inversions identified in this study; INV-1 and INV-2, and 289 kbp inversion. *A - The proportion of inversion haplotypes among all studied samples, B - The proportion of inversion haplotypes in AFR population, C - The proportion of inversion haplotypes in AMR population, D - The proportion of inversion haplotypes in EAS population, E - The proportion of inversion haplotypes in EUR population, F - The proportion of inversion haplotypes in SAS population. Proportions were represented as: Haplotype ID/Count/Percentage.*

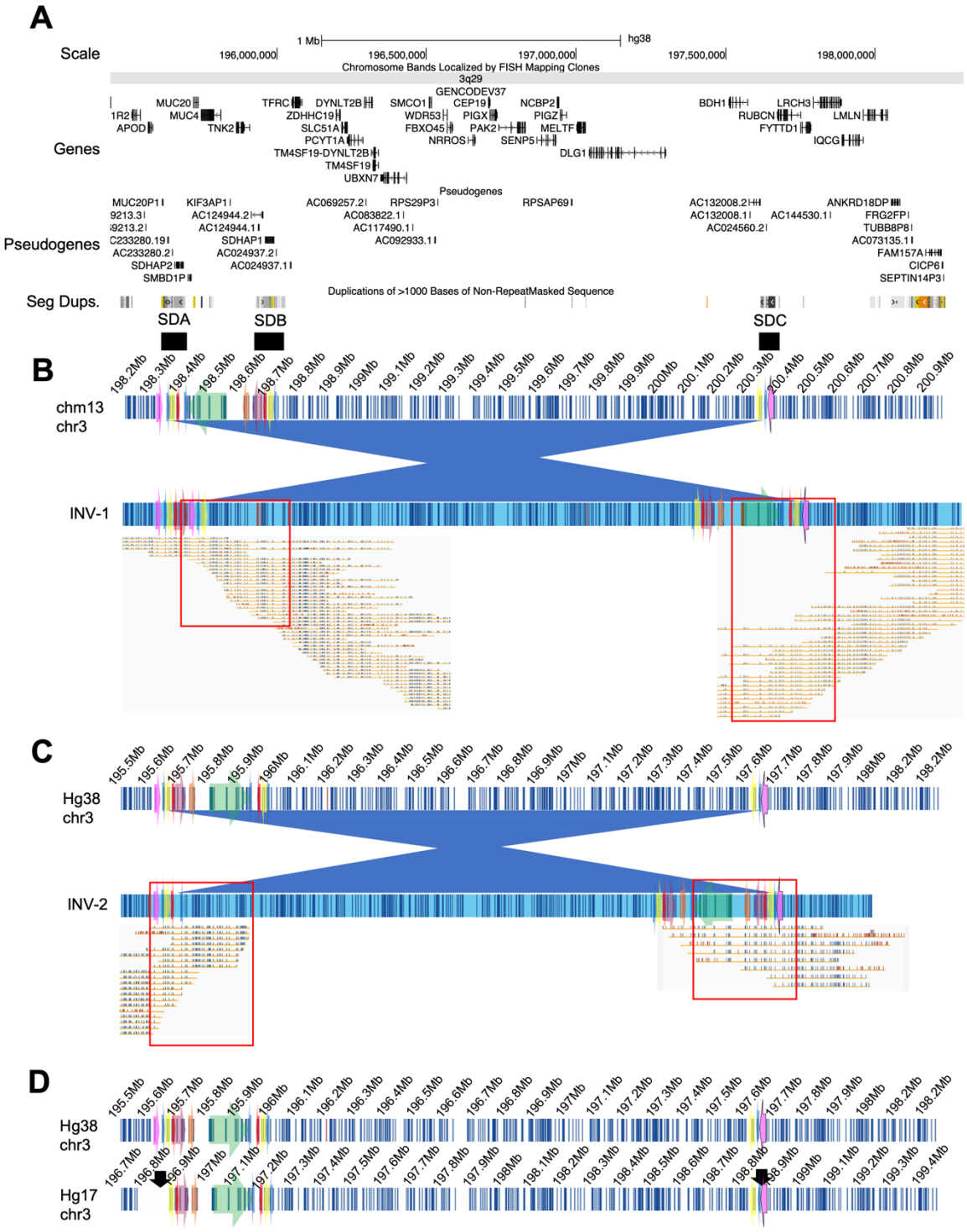

Figure S8: Large inversions identified in this study; INV-1 (~2.03 Mbp) and INV-2 (~2.13 Mbp). *A - UCSC Genome Browser Genes, Pseudogenes, and segmental duplication tracks are represented. B - Upper rectangle with white background and vertical lines represents chm13/T2T (Telomere-to-Telomere) in silico map of the 3q29 region. The rectangle at the bottom represents the structure of INV-1. Red rectangle: highlighting the molecules supporting the inversion breakpoints. C - Upper rectangle with white background and vertical lines represents GRCh38/hg38 in silico map of the 3q29 region. The rectangle at the bottom represents the structure of INV-2. Red rectangle: highlighting the molecules supporting the inversion breakpoints. D - GRCh38/hg38 and NCBI35/hg17 reference assemblies represented with white backgrounds and vertical blue lines. Black arrows in NCBI35/hg17 indicate inversion breakpoints identified previously by Antonacci and colleagues in 2009. Colored arrows in each panel represent 3q29 segments.*

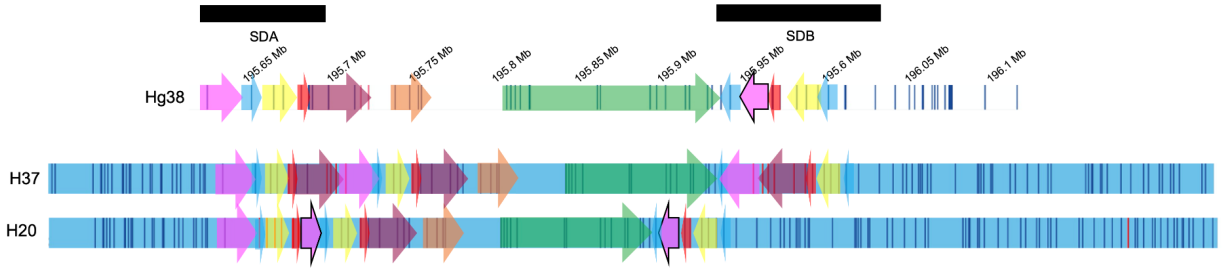

Figure S9: Novel haplotype H37 identified in one of the parents of origin (Father, Family 18) from the 3q29 Project. *Optical mapping contigs in blue rectangles with vertical dark lines. Colored arrows represent the 3q29 segments in each haplotype.*