Local and global chromatin interactions are altered by large genomic deletions associated with human brain development

Xianglong Zhang[1,2*], Ying Zhang[3*], Xiaowei Zhu[1,2*], Carolin Purmann[1,2], Michael S Haney[2], Thomas Ward[1,2], Jie Yao[4], Sherman M Weissman[5#], Alexander E Urban[1,2#]

1 Department of Psychiatry and Behavioral Sciences, 2 Department of Genetics, Stanford University School of Medicine, 3 Department of Genetics and Genomic Sciences, Icahn School of medicine at Mount Sinai, 4 Department of Cell Biology, Yale University School of Medicine, 5 Department of Genetics, Yale University

Email addresses:

Xianglong Zhang, xzhang8@stanford.edu;

Ying Zhang, zzhang006@gmail.com;

Xiaowei Zhu, xwzhu@stanford.edu;

Carolin Purmann, purmann@stanford.edu;

Michael S Haney, haney@stanford.edu;

Thomas Ward, wardtr@stanford.edu;

Jie Yao, yao.j@yale.edu;

Sherman M Weissman, sherman.weissman@yale.edu;

Alexander E Urban, aeurban@stanford.edu.

*These authors contributed equally to this work.

#To whom correspondence should be addressed.

Abstract

Background: Large copy number variants (CNVs) in the human genome are strongly associated with common neurodevelopmental, neuropsychiatric disorders such as schizophrenia and autism. Using Hi-C analysis of long-range chromosome interactions, including haplotype-specific Hi-C analysis, and ChIP-Seq analysis of regulatory histone marks, we studied the epigenomic effects of the prominent heterozygous large deletion CNV on chromosome 22q11.2 and also replicated a subset of the findings for the heterozygous large deletion CNV on chromosome 1q21.1.

Results: There are local and global gene expression changes as well as pronounced and multilayered effects on chromatin states, chromosome folding and topological domains of the chromatin, that emanate from the large CNV locus. Regulatory histone marks are altered in the deletion flanking regions, and in opposing directions for activating and repressing marks. Histone marks are changed along chromosome 22q and genome wide. Chromosome interaction patterns are weakened within the deletion boundaries and strengthened between the deletion flanking regions. The long-range folding contacts between the telomeric end of chromosome 22q and the distal deletion-flanking region are increased. On the chromosome 22q with deletion the topological domain spanning the CNV boundaries is deleted in its entirety while neighboring domains interact more intensely with each other. Finally, there is a widespread and complex effect on chromosome interactions genome-wide, i.e. involving all other autosomes, with some of the effect directly tied to the deletion region on 22q11.2.

Conclusions: These findings suggest novel principles of how such large genomic deletions can alter nuclear organization and affect genomic molecular activity.

Keywords: Germline CNVs, Chromatin States, Chromosome Interactions, Topological Domains

Introduction

Two of the most exciting discoveries in human genetics of the past decade are that small- to medium-sized Copy Number Variants (CNVs) are very common in the human genome, and that there is a group of large CNVs that are strongly associated with brain development and neuropsychiatric disorders such as schizophrenia and the autism spectrum disorders (ASDs) [1,2]. These large CNVs are understood to be providing enticing points of entry to the analysis of the strong but complex genetic and molecular (and possibly even cellular) basis of these common disorders.

Large CNVs, typically sized from hundreds of thousands to millions of basepairs of genomic DNA sequence, were previously known to be in strong association with often-severe but rare congenital malformations, or found in cancer genomes. It was a striking discovery when a series of studies [1,2] showed that there is a group of more than ten large CNVs that are strongly associated with aberrant brain development and a resulting neuropsychiatric phenotype such as schizophrenia or ASD. These large neuropsychiatric CNVs can encompass many genes and their effects across the various molecular levels of gene activity and regulation and the clinical phenotype are complex and only poorly understood.

On the molecular level these large neuropsychiatric CNVs have been mostly studied by focusing on the effects of individual genes or small groups of genes from within the CNV boundaries. Many very interesting insights have been gained using this approach.

67 However, these findings about individual genes fall short of explaining the full effects of
68 the large CNVs. There already have been a number of transcriptome wide studies that at
69 least hint at certain network effects emanating from the large CNVs [3-7]. Which mechanisms
70 mediate such transcription network effects is then the question. Furthermore there are an
71 increasing number of studies that show a potentially very important role of chromatin
72 regulation in the molecular etiology of neuropsychiatric disorder [8-14].

73 Against this backdrop we reasoned that it was worthwhile testing whether large CNVs
74 with association to brain development might cause a disruption or at least alteration of one
75 or several aspects of chromatin conformation, such as the distribution of regulatory
76 chromatin marks, the long-range direct physical interactions between distant regions on
77 one chromosome or between different chromosomes or the higher-order chromatin domain
78 structures that are defined by such marks or interactions. Such effects on these important
79 layers of molecular regulation of gene activity would then constitute a novel principle by
80 which large CNVs could transmit their presence to the machinery of cellular physiology.

81 Here we show in a cohort of lymphoblastoid cell lines (LCLs) derived from patients
82 with 22q11 Deletion Syndrome that indeed chromatin marks, chromatin domains and long-
83 range chromosome interactions are affected in several distinct ways by the large, common
84 and strongly disease-associated CNV on chromosome 22q11.2. We then go on to show in a
85 smaller number of LCLs from different patients that at least some of the same observations
86 can also be made for another neuropsychiatric large CNV, on chromosome 1q21.1.

87 22q11 Deletion Syndrome (22q11DS) is a disorder caused in the vast majority of cases
88 by a heterozygous deletion of about 3 million basepairs spanning about 60 known genes on
89 chromosome 22q11.2. It occurs in 1 per 3,000-6,000 live births [15]. The common phenotypes
90 of 22q11DS include a large spectrum of congenital anomalies, for example of the facial

91 structures and the cardiovascular system – and notably there is a strong association with
92 several neurodevelopmental psychiatric disorders, in particular schizophrenia and ASDs
93 [2,16-19]. In this study we are using the large CNV on 22q11.2 as a model to determine the
94 generalizable principles along which large CNVs of this important category can lead to
95 changes to the various ways in which chromatin is ordered, using unbiased, genome-wide,
96 sequencing-based assays for discovery.

97 **Results**

98 **Producing chromosomal contact maps with Hi-C, deep-coverage capture-Hi-C and**
99 **chromosomal haplotype phasing data**

100 To determine possible effects of the 22q11.2 deletion on chromosomal interactions, we
101 generated Hi-C contact maps for 11 human LCLs (5 patient cell lines with 22q11.2 deletion
102 and 6 control cell lines without), with a total of 3.1 billion Hi-C contact-reads of which 680
103 million read-pairs were of high quality and used for the downstream analyses
104 (**Supplementary Table 1**). The presence and boundaries of the heterozygous 3 Mbp
105 deletion in the patient cell lines were validated by whole genome sequencing
106 **(Supplementary Fig. 1)**. To rule out an increase in overall genome instability in the patient
107 LCLs we carried out whole-genome sequencing at a coverage comparable to that used for
108 the majority of samples by the 1000 Genomes Project (i.e. <10x genome wide coverage).
109 Almost all of our control cell lines had been part of the 1000 Genomes Project and we
110 compared the genome-wide CNV load between our patient LCLs and LCLs used by the 1000
111 Genomes Project. We found no elevated genome-wide CNV burden in our patient cell lines
112 **(Supplementary Table 2)**.

113 As a means of quality control our study included the cell line GM06990, which was the

114 cell line used in the original Hi-C paper [20]. Inter-chromosomal contacts of GM06990 as
115 determined by our own Hi-C data for this line (**Supplementary Fig. 2a**) show the same
116 patterns of chromosomal interactions across the nucleus as in [20]; i.e. small chromosomes
117 generally have more interactions with each other than larger chromosomes with each other,
118 and many more than chromosomes in the medium size range. Specific interaction pairs, e.g.
119 between chromosomes 17 and 19 were also replicated in our GM06990 data. Global inter-
120 chromosomal contact maps resulting from combining all our controls and cases,
121 respectively, again replicated these global interaction patterns (**Supplementary Fig. 2b,c**).

122 In order to be able to carry our haplotype-specific analysis of the chromatin and
123 transcriptome data we conducted haplotype phasing for the two related patients in our
124 cohort, a mother and son duo with the 22q11.2 deletion, where this was made possible by
125 combining the sequencing based approach of statistically-aided long-read haplotyping
126 (SLRH) [21] with an analysis of Mendelian inheritance of single nucleotide variants (SNVs). We
127 were able to phase 1,868,316 and 1,870,948 heterozygous SNVs on the autosomes of
128 ID00014 and ID00016, respectively (see details in Methods and Supplementary
129 Information). Two control samples from related donors, GM12878 and GM12892, were also
130 haplotype-phased using the same approach. For GM12892 we phased 1,929,967 autosomal
131 SNVs and for GM12878 we phased 1,874,181 autosomal SNVs. These phased SNVs were
132 then used to create haplotype-specific Hi-C interaction maps and to carry out allele-specific
133 analyses of the ChIP-Seq and RNA-Seq data.

134 Haplotype-specific Hi-C analysis requires very deep sequencing coverage of Hi-C
135 libraries since only paired-end reads that cover at least one informative (heterozygous) SNV
136 can be used for that purpose. To increase the Hi-C coverage for chromosome 22q to these
137 required very deep levels we carried out custom-designed chromosome-wide targeted

138 capture Hi-C. We performed capture Hi-C by using a custom-designed set of 2.1 million
139 NimbleGen oligomer capture probes representing the entire sequence of chromosome 22q
140 and Hi-C libraries prepared using the in situ Hi-C protocol from the four cell lines for which
141 we generated haplotype phasing data, followed by Illumina sequencing (**Supplementary**
142 **Table 1**). With this novel approach we achieved 10-16 fold levels of chromosome 22q
143 specific enrichment of Hi-C paired-end reads over the standard non-capture Hi-C data
144 (**Supplementary Fig. 3**).

145 **Normalization of Hi-C data**

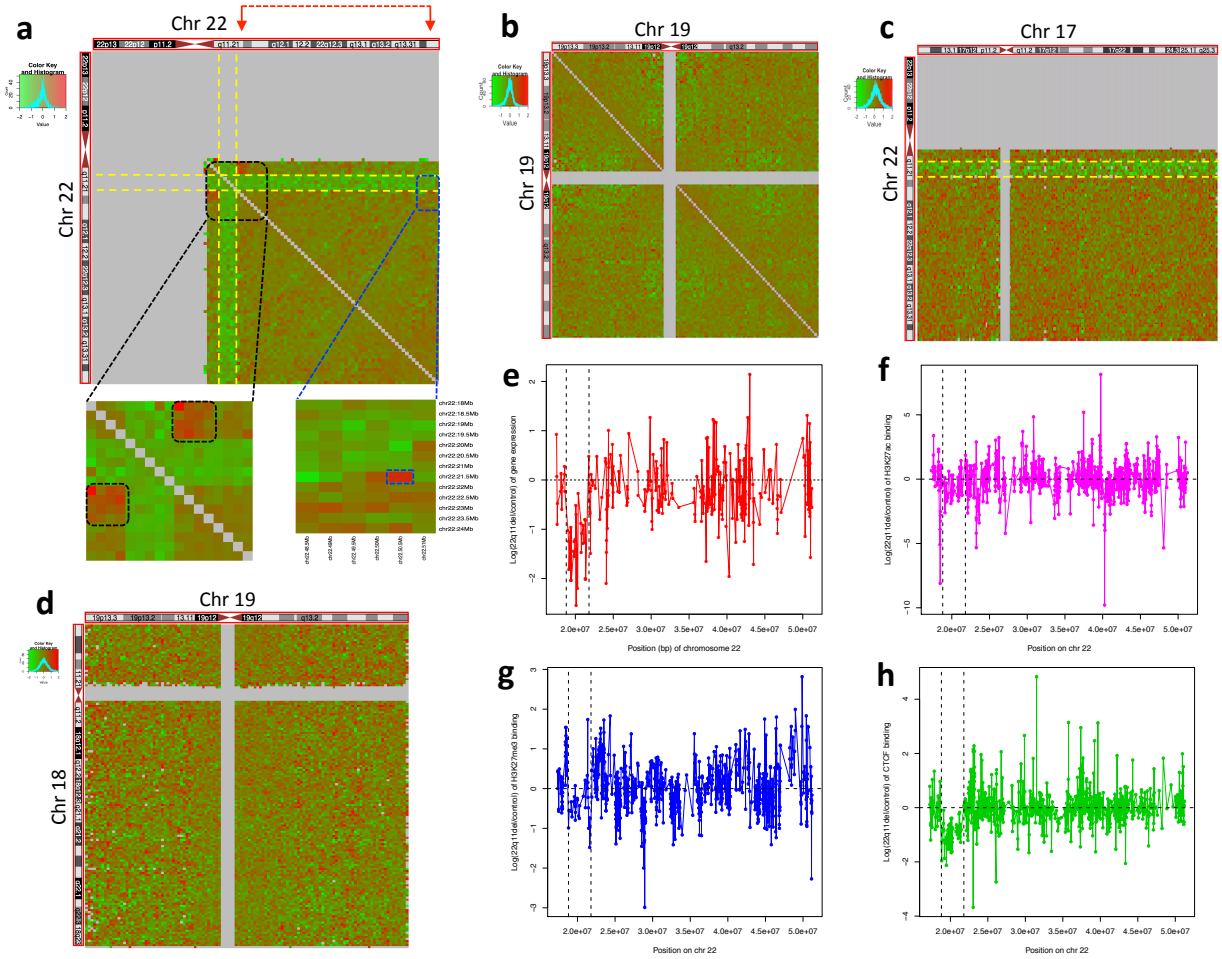
146 Many factors, such as mappability of sequencing reads, GC content, length of the restriction
147 enzyme fragment etc., can lead to biases in Hi-C data. Several alternative computational
148 approaches for Hi-C data analysis have been developed that allow for the removal of these
149 biases and data [22-31]. When using these alternative algorithmic approaches on Hi-C data that
150 was generated for genomes without large CNVs, such as GM12878, the resulting
151 normalization metrics will be highly correlated [32]. For our Hi-C data however, we needed to
152 be certain to use a normalization method that is not thrown off from the outset by the
153 presence of the heterozygous 3 Mbp deletion in 22q11.2.

154 We tested three different commonly used normalization algorithms that were
155 developed for Hi-C data, following the rationale that the normalization methods should not
156 change the general patterns of interaction we can see in the raw data.

157 We found that not all of the available normalization methods are robust for the use
158 with Hi-C data coming from genomes with large CNVs. However, the hicpipe algorithm [22] is
159 quite suitable for this purpose (details of the comparison between normalization methods
160 in the **Supplementary Information** and **Supplementary Fig. 4**).

161 **Chromosomal contacts decrease within the 22q11.2 deletion boundaries and**
162 **between the deletion region and the remainder of chromosome 22q**

163 We observed that in 22q11DS cell lines the chromosomal contacts within the 22q11.2
164 deletion region are decreased significantly compared to control cell lines (**Fig. 1a**). Also
165 strongly reduced are the chromosomal contacts between the 22q11.2 deletion region with
166 the entire remainder of chromosome 22 (**Fig. 1a; Supplementary Fig. 5a**). This decrease is
167 consistent with the copy number of the 22q11.2 deletion region in the patient cell lines, as
168 all of the 22q11DS cell lines are heterozygously deleted for this region. No such decrease of
169 chromosomal contacts that involved an extended and contiguous chromosomal region was
170 observed that did not involve the 22q11.2 region. Likewise there was no such strong
171 decrease of intra-chromosomal contacts over an extended and contiguous chromosomal
172 region on any of the other autosomes (for example chromosome 19 in **Fig. 1b;**
173 **Supplementary Fig. 5b**).



174

175 ***Figure 1. Effects of the 22q11.2 deletion on chromosome contacts, gene expression and***

176 ***epigenetic profiles.*** *Panels a-d: each pixel in the heat-maps represents the intra- or inter-*

177 *chromosomal contact frequency in Hi-C data from 22q11.2del cell lines versus control cell lines*

178 *for a 500 kbp region. Yellow dashed lines indicate the 3 Mbp deletion on chromosome 22q. (a)*

179 *Fold change of cis-contacts along chromosome 22 in 22q11.2del versus control cell lines. Black*

180 *boxes indicate increased contacts between the deletion-flanking regions in 22q11.2del cell*

181 *lines. Blue box: the signal for increased contacts between the centromere-distal deletion-*

182 *flanking region and the telomeric end of chromosome 22q (red arrows and dashed red line*

183 *indicate the corresponding chromosome folding event). (b) Lack of intra-chromosomal fold*

184 *change of contacts for chromosome 19. (c) Fold change of inter-chromosomal contacts*

185 *between chromosome 22 and chromosome 17. (d) Lack of intra-chromosomal fold change of*
186 *contacts between chromosome 18 and chromosome 19. (e) Log2 transformed fold change of*
187 *gene expression for genes on chromosome 22q in RNA-Seq data from 22q11.2del versus cell*
188 *lines. Each point represents a gene. Panels f-h: Log2 transformed fold change in ChIP-Seq*
189 *signals in 22q11.2del versus control cell lines. (f) H3K27ac histone modifications. (g)*
190 *H3K27me3 histone modifications. (h) CTCF binding sites. Black dashed lines indicate the 3*
191 *Mbp deletion on chromosome 22q in e-h.*

192 We then investigated whether this effect of a decrease in chromosomal interactions
193 between the 22q11.2 deletion and *cis*-contacts (i.e. within the deletion boundaries and
194 between the deletion region and elsewhere on chromosome 22q) also holds for *trans*-
195 contacts (i.e. for contacts between the region within the deletion boundaries and the rest of
196 the genome). We found the *trans*-contacts involving the 22q11.2 deletion in the patient cell
197 lines and any other chromosome also decreased compared to control cell lines (**Fig. 1c;**
198 **Supplementary Fig. 5c**). No other regions of chromosome 22q showed such an effect, and
199 neither did any other pair of autosomes that did not include 22q (**Fig. 1d; Supplementary**
200 **Fig. 5d**).

201 **The 22q11.2 deletion has an effect on gene expression and epigenetic profiles in the**
202 **22q11.2 region**

203 To determine the effects of the 22q11.2 deletion on gene expression patterns and chromatin
204 marks, we performed RNA-Seq on 14 cell lines (**Supplementary Table 1**), ChIP-seq for
205 H3K27ac and H3K27me3 histone modifications in 6 cell lines as well as ChIP-Seq for CTCF
206 binding in 5 cell lines, respectively (**Supplementary Table 3**). RNA-Seq analysis showed
207 that if a gene that is located within the boundaries of the large CNV in 22q11.2 is expressed,
208 the level of that expression is decreased in 22q11.2DS patient lines relative to control cell

209 lines (**Fig. 1e**), consistent with a previous study [3].

210 Differential pattern analysis of the ChIP-Seq data for H3K27ac, H3K27me3 and CTCF
211 showed that for the majority of binding sites within the 22q11.2 deletion boundaries the
212 binding signal decreased in 22q11.2DS patient lines compared to control cell lines (**Fig.**
213 **1f,g,h**). This phenomenon of decreased binding signals over a large chromosomal region is
214 specific to chromosome 22q11.

215 **Chromosome contacts increase across the 22q11DS breakpoint junction and**
216 **chromatin marks in its flanking regions are affected in a concerted manner**

217 The distal and proximal flanking regions of the 22q11.2 deletion are brought into close
218 proximity to each other by the formation of the deletion breakpoint junction. We
219 hypothesized that since Hi-C contacts between two given regions will increase with
220 decreasing genomic distance, the contacts between the distal and proximal deletion-
221 flanking regions in 22q11DS patient cells would be markedly enhanced. We indeed found
222 such stronger contacts between the deletion-flanking regions in 22q11.2DS cell lines (**Fig.**
223 **1a; Supplementary Fig. 4a**).

224 We used the haplotype phasing information and the deep capture Hi-C data for two
225 patient and two control cell lines to generate haplotype-specific contact maps for
226 chromosome 22q (**Fig. 2a**). The chromosome contacts between proximal and distal flanking
227 regions of the 22q11.2 deletion increased on the chromosome 22q with the deletion when
228 compared with the intact chromosome 22q within the same patient cell line (**Fig. 2**). The
229 chromosome contacts between proximal and distal deletion-flanking regions in 22q11.2 on
230 the intact chromosome in the patient cell lines were not affected compared to the
231 chromosomes in the controls (**Fig. 2**).

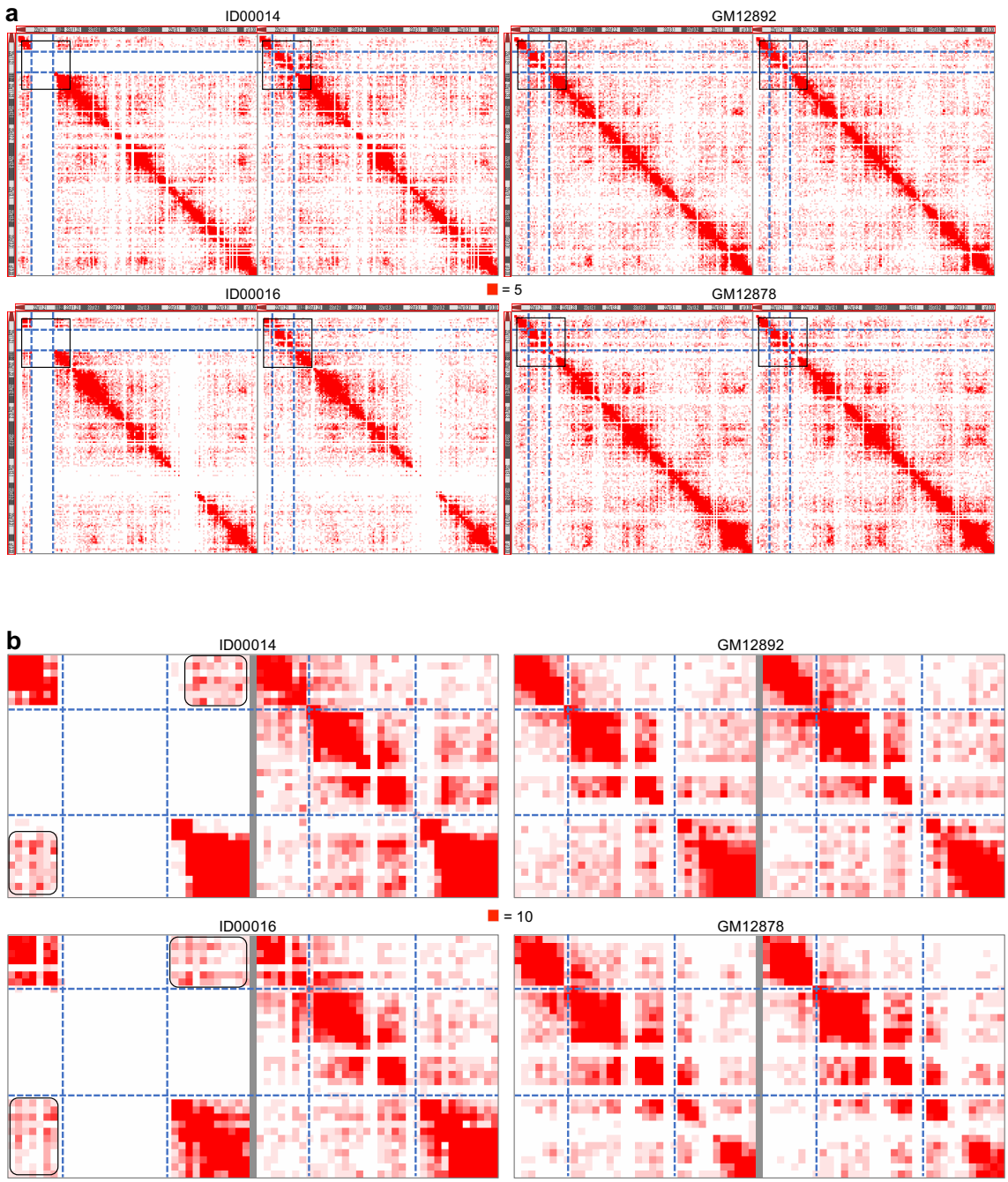
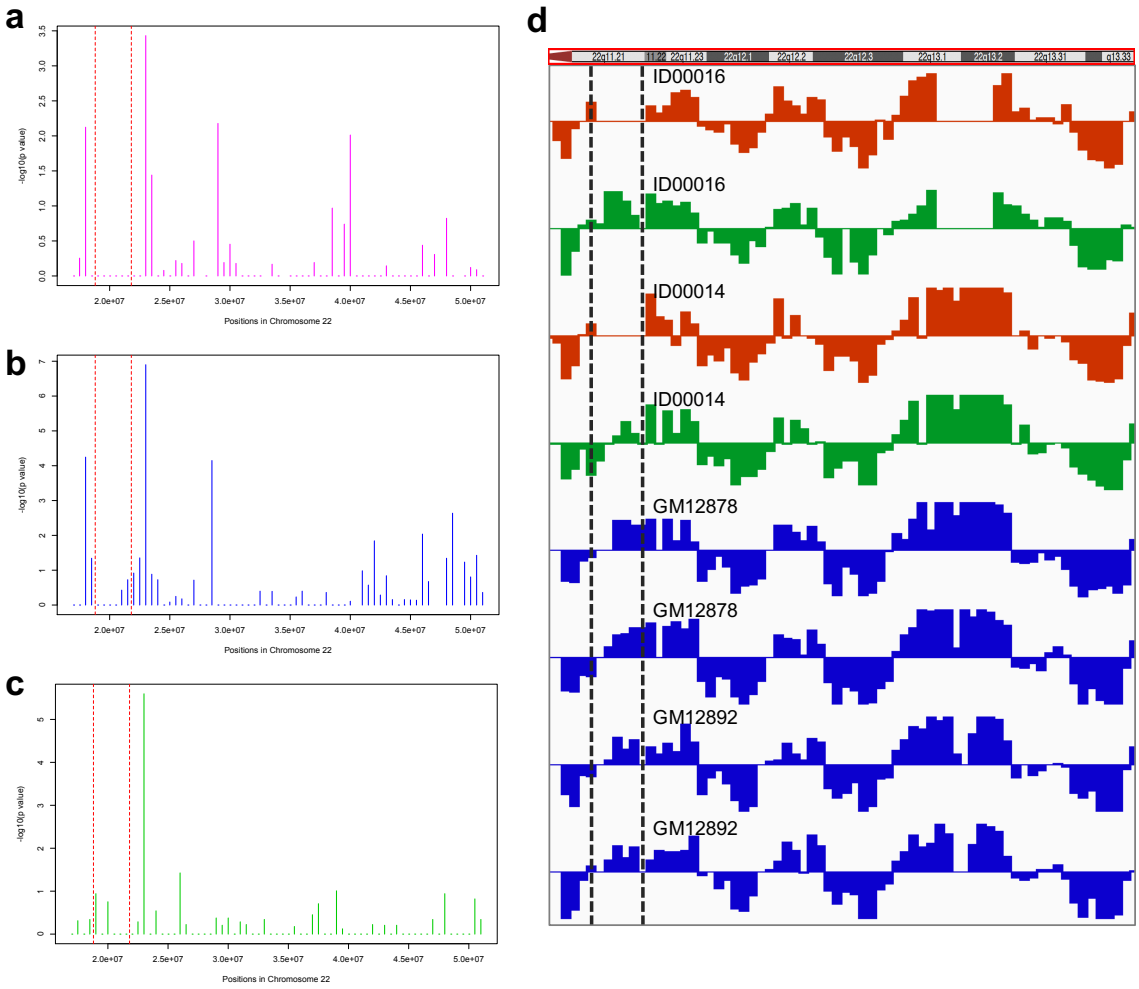


Figure 2. Haplotype-specific intra-chromosomal contacts of chromosome 22q. (a) Contacts within each homologous chromosome 22q for patients (ID00014 and ID00016) and controls (GM12878 and GM12892) at 200 kbp resolution. The chromosome 22q with 22q11.2 deletion is shown on the left for the two patient cell lines. Intensity of contacts is represented by a scale from 0 (white) to 5 (deep red). The gap in signal for the region from 39 Mbp to 42 Mbp in

239 *ID00016 is caused by a very low density of heterozygous SNVs. (b) Zoomed-in view of the*
240 *regions in black boxes in (a). The color scale goes from 0 (white) to 10 (red).*

241 Following this observation we wanted to examine whether there is an effect on the
242 chromatin marks concurrent to these increased chromosomal contact patterns. To do so we
243 performed enrichment analysis in our ChIP-Seq data for significantly differential signal
244 patterns for H3K27ac, H3K27me3 and CTCF, respectively. We found that both the distal and
245 proximal flanking regions of the 22q11.2 deletion were enriched with significantly
246 differentially bound sites of both H3K27ac and H3K27me3 (**Fig. 3a,b**). For CTCF the distal
247 deletion-flanking region was enriched with significantly differentially bound sites (Fisher's
248 exact test p = 2.53E-06) (**Fig. 3c**).



249

250 ***Figure 3. Distribution of significantly differentially enriched H3K27ac histone marks***
251 ***(a), H3K27me3 histone marks (b), CTCF binding sites (c) along chromosome 22q, and***
252 ***haplotype-specific A/B compartments for each homologous chromosome 22q (d).*** *Each*
253 *bar represents a 500 kbp bin. Dashed lines mark the boundaries of the 3 Mbp deletion in*
254 *22q11.2. Shown in (d) is the first eigenvector for the principal component analysis of the*
255 *chromosomal contact matrix. X-axis: position on chromosome 22q. Y-axis: value of the first*
256 *eigenvector. The signal from the homologous chromosome 22q with 22q11.2 deletion and the*
257 *intact chromosome 22q are shown in red and green, respectively, for patients (ID00016 and*
258 *ID00014). Signals from the homologous chromosomes 22q of the controls are in blue.*

259 More specifically, within the proximal deletion-flanking region, from 18 to 18.5 Mbp,
260 we found that 5 out of 24 sites with ChIP-Seq signal for H3K27ac and 6 out of 15 sites with
261 ChIP-Seq signal for H3K27me3 showed significantly differential binding (Fisher's exact test
262 p = 0.0075 for H3K27ac and p = 5.67E-05 for H3K27me3 respectively). Interestingly, for
263 H3K27ac all of the 5 sites with significant differential binding were bound less strongly
264 while for the same region for H3K27me3 all of the 6 sites with significant differential
265 binding were bound more strongly in the 22q11DS patient lines. For the distal region, from
266 23 Mbp to 23.5 Mbp, 5 out of 13 sites with H3K27ac binding and 11 out of 30 sites with
267 H3K27me3 binding showed significantly differential binding (Fisher's exact test p = 0.0004
268 for H3K27ac and p = 1.26E-07 for H3K27me3 respectively). Again we observed the
269 reciprocity between significantly differential changes for the two different histone marks: 4
270 out of 5 sites for H3K27ac gave a less strong signal while all of the 11 of such sites for
271 H3K27me3 showed a stronger signal in the 22q11.2DS patient cell lines.

272 Such reciprocity in signal strengths between these two histone marks is a general
273 feature of their principle of action and indicates that the observed changes in chromatin
274 marks are of physiological relevance.

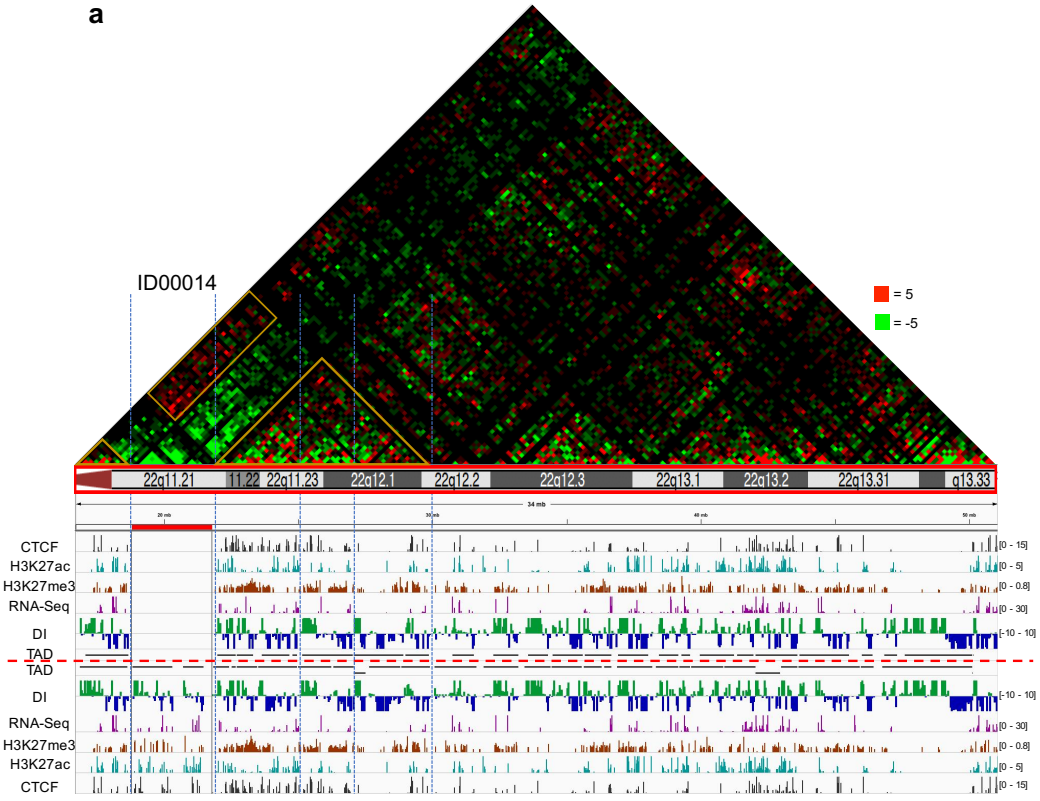
275 **Haplotype-specific analysis of the A/B compartments and of the topological domain**
276 **structure**

277 Previous studies of cell lines without large CNVs have revealed that there is a level of
278 organization in the cell's nucleus where the chromatin is partitioned into two states, termed
279 A and B compartments, which broadly correspond to large regions that are overall active
280 and overall inactive, respectively [20]. These A/B compartments can be derived from Hi-C
281 data. We sought to determine whether the 22q11.2 deletion might lead to changes in the
282 A/B compartments. We computed the haplotype-specific A/B compartments for the four

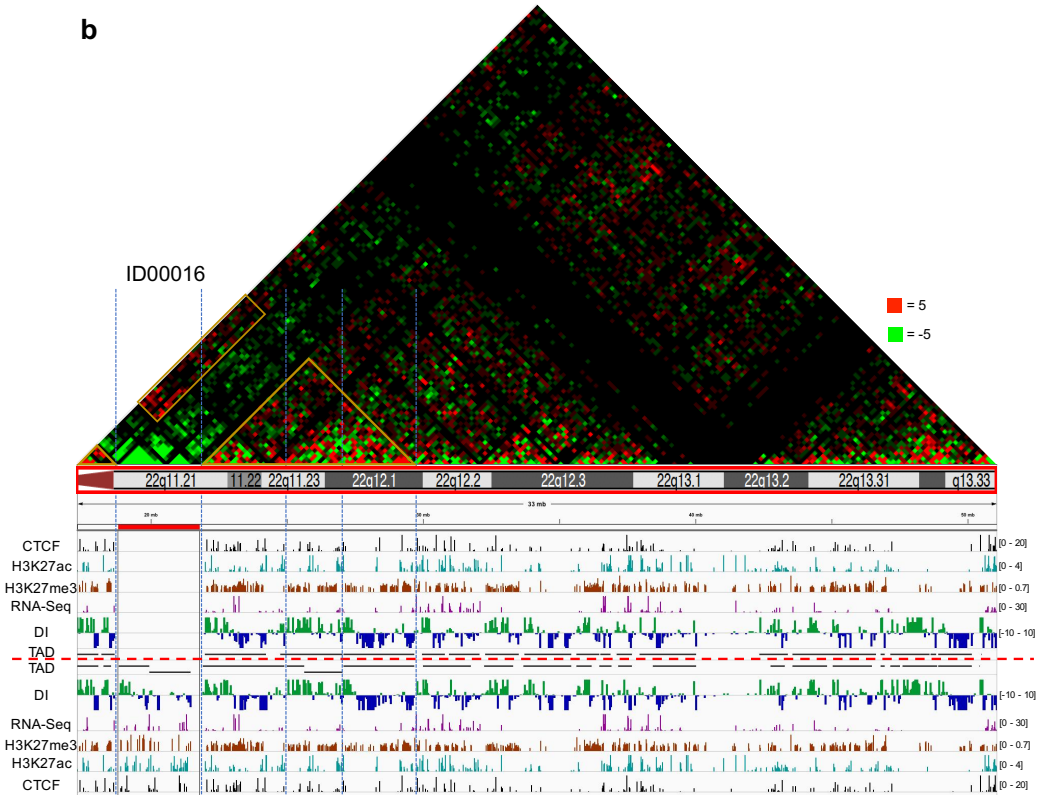
283 cell lines for which both deep capture Hi-C and genome haplotype phasing data is available
284 (**Fig. 3d**). We did not observe differences in the A/B compartments between cell lines from
285 patients and from controls and neither between the two homologous chromosomes 22q in
286 any individual cell line, patient or control. Our results indicated that A/B compartments of
287 the homologous chromosome with the 22q11.2 deletion were not affected by the deletion.

288 Topological domains (also known as topologically associating domains, TADs) are a
289 megabase-sized structural feature of the genome organization that is constituted of highly
290 self-interacting chromosome regions [33]. We calculated the haplotype-specific topological
291 domains on the four cell lines with both capture Hi-C and haplotype phasing data (**Fig. 4,**
292 **Supplementary Fig. 6-7**). Although there were variations in the calling of topological
293 domains across individuals and between the homologous chromosomes, the direction
294 indexes were highly consistent between the two homologous chromosomes in the controls
295 (**Supplementary Fig. 6c-d, Supplementary Fig. 7**) as well as in the patients (**Fig. 4,**
296 **Supplementary Fig. 6a-b**), suggesting no changes of topological domains on the
297 homologous chromosome with the 22q11.2 deletion. In spite of the increased contacts
298 between the flanking regions of the deletion, they were not strong enough to incur fusion of
299 the topological domains proximal and distal of the 22q11.2 deletion. However, the
300 increased contacts between the flanking regions of the deletion extended to multiple
301 deletion-distal topological domains.

302


a
ID00014
= 5
= -5
22q11.21
11.22
22q11.23
22q12.1
22q12.2
22q12.3
22q13.1
22q13.2
22q13.31
q13.33
34 mb
CTCF
H3K27ac
H3K27me3
RNA-Seq
DI
TAD
TAD
DI
RNA-Seq
H3K27me3
H3K27ac
CTCF
[0 - 15]
[0 - 5]
[0 - 0.8]
[0 - 30]
[-10 - 10]
[-10 - 10]
[0 - 30]
[0 - 0.8]
[0 - 5]
[0 - 15]

303


b
ID00016
= 5
= -5
22q11.21
11.22
22q11.23
22q12.1
22q12.2
22q12.3
22q13.1
22q13.2
22q13.31
q13.33
33 mb
CTCF
H3K27ac
H3K27me3
RNA-Seq
DI
TAD
TAD
DI
RNA-Seq
H3K27me3
H3K27ac
CTCF
[0 - 20]
[0 - 4]
[0 - 0.7]
[0 - 30]
[-10 - 10]
[-10 - 10]
[0 - 30]
[0 - 0.7]
[0 - 4]
[0 - 20]

304 ***Figure 4. Differential chromosome contacts for the two chromosomes 22q and***
305 ***haplotype-specific topological domains for two patient cell lines, ID00014 and ID00016.***
306 *Panel a: patient ID00014, panel b: patient ID00016. Heatmaps with bin size 200 kbp are*
307 *showing the intra-chromosomal contacts on the homologous chromosome 22q with the*
308 *22q11.2 deletion minus the intra-chromosomal contacts on the intact chromosome 22q***.** *The*
309 *color scale of the heatmap goes from -5 (green) to 5 (red) with 0 represented by black. Yellow*
310 *rectangles mark the increased contacts between the topological domains proximal and distal*
311 *of the 22q11.2 region on the homologous chromosome 22 with the deletion (the topological*
312 *domain on the proximal side of the deletion increases its contact not just with the topological*
313 *domain immediately distal of the deletion, as predicted, but also with the two following*
314 *topological domains. Yellow triangles and blue dashed lines mark the topological domains*
315 *involved in the increased contacts. Black lines and red bar mark the boundaries and extent of*
316 *the 22q11.2 deletion. Signal tracks for CTCF binding, histone marks, RNA-Seq, direction*
317 *indexes (DI) and topological domains are shown twice, for the chromosome 22q with deletion*
318 *(above dashed red line) and for the intact chromosome 22q (below dashed red line). Direction*
319 *indexes (DI) are calculated at 40 kbp resolution for each homologous chromosome 22q.*
320 *Haplotype-specific gene expression, histone modifications and CTCF binding, respectively, are*
321 *shown after normalizing the reads to 10 million read pairs. RNA-Seq and ChIP-Seq read-count*
322 *bars within the boundaries of the deletion but from the intact chromosome 22q are shown for*
323 *loci where there are the phased SNVs in GM12878.*

324 **Change of chromosomal *cis*-contacts involving the distal flanking region of**
325 **22q11.2del**

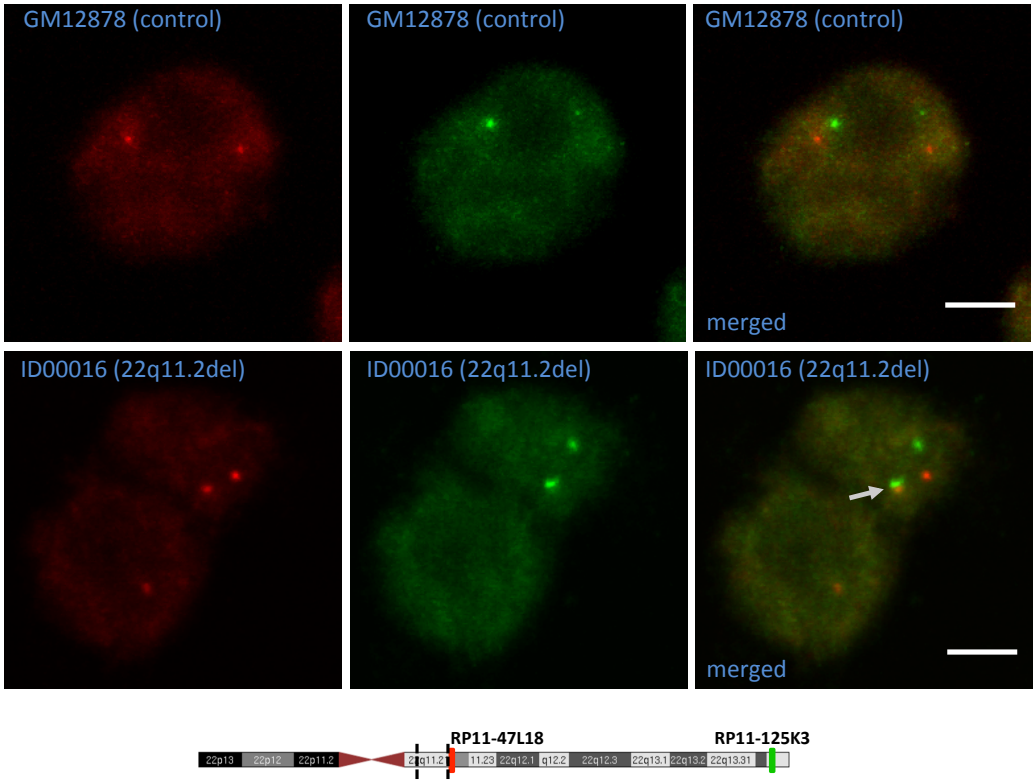
326 Based on the above findings that the genomic region distal of the 22q11.2 deletion was
327 affected on different levels of molecular regulation in 22q11.2DS patient cell lines, we next

328 sought to investigate whether the presence of the 22q11.2 deletion also affected the
329 chromosomal *cis*-contacts of the distal deletion-flanking region with any other region on
330 chromosome 22q. To do so, we analyzed our Hi-C data by calculating the fold change of
331 contacts between the 21.5-22 Mbp window, which is situated right distal of the 22q11.2
332 deletion, and all the other 1 Mbp-sized regions on chromosome 22q. We observed that
333 region 17-18 Mbp, i.e. the breakpoint-proximal region right *proximal* of the deletion,
334 showed the largest fold change (2.04) of contact with region 21.5-22 Mbp between
335 22q11DS and control cell lines. This was as expected given that these chromosomal regions
336 were brought into close proximity to each other by the 22q11.2 deletion (**Fig. 1a**).

337 The second largest fold change (1.96) of *cis*-contacts involving region 21.5-22 Mbp was
338 for contacts with region 50-51 Mbp, i.e. towards the very telomeric end of chromosome 22q
339 (**Fig. 1a**). Intriguingly, we also observed strong positive correlation between region 22-22.5
340 Mbp and region 50-50.5 Mbp for CTCF binding (Pearson's $r = 0.933$, $p = 0.02$)
341 (**Supplementary Fig. 8a**) and between region 22-22.5 Mbp and region 50.5-51 Mbp for
342 H3K27ac enrichment (Pearson's $r = 0.811$, $p < 0.05$) (**Supplementary Fig. 8b**).
343 Furthermore there was weak correlation between region 22-22.5 Mbp and 50.5-51 Mbp for
344 H3K27me3 enrichment (Pearson's $r = 0.74$, $p = 0.090$) (**Supplementary Fig. 8c**). Given that
345 both region 21.5-22 Mbp and region 22-22.5 Mbp are in the distal flanking region of the
346 22q11.2 deletion, our results indicate that increased chromosomal contacts between the
347 distal deletion-flanking region and the telomeric region 50-51 Mbp may be associated with
348 the differential changes of histone modifications and CTCF binding that we found to be in
349 correlation between these two regions.

350 To validate the increased chromosomal contact between the deletion-distal region and
351 the telomeric end of 22q we performed 3D FISH on 5 cell lines with the 22q11.2 deletion

352 and on 6 control cell lines, using two FISH probes specific to loci in region 21.8-22.5 Mbp
353 (RP11-47L18: chr22:21,931,796–22,118,344) and in region 50-51 Mbp (RP11-125K3:
354 chr22:50,149,996-50,287,311), respectively. FISH showed that the distal flanking region of
355 the 22q11.2 deletion is significantly closer to region 50-51 Mbp in nuclear space in
356 22q11DS cell lines than in control cell lines (p-value 0.008) (**Fig. 5; Supplementary Fig. 9**).
357 Taken together, our results strongly indicate that the 22q11.2 deletion causes
358 conformational changes on several levels on chromosome 22q.



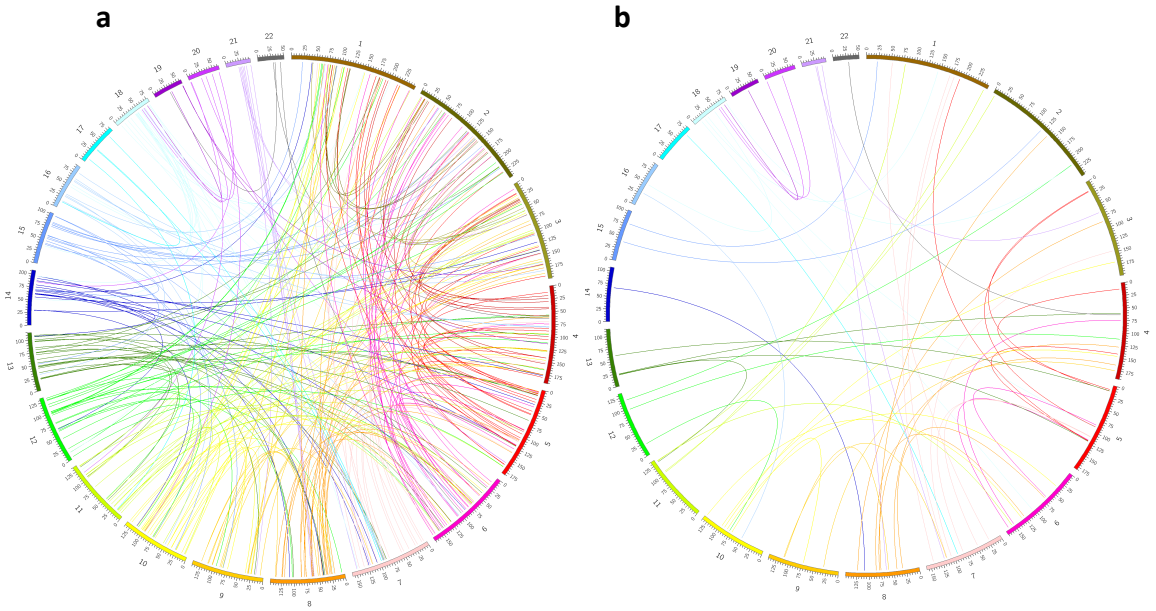
359

360 ***Figure 5. Examples of 3D FISH visualization of intra-chromosomal interaction changes.***
361 *The regions for which Hi-C predicted changes in interaction, regions 21.8-22.5 Mbp and 50-51*
362 *Mbp on chromosome 22q, were visualized by 3D DNA FISH using BAC probes RP11-47L18 and*
363 *RP11-125K3, labeled with biotin (red) or digoxigenin (green). Arrow: an example for the red*

364 *and green FISH probes in close proximity in a cell carrying the 22q11.2 deletion. Scale bars are*
365 *5 μm. Red and green bars on chromosome 22q indicate the locations of the biotin and*
366 *digoxigenin FISH probes, respectively. Black dashed lines indicate the position of the 3 Mbp*
367 *deletion.*

368 **Global changes of inter-chromosomal contact patterns**

369 To explore the effect of the 22q11.2 deletion on *trans*-contacts, i.e. between any non-
370 homologous autosomes in the nucleus of 22q11.2DS patient lines, we analyzed our Hi-C
371 data for significantly different *trans*-contacts between 22q11DS and control cell lines. On
372 the genome-wide level, we found 272 *trans*-contacts with a p-value of less than 0.0001 (**Fig.**
373 **6a**). Interestingly the majority of these chromosomal *trans*-contacts did not involve
374 chromosome 22q as one of the interacting partners. Notably, 56 of these inter-chromosomal
375 contact signals are among the top 5% of the strongest *trans*-contacts (**Fig. 6b**). This
376 enrichment of significantly different genome-wide chromosomal *trans*-contacts within the
377 strongest *trans*-contacts is again highly statistically significant (Fisher's exact test $p < 2.2e-$
378 16). We found far fewer significantly different genome-wide chromosomal *trans*-contacts by
379 randomly swapping the assignment of 22q11.2 deletion and control status across our Hi-C
380 data sets, by comparing within control cell lines and by comparing within 22q11.2del cell
381 lines (**Supplementary Table 4**). None of these swapping analyses achieved the same
382 enrichment. This indicates that a relevant amount of the significantly different genome-
383 wide chromosomal *trans*-contacts is not due to random chromosomal motion or to as-of-yet
384 unknown factors such as cell-culture variations across the LCLs. Rather, our analysis points
385 to a genome-wide disturbance of the network of chromosomal *trans*-contacts that is at least
386 in part attributable to the presence of the 22q11.2 deletion on chromosome 22q.



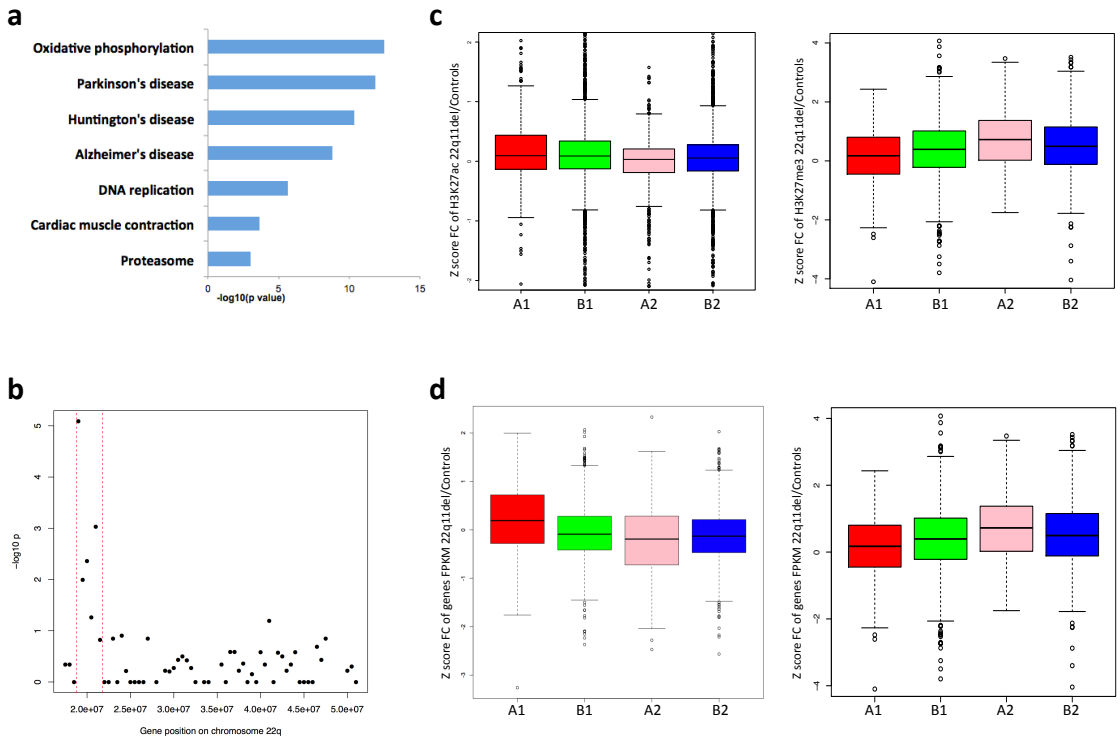
387

388 ***Figure 6. Genome-wide inter-chromosomal contact changes (either more or less***
389 ***frequent) determined by Hi-C analysis in 22q11.2del versus control cell lines.*** *(a) Circos*
390 *plot of the inter-chromosomal contacts exhibiting differential interaction in 22q11.2del versus*
391 *control cell lines at significance level of 0.0001. (b) Circos plot of the inter-chromosomal*
392 *contacts exhibiting differential interaction in 22q11.2del versus control cell lines at*
393 *significance level of 0.0001 and showing only the top 5% strongest inter-chromosomal*
394 *contacts.*

395 **Global changes of gene expression patterns**

396 To investigate the global effect of the 22q11.2 deletion on gene expression, we performed
397 differential expression analysis between the 22q11.2DS and control cell lines. Of the 11,374
398 genes with detectable levels of expression (FPKM > 0.5), 1,610 genes showed significantly

399 differential expression (FDR < 0.05). Gene ontology analysis indicated that these
400 differentially expressed genes are enriched for genes involved in mitochondrial pathways
401 such as the respiratory chain (n = 32, p-value = 2.96E-11). KEGG pathway analysis
402 demonstrated enrichment of genes involved in oxidative phosphorylation and
403 neurodegenerative diseases such as Parkinson's disease (**Fig. 7a**). Earlier studies have
404 noted that there are several genes related to mitochondrial function that are located within
405 the 22q11 deletion boundaries [34], therefore the pathways associated with this cellular
406 function would have a high likelihood to be affected by the change in copy number. The
407 enrichment for pathways related to neurodegenerative disorder is unexpected in LCLs but
408 notable, as an association between 22q11DS and Parkinson's Disorder has been reported
409 previously [35-37]. This may indicate that the LCL cell culture may in certain instances be of
410 relevance for the molecular study of 22q11DS even beyond fundamental and general levels.



411

412 ***Figure 7. Differential expression analysis and correlation between gene expression and***
413 ***epigenetic profiles.*** *(a) KEGG pathway analysis for genes differentially expressed between*
414 *22q11.2del and control cell lines. (b) Enrichment analysis for chromosome 22q for*
415 *differentially expressed genes. Red dashed lines mark the boundaries of 22q11.2del region. (c)*
416 *Differentially expressed genes correlated with different H3K27ac and H3K27me3 signal*
417 *strengths in 22q11.2del versus control cell lines. A1 represents genes significantly up-regulated*
418 *in 22q11.2del cell lines while B1 represents non-significantly up-regulated genes; A2*
419 *represents genes significantly down-regulated in 22q11.2del cell lines while B2 represents*
420 *non-significantly down-regulated genes. The y-axis shows the z-score transformed fold change*
421 *of H3K27ac signals (left) or H3K27me3 signals (right) between 22q11.2del and control cell*
422 *lines. (d) Differential H3K27ac and H3K27me3 signal strengths exhibited correlation with*
423 *differential gene expression between 22q11.2del and control cell lines. A1 represents genes*
424 *whose TSSs were significantly more marked by H3K27ac (left) or by H3K27me3 (right) in*
425 *22q11.2del cell lines. B1 represents genes with TSSs non-significantly more marked by*
426 *H3K27ac (left) or by H3K27me3 (right). A2 represents genes whose TSSs were significantly*
427 *less marked by H3K27ac (left) or by H3K27me3 (right) in 22q11.2del cell lines. B2 represents*
428 *genes with TSSs non-significantly less marked by H3K27ac (left) or by H3K27me3 (right). The*
429 *y-axis shows the z-score transformed fold change of genes' FPKM in 22q11.2del versus control*
430 *cell lines.*

431 We also carried out genome-wide analysis to determine if there are entire genomic
432 regions that are enriched for differentially expressed genes. As expected, the most
433 significant signals for this analysis were located in the 22q11.2 region (**Fig. 7b;**
434 **Supplementary Fig. 10**). No other regions genome-wide achieved FDR corrected
435 significance (**Supplementary Fig. 10**).

436 **Correlation between histone modification and gene expression**

437 To examine whether the gene expression changes are consistent with the histone
438 modification changes, we assigned the H3K27ac and H3K27me3 peaks to their nearest
439 genes based on the distance to their transcription start sites and for each gene, we only
440 retained the closest peak for both histone marks. We observed that significantly
441 upregulated genes in 22q11.2DS cell lines exhibited significantly higher fold change of
442 H3K27ac enrichment (permutation test p = 0.0140) and significantly lower fold change of
443 H3K27me3 enrichment (permutation test p = 0.0018) than genes non-significantly
444 upregulated in 22q11.2DS cell lines (**Fig. 7c**). Consistently, significantly downregulated
445 genes in 22q11.2DS cell lines exhibited significantly lower fold change of H3K27ac
446 enrichment (permutation test p = 0.0002) but significantly higher fold change of H3K27me3
447 enrichment (permutation test p = 0.0227) than genes non-significantly down regulated in
448 22q11.2DS cell lines (**Fig. 7c**).

449 Moreover, genes whose transcription start sites (TSSs) showed significantly
450 upregulated binding by H3K27ac in 22q11DS cell lines exhibited significantly higher fold
451 change of expression (permutation test p < 2.2e-16) than those with non-significantly
452 upregulated binding, while those TSSs showing significantly downregulated binding
453 exhibited a significantly lower fold change of expression (permutation test p = 0.0202) than
454 those with non-significantly downregulated binding (**Fig. 7d**). Consistently, genes whose
455 TSSs showed significantly upregulated binding by H3K27me3 in 22q11DS showed
456 significantly lower fold change of expression (permutation test p = 0.0037) while those with
457 significantly downregulated binding exhibited significantly higher fold change of expression
458 (permutation test p = 0.0027) between 22q11DS and control cell lines (**Fig. 7d**).

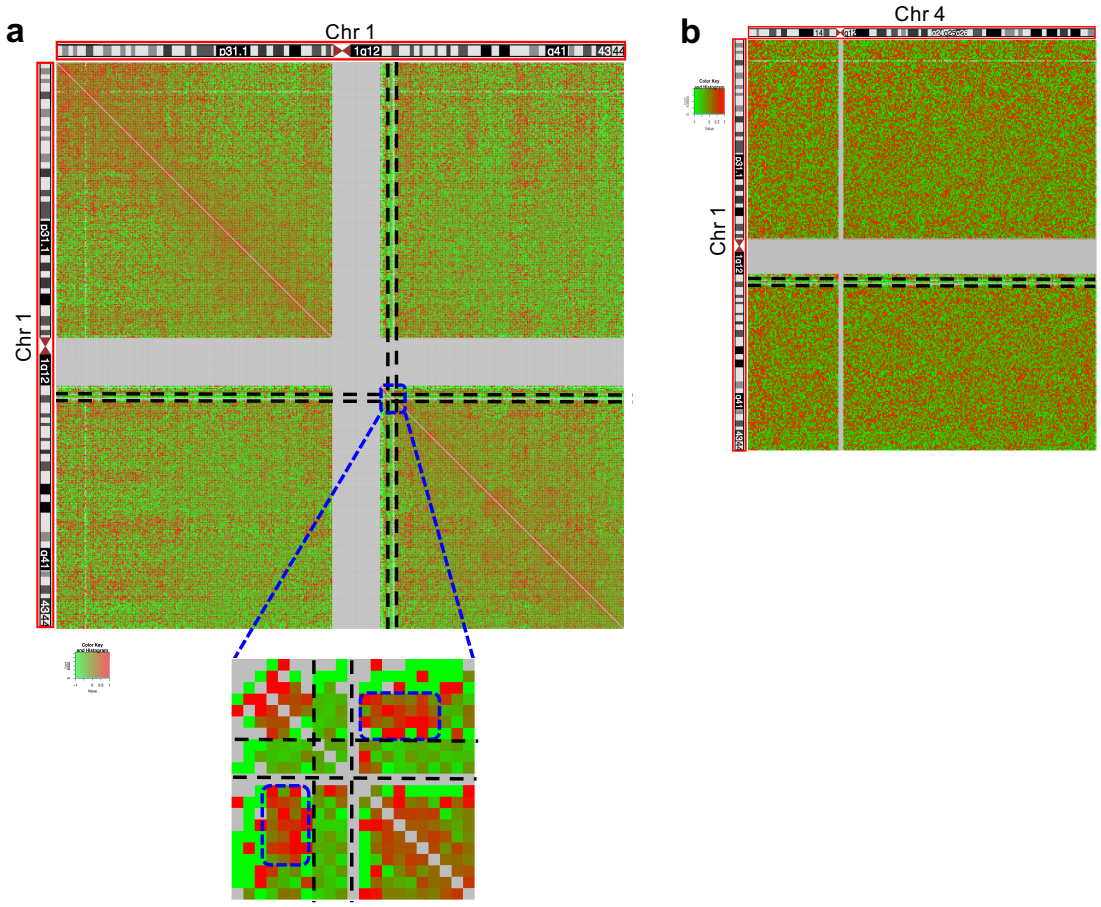
459 We also calculated the correlation between gene expression and H3K27ac binding

460 affinity across the individuals for all the genes with TSSs bound by H3K27ac. In line with the
461 above results, we found significantly higher Pearson's correlation coefficients than the
462 coefficients obtained from permutations (Wilcoxon rank sum test $p < 2.2e-16$)
463 (**Supplementary Fig. 11a**). This difference was even more significant when we included
464 the differential expressed genes (absolute fold change > 2) only (Wilcoxon rank sum test $p <$
465 $2.2e-16$) or the differential H3K27ac bound genes only into our analysis (Wilcoxon rank
466 sum test $p < 2.2e-16$) (**Supplementary Fig. 11a**). We also observed significantly higher
467 Pearson's correlation coefficients between gene expression and H3K27me3 binding than
468 the coefficients obtained from permutations for the differential expressed genes (absolute
469 fold change > 2) only (Wilcoxon rank sum test $p = 9.72e-04$) and the differential H3K27me3
470 bound genes only (Wilcoxon rank sum test $p = 3.71e-06$) (**Supplementary Fig. 11b**).
471 Together, our results demonstrate that gene expression changes are associated with histone
472 modification changes in 22q11DS cell lines compared with control cell lines and that this
473 phenomenon occurs across the entire genome in cells with the deletion in 22q11.2.

474 **Effects of a large CNV on chromosome 1q21.1 on chromosome folding**

475 To explore whether large CNVs other than the one on chromosome 22q11.2 can lead to
476 changes in the patterns of chromosome folding, we performed Hi-C on two LCLs with a
477 heterozygous deletion of approximately 1.35 Mbp in size on chromosome 1q21.1
478 (1q21.1del). This deletion is strongly associated with the development of schizophrenia [38-40].
479 Similar to what we had found for the 22q11.2 deletion, we observed that both *cis*-contacts
480 and *trans*-contacts between the 1q21.1 deletion regions and other regions were decreased
481 in 1q21.1del cell lines relative to control cell lines (**Fig. 8a,b**). As observed in the 22q11DS
482 cell lines, in the 1q21.2del cell lines there was an increase of intra-chromosomal contacts
483 between the regions directly flanking the main CNV of 1q21.1 (**Fig. 8a**). Global changes of

484 inter-chromosomal contacts were also observed (**Supplementary Fig. 12**), again consistent
485 with the inter-chromosomal contact changes in 22q11.2 LCLs. Taken together cells with the
486 large deletion CNV on chromosome 1q21.1 exhibited similar effects on chromosomal *cis*-
487 contacts and *trans*-contacts as found in cells with the large deletion CNV on chromosome
488 22q11.2, which points towards our findings in 22q11DS being generalizable across large
489 neuropsychiatric CNVs.



490

491 ***Figure 8. Effect of the large deletion CNV on 1q21.1 on chromosome conformation.*** *(a)*
492 *Fold change of cis-contacts of chromosome 1 in 1q21.1del versus control cell lines. Blue boxes*
493 *mark the regions of increased contacts between the proximal and distal flanking regions of the*
494 *1q21.1 deletion. (b) Fold change of trans-contacts between chromosome 1 and 4 in 1q21.1del*

495 *cell lines versus control cell lines. Black dashed lines mark the boundaries of the 1q21.1*
496 *deletion. Each cell in the heatmap represents the inter-chromosomal contact level between*
497 *two 500 kbp regions.*

498 **Discussion**

499 Large CNVs are an important feature of the genetic architecture of several major
500 neurodevelopmental psychiatric disorders as well as of conditions involving aberrant
501 morphology of many organ systems. Their effects on the level of phenotype are complex and
502 the molecular mechanisms mediating these effects are very incompletely understood. While
503 it is still a good assumption that a considerable portion of these mechanisms are a direct
504 consequence of the copy number change of the genes within a given CNV's boundaries and
505 the resulting changes in expression levels for these genes, it also seems plausible to
506 investigate whether additional levels of complexity exist regarding the effects of a large CNV
507 across multiple layers of the control of gene activity. This plausibility stems from several
508 observations and lines of reasoning. There are large numbers of genes affected genome
509 wide and far distal from the main CNV. This leads us to consider the basic principles that
510 govern organizational features of the nucleus and the chromatin such as regulatory
511 domains being bounded by protein factors that recognize binding sites in the DNA sequence
512 which in turn could be affected (i.e. deleted or duplicated) by a CNV, or long-range
513 chromosome contacts being also influenced by the length of the involved chromosomes
514 (and that length being changed by large CNVs). Finally there have been several recent
515 reports about mutations in chromatin remodeling genes in the context of neuropsychiatric
516 disorders pointing to the importance of proper molecular management on the epigenomic
517 level in these conditions.

518 Here we studied the effects of the important large CNV on chromosome 22q11.2, and in

519 a more limited fashion the effects of the large CNV on chromosome 1q21.2, on chromatin
520 conformation including long-range chromosome contacts and domain formation, epigenetic
521 profiles and gene expression. We used a novel approach that combined genome phasing and
522 targeted fragment enrichment for very deep chromosome-wide Hi-C to investigate the
523 effects of the 22q11.2 deletion in a haplotype-specific fashion. To our knowledge this is the
524 first study of changes in chromatin conformation and epigenetic profile changes caused by
525 deletions on 22q11.2, and 1q21.2, or any such large CNVs with strong relevance to organ
526 development and in particular brain development and neuropsychiatric disorders using
527 haplotype-specific analysis. We observed multiple effects of the large CNVs on long-range
528 chromosome contacts, chromatin domains and epigenetic profiles as well as on gene
529 expression. More specifically we found increased contacts between distal and proximal
530 flanking regions of the 22q11.2 deletion in 22q11DS cell lines in contrast with control cell
531 lines. Our haplotype-specific analyses of chromosomal interactions demonstrated that these
532 increased contacts only occurred on the chromosome 22q carrying the deletion.
533 Interestingly, both distal and proximal flanking regions of the 22q11.2 deletion were
534 enriched with differential signals for the histone marks H3K27ac and H3K27me3 while
535 distal flanking regions were also enriched with differentially binding sites of CTCF.

536 We did not observe significant correlation either between chromosome contact
537 changes and gene expression changes or between chromosome contact changes and
538 epigenetic profile changes (data not shown) on the genome-wide level. Also, we did not
539 observe the end of chromosome 22 (50-51 Mbp) to be enriched with differentially
540 expressed genes (2 out of 24 genes) despite the Hi-C finding, validated by FISH, of
541 significantly increased contacts with the distal flanking region of the 22q11.2 deletion in
542 22q11.2DS cell lines. Allele-specific expression (ASE) analysis did not provide evidence for
543 the existence of differential expression between the intact chromosome 22q and the

544 chromosome with the 22q11.2 deletion (**Supplementary Fig. 13**), outside the CNV
545 boundaries. Taken together our results indicate that chromosome conformation changes
546 caused by the 22q11.2 deletion contribute to gene expression changes and epigenetic
547 profile changes between 22q11.2DS and control cell lines but not always in a linear and
548 clearly deterministic way.

549 However, we did observe significant positive correlation between H3K27ac changes
550 and gene expression changes but negative correlation between H3K27me3 and gene
551 expression changes on the genome-wide level. These findings demonstrated that the
552 epigenetic profiles were reshaped genome-wide in correlation with extensive gene
553 expression changes in 22q11.2DS cell lines.

554 Our pathway enrichment analysis of differentially expressed genes revealed candidate
555 pathways that might be associated with various clinical phenotypes of 22q11DS such as
556 cardiac symptoms and neurodegenerative disorders **(Fig. 7a)**, in addition to general
557 molecular pathways, some with obvious potential association to 22q11DS such as
558 mitochondrial energy metabolism. It is important to remain cautious when interpreting
559 these disease-specific results as they are based on data from a cell type that is not directly
560 relevant to the organ systems in question. At the same time many genes will of course be
561 relevant in more than one cell type and organ, and their expression levels may be regulated
562 by similar molecular mechanisms across cell types. With the proper caution in place the
563 findings about potential relevant pathways may be considered as an unexpected side
564 product of this study that may point towards additional routes for further analysis. In any
565 case the pathway analyses do not affect the findings about effects on chromatin states and
566 chromosome folding in 22q11DS.

567 Our study does not address the question of what the molecular causes for phenotypic

568 variance between individual carriers of the 22q11.2 deletion might be. This is a highly
569 important question but one that, given the vast number in each individual genome of
570 potential genetic modifiers that together with environmental factors are the prime suspects
571 for such variance, will need far larger cohort sizes of cell lines modeling the tissues where
572 such variance is observed. Such cohorts are currently being assembled and will be the basis
573 of highly interesting work in the coming years.

574 The effects of decreasing *cis*- and *trans*-contacts and the increased chromosomal
575 contacts between distal and proximal deletion-flanking regions in 22q11Ds cell lines were
576 also observed in cell lines with the 1q21.1 deletion. That similar results that were obtained
577 on that level of observation for the 22q11.2 deletion and the 1q21.1 deletion could suggest
578 that at least some of the changes caused by a large deletion CNV are generalizable for such
579 chromosomal aberrations instead of specific to the 22q11.2 deletion.

580 As a cautionary note on the technical level, we demonstrated here that for genomes
581 with a large deletion CNV the appropriate normalization methods for Hi-C data have to be
582 chosen with great care to avoid false findings. For instance we would have reached the
583 conclusion that the chromosomal contacts within the deletion regions are not decreased in
584 cell lines with deletion compared with control cell lines if the hiclib software package [23] had
585 been used for normalization of Hi-C data, instead of the hicpipe package [22].

586 Lastly, we were able to generate haplotype-specific maps of chromosomal contacts by
587 using a novel approach that combined genome-sequencing based phasing information and
588 chromosome-wide capture of Hi-C fragments before sequencing. These haplotype-specific
589 contact analyses enabled us to distinguish between the intact chromosome 22q and the
590 chromosome 22q with deletion. Based on this we derived chromosomal A/B compartments
591 and topological domains for both homologous chromosomes 22q separately. There was no

592 switch of A/B compartments on the chromosome 22q with the 22q11.2 deletion compared
593 to the intact chromosome 22q in the patients. There also was no evidence for newly fused
594 topological domains or changes of topological domain boundaries on the chromosome 22q
595 with the 22q11.2 deletion. Earlier, Lupiáñez et al. [41], using the 4C assay, demonstrated in a
596 mouse model as well as in human lines from patients with rare malformations of the limbs,
597 that large structural changes in the genomic sequence on mouse chromosome 1 and human
598 chromosome 2 can affect the topological domain architecture that is situated directly on top
599 of the sequence change. However, our results did not show comparable effects of the
600 22q11.2 deletion. Further studies will determine which factors, such as for example the size
601 of a given CNV or its position relative to topological domain-defining sequence elements,
602 might determine whether topological domains are altered or not. Interestingly, although the
603 topological domains on or nearby the 22q11.2 deletion themselves were not altered (rather,
604 they were either deleted entirely or left unchanged) there was an unexpected pattern of
605 increase in contact between the topological domains flanking the deletion, where the
606 proximal topological domain not only increased its frequency of interaction with the
607 nearest distal topological domain but also with the two topological domains following that.

608 In summary, we found multiple effects of the large deletion CNV on chromosome
609 22q11.2 on long-range chromosome contacts, chromatin organization, epigenetic profiles
610 and gene expression. Extensive changes on these levels caused by the 22q11.2 deletion are
611 global and seem to be rippling along the entire chromosome carrying the deletion as well as
612 across the entire nucleus rather than being confined to the deletion region only. Such effects
613 had never been shown before outside of cancer cell lines [42]. Furthermore, in contrast to the
614 findings in cancer cells we used a larger cohort of individual patient cell lines for most of
615 our analyses, with all lines carrying only one main large CNV that is clearly and strongly
616 associated in a causative manner with a neurodevelopmental phenotype. The earlier study

617 in cancer used only two cancer cell lines that each carried multiple large CNVs which could
618 also have been a consequence rather than a cause of the disease phenotype. This makes it
619 much more likely that the higher-order effects of the large CNVs that we observed may be
620 contributing to the molecular etiology of the developmental disorders in question, a point
621 which is further strengthened by another recent paper [4], where the authors describe
622 studying the effects of large CNVs on chromosome 16p11 on chromosome interactions. The
623 large CNVs on 16p11 are almost as strongly associated with neurodevelopmental disorders
624 than the large CNV on 22q11.2. While the study on the 16p11 CNVs used a somewhat
625 smaller number of cell lines than our study and also used 4C as a method of discovery,
626 which is, unlike the Hi-C method used by us, not able to detect changes in a global and
627 unbiased fashion, it is one more independent piece of supporting evidence for the biological
628 validity and general relevance of the findings which we describe here.

629 While we were able to show possible correlations across several pairs of the molecular
630 levels that we assayed in this study there are other combinations of molecular levels that
631 show no obvious connection to each other in our data. We believe that this could be a
632 function of either the developmental time point or the cell type, or both, being removed
633 from those where the 22q11.2 and 1q21.1 deletions most likely exert some of their
634 strongest effects (e.g. during embryonal development and in cells of the developing central
635 nervous system). The clear effects on several genomic and epigenomic levels that we were
636 able to observe in LCLs could represent the afterglow or reflection of a molecular tragedy
637 that played out earlier in the development of various organs in the patients carrying these
638 deletions. At the same time the strength of this afterglow would hint at the strength of the
639 effects that impacted across various levels of molecular control and gene regulation by
640 CNVs of this size.

641 **Methods**

642 **Cell lines and data generation for this study.** All cell lines were either acquired from the
643 Coriell cell repository (cell lines IDs starting with GM or ID) or were taken from the
644 Molecular Genetics of Schizophrenia (MGS) cohort (dbGaP Study Accession:
645 phs000167.v1.p, cell lines 52425 and 82699), and were appropriately consented. The Hi-C
646 assay was carried out according to the original protocol [20], with several modifications. For
647 capture Hi-C of chromosome 22q we prepared separate in situ Hi-C libraries using the
648 previously reported protocol [32]. 2.1 million oligomer capture probes that densely tile the
649 sequence of chromosome 22q were designed using the NimbleDesign Software and ordered
650 from Roche Sequencing (formerly NimbleGen-Roche, Pleasanton, CA). Roche Sequencing's
651 SeqCap EZ Choice XL Enrichment Kit was used to capture fragments from the in situ Hi-C
652 libraries using the chromosome 22q specific capture oligomers. ChIP-Seq was performed as
653 described in previous studies [43,44]. The RNA-seq libraries were generated according as in [45].
654 All experimental procedures are described in the **Supplementary Information**.

655 **Hi-C data analysis.** All Hi-C data were produced using Illumina paired-end sequencing with
656 a read length 2 x 101 bp. As there might be ligation junctions in the reads, we performed
657 iterative mapping using bowtie2 as in Imakaev et al. 2012 [20]. Briefly, we computationally cut
658 all the reads to 25 bp first and mapped them to human genome (hg19). Then we extended
659 the non-uniquely mapped reads by 5 bp to 30 bp and mapped them again. This process was
660 repeated until the read length was extended to 101 bp. This iterative mapping did improve
661 the mapping rate (**Supplementary Table 5**). Each read end was mapped separately using
662 the single ends mode. Only uniquely mapping reads were used and PCR duplicate read pairs
663 were removed. We only included autosomes in our study. The filtered contact number is
664 listed in **Supplementary Table 6**.

665 We compared three different data normalization methods: hiclib [23], hicpipe [22] and
666 HiCNorm [24]. All of the three tools were run using the default parameters except for the
667 segment length threshold being set to 600 bp. We chose normalized metrics on hicpipe for
668 the following analyses, using a bin size of 40 kbp for topological domain analysis and of 500
669 kbp for the other analyses. The total number of contacts was normalized for each sample
670 before combining cell lines in each category (control, 22q11del and 1q21del, respectively).
671 Fold changes of log2 transformed mean contacts between deletion cell lines and control cell
672 lines were calculated by (deletion – control)/control.

673 To identify the differential inter-chromosomal contacts, we only included contacts with
674 at least 1 supporting read pair in each of the cell lines. Differential contacts analysis was
675 conducted by two-sided Student's t-test with the "t.test" function in R using the normalized
676 metrics. Fisher's exact test was used to assess the enrichment of differential contacts within
677 the top 5% strongest contacts. We also performed the same analysis by permuting the
678 control and 22q11del status of the cell lines ten times. Comparison within control cell lines
679 and within 22q11del cell lines were performed by randomly dividing the cell lines into two
680 groups three times.

681 **Generation of haplotype-specific chromosomal contact maps**. To phase the homologous
682 chromosomes of two patient cell lines we first carried out deep whole genome sequencing
683 (37.4x genome-wide coverage for ID00014 and 32.8x genome-wide coverage for ID00016),
684 which yielded the required heterozygous SNV information. Haplotype phasing was then
685 carried out by statistically aided, long-read haplotyping (SLRH) [21] in combination with
686 Mendelian inheritance patterns of informative SNVs, based on the knowledge that ID00014
687 is a parent of ID00016. Briefly, genomic DNA was sheared to fragments of about 10 kbp
688 length, which were diluted as needed and pipetted into a 384-well plate with 3,000–6,000

689 gDNA fragments in each well. Within each well fragments were amplified, fragmented
690 further and ligated to barcodes unique to each well. Fragments from across all wells were
691 pooled together and sequenced on Illumina HiSeq. Sequencing reads were mapped to the
692 reference genome and assigned to their unique well based on their barcodes. For each well
693 reads were assembled into haplotype blocks based on their overlapping heterozygous SNVs.
694 The haplotype blocks were then used to construct long haplotype contigs using Illumina's
695 haplotyping algorithm Prism (v2.2). The N50 length of the haplotype contigs was 492,634
696 bp and 453,807 bp for ID00014 and ID00016 respectively. SNVs which could be phased by
697 Mendelian inheritance analysis (heterozygous in one individual but homozygous in the
698 other) were assigned to the phased haplotype contigs to assemble these haplotype contigs
699 into whole phased chromosomes. The phased genome data of two control cell lines from
700 related donors had been previously reported [21] and were included in the analysis. More
701 details on this approach can be found in the **Supplementary Information.**

702 Capture Hi-C reads were mapped by BWA-MEM [46] in single end mode after ligation
703 junction removal using Cutadapt [47]. Reads spanning the position of phased SNVs were
704 assigned to the corresponding homologous chromosome using custom-written scripts. Then
705 HOMER [27] was used to pair up the reads, filter the read pairs assigned to haplotypes
706 (duplicates, self-ligations, read pairs less than 1.5x the sequencing fragment length or
707 distance to restriction site more than 1.5x the fragment length), and generate contact
708 matrices for each homologous chromosome separately. To normalize the contacts involving
709 the 22q11.2 region to the comparable level with contacts involving other regions on the
710 intact homologous chromosome 22 in the patients we calculated the expected number of
711 phased heterozygous SNVs in this 3 million bp region, based on the density of phased
712 heterozygous SNVs on chromosome 22q in each patient. Random single base pair positions
713 in the 22q11.2 region were sampled to the calculated expected number. Only the contacts

714 involving these sampled positions were included in the analysis. We performed this random
715 sampling for ten times and the pattern remained consistently similar. We adjusted for
716 recombination events for the haplotypes of ID00014 on the grounds that intra-
717 chromosomal contacts should be more prevalent than inter-chromosomal contacts.

718 Identification of A and B compartments was performed as in Lieberman et al. 2009 [20]
719 and topological domains were identified as in Dixon et al. 2012 [33].

720 **3D FISH.** Two human DNA BACs (clones RP11-47L18 and RP11-125K3) covering two
721 distinct regions of chromosome 22 were labeled with biotin or digoxigenin (DIG) by nick
722 translation kit (Roche Applied Sciences) to make FISH probes (Roche). In-situ hybridization
723 was performed according to the method published by [48], with several modifications.
724 Experimental details and data analysis could be found in the **Supplementary Information**.

725 **Differential expression analysis.** All RNA-seq data were generated using Illumina paired-
726 end sequencing with read length 101 bp. Reads were mapped to hg19 and transcriptome
727 reference with TopHat 2 [49]. TopHat 2 was run using default parameters but with the
728 coverage search being turned off. The mapped reads were analyzed by Cufflinks [50].
729 Differential expression was estimated with Cuffdiff 2 [51]. We excluded the genes with low
730 expression (FPKM < 0.5) from downstream analysis.

731 Pathway analysis of significantly differential expressed genes was conducted with
732 DAVID [52] using all the expressed genes as background.

733 **Allele-specific gene expression analysis.** Samtools mpileup (v0.1.19) and BCFtools
734 (v0.1.19) [53] were used to count the number of reads mapped from bam files of RNA-Seq data
735 to each allele of the heterozygous single nucleotide variants (SNVs). Binomial tests were
736 performed to determine if the percentage of RNA-Seq reads mapped to the alternative allele

737 was significantly different from the mean frequency of the alternative allele of all
738 heterozygous SNVs within each individual. Only heterozygous SNVs with read-coverage >
739 10 were included in the analysis.

740 **ChIP-Seq data analysis.** All ChIP-Seq data were generated using Illumina paired-end
741 sequencing with read length 101 bp. Reads were aligned to hg19 with BWA-MEM using
742 default parameters [46]. Reads with low mapping quality (< 30) were removed. PCR duplicate
743 reads were removed using Picard (http://broadinstitute.github.io/picard). As a quality
744 control, we calculated the normalized strand cross-correlation coefficient (NSC) and relative
745 strand correlation (RSC) [54] to assess the signal-to-noise ratios. All the data showed higher
746 NSC than RSC (**Supplementary Fig. 14**). Replicates for the same cell line on average
747 showed higher correlation than datasets from different cell lines (**Supplementary Fig. 15**).

748 For CTCF and H3K27ac, we used MACS2 [55] to call narrow peaks with default
749 parameters. For H3K27me3, we used the broad peak calling in MACS2. For all peak calling,
750 we used the corresponding whole-cell extract input library as background. For differential
751 bound analysis, we used the R package DiffBind [56] with the effective library size for read
752 counts normalization. Then DBA_DESEQ2 method was employed to conduct the differential
753 bound analysis. Signal artifact blacklist regions were excluded from our analysis
754 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability).

755 **Enrichment analysis.** For the enrichment analysis of differentially expressed genes we
756 divided each chromosome into 500 kbp bins. Within each bin, we calculated the total
757 number of expressed genes and the number of genes with significantly differential
758 expression between 22q11del and control cell lines. Then we conducted Fisher's exact test
759 to identify bins enriched with significantly differentially expressed genes against the
760 background of the whole genome.

761 For the analysis of differentially enriched sites for CTCF, H3K27ac and H3K27me3, we
762 also used 500 kbp bins. Log2 transformed fold changes of normalized read numbers in
763 binding sites between 22q11del and control cell lines were further transformed to Z-scores.
764 We considered binding sites with Z-score >2 or < -2 as significantly bound sites. Then
765 within each 500 kbp bin, we calculated the total number of binding sites and the number of
766 significantly differentially enriched sites between 22q11del and control cell lines. Then we
767 conducted Fisher's exact test to identify bins with significantly differentially enriched sites
768 against the background of the whole genome.

769 **Correlation analysis of gene expression and histone modification.** To estimate the
770 correlation between gene expression and histone modification, we assigned each binding
771 site of H3K27ac and H3K27me3 to its nearest ENSEMBL TSS using the R package
772 ChIPpeakAnno [57]. If a TSS was associated with multiple binding sites, only the nearest
773 binding site was retained. We then used this assignment for downstream correlation
774 analysis. To determine the cutoff for the distance in which binding sites are associated with
775 TSSs, we plotted the distribution of distances between binding sites and their assigned TSSs
776 (**Supplementary Fig. 16**). Based on the distribution, we set the cutoff to distance to TSS to
777 ± 1 kbp for H3K27ac and to ± 5 kbp for H3K27me3.

778 To investigate the effects of histone modifications on gene expression, we divided the
779 genes into two categories based on the differential expression analysis: differentially
780 expressed genes (FDR < 0.05) (category A) and not differentially expressed genes (FDR >
781 0.05) (category B). Within each category, we further categorized the genes into two groups:
782 up-regulated expressed genes (A1, B1) and down-regulated expressed genes (A2, B2) in
783 22q11del cell lines relative to control cell lines. Then for the genes within each of the four
784 groups (A1, B1, A2, B2), we calculated the Z-score transformed fold changes of the

785 normalized read counts in the TSS binding sites of histone marks between 22q11del and
786 control cell lines. To obtain the statistical significance of the fold change differences
787 between A1 and B1 and between A2 and B2, we performed permutation tests with 9,999
788 permutations.

789 We also carried out the reverse analysis. TSS binding sites of histone marks were
790 partitioned into two categories based on the differential bound analysis: differentially
791 bound sites (|Z-score| > 1 for H3K27ac, |Z-score| > 2 for H3K27me3) (category A) and non-
792 differentially bound sites (|Z-score| < 1 for H3K27ac, |Z-score| < 2 for H3K27me3) (category
793 B). Within each category, we further categorized the binding sites into two groups: up-
794 regulated bound sites (A1, B1) and down-regulated bound sites (A2, B2) in 22q11del cell
795 lines. Then we calculated the fold changes of the genes' FPKM between 22q11del and
796 control cell lines within each group. Permutation tests were performed with 9,999
797 permutations to obtain statistical significance.

798 To estimate the correlation between gene expression and histone modification in a
799 direct way, we calculated Pearson's correlation coefficient between gene's FPKM and
800 normalized read counts in the corresponding TSS binding site for each gene across all of the
801 cell lines. To obtain statistical significance, we first permuted genes' FPKM across the cell
802 lines for each TSS 10 times to assess the background correlation levels, and then performed
803 the Wilcoxon rank sum test between the observed correlation coefficients and the
804 background correlation coefficients. We also performed the same analysis using the
805 differentially expressed genes only and differentially enriched binding sites only.

806 **Correlation analysis for epigenetic marks.** To assess the correlation of epigenetic marks
807 binding between different regions on the same chromosomes, we divided the chromosomes
808 into 500 kbp bins. Within each bin, we calculated the mean value of the normalized read

809 counts for all the binding sites of each epigenetic mark. Then we calculated the Pearson's
810 correlation coefficients of obtained mean values across the cell lines between any two bins
811 on the same chromosomes.

812 **Data availability**. Hi-C, ChIP-Seq and RNA-Seq data from this study have been submitted to
813 the NCBI Gene Expression Omnibus (GEO; http://www.ncbi. nlm.nih.gov/geo/) under
814 accession number GSE76922.

815 **Competing interests**

816 The authors declare that no competing interests exist.

817 **Authors' contributions (in order of position in Authors List)**

818 XZ contributed to study design, carried out data analysis and contributed to writing the
819 paper, YZ contributed to study design and produced experimental data, XZ contributed to
820 study design and carried out data analysis, CP produced experimental data and contributed
821 to data interpretation, MSH produced experimental data, TW produced experimental data,
822 JY produced experimental data, SMW and AEU conceived of and designed the study,
823 directed data production and data analysis and wrote the paper.

824 **Acknowledgments**

825 This research was supported by external funding from the National Institute of Mental
826 Health (NIMH) (grant number: MH100010), NARSAD and the March of Dimes, as well as
827 funds from Stanford Pediatric Research Fund (Grant number: UL1 RR025744) and
828 departmental funds from the Stanford Department of Psychiatry and Behavioral Sciences.

829 We thank Dr. Eitan Yaffe (Stanford University) for his advice on the use of hicpipe.

830 **References**

831 1 Kirov, G. CNVs in neuropsychiatric disorders. *Hum Mol Genet* **24**, R45-49,
832 doi:10.1093/hmg/ddv253 (2015).
833 2 Torres, F., Barbosa, M. & Maciel, P. Recurrent copy number variations as risk
834 factors for neurodevelopmental disorders: critical overview and analysis of
835 clinical implications. *J Med Genet* **53**, 73-90, doi:10.1136/jmedgenet-2015-
836 103366 (2016).
837 3 Rees, E. *et al.* Evidence that duplications of 22q11.2 protect against
838 schizophrenia. *Mol Psychiatry* **19**, 37-40, doi:10.1038/mp.2013.156 (2014).
839 4 Loviglio, M. N. *et al.* Chromosomal contacts connect loci associated with
840 autism, BMI and head circumference phenotypes. *Mol Psychiatry*,
841 doi:10.1038/mp.2016.84 (2016).
842 5 Le Pichon, J. B., Yu, S., Kibiryeva, N., Graf, W. D. & Bittel, D. C. Genome-wide
843 gene expression in a patient with 15q13.3 homozygous microdeletion
844 syndrome. *Eur J Hum Genet* **21**, 1093-1099, doi:10.1038/ejhg.2013.1 (2013).
845 6 Blumenthal, I. *et al.* Transcriptional consequences of 16p11.2 deletion and
846 duplication in mouse cortex and multiplex autism families. *Am J Hum Genet*
847 **94**, 870-883, doi:10.1016/j.ajhg.2014.05.004 (2014).
848 7 Harewood, L., Chaignat, E. & Reymond, A. Structural variation and its effect
849 on expression. *Methods Mol Biol* **838**, 173-186, doi:10.1007/978-1-61779-
850 507-7_8 (2012).

851 8 Gavin, D. P. & Floreani, C. Epigenetics of schizophrenia: an open and shut
852 case. *Int Rev Neurobiol* **115**, 155-201, doi:10.1016/B978-0-12-801311-
853 3.00005-6 (2014).
854 9 Fullard, J. F. *et al.* Understanding the genetic liability to schizophrenia
855 through the neuroepigenome. *Schizophr Res*,
856 doi:10.1016/j.schres.2016.01.039 (2016).
857 10 Nestler, E. J., Peña, C. J., Kundakovic, M., Mitchell, A. & Akbarian, S. Epigenetic
858 Basis of Mental Illness. *Neuroscientist*, doi:10.1177/1073858415608147
859 (2015).
860 11 Mitchell, A. C. *et al.* The genome in three dimensions: a new frontier in human
861 brain research. *Biol Psychiatry* **75**, 961-969,
862 doi:10.1016/j.biopsych.2013.07.015 (2014).
863 12 Barnard, R. A., Pomaville, M. B. & O'Roak, B. J. Mutations and Modeling of the
864 Chromatin Remodeler CHD8 Define an Emerging Autism Etiology. *Front*
865 *Neurosci* **9**, 477, doi:10.3389/fnins.2015.00477 (2015).
866 13 Vallianatos, C. N. & Iwase, S. Disrupted intricacy of histone H3K4 methylation
867 in neurodevelopmental disorders. *Epigenomics* **7**, 503-519,
868 doi:10.2217/epi.15.1 (2015).
869 14 Zhubi, A., Cook, E. H., Guidotti, A. & Grayson, D. R. Epigenetic mechanisms in
870 autism spectrum disorder. *Int Rev Neurobiol* **115**, 203-244,
871 doi:10.1016/B978-0-12-801311-3.00006-8 (2014).
872 15 McDonald-McGinn, D. M. *et al.* 22q11.2 deletion syndrome. *Nat Rev Dis*
873 *Primers* **1**, 15071, doi:10.1038/nrdp.2015.71 (2015).

874 16 Biswas, A. B. & Furniss, F. Cognitive phenotype and psychiatric disorder in
875 22q11.2 deletion syndrome: A review. *Res Dev Disabil* **53-54**, 242-257,
876 doi:10.1016/j.ridd.2016.02.010 (2016).
877 17 Swillen, A. & McDonald-McGinn, D. Developmental trajectories in 22q11.2
878 deletion. *Am J Med Genet C Semin Med Genet* **169**, 172-181,
879 doi:10.1002/ajmg.c.31435 (2015).
880 18 Schneider, M. *et al.* Psychiatric disorders from childhood to adulthood in
881 22q11.2 deletion syndrome: results from the International Consortium on
882 Brain and Behavior in 22q11.2 Deletion Syndrome. *Am J Psychiatry* **171**, 627-
883 639, doi:10.1176/appi.ajp.2013.13070864 (2014).
884 19 Urban, A. E. *et al.* High-resolution mapping of DNA copy alterations in human
885 chromosome 22 using high-density tiling oligonucleotide arrays. *Proc Natl*
886 *Acad Sci U S A* **103**, 4534-4539, doi:10.1073/pnas.0511340103 (2006).
887 20 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range
888 interactions reveals folding principles of the human genome. *Science* **326**,
889 289-293, doi:10.1126/science.1181369 (2009).
890 21 Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical
891 methods. *Nat Biotechnol* **32**, 261-266, doi:10.1038/nbt.2833 (2014).
892 22 Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates
893 systematic biases to characterize global chromosomal architecture. *Nat Genet*
894 **43**, 1059-1065, doi:10.1038/ng.947 (2011).

895 23 Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of
896 chromosome organization. *Nat Methods* **9**, 999-1003,
897 doi:10.1038/nmeth.2148 (2012).
898 24 Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression.
899 *Bioinformatics* **28**, 3131-3133, doi:10.1093/bioinformatics/bts570 (2012).
900 25 Schmid, M. W., Grob, S. & Grossniklaus, U. HiCdat: a fast and easy-to-use Hi-C
901 data analysis tool. *BMC Bioinformatics* **16**, 277, doi:10.1186/s12859-015-
902 0678-x (2015).
903 26 Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data
904 processing. *Genome Biol* **16**, 259, doi:10.1186/s13059-015-0831-x (2015).
905 27 Heinz, S. *et al.* Simple combinations of lineage-determining transcription
906 factors prime cis-regulatory elements required for macrophage and B cell
907 identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
908 28 Li, W., Gong, K., Li, Q., Alber, F. & Zhou, X. J. Hi-Corrector: a fast, scalable and
909 memory-efficient package for normalizing large-scale Hi-C data.
910 *Bioinformatics* **31**, 960-962, doi:10.1093/bioinformatics/btu747 (2015).
911 29 Servant, N. *et al.* HiTC: exploration of high-throughput 'C' experiments.
912 *Bioinformatics* **28**, 2843-2844, doi:10.1093/bioinformatics/bts521 (2012).
913 30 Shavit, Y. & Lio', P. Combining a wavelet change point and the Bayes factor for
914 analysing chromosomal interaction data. *Mol Biosyst* **10**, 1576-1585,
915 doi:10.1039/c4mb00142g (2014).

916 31 Sauria, M. E., Phillips-Cremins, J. E., Corces, V. G. & Taylor, J. HiFive: a tool
917 suite for easy and efficient HiC and 5C data analysis. *Genome Biol* **16**, 237,
918 doi:10.1186/s13059-015-0806-y (2015).
919 32 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals
920 principles of chromatin looping. *Cell* **159**, 1665-1680,
921 doi:10.1016/j.cell.2014.11.021 (2014).
922 33 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by
923 analysis of chromatin interactions. *Nature* **485**, 376-380,
924 doi:10.1038/nature11082 (2012).
925 34 Maynard, T. M. *et al.* Mitochondrial localization and function of a subset of
926 22q11 deletion syndrome candidate genes. *Mol Cell Neurosci* **39**, 439-451,
927 doi:10.1016/j.mcn.2008.07.027 (2008).
928 35 Butcher, N. J. *et al.* Association between early-onset Parkinson disease and
929 22q11.2 deletion syndrome: identification of a novel genetic form of
930 Parkinson disease and its clinical implications. *JAMA Neurol* **70**, 1359-1366,
931 doi:10.1001/jamaneurol.2013.3646 (2013).
932 36 Boot, E. *et al.* Movement disorders and other motor abnormalities in adults
933 with 22q11.2 deletion syndrome. *Am J Med Genet A* **167A**, 639-645,
934 doi:10.1002/ajmg.a.36928 (2015).
935 37 Mok, K. Y. *et al.* Deletions at 22q11.2 in idiopathic Parkinson's disease: a
936 combined analysis of genome-wide association data. *Lancet Neurol* **15**, 585-
937 596, doi:10.1016/S1474-4422(16)00071-5 (2016).

938 38 Stefansson, H. *et al.* Large recurrent microdeletions associated with
939 schizophrenia. *Nature* **455**, 232-236, doi:10.1038/nature07229 (2008).
940 39 Consortium, I. S. Rare chromosomal deletions and duplications increase risk
941 of schizophrenia. *Nature* **455**, 237-241, doi:10.1038/nature07239 (2008).
942 40 Levinson, D. F. *et al.* Copy number variants in schizophrenia: confirmation of
943 five previous findings and new evidence for 3q29 microdeletions and VIPR2
944 duplications. *Am J Psychiatry* **168**, 302-316,
945 doi:10.1176/appi.ajp.2010.10060876 (2011).
946 41 Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause
947 pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025,
948 doi:10.1016/j.cell.2015.04.004 (2015).
949 42 Taberlay, P. C. *et al.* Three-dimensional disorganization of the cancer genome
950 occurs coincident with long-range genetic and epigenetic alterations. *Genome*
951 *Res* **26**, 719-731, doi:10.1101/gr.201517.115 (2016).
952 43 Kasowski, M. *et al.* Extensive variation in chromatin states across humans.
953 *Science* **342**, 750-752, doi:10.1126/science.1242510 (2013).
954 44 Raha, D., Hong, M. & Snyder, M. ChIP-Seq: a method for global identification
955 of regulatory elements in the genome. *Curr Protoc Mol Biol* **Chapter 21**, Unit
956 21.19.21-14, doi:10.1002/0471142727.mb2119s91 (2010).
957 45 Abyzov, A. *et al.* Somatic copy number mosaicism in human skin revealed by
958 induced pluripotent stem cells. *Nature* **492**, 438-442,
959 doi:10.1038/nature11629 (2012).

960 46 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-
961 Wheeler transform. *Bioinformatics* **25**, 1754-1760,
962 doi:10.1093/bioinformatics/btp324 (2009).
963 47 *Martin, M.* Cutadapt Removes Adapter Sequences From High-Throughput
964 Sequencing Reads. *EMBnet.journal* **17**, 10-12, doi:10.14806/ej.17.1.200
965 (2011).
966 48 Yao, J., Fetter, R. D., Hu, P., Betzig, E. & Tjian, R. Subnuclear segregation of
967 genes and core promoter factors in myogenesis. *Genes Dev* **25**, 569-580,
968 doi:10.1101/gad.2021411 (2011).
969 49 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence
970 of insertions, deletions and gene fusions. *Genome Biol* **14**, R36,
971 doi:10.1186/gb-2013-14-4-r36 (2013).
972 50 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals
973 unannotated transcripts and isoform switching during cell differentiation.
974 *Nat Biotechnol* **28**, 511-515, doi:10.1038/nbt.1621 (2010).
975 51 Trapnell, C. *et al.* Differential analysis of gene regulation at transcript
976 resolution with RNA-seq. *Nat Biotechnol* **31**, 46-53, doi:10.1038/nbt.2450
977 (2013).
978 52 Huang, d. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment
979 tools: paths toward the comprehensive functional analysis of large gene lists.
980 *Nucleic Acids Res* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).
981 53 Li, H. A statistical framework for SNP calling, mutation discovery, association
982 mapping and population genetical parameter estimation from sequencing

983 data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509
984 (2011).
985 54 Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and
986 modENCODE consortia. *Genome Res* **22**, 1813-1831,
987 doi:10.1101/gr.136184.111 (2012).
988 55 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**,
989 R137, doi:10.1186/gb-2008-9-9-r137 (2008).
990 56 Stark, R. & Brown, G. Accessed 15 Aug 2015
991 (http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/d
992 oc/DiffBind.pdf, 2011).
993 57 Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq
994 and ChIP-chip data. *BMC Bioinformatics* **11**, 237, doi:10.1186/1471-2105-11-
995 237 (2010).
996