1	Investigation of a Pathogenic Inversion in UNC13D and Comprehensive Analysis
2	of Chromosomal Inversions Across Diverse Datasets
3	Tugce Bozkurt-Yozgatli ^{1,2} , Ming Yin Lun ³ , Jesse D. Bengtsson ³ , Ugur Sezerman ^{1,4} , Ivan K. Chinn ^{5,6} ,
4	Zeynep Coban-Akdemir ^{2*} , Claudia M.B. Carvalho ^{3*}
5	¹ Department of Biostatistics and Bioinformatics, Institute of Health Sciences, Acibadem Mehmet
6	Ali Aydinlar University, Istanbul, Turkey
7	² Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental
8	Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston,
9	TX, USA
10	³ Pacific Northwest Research Institute, Seattle, WA 98122, USA.
11	⁴ Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali
12	Aydinlar University, Istanbul, Turkey.
13	⁵ Department of Pediatrics, Division of Immunology, Allergy, and Retrovirology, Baylor College of
14	Medicine and Texas Children's Hospital, Houston, TX, USA.
15	⁶ Center for Human Immunobiology of Texas Children's Hospital/Department of Pediatrics, Baylor
16	College of Medicine, Houston, TX, USA.
17	*Co-Corresponding authors:
18	Zeynep Coban-Akdemir, PhD, Assistant Professor, UTHealth School of Public Health, 1200
19	Pressler Street, Houston, TX 77030-3900, Email: <u>Zeynep.H.CobanAkdemir@uth.tmc.edu</u> , Office:
20	+1 512 391 2536
21	Claudia M.B. Carvalho, PhD, Assistant Investigator, Pacific Northwest Research Institute, 720
22	Broadway, Seattle, WA 98122, Email: <u>ccarvalho@pnri.org</u> , Office: +1 206 338 0694

23 ABSTRACT

24 Inversions are known contributors to the pathogenesis of genetic diseases. Identifying inversions 25 poses significant challenges, making it one of the most demanding structural variants (SVs) to 26 detect and interpret. Recent advancements in sequencing technologies and the development of 27 publicly available SV datasets have substantially enhanced our capability to explore inversions. 28 However, a cross-comparison in those datasets remains unexplored. In this study, we reported a 29 proband with familial hemophagocytic lymphohistiocytosis type-3 carrying c.1389+1G>A in trans 30 with NC 000017.11:75576992 75829587inv disrupting UNC13D, an inversion present in 31 0.006345% of individuals in gnomAD(v4.0). Based on this result, we investigate the features of 32 potentially pathogenic inversions in public datasets. 98.9% of inversions are rare in gnomAD, and 33 they disrupt 5% of protein-coding genes associated with a phenotype in OMIM. We then 34 conducted a comparative analysis of the datasets, including gnomAD, DGV, and 1KGP, and two 35 recent studies from the Human Genome Structural Variation Consortium revealed common and 36 dataset-specific inversion characteristics suggesting methodology detection biases. Next, we 37 investigated the genetic features of inversions disrupting the protein-coding genes by classifying 38 the intersections between them into three categories. We found that most of the protein-coding 39 genes in OMIM disrupted by inversions are associated with autosomal recessive phenotypes 40 regardless of categories supporting the hypothesis that inversions in trans with other variants are 41 hidden causes of monogenic diseases. This effort aims to fill the gap in our understanding of the 42 molecular characteristics of inversions with low frequency in the population and highlight the 43 importance of identifying them in rare disease studies.

Keywords: Structural variants, genomic disorders, Mendelian diseases, genome sequencing,
 optical genome mapping.

46 **INTRODUCTION**

47 Inversions are defined as a type of structural variant (SV) that refers to orientation 48 changes in DNA segments. They can be copy-number neutral (classical/simple/balanced) with 49 two breakpoint junctions or be part of complex genomic rearrangements (CGRs) with other copy-50 number variations (CNVs) [1]. The main mechanism for the formation of classical inversions has 51 been proposed to be non-allelic homologous recombination (NAHR) between inverted repeats 52 [2-4]. Other biological mechanisms may result in inversion formation, including DNA repair-53 associated events (non-homologous end joining (NHEJ), and microhomology-mediated end 54 joining (MMEJ)) and DNA replication-associated events (e.g., fork stalling and template switching) 55 [1,5].

56 Inversions may have an impact on disease phenotypes, often by directly disrupting a 57 particular gene [6]. They may occur within a gene, leading to the disease manifestation by causing 58 the skipping of exonic regions [7]. Mor-Shaked et al. reported a pathogenic inversion in PRKN, 59 leading to the skipping of exon 5 in individuals with early-onset Parkinson's disease (PARK2, 60 OMIM #600116) [7]. Alternatively, one of the inversion breakpoints can disrupt a gene and result 61 in a disease phenotype [8]. For instance, one of the breakpoints of a 253-kb inversion mapping 62 to intron 30 of UNC13D contributes to the manifestation of familial hemophagocytic 63 lymphohistiocytosis 3 (FHL3, OMIM #60898) [9,10]. In addition to Mendelian disorders, 64 inversions are also recognized as significant contributors to common complex disease traits [11– 65 13] and disease prognosis [14]. Additionally, they can also play a role as genetic modifiers

66 affecting disease phenotypes [15]. Moreover, some inversions have no direct effect on disease 67 phenotype by themselves, but they may predispose the loci to further genomic rearrangements 68 with pathogenic consequences [2,16] including the formation of recombinant chromosomes [1]. 69 Inversion detection is challenging due to their balanced nature and the fact that 70 breakpoints often map to repeats. Those features make them undetectable by comparative 71 genomic hybridization (aCGH) and exome sequencing (ES) [17]. Although short-read whole 72 genome sequencing (WGS) enables the detection of some inversions, it also introduces the issues 73 of false positives and the inability to sequence breakpoint junctions in the repetitive parts of the 74 genome [18,19]. Long-read WGS technologies, including Pacific Biosciences (PacBio) and Oxford 75 Nanopore (ONT), single-cell template strand sequencing (Strand-seq) [20], and optical genome 76 mapping [21] have improved our ability to detect inversions since these methodologies are more 77 suitable to detect changes in DNA orientation including within complex repeat regions [4,22].

78 Published population datasets using different sequencing technologies like those in Ebert 79 et al. [22], and Porubsky et al. [4], and publicly available databases such as Genome Aggregation 80 Database (gnomAD) [23], The Database of Genomic Variants (DGV) [24], and 1000 Genomes 81 Project (1KGP) [25] provide valuable resources for SV analysis. The recent release of gnomAD 82 dataset version 4 (v4.0) includes short-read genome sequencing data from 63,046 unrelated 83 human samples across the world [23]. The DGV dataset is derived from different methodologies 84 such as sequencing, aCGH, and Fluorescence in situ hybridization (FISH) [24]. Byrska-Bishop et al. 85 released expanded short-read WGS of 1KGP consisting of 3,202 samples, including 602 trios 86 across diverse global populations [25]. Porubsky et al. [4] reported inversions from 41 human 87 samples by integrating Strand-seq [20], haplotype-resolved de novo sequence assemblies

generated from PacBio long-reads, and Bionano genomics single-molecule optical mapping [21].
Ebert *et al.* published 64 assembled haplotypes from 32 diverse human genomes using long-read
WGS and strand-seq [22].

Here, we report a proband carrying a pathogenic inversion *in trans* with a singlenucleotide variant (SNV) affecting *UNC13D*. Then, we comprehensively compare inversions disrupting genes reported in various datasets, gnomAD (v4.0) [23], DGV (release date: 2020-02-25) [24], 1KGP (release date: 2021-10-05) [25], inversions released by Ebert *et al.* [22] and Porubsky *et al.* [4] (Figure 1). Our goal is to provide insights into the features of inversions present in population datasets to genomic disorders.

97 **METHODS**

98 Case presentation

99 The proband (SEA110) is a Caucasian white non-Hispanic, non-Latino male. He was 100 diagnosed with VACTERL (vertebral defects, anal atresia, cardiac defects, tracheoesophageal 101 fistula, renal anomalies, and limb abnormalities) after birth. He did not meet early developmental 102 milestones on time. The patient had frequent respiratory infections that required supplemental 103 oxygen, including respiratory syncytial virus infection. He presented with pancytopenia at the 104 range of 11-15 months of life, which was initially felt to be likely viral-mediated. He was 105 hospitalized and discharged. He seemed well but then developed daily fevers and increased stool 106 output. He was re-hospitalized and found to have hepatosplenomegaly by abdominal ultrasound. 107 He then developed acute respiratory failure and required intubation with pressor support. 108 Laboratory testing ultimately confirmed a diagnosis of hemophagocytic lymphohistiocytosis 109 (HLH) by HLH-2004 criteria [26]: fever, splenomegaly, anemia with thrombocytopenia,

hypofibrinogenemia, hypertriglyceridemia, hyperferritinemia, elevated soluble interleukin-2 receptor levels, and impaired CD107A mobilization. Further clinical information can be provided upon request from the corresponding author. Initial genetic testing consisted of proband ES and chromosomal microarray testing, both of which were performed by a commercial clinical laboratory. Results were reported as negative for both tests. Upon re-hospitalization, clinical targeted gene panel testing was ordered for inborn errors of immunity and cytopenias, which identified a pathogenic variant at *UNC13D* c.1389+1G>A.

117 Patient sample collection

As a result of clinical targeted gene panel findings, SEA110 was tested by the Baylor Genetics Clinical Diagnostic Laboratory using rapid short-read WGS. Informed consent was obtained for research participation under Pacific Northwest Research Institute approved WCG IRB Protocol #H-47127 20202158.

122 **DNA Extraction**

DNA was extracted from whole blood using the QIAGEN Puregen DNAeasy kit following the manufacturer's direction with modification of the centrifugation steps, which were extended to 10 minutes. Ultrahigh molecular weight DNA was extracted from whole blood with the Bionano SP-G2 Blood and CellCulture DNA Isolation Kit (#80060) following the manufacturer's direction.

128 ONT-library preparation and sequencing run

DNA from SEA110 was sheared to an N50 of approximately 10 kb using a Covaris g-TUBE and an Eppendorf 5424 rotor at 5000 rpm. End repair and ligation of adapters for Oxford nanopore sequencing followed the manufacturer's direction for kit LSK114. Sequencing used

132 Minknow version 23.07.12, with adaptive sampling to enrich for the region of interest. The 133 enrichment region (chr17:75526717-75896404, GRCh38) and reference as a minimap2 index file 134 were provided [27]. Following sequencing, passed reads were re-called using guppy 6.0.1 and the 135 super high accuracy model. Passed reads were mapped to GRCh38 using minimap2 (-Y --136 secondary=no map-ont). After SN -a mapping, -X 137 Vs were called using Clair3 [28] and reads were haplotagged by Whatshap [29].

138 Breakpoint junction amplification and Sanger sequencing

Inversion junctions were amplified using primers reported previously with one additional
sequencing primer (Supplementary Table 1) [9]. Amplification used the Q5 Polymerase (NEB),
and PCR products were gel extracted with the Monarch DNA Gel Extraction kit (NEB) following
the manufacturer's direction. Purified products were sent for Sanger sequencing by GENEWIZ.
Sanger sequencing was analyzed using Geneious Prime software (Dotmatics).

144 **Optical Genome Mapping**

145 Ultrahigh molecular weight DNA (UHMW) was labeled with the Bionano Direct Label and 146 Stain-G2 (DLS2-G2) Kit (#80046) following the manufacturer's direction. In brief, 750 ng of UHMW 147 DNA was labeled with proprietary green fluorophore (DL-Green), and after purification, the DNA 148 back bone was stained with a proprietary DNA stain. After staining, the sample was run on a 149 Bionano Saphyr instrument. A de novo assembly was generated in Bionano access version 1.8.1, 150 with a molecule N50 of 150.38 kb in length and 15.61 labels per 100 kb. The resulting assembly 151 was compared to the hg38 reference genome, variants were called using Bionano solve version 152 1.8.1.

153 Datasets utilized in this study

154	We analyzed the inversions mapped to the reference human genome of hg38 from three							
155	publicly accessible databases, gnomAD (v4.0) [23], DGV (release date: 2020-02-25) [24] and 1KGP							
156	(release date: 2021-10-05) [25], and two recent studies of Ebert <i>et al.</i> [22] and Porubsky <i>et al</i> . [4]							
157	(Figure 1). We extracted inversion calls in autosome (chr1-22) and sex (chrX and chrY)							
158	chromosomes from the datasets. The gnomAD (v4.0) [23] SV dataset was downloaded from							
159	https://gnomad.broadinstitute.org/downloads. The DGV [24] SV dataset was downloaded from							
160	the link: http://dgv.tcag.ca/dgv/docs/GRCh38 hg38 variants 2020-02-25.txt. DGV [24] includes							
161	inversions from several studies (Supplementary table 2) derived from different methodologies,							
162	including sequencing, oligo aCGH, and FISH. We included inversions detected by all of these							
163	studies. SV data in the 1KGP was downloaded from the following link:							
164	https://www.internationalgenome.org/data-portal/data-collection/30x-grch38. The updated							
165	callset to the original release of the inversions reported by Ebert et al. [22] was downloaded from							
166	the following link:							
167	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated							
168	callset/. Lastly, we included the inversions reported by Porubsky et al. [4].							

169 Gene Annotations

We downloaded the gene regions with their canonical transcripts present in the hg38 version of the GENCODE (v46) database (Data update date: 2024-04-02) through the University of California Santa Cruz (UCSC) [30] to identify the inversions intersecting with the human protein-coding genes. We filtered the dataset to extract only the genes with protein-coding transcripts, excluding those with other transcript types. (Supplementary figure 1). Then, we retained the genes in human autosome chromosomes (chr1-22) and sex chromosomes (chrX and

176 chrY). We also downloaded the dataset of the Online Mendelian Inheritance in Man (OMIM)

- 177 (data freeze date: 06-18-2024) [31] (<u>https://www.omim.org/downloads</u>) as well as rare disease-
- 178 related genes in Orphanet data (<u>https://www.orphadata.com/genes/</u>).

179 Analysis of inversions intersecting inversions in other datasets and protein-coding genes

180 We used the Bedtools (v2.30.0) [32] intersect function with the fraction option 0.5 to 181 detect the overlap between inversion locations in different datasets. Bedtools intersect function 182 takes a genomic feature as the first input and finds overlapped regions between another genomic 183 feature as the second input. The fraction option 0.5 allows us to find the overlap, including at 184 least 50% of the sequence length of inversions. We also implemented the Bedtools (v2.30.0) [32] 185 intersect function with the default parameters to detect the overlap between inversions and 186 protein-coding genes. The intersections between inversions and human protein-coding genes 187 were classified into three distinct categories. In category 1, inversions cover genes; in category 2, 188 one of the inversion breakpoints maps within a gene; in category 3, inversions map entirely within 189 genes (Figure 1).

190 Enrichment analysis of the genes intersecting inversions

We performed gene set enrichment analysis with the protein-coding genes overlapping with inversions in categories 2 and 3 by applying Enrichr [33]. The list of the genes intersecting inversions in each intersection category was given as input to Enrichr [33]. Then, we reported the Human Phenotype Ontology (HPO) terms enriched by these genes.

195 **Computational Analysis**

Computational analyses were carried out using R (v.4.2.0) [34]. The plots were generated
using the package ggplot2 [35] and the UpSet R package [36].

198 **RESULTS**

199 A pathogenic UNC13D inversion is present in gnomAD

200 We identified an inversion accompanied by the canonical donor splice site SNV in UNC13D 201 in SEA110 (Figure 2 and Supplementary figure 2). The 253-kb inversion has been documented in 202 individuals with Swedish ancestry and reported to cause FHL3 when inherited as homozygous or 203 in trans with pathogenic SNVs and small indels in UNC13D [9,10]. We observed an almost 204 identical inversion reported in gnomAD at coordinates chr17:75576924-75829482 205 (INV CHR17 66182818), which is present in 0.006345%, exclusively in heterozygous state in 206 individuals from European Finnish and Admixed American populations (Supplementary Figure 3). 207 The SEA110 inversion shows two breakpoint junctions with 111 (junction 1) and 23 (junction 2) 208 nucleotides similarity generated by Alu-Alu mediated rearrangement (AAMR) (Figure 2). Parental 209 samples are not available to test for inheritance; therefore, we do not have information about 210 ancestry and cannot investigate whether this inversion is the same reported in gnomAD (a 211 potential founder event) or if it is a recurrent inversion generated independently via AAMR in 212 this proband. Optical Genome Mapping supports the breakpoint junctions of the inversion 213 obtained by Sanger sequencing. The detected inversion has multiple molecules spanning both 214 breakpoints and several molecules spanning the entire inversion supporting the inversion call. 215 Bionano solve software called the inversion as heterozygous, but lack of label density in UNC13D 216 results in the exclusion of UNC13D from the called inversion. ONT sequencing was applied to 217 confirm heterozygosity, and manual phasing indicated the pathogenic SNV and inversions are in 218 trans (Figure 2C).

219 Inversions in gnomAD (v4.0) are rare and affect protein-coding genes

We hypothesized that pathogenic inversions are present as rare alleles in the general population. To investigate this concept, we categorized 2185 inversions in gnomAD into two groups: Rare (allele frequency <5%) and common (allele frequency ≥5%). Altogether, 2,161 (98.9%) inversions are rare; 24 inversions (1.1%) are common in gnomAD (Supplementary figure 4A).

225 We investigated the human protein-coding genes affected by rare and common 226 inversions in gnomAD. We analyzed 19,697 protein-coding genes in GENCODE (v46); 4,921 are 227 related to a phenotype in OMIM, 11,306 are not yet linked with a phenotype in OMIM, and 3,470 228 genes are not cataloged in OMIM. We overlapped inversions in gnomAD and protein-coding 229 genes and categorized the intersections into three groups (Category 1, category 2, and category 230 3). Next, we focused on the inversions in categories 2 and 3 since they can be critical mechanisms 231 for disease pathology (Supplementary table 3). 279 rare gnomAD inversions affect 5% of genes 232 associated with a phenotype in OMIM (247 out of 4,921; Supplementary figure 4C) in contrast 233 with 4.6% of genes not associated with a phenotype in OMIM (521 out of 11,306; Supplementary 234 Figure 4C) based on categories 2 and 3. Furthermore, 254 out of 279 rare gnomAD inversions 235 have not been found in the homozygous state and affect 106 autosomal recessive (AR) disease 236 genes (Supplementary table 4).

237 Features of the inversions reported in distinct datasets

To compare the characteristics of inversions in gnomAD [23] with other publicly available datasets, we conducted a comparative analysis using inversion data from DGV [24], 1KGP [25], and two recent publications of Ebert *et al.* [22] and Porubsky *et al.* [4] (Figure 1).

241 We extracted 2,185 inversions from gnomAD, 3,468 inversions from DGV, 920 inversions 242 from 1KGP, 414 inversions from the data released by Ebert et al., and 339 inversions from the 243 callset published by Porubsky et al. The summary statistics of inversion length in each dataset are provided in Table 1. gnomAD shows a more even distribution regarding size and displays the 244 245 largest events (Supplementary figure 5), including a 118.67 Mb pericentric inversion 246 (INV CHR5 77480914). Most of DGV inversions (75%) are between 0.035 kb and 24.22 kb. 1KGP 247 inversions tend to be smaller as the median length of 0.831 kb, whereas Ebert *et al.* and Porubsky 248 et al. show the highest median length of 293.19 kb and 251.71 kb, respectively.

249 Estimating redundancy among the inversions available from different datasets

250 We investigated the number of common and dataset-specific inversions across different 251 datasets using very stringent criteria based on the start and end locations of the inversions 252 (Supplementary figure 6). Redundancies in the datasets are expected due to the overlap of 253 samples reported in distinct publications (e.g., Ebert et al. and Porubsky et al.) or inclusion of 254 datasets into publicly shared ones (e.g., gnomAD v.2 is included in DGV). We observed very little 255 redundancy for inversions among the individual datasets (Supplementary figure 6) because the 256 different applied sequencing technologies provided distinct resolutions concerning breakpoint 257 junctions. We then decreased the stringency to intersect inversions in each dataset with at least 258 50% of their sequence (Supplementary figure 7). The inversions in gnomAD and DGV share (49.4% 259 and 77.2%) more inversions with each other compared to other datasets. 78.3% of 1KGP 260 inversions overlap with at least one inversion in gnomAD. Around 70% of inversions in Ebert et 261 al. and Porubsky et al. overlap with each other.

262 Inversions disrupting genes

We overlapped the inversions in the datasets with the protein-coding genes. Then, we classified the overlaps between inversions and protein-coding genes into three categories, as defined previously defined in this manuscript (Figure 1). The majority of the overlaps from all datasets, except 1KGP, map with category 1 (76.8% in DGV to 97.1% in gnomAD). 65.9% of inversion-gene intersections belong to category 3 in 1KGP (Figure 3).

268 Next, we focused on the inversions in categories 2 and 3 since they can be critical 269 mechanisms for disease pathology (Supplementary table 5 and 6). We delved deep into the 270 protein-coding genes associated with clinical phenotypes in OMIM disrupted by inversions in this 271 intersection between categories 2 and 3. In total, 847 inversions have one breakpoint junction 272 mapping to 830 protein-coding genes based on category 2 and can be potentially relevant to 273 genetic disorders (Supplementary table 5). On the other hand, in total, breakpoint junctions of 274 1,586 inversions are within 1030 protein-coding genes based on category 3 and can also be 275 potentially relevant to genetic disorders (Supplementary table 6). Interestingly, both DGV and 276 gnomAD inversions show higher frequencies of disrupting genes associated with disease 277 compared to other datasets (1.6% and 2.1%, respectively) in Category 2 (Supplementary figure 278 8A). Importantly, inversions in both datasets also disrupt a higher proportion of OMIM 279 phenotype-related genes in Category 3 (3.7% and 3.2%, respectively), while Porubsky et al. has a 280 smaller proportion (0.2%, Supplementary figure 8B). The inheritance pattern of the genes 281 overlapping with inversions for categories 2 and 3 for each dataset is given in Supplementary 282 figure 8C and D. About 40.9% and 50% of the inversions in both categories 2 and 3 regardless of 283 dataset are in AR disease genes (Supplementary figure 8C and D). Autosomal dominant (AD)

inheritance is the second most prominent disease gene pattern (16.7% and 33.8%,
Supplementary figure 8C and D).

We then performed gene set enrichment analysis with all protein-coding genes in the datasets that intersect with the protein-coding genes following categories 2 and 3. All enriched HPO terms belonging to each category are provided in Supplementary table 7 and 8.

289 **DISCUSSION**

290 In this study, we reported a case with c.1389+1G>A and NC 000017.11: 291 75576992 75829587 inv in UNC13D presenting with an FHL3 phenotype. The inversion in the 292 patient disrupts UNC13D following category 2 (Figure 2). The pathogenic inversion is present in 293 heterozygosity in gnomAD (v4.0) in individuals from European Finnish and Admixed American 294 populations (Supplementary Figure 3). To identify more inversions that are likely pathogenic like 295 the one in UNC13D, we delved deep into gnomAD inversions. We extensively investigated 296 gnomAD inversions to gain a comprehensive understanding of inversions in an individual 297 genome. 279 rare inversions in gnomAD affect 247 protein-coding genes associated with a 298 phenotype in OMIM based on categories 2 and 3; 254 of them have not been found in the 299 homozygous state and overlap with 106 AR disease genes (Supplementary table 4), similar to the 300 overlap between INV CHR17 66182818 and UNC13D. For instance, 15,736-bp inversion in 301 gnomAD, INV chr1 04df2580,

302 (https://gnomad.broadinstitute.org/variant/INV_CHR1_04DF2580?dataset=gnomad_sv_r4)

disrupts *DPYD* with the breakpoint junctions in intron 12 and intron 8 of *DPYD*. Van Kuilenburg *et* al. has reported a 115,731-bp inversion with breakpoints in intron 8 and intron 12 of *DPYD* in a

305 patient with Dihydropyrimidine dehydrogenase deficiency (OMIM #274270) [37].

306 Then, we conducted analyses on inversions from diverse datasets. It is important to 307 highlight that these inversions were derived from different sequencing technologies (Table 1). 308 While the inversions in 1KGP and gnomAD were detected using short-read WGS, the inversions 309 reported by Ebert et al. and Porubsky et al. were identified by long-read WGS and Strand-seq. 310 Strand-seq was shown to be the ideal technology to detect inversions, especially those mediated 311 by large segmental duplications or other genomic repeats which often happen as a result of 312 NAHR; 72% of balanced inversions in Porubsky et al. are generated by NAHR [4,22]. In contrast, 313 short-reads are not suitable to identify such inversions, although it can resolve inversions with 314 blunt or microhomology at the breakpoint junctions such as those generated by NHEJ [1]. 315 Therefore, while we expected to detect redundancy among datasets, we also expected to identify 316 unique inversions only identifiable by certain methodologies but invisible to others. While 317 between 11.1% to 49.4% of the inversions in gnomAD overlap with inversions in other datasets, 318 from 21.6% to 76.4% of inversions in Porubsky et al. overlap with inversions in other datasets. 319 Strikingly, gnomAD (v4.0) has inversions with a longer length and a higher number of larger 320 inversions (median length of 7.1 kb), which raises the question of whether Mb size inversions, 321 including pericentric ones, are more often generated by NHEJ (Table 1). In fact, we have 322 investigated large inversions detected by karyotyping (8 Mb to 178 Mb) in a diagnostic setting, 323 and found that none of the resolved inversions (13/18 or 72%) are mediated by repeats [1] which 324 has been confirmed by a second more recent study [38]. Besides, it should be taken into account 325 that these inversions were generated by different SV callers, and these tools exhibit different 326 false positive rates [18,39]. These potential false calls may overlap with protein-coding genes in 327 our analysis. Also, redundancies in these datasets will occur due to the same ancestral inversions

being reported from distinct individuals while identified by distinct technologies, due to analysis of similar samples or due to the incorporation of entire datasets into larger ones, *e.g.*, DGV incorporates 1KGP phase 3 (Supplementary table 2).

331 Next, we examined whether the inversions in all datasets disrupt human protein-coding 332 genes by classifying inversion-gene intersections into three different categories (Figure 1). The 333 majority of the overlaps in all datasets except 1KGP are from category 1 (Figure 3) which is 334 consistent with the small inversion sizes in 1KGP (Supplementary Figure 5, Table 1) but also 335 indicates that inversions in 1KGP often have both breakpoints within genes which potentially can 336 lead to truncated transcripts subjected to nonsense mediated decay (NMD) or to exon skipping. 337 The results also show that most of the inversions intersecting with protein-coding genes in the 338 other datasets are longer than the gene length. Besides, 97.1% of intersections in gnomAD belong 339 to category 1, consistent with gnomAD presenting longer inversions compared to other datasets 340 (Table 1). Next, we focused on the protein-coding genes that are associated with a phenotype in 341 OMIM disrupted by inversions. Upon examining the genes overlapping with the inversions in 342 categories 2 and 3, we found that most genes intersecting with inversions across all datasets 343 belong to the AR group, while AD disease genes are the second most prominent group. (Figure 344 4B and Figure 4C). Inversions that disrupt AD disease genes can also be particularly noteworthy, 345 as they might introduce genomic instability in these regions, potentially leading to the formation 346 of other SVs [16].

The number of inversions involving protein-coding genes associated with one or more phenotypes is markedly distinct in each dataset, with gnomAD and DGV showing a higher overlap rate with OMIM phenotype-related genes than other datasets (Supplementary Figure 8). We

observed that the genes disrupted by inversions in categories 2 and 3 are associated with both
 Mendelian disorders, such as Spinocerebellar ataxia 31 (OMIM #619422), and complex disease
 traits, such as susceptibility to autism (OMIM #618830).

We further performed gene set enrichment analysis on the genes interrupted by inversions in categories 2 and 3. All enriched HPO terms except Autosomal dominant inheritance (HP:0000006) for category 3 are statistically insignificant (Supplementary table 7 and 8). This result might be expected since we used diverse genes that overlap inversions in the whole genome. Nevertheless, we still report the list of HPO terms enriched by the genes disrupted by inversions to be able to gain an insight into these genes and their related phenotypes.

Finally, sequencing technologies, including short-read WGS, long-read WGS, Strand-seq, and optical mapping, have significantly contributed to the discovery of inversions. Publicly accessible datasets using these technologies are important resources that may facilitate discoveries of pathogenic inversions underlying various disease traits. This study sheds light on the possible impact of the inversions in these datasets on revealing disease phenotypes.

364 DATA AVAILABILITY

- 365 Gnomad SV data: https://gnomad.broadinstitute.org/downloads
- 366 DGV SV data: http://dgv.tcag.ca/dgv/docs/GRCh38 hg38 variants 2020-02-25.txt
- 367 1KGP SV data: https://www.internationalgenome.org/data-portal/data-collection/30x-grch38
- 368 The updated callset to the original release of the inversions reported by Ebert *et al.* [22]:
- 369 http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_
- 370 callset/
- 371 GENCODE v46: <u>https://genome.ucsc.edu/cgi-bin/hgTables</u>

- 372 OMIM gene list: <u>https://www.omim.org/downloads</u>
- 373 Orphanet gene list: <u>https://www.orphadata.com/genes/</u>

374 CODE AVAILABILITY

- 375 The script for data analysis in this manuscript is available at <u>https://github.com/Carvalho-</u>
- 376 Lab/Tugce INV/tree/main.

377 **REFERENCES**

- 1. Pettersson M, Grochowski CM, Wincent J, Eisfeldt J, Breman AM, Cheung SW, et al.
- 379 Cytogenetically visible inversions are formed by multiple molecular mechanisms. Human
- 380 Mutation. 2020;41:1979–98.
- 381 2. Flores M, Morales L, Gonzaga-Jauregui C, Domínguez-Vidaña R, Zepeda C, Yañez O, et al.
- 382 Recurrent DNA inversion rearrangements in the human genome. Proc Natl Acad Sci USA.
- 383 2007;104:6099–106.
- 384 3. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and
- 385 sequencing of structural variation from eight human genomes. Nature. 2008;453:56–64.
- 386 4. Porubsky D, Höps W, Ashraf H, Hsieh P, Rodriguez-Martin B, Yilmaz F, et al. Recurrent
- inversion polymorphisms in humans associate with genetic instability and genomic disorders.
- 388 Cell. 2022;185:1986-2005
- 5. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic
 disorders. Nat Rev Genet. 2016;17:224–38.

- 391 6. Puig M, Casillas S, Villatoro S, Cáceres M. Human inversions and their functional
- 392 consequences. Brief Funct Genomics. 2015;14:369–79.
- 393 7. Mor-Shaked H, Paz-Ebstein E, Basal A, Ben-Haim S, Grobe H, Heymann S, et al. Levodopa-
- 394 responsive dystonia caused by biallelic *PRKN* exon inversion invisible to exome sequencing.
- 395 Brain Communications. 2021;3:fcab197.
- 396 8. Jones ML, Murden SL, Brooks C, Maloney V, Manning RA, Gilmour KC, et al. Disruption of
- 397 AP3B1by a chromosome 5 inversion: a new disease mechanism in Hermansky-Pudlak syndrome
- type 2. BMC Medical Genetics. 2013;14:42.
- 399 9. Meeths M, Chiang SCC, Wood SM, Entesarian M, Schlums H, Bang B, et al. Familial
- 400 hemophagocytic lymphohistiocytosis type 3 (FHL3) caused by deep intronic mutation and
- 401 inversion in UNC13D. Blood. 2011;118:5783–93.
- 402 10. Qian Y, Johnson JA, Connor JA, Valencia CA, Barasa N, Schubert J, et al. The 253-kb inversion
- 403 and deep intronic mutations in UNC13D are present in North American patients with familial
- 404 hemophagocytic lymphohistiocytosis 3. Pediatric Blood & Cancer. 2014;61:1034–40.
- 405 11. de Jong S, Chepelev I, Janson E, Strengman E, van den Berg LH, Veldink JH, et al. Common
- 406 inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific
- 407 manner. BMC Genomics. 2012;13:458.
- 408 12. Pilbrow AP, Lewis KA, Perrin MH, Sweet WE, Moravec CS, Tang WHW, et al. Cardiac CRFR1
- 409 Expression Is Elevated in Human Heart Failure and Modulated by Genetic Variation and
- 410 Alternative Splicing. Endocrinology. 2016;157:4865–74.

411 13. González JR, Ruiz-Arenas C, Cáceres A, Morán I, López-Sánchez M, Alonso L, et al.

412 Polymorphic Inversions Underlie the Shared Genetic Susceptibility of Obesity-Related Diseases.

413 The American Journal of Human Genetics. 2020;106:846–58.

414 14. Ruiz-Arenas C, Cáceres A, Moreno V, González JR. Common polymorphic inversions at

415 17q21.31 and 8p23.1 associate with cancer prognosis. Hum Genomics. 2019;13:57.

416 15. Nomura T, Suzuki S, Miyauchi T, Takeda M, Shinkuma S, Fujita Y, et al. Chromosomal

417 inversions as a hidden disease-modifying factor for somatic recombination phenotypes. JCI

418 Insight. 2018;3:e97595.

419 16. Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, et al. A 1.5 million–base pair

420 inversion polymorphism in families with Williams-Beuren syndrome. Nat Genet. 2001;29:321–5.

421 17. Vicente-Salvador D, Puig M, Gayà-Vidal M, Pacheco S, Giner-Delgado C, Noguera I, et al.

422 Detailed analysis of inversions predicted between two human genomes: errors, real

423 polymorphisms, and their origin and population distribution. Human Molecular Genetics.

424 2017;26:567–81.

425 18. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform

426 discovery of haplotype-resolved structural variation in human genomes. Nat Commun.

427 2019;10:1784.

428 19. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of
429 short read general-purpose structural variant calling software. Nat Commun. 2019;10:3240.

430	20. Falconer E, Hills M, Naumann U, Poon SSS, Chavez EA, Sanders AD, et al. DNA template
431	strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat
432	Methods. 2012;9:1107–12.
433	21. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, et al. Genome mapping on
434	nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol.
435	2012;30:771–6.
436	22. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-
437	resolved diverse human genomes and integrated analysis of structural variation. Science.
438	2021;372:eabf7117.
439	23. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation
440	reference for medical and population genetics. Nature. 2020;581:444–51.
441	24. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants:
442	a curated collection of structural variation in the human genome. Nucl Acids Res.
443	2014;42:D986–92.

444 25. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage

445 whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios.

446 Cell BABABAB. 2022;185:3426-3440.e19.

447 26. Henter J-I, Horne A, Aricó M, Egeler RM, Filipovich AH, Imashuku S, et al. HLH-2004:

448 Diagnostic and therapeutic guidelines for hemophagocytic lymphohistiocytosis. Pediatr Blood

449 Cancer. 2007;48:124–31.

450 27. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.

451 2018;34:3094–100.

- 452 28. Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. Symphonizing pileup and full-alignment for
- 453 deep learning-based long-read variant calling. Nat Comput Sci. 2022;2:797–803.
- 454 29. Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, et al. WhatsHap: fast and
- 455 accurate read-based phasing [Internet]. bioRxiv; 2016 [cited 2024 Apr 23]. p. 085050. Available
- 456 from: https://www.biorxiv.org/content/10.1101/085050v2
- 457 30. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC
- 458 Genome Browser database: 2019 update. Nucleic Acids Research. 2019;47:D853–8.
- 459 31. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across
- 460 phenotype–gene relationships. Nucleic Acids Research. 2019;47:D1038–43.
- 461 32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
- 462 Bioinformatics. 2010;26:841–2.
- 463 33. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and
- 464 collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128.
- 465 34. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna,
- 466 Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-
- 467 project.org/

- 468 35. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag
- 469 New York; 2016. Available from: https://ggplot2.tidyverse.org
- 470 36. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting
- 471 sets and their properties. Bioinformatics. 2017;33:2938–40.
- 472 37. Van Kuilenburg ABP, Tarailo-Graovac M, Meijer J, Drogemoller B, Vockley J, Maurer D, et al.
- 473 Genome sequencing reveals a novel genetic mechanism underlying dihydropyrimidine
- 474 dehydrogenase deficiency: A novel missense variant c.1700G>A and a large intragenic inversion
- in *DPYD* spanning intron 8 to intron 12. Human Mutation. 2018;39:947–53.
- 476 38. Bilgrav Saether K, Eisfeldt J, Bengtsson J, Lun MY, Grochowski CM, Mahmoud M, et al. Mind
- 477 the gap: the relevance of the genome reference to resolve rare and pathogenic inversions.
- 478 medRxiv. 2024;2024.04.22.24305780.
- 479 39. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of
- 480 structural variation detection algorithms for whole genome sequencing. Genome Biol.
- 481 2019;20:117.

482 **ACKNOWLEDGEMENTS**

483 We thank the patient and family for participation in this study.

484 **AUTHOR CONTRIBUTIONS**

- 485 Conceptualization: CMBC, ZCA; Data Analysis: TBY and JDB; Funding acquisition: CMBC; Clinical
- 486 data: IKC; Supervision: CMBC and ZCA; Writing, review, and editing: TBY, MYL, JDB, US, IKC, ZCA,
- 487 and CMBC. All authors have read and approved the final manuscript.

488 FUNDING

- 489 This work was supported in part by the United States National Institute of General Medical
- 490 Sciences NIGMS R01 GM132589 (CMBC). IKC was supported by the Jeffrey Modell Foundation at
- 491 Texas Children's Hospital. TBY was supported by the Turkish Scientific and Technological
- 492 Research Council (TUBITAK) 2214-A Program.

493 **COMPETING INTERESTS**

494 The authors declare no competing interests.

495 ETHICAL APPROVAL

- 496 This study is approved by the Baylor College of Medicine (BCM) Institutional Review Board and
- 497 WIRB for the Pacific Northwest Research Institute (IRB Protocol #H-47127/20202158).

498 **TABLES**

499 Table 1. Summary statistics of the datasets analyzed in this study.

	Sequencing technology	Number	Minimum	1 st	Median	Mean	3 rd	Maximum
Dataset		of	length	Quartile	length	length	Quartile	length
		inversions	(kb)	(kb)	(kb)	(kb)	(kb)	(kb)
gnomAD- SV (v4.0)	Short-read WGS	2185	0.052	0.896	7.1	2402.4	323.63	118667.16
DGV (release date: 2020-02- 25)	Mixed	3468	0.035	0.395	2.67	168.5	24.3	9734
1KGP (release date:	Short-read WGS	920	0.052	0.238	0.831	9.41	6.16	98.73

2021-10- 05)								
Ebert <i>et</i> <i>al.</i> 2021	Long-read WGS, Strand-seq	414	0.3	8.13	23.94	293.19	87.63	57207.41
Porubsky <i>et al.</i> 2022	Long-read WGS, Strand-seq, Single- molecule optical mapping	399	0.236	4.67	20.73	251.71	114.28	23268.23

500

501 **FIGURES**



Figure 1. Overview of the datasets and the study design. We extracted inversions from publicly available databases, gnomAD (v4.0) [23], DGV (release date: 2020-02-25) [24], 1KGP (release date: 2021-10-05) [25] and two recent publications of Ebert *et al.* [22] and Porubsky *et al.* [4] We

506 then intersect inversions with OMIM genes and grouped inversion-gene intersections into three

507 categories.



509 Figure 2. UNC13D variants in the patient. (A) Pedigree of patient SEA110 and IGV screenshot 510 displaying nanopore sequencing reads that detected the pathogenic SNV in UNC13D 511 (NM 199242). Chr17 Hap1 represents the haplotype carrying the SNV in UNC13D, the blowout 512 of UNC13D point to the approximate location of the SNV. (B) Diagram of Chr17 Hap2, showing 513 the inversion junction sequencing alignments of each breakpoint. Arrows point to the alignments 514 for junctions 1 and 2 (jct1/2). PCR primers used to obtain the breakpoint junctions for Sanger 515 sequencing are indicated by arrows. Arrows are not to scale. (C) Optical Genome Mapping 516 showing the inversion in Chr17 Hap2, panels show molecules spanning each junction and the 517 location of UNC13D relative to the inversion call.



Figure 3. The categories of the intersections between inversions and protein-coding genes and percentages of intersections belonging to these categories. We grouped the intersections between inversions and OMIM phenotype-related genes into three categories. The first category comprises genes covered by inversions, the second category includes intersections where one of

523 the inversion breakpoints is located within a gene region, and the third category involves





525

Figure 4. (A) Number of OMIM phenotype-related genes (protein-coding) overlapping with the inversions in all datasets in category 2 and 3. (B) Inheritance pattern of the genes overlapping with all inversions in category 2. (C) Inheritance pattern of the genes overlapping with all inversions in category 3.

530 SUPPLEMENTARY INFORMATION

- 531 Supplementary Figure 1. Barplot of the transcript types of the genes in GENCODE v46.
- 532 Supplementary Figure 2. IGV visualization of the detected variants in SEA110. (A) Illustration of
- 533 which mapped reads correspond to each junction, reads mapping to jct2 are the reads

534 highlighted in red with soft clipping extending to green, while reads highlighted in green with soft 535 clipping extending into red map to junction 1. (B) Manual phasing of c.1389+1G>A to the non-536 inverted haplotype. Black boxes highlight SNPs that are represented unique to the SNV 537 haplotype. The dashed black light indicates a read that extends past jct2 from the inversion and 538 contains c.1389+1G>A. 539 Supplementary Figure 3. The UNC13D inversion in gnomAD. 540 Supplementary Figure 4. Rare and common inversions in gnomAD. (A) 99% of inversions in 541 gnomAD v4.0 are rare with <0.5% frequency. (B) Number of genes intersecting with common 542 inversions in gnomAD v4.0 based on category 2 and 3. (C) Number of genes intersecting with rare 543 inversions in gnomAD v4.0 based on category 2 and 3. 544 Supplementary Figure 5. The log2 transformed length plot of the inversion sizes in the datasets. 545 Supplementary Figure 6. Plot of number of common and dataset-specific inversions in the 546 datasets. 547 Supplementary Figure 7. Comparison of inversions in each dataset based on overlapping at least 548 50% of sequence length. 549 Supplementary Figure 8. The protein-coding genes that are related to a phenotype in OMIM 550 overlapping with inversions. (A) The percentage of the OMIM phenotype-related genes 551 overlapping inversions in category 2. (B) The percentage of the OMIM phenotype-related genes 552 overlapping inversions in category 3. (C) The pie charts of inheritance patterns of genes 553 overlapping with inversions in each dataset based on category 2. (D) The pie charts of inheritance

554 patterns of genes overlapping with inversions in each dataset based on category 3.

555 Supplementary Table 1. Primer sets used in the study.

- 556 Supplementary Table 2. References of DGV inversions.
- 557 Supplementary Table 3. Rare gnomAD inversions-genes intersections in category 2 and 3.
- 558 Supplementary Table 4. Rare gnomAD inversions with homozygous frequency 0 OMIM AR genes
- intersections in category 2 and 3.
- 560 Supplementary Table 5. Inversion-gene intersections from all datasets in category 2.
- 561 Supplementary Table 6. Inversion-gene intersections from all datasets in category 3.
- 562 Supplementary Table 7. Enriched HPO terms for the protein-coding genes intersecting with the
- 563 inversions in all datasets in category 2.
- 564 Supplementary Table 8. Enriched HPO terms for the protein-coding genes intersecting with the
- 565 inversions in all datasets in category 3.
- 566
- 567
- 568
- 569

570