

1 Title

2 Profiling copy number variation and disease associations from 50,726 DiscovEHR Study
3 exomes

5 Authors

6 Evan K. Maxwell¹, Jonathan S. Packer¹, Colm O'Dushlaine¹, Shane E. McCarthy¹, Abby Hare-
7 Harris², Jeffrey Staples¹, Claudia Gonzaga-Jauregui¹, Samantha N. Fetterolf², W. Andrew
8 Faucett², Joseph B. Leader², Andres Moreno-De-Luca², Giusy Della Gatta¹, Margaret Scollan¹,
9 Trikaladarshi Persaud¹, John Penn¹, Alicia Hawes¹, Xiaodong Bai¹, Sarah Wolf¹, Alexander E.
10 Lopez¹, Rick Ulloa¹, Christopher Sprangel¹, Rostislav Chernomorsky¹, Ingrid B. Borecki¹,
11 Frederick E. Dewey¹, Aris N. Economides¹, John D. Overton¹, H. Lester Kirchner², Michael F.
12 Murray², Marylyn D. Ritchie², David J. Carey², David H. Ledbetter², George D. Yancopoulos¹,
13 Alan R. Shuldiner¹, Aris Baras¹, Omri Gottesman¹, Lukas Habegger¹, Christa Lese Martin²,
14 Jeffrey G. Reid^{1*}

16 Affiliations

17 ¹Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, 10591

18 ²Geisinger Health System, Danville, PA, 17822

19 *Corresponding author: jeffrey.reid@regeneron.com

21 Abstract

22 Copy number variants (CNVs) are a substantial source of genomic variation and contribute to a
23 wide range of human disorders. Gene-disrupting exonic CNVs have important clinical
24 implications as they can underlie variability in disease presentation and susceptibility. The
25 relationship between exonic CNVs and clinical traits has not been broadly explored at the

26 population level, primarily due to technical challenges. We surveyed common and rare CNVs in
27 the exome sequences of 50,726 adult DiscovEHR study participants with linked electronic
28 health records (EHRs). We evaluated the diagnostic yield and clinical expressivity of known
29 pathogenic CNVs, and performed tests of association with EHR-derived serum lipids, thereby
30 evaluating the relationship between CNVs and complex traits and phenotypes in an unbiased,
31 real-world clinical context. We identified CNVs from megabase to exon-level resolution,
32 demonstrating reliable, high-throughput detection of clinically relevant exonic CNVs. In doing so,
33 we created a catalog of high-confidence common and rare CNVs and refined population
34 frequency estimates of known and novel gene-disrupting CNVs. Our survey among an
35 unselected clinical population provides further evidence that neuropathy-associated duplications
36 and deletions in 17p12 have similar population prevalence but are clinically under-diagnosed.
37 Similarly, adults who harbor 22q11.2 deletions frequently had EHR documentation of
38 neurodevelopmental/neuropsychiatric disorders and congenital anomalies, but not a formal
39 genetic diagnosis (i.e., deletion). In an exome-wide association study of lipid levels, we
40 identified a novel five-exon duplication within *LDLR* segregating in a large kindred with features
41 of familial hypercholesterolemia. Exonic CNVs provide new opportunities to understand and
42 diagnose human disease.

43

44 **Introduction**

45 Copy-number variants (CNVs), a type of structural variation, are regions of the genome
46 that deviate from the expected normal diploid state as a result of deletion or duplication events.
47 CNVs have been shown to cause or increase risk of a number of diseases¹. Thus a
48 comprehensive view of CNVs may further our understanding of the genetic bases of human
49 health and disease in service of precision medicine. CNV detection in large clinical and research
50 cohorts is generally performed using array-based technologies given considerations of cost,
51 reliability, and throughput². However, commercially available array-based technologies, or
52 custom-designed arrays used in some clinical laboratories, are limited by probe design and
53 density, achieving a maximum resolution of 5 kb or larger, with calls spanning 50-250 kb being
54 the most reliable. This has additional technical biases and leaves a major class of smaller,
55 exonic CNVs undetected by clinical genomic testing.

56 While genomic sequencing (whole-exome and whole-genome) is becoming a common
57 diagnostic tool, CNV ascertainment from sequence data is not standard practice due to
58 complexities in reliable detection. Whole-genome sequencing can achieve base-pair
59 resolution^{3,4}, but the cost and analytical challenges have thus far limited its application. Exome
60 sequencing provides a balance of resolution, throughput, and cost, making it a more cost-
61 effective choice for clinical labs and population-scale cohort studies, however exome-based
62 CNV detection methods have faced significant technical challenges limiting their adoption^{5,6}. We
63 recently developed a high-throughput exome-CNV detection method CLAMMS⁷ that enables
64 reliable common and rare CNV detection across the size spectrum at population-scale. In this
65 study, we sought to explore the phenotypic consequences of exonic CNVs identified in a large
66 clinical care population.

67

68 **RESULTS**

69

70 **Spectrum of CNVs in DiscovEHR Participants**

71 We applied CLAMMS⁷ to the sequenced exomes of 50,726 adult patient-participants who
72 receive health care through the Geisinger Health System (GHS), all of whom consented to
73 participate in the MyCode® Community Health Initiative biobank⁸ and contributed DNA samples
74 for genomic sequencing and analysis as part of the Regeneron-GHS DiscovEHR Study⁹. Each
75 exome is linked to de-identified data extracted from the individual's electronic health record
76 (EHR), with a median of 13.4 years of longitudinal clinical data available. The sequenced cohort
77 is highly homogeneous in terms of European ancestry (~97%). While the DiscovEHR Study is
78 not targeted towards any particular age or phenotype, the first 50,000 individuals recruited are
79 enriched for adults (mean age = 61 years) with chronic health problems who interact frequently
80 with the healthcare system, as well as patient-participants sequenced from the Coronary
81 Catheterization Laboratory and the Bariatric Surgery Clinic.

82 Following application of quality control filters (**Methods**), our high-confidence CNV call
83 set includes 47,349 participants (>93%) and 475,664 CNV events at 13,782 loci
84 (**Supplementary Table S2**). The median size of observed common CNV loci (allele frequency
85 (AF) $\geq 1\%$) is 7.1 kb (deletions 4.4 kb, duplications 13.4 kb), compared to 17.8 kb (deletions
86 8.4 kb, duplications 32.8 kb) in rare CNV loci (AF < 1%). On average, we observed ~10 high-
87 confidence exonic CNV alleles affecting 14 genes per individual, most of which are common
88 (6.6 deletions and 1.7 duplications, considering only highly mappable regions of the exome). We
89 found an average of one very rare CNV allele per individual (AF < 0.1%; 0.6 duplications and
90 0.4 deletions; **Fig. 1B, Supplementary Table S2**), and roughly one out of seven individuals
91 possesses at least one CNV that is unique to their exome, relative to the cohort. The vast
92 majority (91%) of CNV loci are extremely rare, observed in fewer than 10 individuals (AF <
93 0.01%) in our sequenced cohort, with over half representing singletons (**Supplementary Fig.**
94 **S2**). In total, 13,170 genes are intersected by at least one CNV of less than 2 Mb in length

95 (~73% of the total callable gene set; **Supplementary Table S2**), and, concordant with other
96 surveys^{3,10}, loss-of-function intolerant genes are heavily depleted for observed deletions
97 compared to duplications (**Fig. 1C, Supplementary Methods**). In our accompanying linkage
98 disequilibrium map, only half of common CNVs and 30 rare CNVs (0.2%) are well tagged ($r^2 \geq$
99 0.3) by a SNP marker, suggesting these CNVs cannot be captured or imputed from genotyping
100 arrays (**Supplementary Methods**).

101 **Segregation of CNVs in Computationally Inferred Kindreds**

102 Cross-referencing CNVs with computationally inferred pedigrees (**Supplementary Methods**)
103 enables tracking of familial variants as well as analysis of transmission rates, providing
104 estimates of cohort-wide call sensitivity and specificity. The call set exhibits an average parent-
105 child transmission rate of ~46.5% for rare heterozygous CNVs, consistent with a theoretical rate
106 of 50% assuming equal probability of transmission and lack of transmission disequilibrium. This
107 includes many rare small CNVs (≤ 3 exons, 0.54 per sample cohort-wide) transmitted at
108 42.17%. We validated, using qPCR, a sample of 100 high-confidence CNVs identified in 259
109 parent-child duos (**Methods**), estimating precision and recall to be 98% and 92%, respectively,
110 in the smallest class of exonic CNVs (1-3 exons). Comparatively, an alternative (and widely
111 used) exome CNV calling method, XHMM¹¹, achieved only 36.1% recall (1% for single-exon
112 calls) with 96.8% precision in this size range (**Supplementary Table S3 & Methods**).

113 **Prevalence and Clinical Impact of Disease-associated Exonic CNVs**

114 We surveyed frequencies of a selected set of known disease-associated CNVs likely reflecting
115 the spectrum of coding CNVs expected in a broad, predominately European hospital-based
116 population (**Table 1**). This EHR-linked resource provided the opportunity to evaluate penetrance
117 and expressivity of Mendelian disease-associated CNVs and to characterize the clinical
118 consequences of other exonic CNVs among a broadly sampled adult population.

119 We identified 12 patient-participants with 22q11.2 deletion syndrome (MIM #188400),
120 two of which are from the same family. This deletion is one of the most common and clinically

121 documented recurrent CNVs; our observed prevalence of 1 in 3,946 in the DiscovEHR cohort is
122 in line with previous estimates of 1 in 4,000 live births¹². Of the 12 participants with a 22q11.2
123 deletion, 10 (83%) had ICD-9 codes consistent with the 22q11.2 deletion syndrome phenotype,
124 including 8 (67%) with a diverse set of neurodevelopmental/neuropsychiatric disorders (NDD)
125 and 6 (50%) with congenital anomalies, with 5/12 (42%) participants having both NDD and
126 congenital anomaly phenotypes reported (**Supplementary Methods**). However, only one
127 patient-participant (age 50 years old) with a 22q11.2 deletion had a specific genetic diagnosis of
128 velocardiofacial syndrome (ICD 758.32) documented in their EHR, suggesting that most
129 individuals are likely unaware of the genetic etiology of their clinical phenotypes.

130 We also found 25 patient-participants with 17p12 duplications that include the *PMP22*
131 gene; 21 having duplications spanning the common recurrent CNV locus associated with the
132 most common form of peripheral neuropathy, Charcot-Marie-Tooth disease type 1A (CMT1A;
133 MIM #118220)¹³, three with atypical duplications in two families, and one with breakpoints that
134 could not be reliably estimated (**Supplementary Fig. S12**). Fourteen out of 21 of the patient-
135 participants with a common recurrent duplication and 2/3 with an atypical duplication (a parent-
136 child duo) had a clinical diagnosis of CMT1A. Seven out of 21 individuals with a recurrent
137 duplication had non-specific diagnoses consistent with CMT1A recorded in their EHR
138 (**Supplementary Methods**). Three patient-participants had neuropathy phenotypes attributed to
139 type-2 diabetes, suggesting a possible misdiagnosis of the primary hereditary neuropathy.
140 Similarly, we identified and validated 26 patient-participants with the reciprocal 17p12 deletion
141 associated with Hereditary Neuropathy with liability to Pressure Palsies (HNPP; MIM
142 #162500)¹³, all of which spanned the common recurrent CNV region. However, only 1/26
143 patient-participants had a clinical diagnosis of HNPP in their EHR and 6 had consistent, but
144 non-specific diagnoses. Our observed population frequencies for the recurrent CMT-associated
145 duplication (1/2,255 total, 1/2,785 *de novo*; **Supplementary Methods**) and the related HNPP-
146 associated deletion (1/1,821 total, 1/2,255 *de novo*) are roughly equal, but their EHR-based

147 diagnostic rates are vastly different, suggesting that the deletion is largely clinically under-
148 diagnosed. Forty-nine patient-participants with 17p12 CNVs and five non-carrier relatives were
149 validated with qPCR, demonstrating accurate identification of this clinically relevant CNV from
150 the exome data (**Supplementary Fig. S12 & Methods**).

151 Finally, among 323 patient-participants with 2q13 deletions encompassing the entirety of
152 the nephronophthisis-1 (*NPHP1*) gene, we identified two individuals with homozygous deletions.
153 Both of these individuals (age 31 and 38 years) had a documented diagnosis of medullary cystic
154 kidney (ICD 753.16) and end stage renal disease (ICD 585.6) requiring transplantation,
155 consistent with autosomal recessive juvenile nephronophthisis (OMIM #256100). These results
156 exemplify the importance of differentiating heterozygous versus homozygous deletions for
157 clinical correlations.

158 **Exome-wide analysis of associations of exonic CNVs with serum lipid levels**

159 We next tested the ability to identify novel disease susceptibility loci through the intersection of
160 exome CNVs and EHR-derived phenotypes. We performed an exome-wide association study of
161 our CNV loci with fasting serum lipid levels (**Methods**), which are heritable risk factors for
162 ischemic cardiovascular disease, in a subset of 39,087 individuals with available lipid data. At a
163 Bonferroni-corrected significance threshold of $p < 1.2 \times 10^{-5}$, three CNV loci significantly
164 associated with lipid levels (**Table 2**). We then evaluated the penetrance of these lipid-
165 associated variants to coronary artery disease (CAD).

166 A novel duplication of g.chr19:11230767-11241993 (hg19/GRCh37) encompassing
167 exons 13-17 of the low-density lipoprotein receptor gene *LDLR* was identified in 29 individuals
168 and found to be associated with high LDL cholesterol ($\beta = 1.73$ [76 mg/dl], $p = 1.3 \times 10^{-13}$) and high
169 total cholesterol ($\beta = 1.38$ [61 mg/dl], $p = 3.8 \times 10^{-9}$; **Methods**). Breakpoints were identified in *Alu*
170 repeats within introns 12 and 17 using whole-genome sequencing, confirming a tandem
171 insertion site predicted to occur in-frame (**Fig. 2A**). The duplication was validated in all carriers
172 by Sanger sequencing (**Supplementary Fig. S6**). CNVs have previously been reported in

173 *LDLR*¹⁴; however, this duplication appears to be novel. In overexpression experiments in
174 HEK293 cells (**Supplementary Methods**), *LDLR* containing the exon 13-17 tandem duplication
175 (Dup13-17) produced a larger protein product than the wild-type (WT) receptor that accumulates
176 in the endoplasmic reticulum. As a result, the mutant receptor is not expressed on the cellular
177 surface and has no LDL uptake activity. Thus, the duplication behaves similarly to the known
178 FH-causing p.G549D point mutation, classified as a transport-inhibiting (class-2) mutation (**Figs.**
179 **2B-E & S13**). Taken together, this evidence suggests that the duplication causes loss-of-
180 function of LDL receptor activity.

181 The *LDLR* duplication associates with markedly increased CAD risk (OR=5.01, Logistic
182 regression, $p=6 \times 10^{-4}$). Using pedigree reconstruction and distant-relatedness analysis
183 (**Supplementary Methods**), we connected 27/29 carriers into a single large estimated
184 pedigree, dating their common ancestor to at least six generations ago and identifying 10
185 related individuals not harboring the *LDLR* duplication (**Fig. 2C**). In this extended pedigree, the
186 duplication segregated with high LDL levels and 15/29 duplication carriers had ischemic heart
187 disease (IHD) diagnoses, 11 of who presented with early-onset IHD (**Supplementary**
188 **Methods**). Given the genetic and functional evidence, and the observation that familial
189 hypercholesterolemia (FH) cases are frequently attributed to *LDLR* mutations¹⁴, we conclude
190 that this is a novel pathogenic FH-causing CNV. Present in 1/1,749 samples, this novel *LDLR*
191 duplication represents roughly 13% of the pathogenic FH variants (30% of those specific to
192 *LDLR*) observed in our cohort, which have an overall prevalence of 1:256 unselected individuals
193 in sequenced participants¹⁵. Notably, only eight patient-participants carrying the *LDLR*
194 duplication and one relative without the duplication have a probable FH diagnosis coded in their
195 EHR (**Supplementary Methods**).

196 We also identified a common deletion encompassing the last 6/7 exons of the leukocyte
197 immunoglobulin (Ig)-like receptor A3 gene (*LILRA3*; AF \cong 17%, consistent with previous
198 estimates in Europeans¹⁶), that was associated with increased HDL levels ($\beta=0.05$ [0.65 mg/dl],

199 $p=4.5 \times 10^{-7}$, **Methods**). No association with the prevalence of CAD was observed. While the size
200 of this variant (~2.6 kb exonic, 6.7 kb genomic) is not amenable to detection with standard
201 clinical CNV arrays, CLAMMS calls at this locus were previously qPCR-validated in 165
202 samples (69 CNV carriers) with perfect sensitivity and specificity⁷. Genome-wide association
203 studies (GWAS) have identified SNVs adjacent to *LILRA3* that associate with HDL levels¹⁷; our
204 analysis revealed that the deletion and GWAS SNV (rs386000) are in linkage disequilibrium
205 ($r^2=0.77$, $D'=0.959$; **Supplementary Dataset S2**). Multiple expression quantitative trait loci
206 (eQTLs) have been mapped to *LILRA3*, but the deletion is likely to be driving the effect¹⁸ and
207 contributing to the variation in HDL levels. This deletion has previously been investigated for
208 association with other diseases¹⁶; we observed nominally increased risk for rheumatoid arthritis
209 ($p=0.041$, OR=1.1, 95% confidence interval=[1.004,1.223], logistic regression) and decreased
210 risk for prostate malignancies ($p=0.045$, OR=0.9, 95% confidence interval=[0.797,0.997], logistic
211 regression) among carriers (**Supplementary Methods**). Further testing of this *LILRA3* CNV in
212 other well-powered cohorts will shed light into the true genotype-phenotype associations. In
213 addition to *LILRA3* deletions, we demonstrate the ability to detect a small, complex CNV in the
214 haptoglobin gene (*HP*) through a single mappable exon well enough to recapitulate the
215 directionality of previously observed associations with increased LDL and total cholesterol¹⁹
216 (**Supplementary Methods**).

217

218 **DISCUSSION**

219 Our study demonstrates the utility of exome sequencing data for analyzing common and
220 rare CNVs in the emerging paradigm of precision medicine using EHR-derived health
221 information. We show robust detection of small exonic CNVs well below the size threshold of
222 array-based technologies (**Supplementary Methods**), providing further evidence that the

223 density of markers on genotyping arrays is insufficient for characterizing the full CNV spectrum
224 in humans and necessitates higher resolution methods (**Supplementary Fig. S4**).

225 Our catalog of exonic CNVs represents a substantial source of genomic variation in the
226 DiscovEHR study population and exemplifies the clinical value of CNVs. We survey the clinical
227 expressivity of known pathogenic CNVs, and via exome-wide associations with serum lipid
228 traits, discover a novel duplication in *LDLR* causing familial hypercholesterolemia (FH). Among
229 both known and novel pathogenic CNVs, the genetic basis for disease is frequently under-
230 diagnosed. Clinical variability is common among individuals with known recurrent CNVs (e.g.
231 17p12 CNVs and 22q11.2 deletion syndrome), presenting a diagnostic challenge to clinicians,
232 especially in patients with more mild symptoms and those born before the widespread use of
233 clinical genetic testing. In the case of the *LDLR* duplication, this novel CNV is undetectable with
234 standard genetic screening platforms; thus, the handful of carriers with FH-consistent EHR
235 documentation were diagnosed based on clinical features, not the underlying genetic cause.
236 Combined with other reports of *LDLR* rearrangements^{14,20} and the density of *Alu* repeat
237 elements (2.24×10^{-3} per base intronic sequence, 99.4th percentile for intron-containing genes;
238 **Supplementary Methods**), *LDLR* may be particularly prone to genomic rearrangements²¹.
239 These data highlight the potential for high-throughput CNV screening to improve diagnostic
240 rates for FH and inform patient treatment.

241 With the growing prevalence of whole-exome and whole-genome sequencing, the
242 substantial body of literature implicating CNVs in both rare and common disease, and the
243 magnitude of copy-number variation we observe in the exome, the inclusion of methods to
244 identify CNVs and other structural variants within standard sequence-based informatics
245 pipelines is long overdue. In addition, the ability to identify rare homozygous deletions (e.g.
246 2q13) in large cohort studies, like DiscovEHR, will contribute to human gene knockout catalogs
247 and the identification of autosomal recessive conditions. As we find that over 90% of distinct
248 CNVs are present in less than 1/5,000 individuals; large sample sizes or targeted recruitment of

249 additional family members are essential to establish phenotypic associations. Furthermore,
250 detection of exonic and larger clinically relevant CNVs from exome sequence data will
251 streamline genetic testing so that both sequence and copy number variants can be called from
252 one test methodology. The ability to obtain a more comprehensive view of human genetic
253 variation, both per individual and across populations, will facilitate the advent of precision
254 medicine.

255

256 **METHODS**

257 ***Study population***

258 The human genetics studies were conducted as part of the DiscovEHR study of the Regeneron
259 Genetics Center and the Geisinger Health System. Patient-participants who receive health care
260 through Geisinger Health System were consented to participate in the MyCode Community
261 Health Initiative and DiscovEHR cohort following an IRB approved protocol⁸. DNA samples and
262 exome data from 50,726 adult patient-participants were included in this study. Detailed
263 information on the clinical characteristics of this cohort can be found in²². The Regeneron
264 Genetics Center funded study sample collection, sequence data generation, and clinical and
265 sequence data analysis. All participants gave informed written consent.

266 ***Sample preparation and sequencing***

267 Genomic DNA samples were transferred to the Regeneron Genetics Center from the Geisinger
268 Health System in 2D matrix tubes (Thermo Scientific, Waltham, MA), logged into our LIMS
269 (Sapio Sciences, Baltimore, MD), and stored in our automated biobank at -80°C (LiCONiC
270 TubeStore, Woburn, MA). Sample quantity was determined by fluorescence (Life Technologies,
271 Carlsbad, CA) and quality assessed by running 100ng of sample on a 2% pre-cast agarose gel
272 (Life Technologies). The DNA samples were normalized and one aliquot was sent for
273 genotyping (Illumina Inc., San Diego, CA, Human OmniExpress Exome Beadchip) and another

274 sheared to an average fragment length of 150 base pairs using focused acoustic energy
275 (Covaris LE220, Woburn, MA). The sheared genomic DNA was prepared for exome capture
276 with a custom reagent kit from Kapa Biosystems (Wilmington, MA) using a fully-automated
277 approach developed at the Regeneron Genetics Center (Tarrytown, NY). A unique 6 base pair
278 barcode was added to each DNA fragment during library preparation to facilitate multiplexed
279 exome capture and sequencing. Equal amounts of sample were pooled prior to exome capture
280 with NimbleGen (Roche NimbleGen, Madison, WI) probes (SeqCap VCRome). Captured
281 fragments were bound to streptavidin-conjugated beads and non-specific DNA fragments
282 removed by a series of stringent washes according to the manufacturer's recommended
283 protocol (Roche NimbleGen). The captured DNA was PCR amplified and quantified by qRT-
284 PCR (Kapa Biosystems). The multiplexed samples were sequenced using 75 bp paired-end
285 sequencing on an Illumina v4 HiSeq 2500 to a coverage depth sufficient to provide greater than
286 20x haploid read depth of over 85% of targeted bases in 96% of samples (approximately 80x
287 mean haploid read depth of targeted bases).

288 Upon completion of sequencing, raw data from each Illumina HiSeq 2500 run was
289 gathered in local buffer storage and uploaded to the DNAnexus (Mountain View, CA) platform²³
290 for automated analysis. Sample-level read files were generated with CASAVA (Illumina Inc.)
291 and aligned to GRCh37.p13 with BWA-mem^{24,25}. The resultant BAM files were processed using
292 GATK²⁶ and Picard to sort, mark duplicates, and perform local realignment of reads around
293 putative indels.

294 **CNV calling and quality control**

295 CLAMMS⁷ was applied to the exomes of all samples in parallel within a distributed compute
296 environment. CLAMMS calls CNVs from each sample's source BAM file using 1) base-level
297 depth-of-coverage calculations from aligned reads having mapping quality ≥ 30 , and 2) seven
298 sequencing quality control (QC) metrics computed using Picard
299 (<http://broadinstitute.github.io/picard/>). Depth-of-coverage profiles are generated for exon

300 “windows” representing either the entire exon or a 500-1000 base pair contiguous segment of
301 an exon for long exons. Exon window coverage distributions are normalized independently for
302 every sample, adjusting for GC content and overall sequencing depth. CLAMMS detects CNVs
303 by comparing a sample’s normalized coverage profile to a reference panel of 100 samples
304 matched with respect to QC metrics via a k-nearest neighbors search, where expected
305 coverage distributions are modelled with exon-window specific mixture models linked together
306 with a genomic distance-aware Hidden Markov Model (HMM).

307 Extensive quality control procedures were applied to the call set with the goal of
308 producing a set of CNV loci exhibiting high specificity (see **Supplementary Methods**). Briefly,
309 the sample set was filtered of “outlier” samples having inflated CNV rates (>28 CNVs, 2x
310 median) resulting in a high-confidence call set containing 47,349 samples (93%). CNVs in the
311 remaining samples were annotated with: 1) two model-based metrics from CLAMMS ($Q_{\text{non_dip}}$
312 and Q_{exact} , representing confidence measures that the region is non-diploid and consistent with
313 the exact called copy number, respectively), 2) allele balance and zygosity of SNVs within called
314 CNV regions, 3) the frequency of overlapping CNVs among outlier samples, and 4) the quality
315 and zygosity measures for all non-outlier carriers of the locus (representing cohort-wide locus
316 performance). Filtering criteria based on these annotations were trained with respect to the
317 cohort-wide transmission rate of rare heterozygous CNVs among parent-child duos from
318 genetically reconstructed pedigrees generated with PRIMUS²⁷. The training set was targeted to
319 47.5% transmission, achieving 47.36% transmission in the training set, 46.02% in the test set,
320 and 46.59% combined. This includes several small (1-3 exon) CNVs (~29% of loci) transmitted
321 at 42.17%, with single-exon CNVs (~8% of loci) being transmitted at 38.96% (**Supplementary**
322 **Table S1**).

323 **CNV validation data**

324 In a prior study⁷, we validated select common and rare CNV loci identified by CLAMMS using
325 TaqMan qPCR and demonstrated higher performance and resolution than alternative exome

326 CNV callers as well as CNV detection from SNP arrays. In the present study, we have
327 expanded our validation set to include additional TaqMan qPCR validation on a representative
328 set of 100 high-confidence CNV loci identified in 259 parent-child duos, with a focus on small
329 CNVs (1-3 exon; 86/100 loci) having at least one duo in which a non-transmission is predicted
330 from CLAMMS (i.e. present in the parent, absent in the child). This provided an unbiased
331 method to test the sensitivity and specificity of CLAMMS at high-confidence small CNV loci,
332 particularly with respect to interpretation of the cohort-wide transmission rate estimates (42.17%
333 for 1-3 exon CNVs, 38.96% for single-exon CNVs; **Supplementary Table S1**). Among the small
334 CNV loci, we observed 10.5% of non-transmission calls did not qPCR validate, corresponding to
335 a 2.2% FPR, 8% FNR, and a total error rate of 5.3% (1-accuracy). Among single-exon CNVs,
336 we observed a 3.5% FPR, 9.2% FNR, and a total error rate of 6.6% (**Supplementary Table**
337 **S3**).

338 In addition to TaqMan qPCR validations, a large subset of samples (N=34,246) were
339 processed with Illumina HumanOmniExpressExome-8v1-2 genotyping arrays in twelve batches
340 and subsequently tested with PennCNV²⁸. PennCNV calls were quality controlled using a 95th
341 percentile cutoff applied to each batch for samples having extreme Log-R Ratio standard
342 deviation (LRR_SD) and B Allele Frequency (BAF) drift. Calls at neighboring loci were
343 considered for merging using the clean_cnv module with the --bp combine_seg flag. We tested
344 the transmission rate of PennCNV calls from samples in identified parent-child duo relationships
345 having fewer than 50 CNVs (N=21,792), as excessive numbers of CNVs often indicate low
346 sample quality, high rates of *de novo* mutation, or somatic copy number variation that would
347 confound transmission rate calculations. Consistent with our previous observations⁷, we found
348 that PennCNV calls were only reliable (i.e. produced transmission rates close to 50%) at size
349 thresholds approaching >100 Kb or larger (**Supplementary Figure S4**).

350 **Statistical Analysis**

351 Associations between CNVs and lipid traits were performed using a linear mixed model
352 implemented in BOLT-LMM²⁹ with a genetic relationship matrix included as a random effect.
353 Deletions and duplications at the same locus were considered separately. Median values for
354 serially measured laboratory traits, including total cholesterol, low-density lipoprotein cholesterol
355 (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides were calculated for all
356 individuals with two or more measurements in the EHR following removal of likely spurious
357 values that were > 3 standard deviations from the intra-individual median value. For the
358 purposes of exome-wide association analysis of serum lipid levels, total cholesterol, LDL-C,
359 triglycerides, and HDL-C were adjusted for lipid-altering medication use by dividing by 0.8, 0.7,
360 0.82, and 1.044, respectively, to estimate pre-treatment lipid values. Medication-adjusted HDL-
361 C and triglycerides were log₁₀ transformed, and medication-adjusted LDL-C and total cholesterol
362 values were not transformed. We then calculated trait residuals after adjustment for age, age²,
363 sex, and the first ten principal components of ancestry, and rank-inverse-normal transformed
364 these residuals prior to exome-wide association analysis (**Supplementary Methods**).

365

366 **References**

- 367 1. Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy Number Variation in Human
368 Health, Disease, and Evolution. *Annu. Rev. Genom. Human Genet.* **10**, 451–481 (2009).
- 369 2. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes
370 associated with developmental delay. *Nat Genet* **46**, 1063–1071 (2014).
- 371 3. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human
372 copy-number variation. *Science* **349**, aab3761 (2015).
- 373 4. Handsaker, R. E., Van Doren, V. & Berman, J. R. Large multiallelic copy number
374 variations in humans. *Nat Genet* **47**, 296–303 (2015).
- 375 5. Tan, R. *et al.* An Evaluation of Copy Number Variation Detection Tools from Whole-

- 376 Exome Sequencing Data. *Human Mutation* **35**, 899–907 (2014).
- 377 6. Hong, C. S., Singh, L. N., Mullikin, J. C. & Biesecker, L. G. Assessing the reproducibility
378 of exome copy number variations predictions. *Genome Medicine* **8**, 612 (2016).
- 379 7. Packer, J. S. *et al.* CLAMMS: a scalable algorithm for calling common and rare copy
380 number variants from exome sequencing data. *Bioinformatics* **32**, 133–135 (2016).
- 381 8. Carey, D. J. *et al.* The Geisinger MyCode community health initiative: an electronic health
382 record–linked biobank for precision medicine research. *Genetics in Medicine* **18**, 906–913
383 (2016).
- 384 9. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-
385 exome sequences from the DiscovEHR study. *Science* **354**, aaf6814–aaf6814 (2016).
- 386 10. Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in
387 59,898 human exomes. *Nat Genet* **48**, 1107–1111 (2016).
- 388 11. Fromer, M. *et al.* Discovery and Statistical Genotyping of Copy-Number Variation from
389 Whole-Exome Sequencing Depth. *The American Journal of Human Genetics* **91**, 597–
390 607 (2012).
- 391 12. Bassett, A. S. *et al.* Practical Guidelines for Managing Patients with 22q11.2 Deletion
392 Syndrome. *The Journal of Pediatrics* **159**, 332–339.e1 (2011).
- 393 13. Chance, P. F. *et al.* Two autosomal dominant neuropathies result from reciprocal DNA
394 duplication/deletion of a region on chromosome 17. *Human Molecular Genetics* **3**, 223–
395 228 (1994).
- 396 14. Leigh, S., Foster, A. H. & Whittall, R. A. Update and analysis of the University College
397 London low density lipoprotein receptor familial hypercholesterolemia database. *Annals*
398 *of Human Genetics* **72**, 485–498 (2008).
- 399 15. Abul-Husn, N. S. *et al.* Genetic identification of familial hypercholesterolemia within a
400 single U.S. health care system. *Science* **354**, aaf7000–aaf7000 (2016).
- 401 16. Hirayasu, K. & Arase, H. Functional and genetic diversity of leukocyte immunoglobulin-

- 402 like receptor and implication for disease associations. *Journal of Human Genetics* **60**,
403 703–708 (2015).
- 404 17. Teslovich, T. M., Musunuru, K., Smith, A. V. & Edmondson, A. C. Biological, clinical and
405 population relevance of 95 loci for blood lipids. *Nat Genet* **466**, 707–713 (2010).
- 406 18. Chiang, C. *et al.* The impact of structural variation on human gene expression. *bioRxiv*
407 (2016). doi:10.1101/055962
- 408 19. Boettger, L. M. *et al.* Recurring exon deletions in the HP (haptoglobin) gene contribute to
409 lower blood cholesterol levels. *Nat Genet* 1–9 (2016). doi:10.1038/ng.3510
- 410 20. Wald, D. S. *et al.* Child–Parent Familial Hypercholesterolemia Screening in Primary Care.
411 *N Engl J Med* **375**, 1628–1637 (2016).
- 412 21. Boone, P. M. *et al.* The Alu-Rich Genomic Architecture of SPAST Predisposes to Diverse
413 and Functionally Distinct Disease-Associated CNV Alleles. *The American Journal of*
414 *Human Genetics* **95**, 143–161 (2014).
- 415 22. Dewey, F. E. *et al.* Inactivating Variants in ANGPTL4 and Risk of Coronary Artery
416 Disease. *N Engl J Med* **374**, 1123–1133 (2016).
- 417 23. Reid, J. G. *et al.* Launching genomics into the cloud: deployment of Mercury, a next
418 generation sequence analysis pipeline. *BMC Bioinformatics* **15**, 30 (2014).
- 419 24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
420 transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 421 25. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
422 *arXiv* (2013).
- 423 26. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing
424 next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).
- 425 27. Staples, J. *et al.* PRIMUS: Rapid Reconstruction of Pedigrees from Genome-wide
426 Estimates of Identity by Descent. *The American Journal of Human Genetics* **95**, 553–564
427 (2014).

- 428 28. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-
429 resolution copy number variation detection in whole-genome SNP genotyping data.
430 *Genome Research* **17**, 1665–1674 (2007).
- 431 29. Loh, P. R., Tucker, G., Bulik-Sullivan, B. K. & Vilhjalmsón, B. J. Efficient Bayesian
432 mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284–
433 290 (2015).
- 434 30. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nat Genet*
435 **536**, 285–291 (2016).

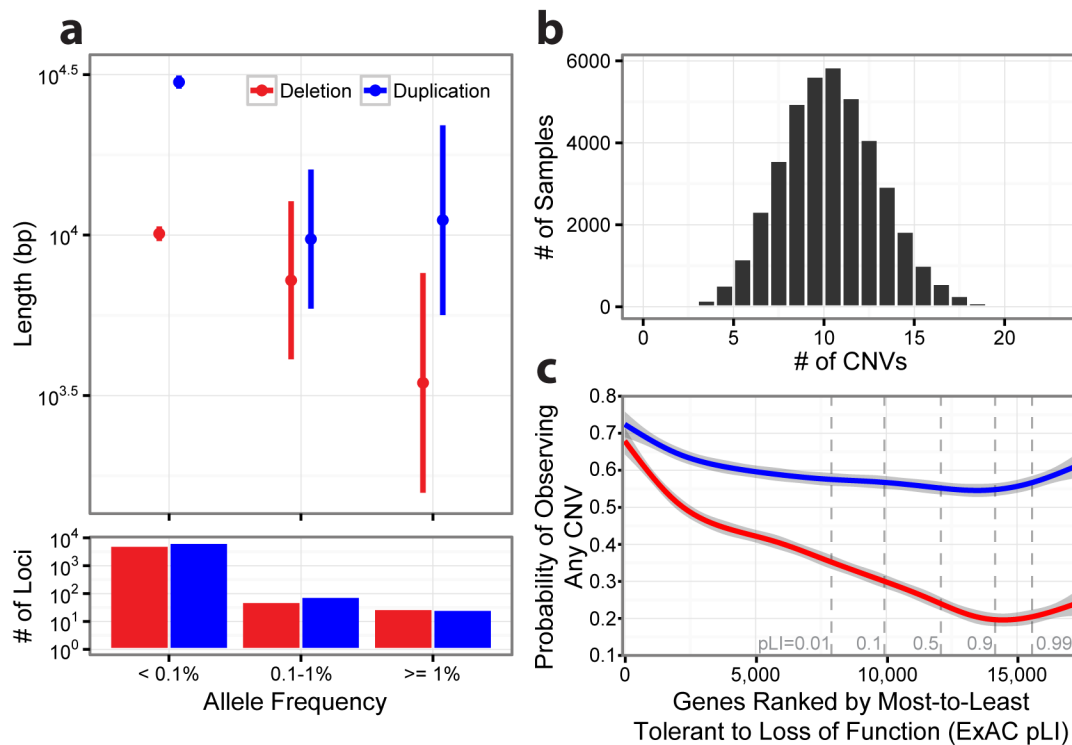
436 **Acknowledgements**

437 The authors would like to thank the MyCode Community Health Initiative participants for their
438 permission to utilize their health and genomics information in the DiscovEHR collaboration and
439 for their ongoing engagement in multiple facets of this research. The authors also thank David
440 D’Ambrosio and Jennifer Espert for performing the TaqMan qPCR validation experiments and
441 Dr. Scott Myers for assisting with clinical correlations. This research was funded by the
442 Regeneron Genetics Center, a wholly-owned subsidiary of Regeneron Pharmaceuticals. This
443 study was also partially supported by grant RO1MH074090 (Drs. Ledbetter and Martin) from the
444 National Institute of Mental Health. Regeneron manufactures and markets Praluent
445 (alirocumab), a *PCSK9* inhibiting antibody, for treatment of familial hypercholesterolemia and
446 other indications. Under the relevant agreement with the Geisinger Health System (GHS),
447 Regeneron is prohibited from making genomic DNA samples and individual level phenotype
448 information obtained from the GHS MyCode-DiscovEHR Project available to another party, and
449 Regeneron is prohibited from disclosing a copy of the agreement to another party. A patent
450 application on the CLAMMS copy number variant calling method has been filed by Regeneron,
451 and a patent application that contains data in the manuscript has been filed by Regeneron.

452

453

454 Figures

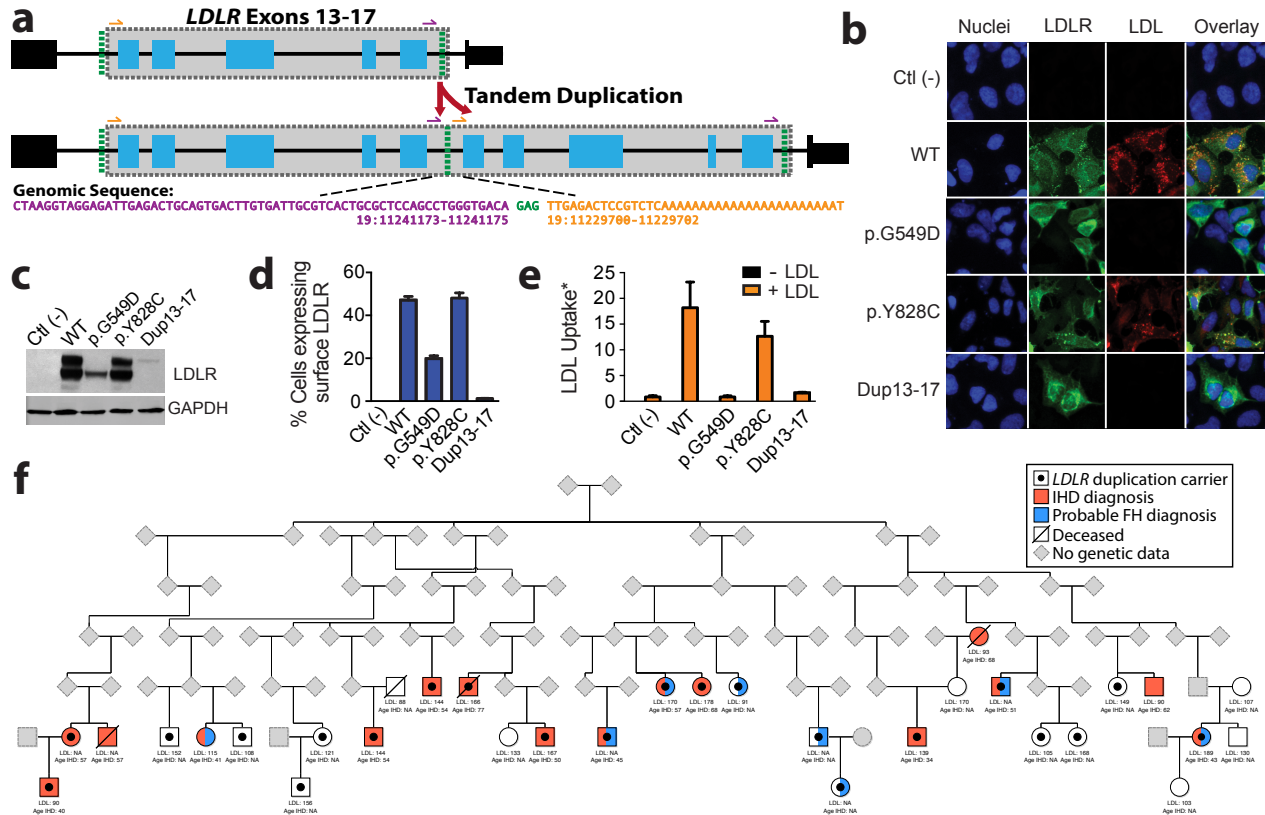


455

456 **Figure 1: Exome-wide survey of high-confidence CNVs from 47,349 individuals.**

457 **A)** Mean length (95% confidence bands) for deletion and duplication loci at varying allele
458 frequency ranges. **B)** Sample-wise distribution of CNV count (**Supplementary Table S2**). **C)**
459 Comparison of CNV intolerance in genes relative to loss-of-function (LOF) intolerance
460 probabilities computed from single nucleotide variants (ExAC pLI metric v0.3³⁰). While
461 duplications are observed consistently in ~60% of genes regardless of LOF intolerance,
462 observed deletion frequencies decrease in concordance with LOF intolerance.

463



464

465 **Figure 2: A familial hypercholesterolemia-associated tandem duplication within *LDLR***

466 **causes loss-of-function and segregates with high LDL cholesterol and heart disease**

467 **A)** Whole-genome sequencing and Sanger validation (**Supplementary Fig. S6**) confirmed the

468 *LDLR* duplication in 29 individuals occurs in tandem encompassing exons 13-17 (“Dup13-17”).

469 Both the breakpoint and insertion loci occur in intronic *Alu* repeat sequences, having a shared

470 three nucleotide microhomology (green). The predicted protein translation is in-frame, creating a

471 larger mutant receptor (**C**) that is not expressed on the cellular surface (**B & D, Supplementary**

472 **Fig. S13**), resulting in significantly reduced LDL uptake activity and loss-of-function (**B & E**). **F)**

473 Reconstructed pedigree estimate containing 22/29 carriers of the duplication and ten unaffected

474 related (first or second degree) individuals from the sequenced cohort. Five additional carriers

475 (not drawn) are included in this pedigree (**Supplementary Methods**). Elevated LDL and total

476 cholesterol as well as increased prevalence of coronary artery disease and early-onset ischemic

477 heart disease (“Age IHD” <55 for males and <65 for females) segregate with duplication

478 carriers; but only eight carriers have a probable FH diagnosis coded in their EHRs

479 (Supplementary Methods). *LDL Uptake = Cytoplasmic LDL Puncti / # GFP+ Cells.

480 **Tables**

| Phenotype | MIM # | Locus | Primary Gene | CNV | Observed Frequency (N Carriers) | Estimated Population Rate |
|---|-------------------|---------|--------------|---------|---------------------------------|---------------------------|
| Autosomal dominant or sporadic Mendelian CNVs | | | | | | |
| Adult-onset leukodystrophy | 169500 | 5q23.2 | <i>LMNB1</i> | DUP | 2.1x10 ⁻⁵ (1) | 1 in 47,349 |
| Spinocerebellar ataxia type 20 | 608687 | 11q12 | | DUP | 0.0 (0) | < 1 in 47,349 |
| Hereditary Neuropathy with Liability to Pressure Palsies (HNPP) | 162500 | 17p12 | <i>PMP22</i> | DEL | 5.5x10 ⁻⁴ (26) | 1 in 2,255 |
| Charcot-Marie-Tooth Disease Type 1A (CMT1A) | 118220 | 17p12 | <i>PMP22</i> | DUP | 4.4x10 ⁻⁴ (21) | 1 in 2,785 |
| DGS/VCFS | 188400/ 192430 | 22q11.2 | <i>TBX1</i> | DEL | 2.5x10 ⁻⁴ (12) | 1 in 3,946 |
| Microduplication 22q11.2 | 608363 | 22q11.2 | | DUP | 0.00139 (66) | 1 in 717 |
| Autosomal recessive Mendelian CNVs | | | | | | |
| Gaucher disease | 230800 | 1q21 | <i>GBA</i> | DEL | 2.1x10 ⁻⁵ (1) | 1 in 47,349 |
| Familial juvenile nephronophthisis | 256100 | 2q13 | <i>NPHP1</i> | DEL | 0.00682 (323) | 1 in 147 |
| beta-thalassemia | 141900 | 11p15 | <i>HBB</i> | DEL | 4.2x10 ⁻⁵ (2) | 1 in 23,675 |
| alpha-thalassemia | 141750 | 16p13.3 | <i>HBA</i> | DEL | 4.2x10 ⁻⁵ (2) | 1 in 23,675 |
| Pituitary dwarfism | 262400 | 17q24 | <i>GH1</i> | DEL | 2.1x10 ⁻⁵ (1) | 1 in 47,349 |
| X-linked Mendelian CNVs | | | | | | |
| Ichthyosis | 308100 | Xp22.31 | <i>STS</i> | DEL | 0.00109 (52) | 1 in 911 |
| Intellectual disability | 300706 | Xp11.22 | <i>HUWE1</i> | DUP | 5.9x10 ⁻⁴ (28) | 1 in 1,691 |
| Pelizaeus-Merzbacher disease | 312080 | Xq22.2 | <i>PLP1</i> | DEL/DUP | 0.0 (0) | < 1 in 47,349 |
| Hemophilia A | 306700 | Xq28 | <i>F8</i> | DEL | 0.0 (0) | < 1 in 47,349 |
| Hunter syndrome | 309900 | Xq28 | <i>IDS</i> | DEL | 0.0 (0) | < 1 in 47,349 |
| Progressive neurological symptoms (MR+SZ) | 300260 | Xq28 | <i>MECP2</i> | DUP | 2.1x10 ⁻⁵ (1) | 1 in 47,349 |

481 **Table 1: Observed frequencies of select known disease-associated CNV loci.**

482

483

| CNV Locus | | | | Allele | Beta | Beta | Standard | |
|-----------------------------|------|---------------|-------|-----------------------|--------|---------|----------|-----------------------|
| Coordinates | Type | Genes | Trait | Frequency | (LMM) | (mg/dl) | Error | P-value |
| (hg19) | | | | | | | (LMM) | |
| chr19:11230767- 11241993 | DUP | <i>LDLR</i> | LDL | 2.53×10^{-4} | 1.734 | 76.17 | 0.234 | 1.3×10^{-13} |
| chr19:11230767- 11241993 | DUP | <i>LDLR</i> | TCHOL | 2.53×10^{-4} | 1.384 | 60.87 | 0.235 | 3.8×10^{-9} |
| chr19:54801926- 54804607 | DEL | <i>LILRA3</i> | HDL | 1.71×10^{-1} | 0.052 | 0.65 | 0.010 | 4.5×10^{-7} |
| chr16:15125591- 16292040 | DUP | Many | LDL | 1.53×10^{-3} | -0.439 | -14.07 | 0.095 | 3.6×10^{-6} |

484 **Table 2: Exome-wide significant associations between high-confidence exonic CNV loci**
485 **and EHR-derived serum lipid traits (LDL-c, HDL-c, total cholesterol “TCHOL”, and**
486 **triglycerides). LMM = Linear Mixed Model (Methods).**