Sequencing individual genomes with recurrent deletions reveals allelic architecture and disease loci for autosomal recessive traits

Bo Yuan [1,2,10], Katharina Schulze[1], Nurit Assia Batzir [1], Jefferson Sinson [1], Hongzheng Dai[1,2], Wenmiao Zhu [1,2], Francia Bocanegra [3], Chin-To Fong [4], Jimmy Holder [5], Joanne Nguyen [6], Christian P. Schaaf [1,7], Yaping Yang [1,2], Weimin Bi [1,2], Christine Eng [1,2], Chad Shaw [1,2], James R. Lupski [1,5,8,9], Pengfei Liu [1,2]

[1] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
[2] Baylor Genetics, Houston, TX, USA
[3] Instituto de Referencia Andino, Bogotá, Colombia
[4] Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA
[5] Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA
[6] Department of Pediatrics, University of Texas Health Science Center, Houston, TX, USA
[7] Institute of Human Genetics, Heidelberg University, Heidelberg, Germany
[8] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX
[9] Texas Children's Hospital, Houston, TX
[10] Present Address: Seattle Children's Hospital, Seattle, WA.

[*]To whom correspondence should be addressed:
Pengfei Liu, PhD
Molecular and Human Genetics
Baylor College of Medicine
One Baylor Plaza
Houston, TX 77030
Phone: (713) 798-5122
Email: pengfeil@bcm.edu

Abstract

In medical genetics, discovery and characterization of disease contributory genes and alleles depends on genetic reasoning, study design, and patient ascertainment. We present novel genomic insights to enhance discovery in the challenging context of autosomal recessive (AR) traits and bi-allelic variation. We demonstrate computationally that new mutation mediated by nonallelic homologous recombination (NAHR), involving recurrent deletions at 30 genomic regions, likely drives recessive disease burden for over 70% of loci within these segmental deletions or 2% of loci genome wide. Meta-analyses of literature-reported patients suggest that NAHR-deletions are depleted from the ascertained pool of AR trait alleles. Exome reanalysis on subjects harboring recurrent deletions uncovered new disease-contributing variants in genes including *COX10*, *ERCC6*, *PRRT2* and *OTUD7A*. Our data implicate that genomic sequencing of personal genomes with NAHR-deletions could dramatically foment allele and gene discovery, enhance clinical molecular diagnosis, and could potentially enable human haploid genetics screens.

Word count: 146

Introduction

Over the last few decades, efforts to decipher molecular and genetic mechanisms underlying Mendelian disorders have repeatedly demonstrated that mutations aggregate in personal genomes and lead to human diseases in a continuum of allelic modalities, ranging from monoallelic (dominant), biallelic (recessive), triallelic, to multiallelic and more complex modes of inheritance.[1] Recent large-scale family-based genomic studies using exome sequencing (ES) have uncovered hundreds of new disease loci, with the majority following traditional Mendelian inheritance, i.e., monoallelic (AD) or biallelic (AR) trait segregation.[2] Although optimism has been increasing towards achieving disease annotation for a substantial portion of the haploinsufficient part of the human genome through dominant disease gene discoveries, statistical analysis from rare disease cohort studies suggests that the trajectory to understand the recessive or biallelic-disease-causing portion of the human genome is less certain.[3]

This apparent discrepancy in discovery between dominant and recessive trait genes can perhaps most parsimoniously be explained by the Clan Genomics Model.[1] The model predicts that dominant diseases are largely caused by emergence of new alleles in recent generations, *i.e. de novo* mutations (DNMs) including genomic changes reflecting structural variant mutagenesis. Differently, recessive diseases arise when a pair of disease alleles at a locus – whether existing in a population or *de novo* - are aggregated in a personal genome in the *trans* configuration by transmission genetics and products of a mating. Individual recessive alleles can emerge and be carried in populations, clans and pedigrees with no impact on phenotype. Assuming a dominant disease allele emerges at a rate of 10^{-8} per locus per generation, the DNM rate at a human locus,

a recessive disease trait with comparable incidence would require at least one of the two alleles of greater than ~10^{-4} population allele frequency for disease ascertainment to be feasible.

Disease causing variants with such high allele frequencies tend to be found in established recessive disorders, often attributed to founder mutation alleles.[4] In contrast, if a yet-to-be-defined recessive disease trait gene does not have pathogenic alleles represented at a sufficient population frequency, disease discovery and annotation of the gene would be greatly hampered due to the extremely low incidence and difficulty in ascertaining affected individuals, even when considering a world-wide population of 7.8 billion. An exception to this rule occurs when disease gene discovery research is conducted in populations with an elevated coefficient of consanguinity and autozygosity. In this circumstance, the allele pool shrinks to the Clan of the patient's extended family, which dramatically escalates the effective disease-allele frequency.[4; 5] Thus, the probability of ascertaining patients with the recessive disorder increases exponentially, even when all existing alleles are ultra-rare in sampling the general population.[6]

Here, we present a previously under-recognized resource of disease alleles to enable discovery of additional alleles for recessive trait disorders. This under-recognized resource are loss-of-function (LoF) alleles caused by large recurrent genomic deletions rendering a locus haploinsufficient. Recurrent genomic deletions are a subset of contiguous gene deletions that are characterized by a special type of mutational mechanism called nonallelic homologous recombination (NAHR).[7] NAHR is mediated by ectopic recombination between highly similar repeat sequences termed low-copy repeats (LCRs) or segmental duplications (SDs).[8] The human

genome is evolutionarily structured to be highly enriched for SDs, which creates a large number of architectural hotspots for genomic disorders and mirror traits to emerge.[9; 10]

It has been shown that the mutation rate of NAHR at a given locus can be as high as ~10^{-4} to 10^{-5}, which is orders of magnitudes higher than new mutation rates from single nucleotide variants and indels (collectively referred to as small variants).[11] The high new mutation rate ensures that these genomic deletions continuously arise *de novo* in the human population among unrelated individuals.[12; 13] This 'recurrent' nature of genomic deletions distinguishes them from the other recessive disease alleles that are more 'stationary' or artifacts of past population history amplified by recent population expansion. Moreover, as genotyping assays are performed on relatives of patients with deletions as well as on individuals without a disease indication, we recognize that many recurrent genomic deletions are incompletely penetrant,[14-21] i.e., the fitness of the deletion allele can be high, at least in certain genomic backgrounds.[22; 23] These findings support the contention that recurrent genomic deletions have two key properties relevant to recessive disease gene discovery: first, they act as highly prevalent recessive alleles in comparison to other small variant recessive disease alleles and second they continue to renew because of high mutation rates of structural variant mutagenesis. We hypothesized that because of these two attributes, population prevalence and new mutation rate, NAHR alleles are contributing to a considerable fraction of recessive diseases and are perhaps among the most important alleles at these recessive trait loci.

To pursue this hypothesis, we formulated a large computational analysis using both genome-wide population data resources as well as clinical genomics information. We demonstrate, using

computational and sequencing analyses, that recurrent genomic deletions contribute to a major fraction of individual disease burden for over 2% of known recessive trait genes or 70% of known recessive diseases in regions encompassed by LCRs. This finding regarding allelic architecture of diseases, leveraging the high incidence of recurrent genomic deletion alleles, can prime powerful future strategies for recessive disease gene and allele discoveries exploiting the concept of 'human haploid genetics' by investigating genomic intervals of recurrent segmental aneusomy.

Results

NAHR loci studies reveal 30 deletions commonly recognized in clinical populations

In order to systematically evaluate the contribution of recurrent genomic deletions to autosomal recessive conditions, we first mapped all possible loci that are susceptible to recurrent deletions caused by NAHR between directly oriented SDs[24] using the GRCh38 human reference genome sequence (Figure S1, Table S1). The collapsed NAHR map contains 717 unique recurrent deletion regions. We enumerated the subset of recurrent deletion events with available data from screening efforts in the literature or clinical testing to substantiate a prevalence estimate, and focused the subsequent analyses on these regions (n=51). We identified 30 deletions with a maximal population prevalence over 1/1,000,000 based on estimates from the UK Biobank, the Icelandic, and the gnomAD SV studies [15-17] (Table 1, Table S2). These 30 deletions span 64 Mb of unique genomic sequences, contribute to an aggregate population allele burden of 1.3%, and encompass 1517 genes, of which 78 are known to cause recessive disorders. An additional 21 deletions, with populational prevalence possibly lower than 1/1,000,000, are also identified to recur in high prevalence if a clinical cohort is ascertained (Table S2). With the 21 ultra-rare

deletions included, the span of genomic coverage increases to 82 Mb; the number of genes involved becomes 1828, with 100 representing established recessive disease trait genes.

NAHR-deletion: the most prevalent allele for interval mapped recessive trait genes

We then catalogued, based on existing knowledge and datasets, a compendium of all reported and predicted carrier alleles for each known recessive trait gene in the human genome. Our objective for this recessive allele catalog is to estimate and dissect the impact of recurrent genomic deletions' contribution to the overall disease burden. Based on mode of inheritance curations from OMIM, DECIPHER, and ClinGen, a totality of 2657 recessive disease trait genes were assembled. The carrier allele burden for each 'recessive trait gene' was calculated by summing up frequencies of unique alleles for all high-quality pathogenic variants from ClinVar, all structural variants (SV) predicted to be LoF from gnomAD SV v2.1, all high-confidence LoF small variants identified in gnomAD v3.1, and the NAHR-mediated recurrent genomic deletions, if applicable. An aggregate of 85,063 small variant and large deletion carrier alleles were identified for the 2657 genes (Table S3). For the 78 known recessive genes in the NAHR deletion regions, the number of per gene pathogenic alleles range from 1 to 308, with a median of 14. As a comparison, the remaining 2510 known recessive genes have a similar median per gene pathogenic allele count, 13, but a wider range, from 1 to 3562.

A drawback of this calculation is that SNV pathogenic missense, in-frame indel, or intronic variants not currently reported in ClinVar are inadvertently omitted. However, we argue that carrier alleles not represented in ClinVar tend to have lower allele frequencies and thus do not have a major impact on the subsequent carrier burden estimates. We further argue that the alleles

that receive an entry and curation in ClinVar have higher frequencies – and therefore greater impact on recessive disease, and these are the alleles more easily ascertained in screening tests of clinical diagnostic laboratories. This latter contention is supported by the aggregate gene-level carrier allele burden from our analysis matching empirical experience in genetic testing carrier screenings results (Table S4).[25]

Nevertheless, to account for potential unrepresented alleles from recessive genes that have not been scrutinized by large-scale systematic clinical or research screening, in the subsequent analyses, we supplemented the disease allele pool for each gene with a 10% extra variant load, comprised of ten hypothetical variants each accounting for 1% of the overall carrier burden (see Methods) for each gene. Of note, the NAHR deletion alleles rank as the most (47/78) or second most (12/78) frequent (highest population allele frequency) carrier alleles together comprising over three quarters of known recessive genes within NAHR regions! Even with the above-mentioned conservative padding to represent ten hypothetical alleles not yet ascertained, the NAHR alleles still contribute to greater than 10% of the total gene-level carrier allele burden for 58 of the 78 genes (Table S5).

NAHR-deletions contribute a major fraction of recessive disease load

It is important to note that, for a recessive trait, the population frequency and relativized frequency of a particular allele from a pool of alleles (**Fa**, fraction of allele burden) is not linearly correlated with the probability of sampling a patient with the specific allele from a group of patients (**Fd**, fraction of the disease burden). The distribution of alleles in affected individuals is determined by the pairwise allele frequency products in a pool. Thus, we calculated allelic

contributions to recessive disease load using an *n* x *n* Punnett square, where *n* is the number of carrier alleles for a recessive trait gene. The calculated NAHR-deletion contribution to disease can be calculated from the matrix. We denote as Fd the modeled probability of sampling one affected individuals carrying at least one recurrent deletion allele from a pool of patients affected with the recessive condition caused by the same gene. We empirically considered a gene to be under significant NAHR-deletion burden for population prevalence of the associated recessive disease, if the recurrent genomic deletion is expected in greater than 20% of all patients with this recessive disorder. By this definition, 71% (56/78) of NAHR-region recessive genes, which account for 2.11% of all known recessive genes, are under significant NAHR-deletion burden for recessive disease prevalence (Table 2)! In the context of the other alleles from the same gene, the disease contribution of the NAHR-deletion (Fd) ranks at the top for 47 genes, and at the second place for 12 more genes. The Fd scores of the top-3 alleles are listed in Table 2 and Table S5 to illustrate a snapshot of the allelic architecture for each recessive trait gene.

We next defined a log-scaled index we termed the NAHR-deletion's Impact to Recessive Disease (NIRD), to depict the gene-level disease load contribution of the NAHR allele relative to an allele with a median level of contribution to the same gene among all population carrier alleles (See Methods section). A positive NIRD score predicts that the NAHR-deletion allele plays a predominant (above the typical allele) role among all carrier alleles of the gene in disease contribution, whereas a negative score predicts a minor (below typical) role. Known recessive genes in the recurrent deletion region tend to have high NIRD scores, with 85% (66/78) scoring above 0, and 74% (58/78) scoring above 2. The two highest NIRD scores are found in *RBM8A*

and *NPHP1*, 9.8 and 7.7, respectively. Both are extremely large values considering the NIRD is log-scaled.

To appreciate the properties of the NIRD scores, we adjusted the algorithm to calculate the disease contribution of any given pathogenic allele for a recessive gene, as an Allelic Impact to Recessive Disease (AIRD) score. The most common carrier allele observed in cystic fibrosis, NM_000492.3(*CFTR*):c.1521_1523delCTT (p.Phe508delPhe), also known as the ΔF508 allele, has an AIRD score of 7.3; the third most common carrier allele for Niemann-Pick disease type A, NM_000543.5(*SMPD1*):c.996del (p.Phe333fs), has an AIRD score of 2; a well-known founder mutation observed in ~10% of patients of Ashkenazi Jewish descent with Tay-Sachs disease, NM_000520.6(*HEXA*):c.1421+1G>C, has an AIRD score of -0.15, due to its lower allele frequency of 1.97×10^{-5} in the general population according to gnomAD v3.1.

<u>Meta-analysis suggests under-representation of the NAHR-deletion allele in currently discovered recessive disease allele pools</u>

Based on the NIRD and related findings, we hypothesized two consequences. First, for the ~2% of known human recessive genes genome-wide or 70% of recessive genes in NAHR regions, one of the most effective but under-utilized strategies for identifying novel disease-causing alleles from human subjects is to sequence affected individuals carrying the heterozygous recurrent genomic deletion encompassing the gene of interest. Second, there likely exist uncharacterized recessive disease genes that may be most effectively identified by sequencing individuals bearing prevalent recurrent genomic deletions -- i.e. any of the 1439 genes within these deletion regions

may have yet to be assigned an autosomal recessive disease trait and could be novel biallelic/recessive disease trait genes.

To test these hypotheses, we analyzed the distributions of a near-complete catalogue of currently discovered disease alleles in 170 patient families affected with one of the recessive disorders whose carrier burden are predicted to be almost exclusively from NAHR-mediated large deletions (Fd > 70% from Table 2). The cohorts are assembled by meta-analysis of all literature reports for patients with the corresponding recessive disorder recorded in HGMD (version 2020.4). *NPHP1*, the top-ranking gene from Table 2, was not included because it is a well-characterized disease gene and therefore literature reported patients may not represent the natural disease allele composition in the world.

It is expected that all patients with biallelic disease variants fall into three categories (1) HMZ: those affected with homozygous small variants possibly from a close- or distant- consanguineous relationship, (2) SNV+SNV: those affected with compound heterozygous small variants, and (3) NAHR+SNV: those affected with a large deletion in *trans* with a small variant. We anticipate that category #1-HMZ accounts for a substantial proportion, demonstrating the well-established robustness of autozygosity mapping as a method for allelic and new recessive gene discovery (as populational rare alleles can be escalated to much higher clan allele frequency).[6] In outbred pedigrees and populations corresponding to categories #2-SNV+SNV and #3-NAHR+SNV, our modeling from the NIRD hypothesis is that #3-NAHR+SNV should account for a higher fraction. The opposing trend would suggest that our current disease gene/allele discovery efforts are not

exploiting the large deletion allele to the fullest extent that a 'human haploid genetics' approach might allow.

The results for *RBM8A* follows a distinctive pattern from the rest of the genes (Table S6). The majority (95%, 73/77) of patients affected with the *RBM8A*- thrombocytopenia absent radius (TAR) syndrome are compound heterozygous for the 1q21.1-TAR deletion and a small variant; no patients with homozygous *RBM8A* pathogenic variants have been reported.[26] This pattern favoring #3-NAHR-SNV is consistent with our computational modeling based on the allele spectra (Table S6), and it is expected because of two characteristics of this gene. First, the NAHR-driven 1q21.1-TAR deletion contributes to an overwhelming fraction of highly penetrant disease alleles for *RBM8A* (Table 2). Second, three hypomorphic disease variants that are observed in most patients with TAR are common in the population, with allele frequencies of 1.79%, 1.74% and 0.57%. These hypomorphic alleles are not disease-causing as homozygotes or in *trans* with each other, but only trigger disease pathogenesis in combination with a highly penetrant allele such as the null allele generated by NAHR-deletion. As a consequence, the *RBM8A* disease alleles that dominate both sets of chromosome 1 homologues are not observed as homozygotes in patients – these homozygotes are either lethal (NAHR-deletion) or non-pathogenic (hypomorphic small variants). The unique molecular allele architecture of *RBM8A* makes it proportionally much more likely to ascertain patients with #3-NAHR+SNV alleles than patients with #1-HMZ or #2-SNV+SNV, regardless of any study design, which possibly led researchers investigating this clinically uniform and genetically homogeneous disorder to focus on the recurrent 1q21.1 deletions in the initial discovery and subsequent characterizations and even confused the initial genetic model.[27]

In contrast, for the remaining genes comprised of data from 93 families, the #3-NAHR+SNV configuration is poorly represented (Table S6). Note that since most of the variants reported in these families are only documented once in affected human subjects, we cannot rule out the possibility that some of these variants are not causative to the clinical presentation. Almost 70% (64/93) of these patients carry homozygous disease alleles. Based on our modeling and the assumption of random mating, patients with homozygous variants are expected to account for a small fraction of the overall cohort, ranging from 0.53% to 5.4% per gene. However, the observed fractions of homozygotes for each gene are 1.9 to 189 (median 47) folds higher than expected. Furthermore, all of the 65 homozygous variants are rare, with 42 being ultra-rare (not observed in gnomAD v3.1). The collective patterns suggest that current efforts investigating these recessive traits tend to ascertain patients from populations with elevated autozygosity or from targeted population groups with their ethnic-specific disease founder alleles.

To avoid potential confounding factors from study designs and patient ascertainment methods, we removed patients with homozygous variants and focused on those with compound heterozygous variants. Our modeling predicts that for these top-ranking genes analyzed, the number of patients with NAHR+SNV should be 2.8 to 20 (median 12) folds higher than the number of patients with compound heterozygous small variants (Table S6). The observed counts from many individual genes are too low to support a meaningful conclusion, but in aggregate, we have identified fewer patients with NAHR+SNV (13) compared with patients with compound heterozygous small variants (16). The recurrent deletions involved are 16p13.11 (n=4), 17p12-HNPP (n=3), DiGeorge 22q11.2 (n=2), 10q11.21q11.23 (n=2), 1q21.1-TAR (n=1), and proximal

16p11.2 (n=1). The poor representation of deletion bearing patients shows a bias that under-represents category #3-NAHR+SNV and deviates from the expectation driven by our analysis using population allele frequencies and the NIRD score.

A human haploid genetics and genomics approach to recessive trait genes

To test the hypotheses delineated above concerning the unrealized discovery potential among individuals bearing genomic segmental deletion alleles, we analyzed clinical exome sequencing (cES) data for patients carrying a prevalent recurrent genomic deletion. The research subjects (patients) were identified from a cohort of 11,091 subjects who were referred for cES at a diagnostic laboratory due to a differential clinical diagnosis including various suspected genetic disorders. We performed an initial screen for patients carrying one of the genomic deletions from Table 1, which resulted in 161 subjects carrying one recurrent deletion and 3 subjects carrying two. The two most frequently observed types of deletions, the 15q11.2 BP1−BP2 deletion (n=41) and the *NPHP1*-2q13 deletion (n=23), are excluded from downstream analysis. This is because none of the coding genes from the 15q11.2 BP1−BP2 deletion have been implicated to be associated with Mendelian diseases,[28] and the critical gene at 2q13, *NPHP1*, has been already extensively studied.[29] We also excluded six subjects harboring the X-linked, hemizygous deletion in the Xp22.31 *STS* locus. After excluding these three groups of deletions, we focused the analysis on the remaining 95 subjects collectively harboring 96 incidences or 26 types of recurrent genomic deletions (Table S7).

Unexpectedly, more than a quarter (26/95) of these subjects were found to have probable small variant molecular diagnostic findings independent from the deletion. This observation suggests

that the cohort of patients studied here are enriched for individuals presenting phenotypes atypical of those associated with the recurrent deletion and perhaps have a 'blended phenotype'.[30] From the remaining 69 subjects with an apparent undiagnostic ES result, we identified 4 subjects with rare variants in coding regions exposed by the deletion as potential molecular diagnoses. The four cases harbored recurrent deletions of 10q11.21q11.23, 17p12, 16p11.2 and 15q13.3, respectively, all with documented incomplete penetrance. The rare variants uncovered by the 10q11.21q11.23 deletion and the 17p12 deletions illustrate the concept of new disease-causing allele identification and characterization, whereas the two cases with the 15q13.3 and the 16p11.2 deletions exemplify the concept of new disease gene/mechanism characterization (Table 3).

Subject #1 has clinical features including ataxia, developmental delay, microcephaly, and short stature. A recurrent 10q11.21q11.23 deletion[31] was identified in this patient. In *trans* with the deletion, a novel missense variant c.1490T>C (p.F497S) in the *ERCC6* gene was identified, which is associated with cerebro-oculo-facio-skeletal syndrome 1 (COFS1, MIM# 214150) or Cockayne syndrome type B (CSB, MIM# 133540). The high allele frequency of the 10q11.21q11.23 deletion (1.412×10^{-4}) increases the probability for a second allele with ultra-low frequency, like the c.1490T>C (p.F497S) *ERCC6* variant, to be correlated with a set of human clinical phenotypes.

Subject #2 presents with a suspected mitochondrial disorder. The patient carries the 17p12 recurrent deletion, associated with hereditary neuropathy with liability to pressure palsies (HNPP, OMIM# 162500), a mild form of peripheral neuropathy due to decreased dosage of the *PMP22*

gene; often only manifested clinically as multifocal neuropathy elicited after sustained trauma to a peripheral nerve that traverses close to the body surface and manifest an entrapment neuropathy[32] or an operative carpal tunnel syndrome co -segregating through multiple generations.[19] We identified a hemizygous variant resulting in an in-frame small duplication of two amino acids, c.1277_1282dup (p.M426_L427dup) in exon 7 of *COX10* (Figure 1). In retrospect, the identification of the first recessive *PMP22* variant allele, T118M, as a compound heterozygote with the HNPP deletion most probably represents an example of such 'haploid genetics' discovery research.[33]

The *COX10* gene, defects of which cause mitochondrial complex IV deficiency (OMIM# 220110) inherited in an autosomal recessive manner, is the only gene in the deletion interval with a known disease association other than *PMP22*. The gene body of *COX10* spans the CMT1A-REP, the repeat sequence mediating the recurrent 17p12 deletion (Figure 1).[34-38] The exon 7 of *COX10* is embedded inside the deletion, and thereby is always deleted in patients with the recurrent 17p12 deletion. The exon 6, located in the CMT1A-REP, may also be deleted in the 17p12 deletion event, depending on the breakpoint of the deletion in the CMT1A-REP. *COX10* is interrupted with HNPP deletion, but two intact wild type copies remain after NAHR reciprocal CMT1A duplication.[11]

Based on our calculation, ~77% of all patients affected with biallelic *COX10* pathogenic alleles in an outbred population carry one HNPP deletion (Table 2). To independently test this hypothesis, we retrospectively investigated results from 596 patients suspected with a mitochondrial disorder who were clinically tested for *COX10* coding region sequencing and

deletion/duplication analysis. The distribution of pathogenic alleles from this cohort is likely to represent that from the entire collection of worldwide *COX10* patients because of the clinical indication-based ascertainment method. Only one patient received a possible diagnostic finding. The patient (subject #3), whose referral indication is COX deficiency, has a rare VUS c.858G>T (p.W286C) in *COX10* in *trans* with the HNPP deletion! Two additional patients were found to have the heterozygous HNPP deletion, but a second hit in *COX10* was not found, although we cannot rule out the possibility of additional findings in intronic or regulatory regions. These findings, though under-powered due to the small sample size, are consistent with our prediction that most patients with COX10 deficiency carry one HNPP deletion allele; either *de novo* or inherited. Considering the high frequency of the HNPP susceptibility allele with absence of selection and late adult onset disease,[39] it is possible that more novel *COX10* disease alleles can be revealed by sequencing individuals with the HNPP deletion and a mitochondrial spectrum of clinical phenotypes, thereby improving our understanding of the biological function of the *COX10* gene.

Subject #4 presented with severe neurodevelopmental diseases and dysmorphic features. We identified a hemizygous *OTUD7A* frameshift variant c.2023_2066del (p.D675Hfs*188) in *trans* with the recurrent 15q13.3 BP4-BP5 deletion, providing evidence for *OTUD7A* as a new disease gene. The recurrent deletion mediated by BP4 and BP5 at the 15q13.3 locus is associated with highly variable phenotypes, ranging from asymptomatic to mild to moderate intellectual disability, epilepsy, behavioral problems, and variable dysmorphic features.[40; 41] While heterozygous deletion causes highly variable phenotypes, reported homozygous 15q13.3 BP4-BP5 deletion consistently manifest disease phenotypes including significant neurodevelopmental

disorders, epilepsy, hypotonia, visual impairments, and other less common phenotypes including autism spectrum disorder, short stature, failure to thrive, microcephaly and variable dysmorphic features (Table S8).[42-46] The critical gene responsible for the clinical presentations of the 15q13.3 BP4-BP5 deletion has been debated, but evidence suggests that *OTUD7A*, encoding a member of a family of deubiquitinating enzymes, may be a plausible candidate.[47; 48] Studies using syntenic heterozygous deletion mouse models suggested a critical role of Otud7a in neuronal development and brain function.[47; 48] Otud7a-null mouse models manifest many cardinal features of the 15q13.3 deletion syndrome.[47] The c.2023_2066del (p.D675Hfs*188) variant identified in subject #4 is located at the last exon of the *OTUD7A* gene, and is thus predicted to not lead to nonsense-mediated decay. However, the variant is predicted to result in substitution of the C-terminal amino acids after aspartic acid with 187 new amino acids and a premature termination of the protein translation. This change may remove the C-terminal Zinc finger A20-type domain and abolish the normal function of the protein. Our finding in Subject #4, together with recent case reports of patients with a homozygous missense *OTUD7A* variant[49] or compound heterozygous 15q13.3 deletion in *trans* with a frameshift *OTUD7A* variant,[50] corroborates that *OTUD7A* may be the critical 'driver gene' in the 15q13.3 deletion syndrome, at least in part through a biallelic autosomal recessive disease trait mechanism. Interestingly, the populational allele pool for *OTUD7A* is depleted for LoFalleles (not considering 15q13.3 deletions), according to estimates from gnomAD (pLI=0.95). Without the 15q13.3 deletion contributing to a major carrier burden, the paucity of small variant disease alleles for *OTUD7A* would make disease association establishment using patient data much more challenging.

In Subject #5 with severe neurodevelopmental disorders, we identified a c.649dup (p.R217fs) pathogenic variant in the *PRRT2* gene in *trans* with the recurrent 16p11.2 BP4-BP5 deletion, providing evidence for a novel disease inheritance mechanism for *PRRT2*. The 16p11.2 BP4-BP5 recurrent deletion is known to be associated with mild dysmorphisms, macrocephaly, and neuropsychiatric phenotypes including DD/ID and autism spectrum disorder (ASD) with incomplete penetrance.[51; 52]. The *PRRT2* gene is highly expressed in mouse brain and spinal cord during early embryonic development.[53] Heterozygous LoFvariants in *PRRT2* cause movement and seizure disorders including familial infantile convulsions with paroxysmal choreoathetosis (OMIM# 602066), episodic kinesigenic dyskinesia 1 (EKD1, OMIM# 128200), or benign familial infantile seizures 2 (BFIS2, OMIM# 605751), with incomplete penetrance documented.[54] The c.649dup (p.R217fs) allele is the most frequent pathogenic variant, occurring at a mutational hotspot with homopolymer of 9 cytosine bases adjacent to 4 guanine bases that are susceptible to DNA replication errors.[55] Currently, autosomal dominant (AD) is considered as the only disease inheritance mode for *PRRT2* in OMIM, although preliminary evidence from case reports suggest that *PRRT2* can cause a more severe neurodevelopmental disorder through a biallelic pathogenic mechanism and an autosomal recessive inheritance model.[56] Our findings in Subject #5 provide further support for a new disease type and inheritance mechanism for *PRRT2*; it may also pinpoint to a potential model that explains penetrance of certain neurological phenotypes observed in patients with the 16p11.2 deletion; a similar biallelic model underlies the penetrance of ~10-12% of congenital scoliosis.[57]

<u>NAHR-deletions contribute to recessive disease burden in population-specific patterns</u>

As suggested earlier, the contribution of a given allele to recessive disease burden is influenced by the composition of other pathogenic alleles from the same gene. Although the field is starting to appreciate inter-individual variabilities in NAHR rates associated with alternative genomic structural haplotypes[29; 58] as well as polymorphisms from *trans* acting factors controlling homologous recombination, such as *PRDM9*,[12] we currently still assume that NAHR mutation rates at a given locus are relatively constant across different ethnic backgrounds. This potentially leaves the remaining alleles, the small variants, as the major driver for any variability in allelic architecture from different population groups. To investigate the degree of inter-population variability for small variant recessive alleles, we used ethnic information from gnomAD, and conducted the modeling described earlier for four population groups, African (AFR), Latino (AMR), East Asian (EAS) and European (EUR) (Table S5). Population specific NIRD scores are compared with the general population to generate ΔNIRD (Figure 2), which can be used to inform the relative odds ratio for NAHR-deletions in recessive disease in the specific population. This analysis provides preliminary computational confirmation for the suspected population variability in NIRD, which implicates that the precision of NIRD can be improved by tuning the model with population specific allelic architecture. In light of these findings, we tentatively propose that a prioritization strategy based on prior knowledge of population allele frequency spectra can be applied to enhance discovery in research study design of genomic sequencing among individuals with large recurrent deletions. We cautiously note that some of the population groups analyzed here may not have a sufficient sample size to allow a complete representation of disease alleles of relatively lower frequency, which may result in overestimation of ΔNIRD when the score is positive. Additional population specific allele frequency data are warranted to improve the accuracy of this analysis.

Discussion

We cataloged the population frequency spectra of pathogenic alleles across all known recessive diseases. Our analyses highlighted the high frequency of SV mutagenesis and NAHR deletion alleles in genomic instability regions susceptible to this structural rearrangement.[59] We subsequently computationally interrogated the impact of allele frequency distributions on recessive disease burden using a modified Punnett square. We used this computational and genetic approach to consider the impact of the NAHR allele on recessive disease traits relative to other recessive variant alleles. Our findings indicate the potentially dramatic disease impact of these NAHR alleles.

Considering current clinical genomics practice and research investigations, we postulate that the role of recurrent genomic deletions contributing to recessive diseases is under-appreciated, therefore potentially impeding discovery of new disease genes and alleles. Traditionally, large genomic deletions have been considered to cause dominant disease traits, primarily by a halpoinsufficiency mechanism, with the driver dosage sensitive gene(s) mapping within the deleted interval. This haploinsufficiency assumption likely arose because these alleles were almost all identified as heterozygotes by screening in symptomatic cohorts and frequently as *de novo* CNV mutations. Our analysis of population frequency implicates that the heterozygous state is often not disease causing, and therefore that these variants act as contributors to recessive disease when they occur in *trans* to a disease allele.

Although many recurrent deletions contribute considerable carrier allele burden to individual recessive Mendelian disease, these deletions are often large enough to include other genes whose homozygous depletions are incompatible with live birth. Two exceptions are the 2q13-*NPHP1* deletion and the *SMN1* deletion (the *SMN1* deletion is predicted by our analysis to be the most frequent NAHR-mediated deletion as shown in Table S1). The fact that other large recurrent deletions are almost never observed in patients as homozygous losses may have led investigators to overlook their equally important role in contributing to recessive disease as compound heterozygotes.[60] Taken in this context, it is also notable how often homozygous duplication at the DiGeorge/Velo-Cardi-Facial-syndrome locus results from one *de novo* NAHR driven duplication allele.[61]

Clinically, there has not been a consensus on whether exome/genome sequencing should be pursued after a recurrent genomic deletion has been identified in a patient. Some may argue that, under the assumption that most Mendelian diseases are caused by one "unifying diagnosis", the identification of a large genomic deletion can be evidence to demote additional candidate molecular diagnoses in the same patient. We argue the opposite: that in patients with a recurrent "contiguous gene deletion syndrome", the possibility of revealing an "additional" recessive disease molecular diagnosis cannot be ignored.

Coincidentally, four of the five hemizygous small variants exposed by the deletion in Table 3 are located in repeat or difficult-to-sequence regions. The *COX10* and the *PRRT2* frameshift variants were incorrectly called as heterozygous changes by the original exome variant calling pipeline due to challenges of being indels and mappability issues. If extra care were not taken to look into

the deleted hemizygous region, these diagnostic variants would have easily been missed during analysis, which might be one of the reasons contributing to the proposed under-detection of recurrent deletion + small variant cases for recessive disorders.

Taken together, we suggest that the future of new recessive disease and allele discovery will greatly benefit from two approaches: (1) sequencing individuals with recurrent deletions and (2) sequencing in population of elevated autozygosity (Figure 3). The autozygosity mapping approach has been the classic method for new disease gene discovery in medical genetics. This approach allows a straightforward strategy to target patient populations (by assessing their degree of autozygosity from social and family histories) to assemble the appropriate cohorts for investigation. However, disease alleles revealed by this approach are usually limited to a specific population.

The NAHR-deletion sequencing approach, on the other hand, has the potential to assign clinical significance to alleles independent of ethnic backgrounds, partly owing to the recurrent and high mutation rate of the NAHR. Nevertheless, this approach requires prior knowledge and screening of individuals for the recurrent deletions. It may have been challenging to collect a cohort of patients with large recurrent deletions two decades ago, but advancements in clinical testing and populational screening have made such a genomic experimental effort feasible now, either from large diagnostic centers or from clinical registries.[17; 22; 28; 51; 52; 62; 63] Even when available subject numbers are limited for recurrent genomic deletions with extremely low penetrance, it is possible to tune the disease gene/allele characterization strategy by targeting specific phenotypes, as demonstrated at the Smith Magenis Syndrome - *MYO15* locus two decades ago.[64] While

researchers are starting to sequence cohorts of individuals with large deletions,[13; 65] it is imperative for clinical and diagnostic genomicists to foster guidelines that facilitate routine genomic sequencing (ES or WGS) on patients who are found to have recurrent genomic deletions, which will benefit both the patients and the research field. Given the great potential in the near future for disease gene discoveries within intervals of genomic deletions, these patients will benefit from routine or perhaps even more prioritized reanalysis of sequencing data.[66]

Performing sequencing on personal genomes with higher population prevalence (those from Table 1) carries additional long-term promise for clinical characterization of common variants, extending the current scope of focus on mono- or bi- allelic inheritance into the more complex spectrum of disease inheritance – including the compound inheritance gene dosage model.[18] High prevalence recurrent genomic deletions are often associated with a high degree of incomplete penetrance of disease phenotype, ranging from 10-90%.[14] The disease causal mechanism described in this study, for example, the 16p11.2 deletion + the *PRRT2* small variant leading to neurodevelopmental diseases, may explain a small portion of the previously-attributed missing heritability for the disease "penetrance" at 16p11.2. However, the totality of the missing heritability is likely not explained by the recessive model alone, because the observed disease penetrance is likely to dominate the aggregation of the recessive disease allele prevalence (individually rare) on the non-deleted chromosome. It is plausible that alternative disease models exist, in which the critical gene triggers disease presentation when one rare LoF allele is combined with one or a set of milder hypomorphic alleles with common population frequency. This compound inheritance model has been demonstrated at the *RBM8A*-1q21.1 locus in association with the TAR syndrome [26], the *TBX6*-16p11.2 locus in association with congenital

scoliosis[18], and the *F12*-5q35 Sotos deletion locus in association with blood clotting.[67] Our analysis in this study was focused on coding sequence changes. Moreover, the limited size of patient cohort ascertained for each recurrent deletion may decrease the power of identifying high frequency hypomorphic alleles. These will be dramatically empowered by genome sequencing in a larger patient cohort. Nevertheless, the unifying theme for both the strictly recessive model and the more complex compound inheritance gene dosage model is that large recurrent genomic deletions provide a unique perspective in characterization of new disease loci and alleles and the emerging field of human haploid genetics.

Methods

Construction of a genome-wide map for NAHR-mediated recurrent genomic deletions

Hotspots of recurrent deletions were characterized with overlapping intervals flanked by clustering SD pairs, which illustrates that distinctive mechanistic events can contribute to each of the same clinically recognized genomic disorders. SD pairs, for which intervening genomic segments are likely to be clinically interpreted as the same genomic deletion, were collapsed. Metrics that could inform estimation of the new mutation rates for each genomic disorder were kept for each SD elements from the merged cluster of repeats, including repeat lengths, distance in between, and sequence similarity.[68] Gene content of the deleted segment is expected to influence the fitness of this allele. We used the number of genes with a high pLI score, i.e. greater intolerant to haploinsufficiency, to estimate the level of selection to the genomic deletion in the population. These metrics were used to calculate a score to estimate the relative populational prevalence of these CNVs. Coordinates for deletion breakpoints are calculated as a weighted average of all the ranges of possible SDs that may mediate the deletion. Thus, the

coordinates are not precise predictions for a specific deletion observed in individual patients, but rather average of all possible types of deletions that were collapsed into the merged deletion. Two separate NAHR-deletion maps were generated using SegDup tracks from GRCh38 and GRCh37.

Prevalence curation for NAHR-mediated recurrent genomic deletions

For genomic deletions with a population prevalence over 1/1,000,000, prevalence estimates were calculated based on the UK Biobank cohort[17] and the Icelandic cohort[16]. If the prevalence estimate was not significantly different between these cohorts (Fisher's exact test, $p>0.05$), the estimate based on the UK Biobank data was taken. Otherwise, prevalence from a third cohort (gnomAD SV[15] or other region specific literature) was compared with estimates from the UK Biobank and the Icelandic cohorts, and the group with a closer match was taken. For genomic deletions with a prevalence lower than 1/1,000,000, we investigated a cohort of 33,452 patients who were referred for clinical Chromosomal Microarray Analysis (CMA) using custom designed Agilent oligo-based Comparative Genomic Hybridization arrays.[69] Of note, deletion prevalence estimates from the CMA cohort do not represent actual prevalences in the general population, but can inform relative prevalence comparison among rare events in the population.

Recessive disease carrier burden calculation

High-quality ClinVar variants were defined as having a pathogenic or likely pathogenic label with at least one-star review status (accessed 01/21/2021). LoF variants from gnomAD SV were defined by variants meeting all the following criteria (1) PASS filter in the VCF file with a quality score over 500, (2) PROTEIN_CODING__LOF flag or

PROTEIN_CODING__DUP_LOF flag in the VCF file, (3) POPMAX allele frequency lower than 1% and no homozygote counts, (4) less than 80% of the span overlapping with segmental duplications, and (6) the LoF consequence affects all RefSeq transcripts of a recessive gene. High-confidence LoF small variants from gnomAD v3.1 were defined by variants that fulfill all the following criteria: (1) PASS filter from the gnomAD v3.1 VCF file, (2) do not fall into a low complexity region, (3) QUALapprox score lower than $1x10^5$, (4) sequenced in over $7.5x10^4$ alleles, (5) population allele frequency lower than 1% with no homozygous counts, and (6) marked as a high-quality LoF variant by LOFTEE[70].

<u>Calculation of NAHR-deletion contribution to disease burden for a specific recessive disorder</u>

All the following calculations are based on the concept of random mating by sampling from a pool of alleles. Suppose, that at an autosomal recessive locus, we have *n* alleles, A_1, A_2, … A_n, with allele frequencies p_1, p_2, p_3, …, p_n. Without loss of generality, we nominate the kth allele as the NAHR-deletion allele -our allele of interest - and let the others index the small variant alleles. For most large recurrent deletions, homozygous loss of the deletion is incompatible with live birth. The two exceptions are the 2q13-*NPHP1* deletion and the 15q13.3 BP4-BP5 deletion, for which homozygous deletions are compatible with live birth. Also, the 15q13.3 BP4-BP5 deletion is encompassed by the 15q11q13 BP3-BP5 deletion, so the enclosed recessive genes have two NAHR-deletions contributing to them. Hypomorphic alleles that cause disease when they co-occur with LoF alleles, which we denote as A_h, are not disease-causing in homozygous states. Such alleles are observed in *RBM8A* and *TBX6*, but at present not in others. The exceptions described above regarding *RBM8A*, *TBX6*, the 2q13-*NPHP1* deletion, and the 15q13.3 BP4-BP5 deletion have been accounted for in the calculations for these special

circumstances. For simplicity, they are not incorporated in the equations below, although the modifications to account for them are simple adjustments to the sums and formulate presented below.

We denote the probability of an individual carrying the NAHR-deletion allele is

$$P\left(A_k\right) = p_k$$

The contribution of the NAHR-deletion to the allele load (**Fa**) is the fraction of the NAHR-deletion allele frequency over the sum of all allele frequencies across all functional alleles - except for the hypomophric alleles that do not in themselves cause disease.

$$Fa_k = \frac{p_k}{\sum_{i=1}^{n} p_i}$$

The Punnett square below models the expected frequencies of an individual to be affected with the recessive trait given the population frequencies of each pair of alleles under a random mating model.

	A_1 p_1	A_2 p_2	…	A_k p_k	…	A_n p_n
A_1 p_1	p_1^2 A_1A_1	p_1p_2 A_1A_2	…	p_1p_k A_1A_k	…	p_1p_n A_1A_n
A_2 p_2	p_1p_2 A_1A_2	p_2^2 A_2A_2	…	p_2p_k A_2A_k	…	p_2p_n A_2A_n
⋮	⋮	⋮	…	…	…	…
A_k p_k	p_1p_k A_1A_k	p_2p_k A_2A_k	…	p_k^2 A_kA_k	…	p_kp_n A_kA_n
⋮	⋮	⋮	…	⋮	…	…
A_n p_n	p_1p_n A_1A_n	p_2p_n A_2A_n	…	p_kp_n A_kA_n	…	p_n^2 A_nA_n

The probability for an individual to be affected with the recessive disorder is the sum of pairwise products of all carrier alleles with contribution from the homozygous NAHR-deletion lethal allele subtracted

$$P\left(D\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j - p_k^2$$

The probability for an individual to be both affected with the recessive disorder and carrying the NAHR-deletion is

$$P\left(A_k \cap D\right) = 2 \cdot \sum_{j=1}^{n} p_k p_j - 2 \cdot p_k^2$$

The contribution of individuals with the NAHR-deletion to the recessive disease load (**Fd**) is

$$Fd_k = P\left(A_k | D\right) = \frac{P\left(A_k \cap D\right)}{P\left(D\right)}$$

The odds for an individual with the recessive disease to carry the NAHR deletion is

$$O_k = \frac{Fd_k}{\sum_{i=1}^{n} Fd_i - Fd_k}$$

Note that the sum of Fd across alleles *1, 2, 3, …, n* equals the sum of the Punnett square matrix plus the lower triangular and the upper triangular. This is equivalent to summing up homozygous allele products once plus compound heterozygous allele products twice. This characteristic arises because the events of recessive disease involving the kth and the jth allele are not disjoint and they overlap for the compound heterozygote entries in the Punnett square.

To calculate the NAHR-deletion Impact to Recessive Disease score, the odds of the NAHR allele is compared to that of the "median" allele from the same gene.

$$NIRD = \log_2 (OR_k) = \log_2 \left(\frac{O_k}{O_{median}} \right)$$

The "median" allele is defined as the midpoint of remaining alleles in the same gene that comprise a cumulative sum of the top 90% of the overall sum of Fd. The alleles consisting the lower 10% of overall Fd sum are disregarded because (1) we find many genes have a long tail of ultra-rare alleles without a frequency estimate from gnomAD, and (2) we supplemented 10% of hypothetical alleles to each gene in our analysis.

Curation of recessive disease alleles

The Human Gene Mutation Database (HGMD, version 2020.4) was queried for recessive disease alleles in genes for which ≥70% of the disease carrier burden was attributed to NAHR-mediated large deletions. The literature linked to each disease allele that was marked as a disease-causing mutation (DM) or possible DM (DM?) was mined for the allelic state of the variant (homozygous, compound heterozygous with another SNV, or compound heterozygous with the NAHR-mediated deletion) as well as the number of unrelated families carrying the allele. Homozygous variants were counted once per family, while compound heterozygous SNVs received one count each per family. Gross deletions and insertions larger than 50 bp were excluded. *RBM8A* was the only gene for which disease-associated polymorphisms with additional supporting functional evidence (DFP) were also included; DM/DM? alleles compound heterozygous with DFP variants were dismissed for all genes other than *RBM8A*. *MYH11* alleles were only added if they were associated with hypoperistalsis. Similarly, only biallelic loss-of-

function *PMP22* alleles were included. The total population frequency of each curated allele was retrieved using gnomAD v3.1.

Retrospective analysis of extant cES for new recessive gene and allele discovery from individuals with recurrent genomic deletions

This study has been performed in accordance with the research protocol approved by Institutional Review Boards at Baylor College of Medicine. A waiver of informed consent has been obtained (H-41191). The patients were evaluated by clinical exome sequencing, with sequencing, data analysis, and interpretation procedures described previously.[71; 72] The identification of deletions was based on a SNP array platform, which was performed concurrently with the exome assay.[73] Sanger sequencing was performed as a validation method for candidate diagnostic small variants.

Code availability

Detailed code for generation of the NAHR-deletion map, preparation of disease alleles, and calculation of disease burden is available at https://github.com/liu-lab/cnvNAHR/.

References

1. Lupski, J.R., Belmont, J.W., Boerwinkle, E., and Gibbs, R.A. (2011). Clan genomics and the complex architecture of human disease. Cell 147, 32-43. 10.1016/j.cell.2011.09.008

2. Posey, J.E., O'Donnell-Luria, A.H., Chong, J.X., Harel, T., Jhangiani, S.N., Coban Akdemir, Z.H., Buyske, S., Pehlivan, D., Carvalho, C.M.B., Baxter, S., et al. (2019). Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet Med 21, 798-812. 10.1038/s41436-018-0408-7

3. Martin, H.C., Jones, W.D., McIntyre, R., Sanchez-Andrade, G., Sanderson, M., Stephenson, J.D., Jones, C.P., Handsaker, J., Gallone, G., Bruntraeger, M., et al. (2018). Quantifying the contribution of recessive coding variation to developmental disorders. Science 362, 1161-1164. 10.1126/science.aar6731

4. Gonzaga-Jauregui, C., Yesil, G., Nistala, H., Gezdirici, A., Bayram, Y., Nannuru, K.C., Pehlivan, D., Yuan, B., Jimenez, J., Sahin, Y., et al. (2020). Functional biology of the Steel syndrome founder allele and evidence for clan genomics derivation of *COL27A1* pathogenic alleles worldwide. Eur J Hum Genet 28, 1243-1264. 10.1038/s41431-020-0632-x

5. Alkuraya, F.S. (2021). A genetic revolution in rare-disease medicine. Nature 590, 218-219.

6. Coban-Akdemir, Z., Song, X., Pehlivan, D., Karaca, E., Bayram, Y., Gambin, T., Jhangiani, S.N., Muzny, D.M., Lewis, R.A., , et al. (2020). *De novo* mutation in ancestral generations evolves haplotypes contributing to disease. bioRxiv. doi.org/10.1101/2020.04.27.064824

7. Lupski, J.R. (2019). 2018 Victor A. McKusick Leadership Award: Molecular Mechanisms for Genomic and Chromosomal Rearrangements. Am J Hum Genet 104, 391-406. 10.1016/j.ajhg.2018.12.018

8. Liu, P., Carvalho, C.M., Hastings, P.J., and Lupski, J.R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. Current opinion in genetics & development 22, 211-220. 10.1016/j.gde.2012.02.012

9. Dennis, M.Y., and Eichler, E.E. (2016). Human adaptation and evolution by segmental duplication. Current opinion in genetics & development 41, 44-52. 10.1016/j.gde.2016.08.001

10. Lupski, J.R. (2015). Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environmental and molecular mutagenesis 56, 419-436. 10.1002/em.21943

11. Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S., and Hurles, M.E. (2008). Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders. Nat Genet 40, 90-95. 10.1038/ng.2007.40

12. Berg, I.L., Neumann, R., Lam, K.W., Sarbajna, S., Odenthal-Hesse, L., May, C.A., and Jeffreys, A.J. (2010). *PRDM9* variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet 42, 859-863. 10.1038/ng.658

13. Beck, C.R., Carvalho, C.M.B., Akdemir, Z.C., Sedlazeck, F.J., Song, X., Meng, Q., Hu, J., Doddapaneni, H., Chong, Z., Chen, E.S., et al. (2019). Megabase Length Hypermutation Accompanies Human Structural Variation at 17p11.2. Cell 176, 1310-1324 e1310. 10.1016/j.cell.2019.01.045

14. Rosenfeld, J.A., Coe, B.P., Eichler, E.E., Cuckle, H., and Shaffer, L.G. (2013). Estimates of penetrance for recurrent pathogenic copy-number variations. Genet Med 15, 478-481. 10.1038/gim.2012.164

15. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alfoldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature 581, 444-451. 10.1038/s41586-020-2287-8

16. Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G.B., Jonsdottir, G.A., Doyle, O.M., et al. (2014). CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature 505, 361-366. 10.1038/nature12818

17. Crawford, K., Bracher-Smith, M., Owen, D., Kendall, K.M., Rees, E., Pardinas, A.F., Einon, M., Escott-Price, V., Walters, J.T.R., O'Donovan, M.C., et al. (2019). Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. J Med Genet 56, 131-138. 10.1136/jmedgenet-2018-105477

18. Wu, N., Ming, X., Xiao, J., Wu, Z., Chen, X., Shinawi, M., Shen, Y., Yu, G., Liu, J., Xie, H., et al. (2015). *TBX6* null variants and a common hypomorphic allele in congenital scoliosis. N Engl J Med 372, 341-350. 10.1056/NEJMoa1406829

19. Potocki, L., Chen, K.S., Koeuth, T., Killian, J., Iannaccone, S.T., Shapira, S.K., Kashork, C.D., Spikes, A.S., Shaffer, L.G., and Lupski, J.R. (1999). DNA rearrangements on both homologues of chromosome 17 in a mildly delayed individual with a family history of autosomal dominant carpal tunnel syndrome. Am J Hum Genet 64, 471-478. 10.1086/302240

20. Ren, X., Yang, N., Wu, N., Xu, X., Chen, W., Zhang, L., Li, Y., Du, R.Q., Dong, S., Zhao, S., et al. (2020). Increased *TBX6* gene dosages induce congenital cervical vertebral malformations in humans and mice. J Med Genet 57, 371-379. 10.1136/jmedgenet-2019-106333

21. Yang, N., Wu, N., Zhang, L., Zhao, Y., Liu, J., Liang, X., Ren, X., Li, W., Chen, W., Dong, S., et al. (2019). *TBX6* compound inheritance leads to congenital vertebral malformations in humans and mice. Hum Mol Genet 28, 539-547. 10.1093/hmg/ddy358

22. Mannik, K., Magi, R., Mace, A., Cole, B., Guyatt, A.L., Shihab, H.A., Maillard, A.M., Alavere, H., Kolk, A., Reigo, A., et al. (2015). Copy number variations and cognitive phenotypes in unselected populations. Jama 313, 2044-2054. 10.1001/jama.2015.4845

23. Lupski, J.R. (2015). Cognitive phenotypes and genomic copy number variations. Jama 313, 2029-2030. 10.1001/jama.2015.4846

24. Dittwald, P., Gambin, T., Szafranski, P., Li, J., Amato, S., Divon, M.Y., Rodriguez Rojas, L.X., Elton, L.E., Scott, D.A., Schaaf, C.P., et al. (2013). NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. Genome Res 23, 1395-1409. 10.1101/gr.152454.112

25. Haque, I.S., Lazarin, G.A., Kang, H.P., Evans, E.A., Goldberg, J.D., and Wapner, R.J. (2016). Modeled Fetal Risk of Genetic Diseases Identified by Expanded Carrier Screening. Jama 316, 734-742. 10.1001/jama.2016.11139

26. Albers, C.A., Paul, D.S., Schulze, H., Freson, K., Stephens, J.C., Smethurst, P.A., Jolley, J.D., Cvejic, A., Kostadima, M., Bertone, P., et al. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit *RBM8A* causes TAR syndrome. Nat Genet 44, 435-439, S431-432. 10.1038/ng.1083

27. Klopocki, E., Schulze, H., Strauss, G., Ott, C.E., Hall, J., Trotier, F., Fleischhauer, S., Greenhalgh, L., Newbury-Ecob, R.A., Neumann, L.M., et al. (2007). Complex inheritance pattern resembling autosomal recessive inheritance involving a microdeletion in thrombocytopenia-absent radius syndrome. Am J Hum Genet 80, 232-240. 10.1086/510919

28. Jonch, A.E., Douard, E., Moreau, C., Van Dijck, A., Passeggeri, M., Kooy, F., Puechberty, J., Campbell, C., Sanlaville, D., Lefroy, H., et al. (2019). Estimating the effect size of the 15q11.2 BP1-BP2 deletion and its contribution to neurodevelopmental symptoms: recommendations for practice. J Med Genet 56, 701-710. 10.1136/jmedgenet-2018-105879

29. Yuan, B., Liu, P., Gupta, A., Beck, C.R., Tejomurtula, A., Campbell, I.M., Gambin, T., Simmons, A.D., Withers, M.A., Harris, R.A., et al. (2015). Comparative Genomic Analyses of the Human *NPHP1* Locus Reveal Complex Genomic Architecture and Its Regional Evolution in Primates. PLoS Genet 11, e1005686. 10.1371/journal.pgen.1005686

30. Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H., Walkiewicz, M., Bi, W., Xiao, R., Ding, Y., et al. (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. N Engl J Med 376, 21-31. 10.1056/NEJMoa1516767

31. Stankiewicz, P., Kulkarni, S., Dharmadhikari, A.V., Sampath, S., Bhatt, S.S., Shaikh, T.H., Xia, Z., Pursley, A.N., Cooper, M.L., Shinawi, M., et al. (2012). Recurrent deletions and reciprocal duplications of 10q11.21q11.23 including *CHAT* and *SLC18A3* are likely mediated by complex low-copy repeats. Hum Mutat 33, 165-179. 10.1002/humu.21614

32. Tyson, J., Malcolm, S., Thomas, P.K., and Harding, A.E. (1996). Deletions of chromosome 17p11.2 in multifocal neuropathies. Ann Neurol 39, 180-186. 10.1002/ana.410390207

33. Roa, B.B., Garcia, C.A., Pentao, L., Killian, J.M., Trask, B.J., Suter, U., Snipes, G.J., Ortiz-Lopez, R., Shooter, E.M., Patel, P.I., et al. (1993). Evidence for a recessive PMP22 point mutation in Charcot-Marie-Tooth disease type 1A. Nat Genet 5, 189-194. 10.1038/ng1093-189

34. Yuan, B., Neira, J., Gu, S., Harel, T., Liu, P., Briceno, I., Elsea, S.H., Gomez, A., Potocki, L., and Lupski, J.R. (2016). Nonrecurrent *PMP22-RAI1* contiguous gene deletions arise from replication-based mechanisms and result in Smith-Magenis syndrome with evident peripheral neuropathy. Hum Genet 135, 1161-1174. 10.1007/s00439-016-1703-5

35. Yuan, B., Harel, T., Gu, S., Liu, P., Burglen, L., Chantot-Bastaraud, S., Gelowani, V., Beck, C.R., Carvalho, C.M., Cheung, S.W., et al. (2015). Nonrecurrent 17p11.2p12 Rearrangement Events that Result in Two Concomitant Genomic Disorders: The *PMP22-RAI1* Contiguous Gene Duplication Syndrome. Am J Hum Genet 97, 691-707. 10.1016/j.ajhg.2015.10.003

36. Reiter, L.T., Murakami, T., Koeuth, T., Gibbs, R.A., and Lupski, J.R. (1997). The human COX10 gene is disrupted during homologous recombination between the 24 kb proximal and distal CMT1A-REPs. Hum Mol Genet 6, 1595-1603. dda195 [pii]

37. Inoue, K., Dewar, K., Katsanis, N., Reiter, L.T., Lander, E.S., Devon, K.L., Wyman, D.W., Lupski, J.R., and Birren, B. (2001). The 1.4-Mb CMT1A duplication/HNPP deletion genomic region reveals unique genome architectural features and provides insights into the recent evolution of new genes. Genome Res 11, 1018-1033. 10.1101/gr.180401

38. Pentao, L., Wise, C.A., Chinault, A.C., Patel, P.I., and Lupski, J.R. (1992). Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. Nat Genet 2, 292-300. 10.1038/ng1292-292

39. DiVincenzo, C., Elzinga, C.D., Medeiros, A.C., Karbassi, I., Jones, J.R., Evans, M.C., Braastad, C.D., Bishop, C.M., Jaremko, M., Wang, Z., et al. (2014). The allelic spectrum of Charcot-Marie-Tooth disease in over 17,000 individuals with neuropathy. Mol Genet Genomic Med 2, 522-529. 10.1002/mgg3.106

40. Ben-Shachar, S., Lanpher, B., German, J.R., Qasaymeh, M., Potocki, L., Nagamani, S.C., Franco, L.M., Malphrus, A., Bottenfield, G.W., Spence, J.E., et al. (2009). Microdeletion 15q13.3: a locus with incomplete penetrance for autism, mental retardation, and psychiatric disorders. J Med Genet 46, 382-388. 10.1136/jmg.2008.064378

41. van Bon, B.W., Mefford, H.C., Menten, B., Koolen, D.A., Sharp, A.J., Nillesen, W.M., Innis, J.W., de Ravel, T.J., Mercer, C.L., Fichera, M., et al. (2009). Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. J Med Genet 46, 511-523. 10.1136/jmg.2008.063412

42. Spielmann, M., Reichelt, G., Hertzberg, C., Trimborn, M., Mundlos, S., Horn, D., and Klopocki, E. (2011). Homozygous deletion of chromosome 15q13.3 including *CHRNA7* causes severe mental retardation, seizures, muscular hypotonia, and the loss of *KLF13* and *TRPM1* potentially cause macrocytosis and congenital retinal dysfunction in siblings. Eur J Med Genet 54, e441-445. 10.1016/j.ejmg.2011.04.004

43. Masurel-Paulet, A., Andrieux, J., Callier, P., Cuisset, J.M., Le Caignec, C., Holder, M., Thauvin-Robinet, C., Doray, B., Flori, E., Alex-Cordier, M.P., et al. (2010). Delineation of 15q13.3 microdeletions. Clin Genet 78, 149-161. 10.1111/j.1399-0004.2010.01374.x

44. Lepichon, J.B., Bittel, D.C., Graf, W.D., and Yu, S. (2010). A 15q13.3 homozygous microdeletion associated with a severe neurodevelopmental disorder suggests putative functions of the *TRPM1*, *CHRNA7*, and other homozygously deleted genes. Am J Med Genet A 152A, 1300-1304. 10.1002/ajmg.a.33374

45. Endris, V., Hackmann, K., Neuhann, T.M., Grasshoff, U., Bonin, M., Haug, U., Hahn, G., Schallner, J.C., Schrock, E., Tinschert, S., et al. (2010). Homozygous loss of *CHRNA7* on chromosome 15q13.3 causes severe encephalopathy with seizures and hypotonia. Am J Med Genet A 152A, 2908-2911. 10.1002/ajmg.a.33692

46. Masurel-Paulet, A., Drumare, I., Holder, M., Cuisset, J.M., Vallee, L., Defoort, S., Bourgois, B., Pernes, P., Cuvellier, J.C., Huet, F., et al. (2014). Further delineation of eye manifestations in homozygous 15q13.3 microdeletions including *TRPM1*: a differential diagnosis of ceroid lipofuscinosis. Am J Med Genet A 164A, 1537-1544. 10.1002/ajmg.a.36471

47. Yin, J., Chen, W., Chao, E.S., Soriano, S., Wang, L., Wang, W., Cummock, S.E., Tao, H., Pang, K., Liu, Z., et al. (2018). Otud7a Knockout Mice Recapitulate Many Neurological Features of 15q13.3 Microdeletion Syndrome. Am J Hum Genet 102, 296-308. 10.1016/j.ajhg.2018.01.005

48. Uddin, M., Unda, B.K., Kwan, V., Holzapfel, N.T., White, S.H., Chalil, L., Woodbury-Smith, M., Ho, K.S., Harward, E., Murtaza, N., et al. (2018). *OTUD7A* Regulates Neurodevelopmental Phenotypes in the 15q13.3 Microdeletion Syndrome. Am J Hum Genet 102, 278-295. 10.1016/j.ajhg.2018.01.006

49. Garret, P., Ebstein, F., Delplancq, G., Dozieres-Puyravel, B., Boughalem, A., Auvin, S., Duffourd, Y., Klafack, S., Zieba, B.A., Mahmoudi, S., et al. (2020). Report of the first

patient with a homozygous *OTUD7A* variant responsible for epileptic encephalopathy and related proteasome dysfunction. Clin Genet 97, 567-575. 10.1111/cge.13709

50. Suzuki, H., Inaba, M., Yamada, M., Uehara, T., Takenouchi, T., Mizuno, S., Kosaki, K., and Doi, M. (2020). Biallelic loss of *OTUD7A* causes severe muscular hypotonia, intellectual disability, and seizures. Am J Med Genet A. 10.1002/ajmg.a.62054

51. Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T., et al. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358, 667-675. 10.1056/NEJMoa075974

52. Shinawi, M., Liu, P., Kang, S.H., Shen, J., Belmont, J.W., Scott, D.A., Probst, F.J., Craigen, W.J., Graham, B.H., Pursley, A., et al. (2010). Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. J Med Genet 47, 332-341. 10.1136/jmg.2009.073015

53. Chen, W.J., Lin, Y., Xiong, Z.Q., Wei, W., Ni, W., Tan, G.H., Guo, S.L., He, J., Chen, Y.F., Zhang, Q.J., et al. (2011). Exome sequencing identifies truncating mutations in *PRRT2* that cause paroxysmal kinesigenic dyskinesia. Nat Genet 43, 1252-1255. 10.1038/ng.1008

54. Meneret, A., Grabli, D., Depienne, C., Gaudebout, C., Picard, F., Durr, A., Lagroua, I., Bouteiller, D., Mignot, C., Doummar, D., et al. (2012). *PRRT2* mutations: a major cause of paroxysmal kinesigenic dyskinesia in the European population. Neurology 79, 170-174. 10.1212/WNL.0b013e31825f06c3

55. Heron, S.E., Grinton, B.E., Kivity, S., Afawi, Z., Zuberi, S.M., Hughes, J.N., Pridmore, C., Hodgson, B.L., Iona, X., Sadleir, L.G., et al. (2012). *PRRT2* mutations cause benign familial infantile epilepsy and infantile convulsions with choreoathetosis syndrome. Am J Hum Genet 90, 152-160. 10.1016/j.ajhg.2011.12.003

56. Labate, A., Tarantino, P., Viri, M., Mumoli, L., Gagliardi, M., Romeo, A., Zara, F., Annesi, G., and Gambardella, A. (2012). Homozygous c.649dupC mutation in *PRRT2* worsens the BFIS/PKD phenotype with mental retardation, episodic ataxia, and absences. Epilepsia 53, e196-199. 10.1111/epi.12009

57. Liu, J., Wu, N., Deciphering Disorders Involving, S., study, C.O., Yang, N., Takeda, K., Chen, W., Li, W., Du, R., Liu, S., et al. (2019). TBX6-associated congenital scoliosis (TACS) as a clinically distinguishable subtype of congenital scoliosis: further evidence supporting the compound inheritance and *TBX6* gene dosage model. Genet Med 21, 1548-1558. 10.1038/s41436-018-0377-x

58. Shaw-Smith, C., Pittman, A.M., Willatt, L., Martin, H., Rickman, L., Gribble, S., Curley, R., Cumming, S., Dunn, C., Kalaitzopoulos, D., et al. (2006). Microdeletion encompassing *MAPT* at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet 38, 1032-1037. 10.1038/ng1858

59. Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. Nat Genet 39, S43-47. 10.1038/ng2084

60. Boone, P.M., Campbell, I.M., Baggett, B.C., Soens, Z.T., Rao, M.M., Hixson, P.M., Patel, A., Bi, W., Cheung, S.W., Lalani, S.R., et al. (2013). Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles. Genome Res 23, 1383-1394. 10.1101/gr.156075.113

61. Bi, W., Probst, F.J., Wiszniewska, J., Plunkett, K., Roney, E.K., Carter, B.S., Williams, M.D., Stankiewicz, P., Patel, A., Stevens, C.A., et al. (2012). Co-occurrence of recurrent duplications of the DiGeorge syndrome region on both chromosome 22 homologues due to inherited and de novo events. J Med Genet 49, 681-688. 10.1136/jmedgenet-2012-101002

62. Edwards, S.D., Schulze, K.V., Rosenfeld, J.A., Westerfield, L.E., Gerard, A., Yuan, B., Grigorenko, E.L., Posey, J.E., Bi, W., and Liu, P. (2021). Clinical characterization of individuals with the distal 1q21.1 microdeletion. Am J Med Genet A. 10.1002/ajmg.a.62104

63. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C.V., Staples, J., Gonzaga-Jauregui, C., et al. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science 354. 10.1126/science.aaf6814

64. Liburd, N., Ghosh, M., Riazuddin, S., Naz, S., Khan, S., Ahmed, Z., Riazuddin, S., Liang, Y., Menon, P.S., Smith, T., et al. (2001). Novel mutations of *MYO15A* associated with profound deafness in consanguineous families and moderately severe hearing loss in a patient with Smith-Magenis syndrome. Hum Genet 109, 535-541. 10.1007/s004390100604

65. Zhao, Y., Diacou, A., Johnston, H.R., Musfee, F.I., McDonald-McGinn, D.M., McGinn, D., Crowley, T.B., Repetto, G.M., Swillen, A., Breckpot, J., et al. (2020). Complete Sequence of the 22q11.2 Allele in 1,053 Subjects with 22q11.2 Deletion Syndrome Reveals Modifiers of Conotruncal Heart Defects. Am J Hum Genet 106, 26-40. 10.1016/j.ajhg.2019.11.010

66. Liu, P., Meng, L., Normand, E.A., Xia, F., Song, X., Ghazi, A., Rosenfeld, J., Magoulas, P.L., Braxton, A., Ward, P., et al. (2019). Reanalysis of Clinical Exome Sequencing Data. N Engl J Med 380, 2478-2480. 10.1056/NEJMc1812033

67. Kurotaki, N., Shen, J.J., Touyama, M., Kondoh, T., Visser, R., Ozaki, T., Nishimoto, J., Shiihara, T., Uetake, K., Makita, Y., et al. (2005). Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating coagulation factor twelve (FXII) deficiency. Genet Med 7, 479-483. 10.109701.GIM.0000177419.43309.37

68. Liu, P., Lacaria, M., Zhang, F., Withers, M., Hastings, P.J., and Lupski, J.R. (2011). Frequency of Nonallelic Homologous Recombination Is Correlated with Length of Homology: Evidence that Ectopic Synapsis Precedes Ectopic Crossing-Over. Am J Hum Genet 89, 580-588. 10.1016/j.ajhg.2011.09.009

69. Yuan, B., Wang, L., Liu, P., Shaw, C., Dai, H., Cooper, L., Zhu, W., Anderson, S.A., Meng, L., Wang, X., et al. (2020). CNVs cause autosomal recessive genetic diseases with or without involvement of SNV/indels. Genet Med 22, 1633-1641. 10.1038/s41436-020-0864-8

70. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434-443. 10.1038/s41586-020-2308-7

71. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. Jama 312, 1870-1879. 10.1001/jama.2014.14601

72. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369, 1502-1511. 10.1056/NEJMoa1306555

73. Dharmadhikari, A.V., Ghosh, R., Yuan, B., Liu, P., Dai, H., Al Masri, S., Scull, J., Posey, J.E., Jiang, A.H., He, W., et al. (2019). Copy number variant and runs of homozygosity detection by microarrays enabled more precise molecular diagnoses in 11,020 clinical exome cases. Genome Med 11, 30. 10.1186/s13073-019-0639-5

Declaration of interests

Baylor College of Medicine (BCM) and Miraca Holdings Inc. have formed a joint venture with shared ownership and governance of Baylor Genetics (BG), which performs clinical exome sequencing and chromosomal microarray genomics assay services. The authors who are affiliated with BG are employees of BCM and derive support through a professional services agreement with the BG. JRL has stock ownership in 23andMe, is a paid consultant for Regeneron Pharmaceuticals and Novartis, and is a co-inventor on multiple United States and European patents related to molecular diagnostics for inherited neuropathies, eye diseases, and bacterial genomic fingerprinting.

ACKNOWLEDGMENTS

This work was supported by the Baylor College of Medicine Precision Medicine Initiative Pilot Award to PL, National Human Genome Research Institute (NHGRI)/ National Heart Lung and Blood Institute (NHLBI) grant number UM1HG006542 to the Baylor Hopkins Center for Mendelian Genomics (BHCMG) to JRL, the National Institute of Neurological Disorders and Stroke (NINDS) R35NS105078 to JRL.

.

Tables and Figures

Table 1. Recurrent genomic deletions that are prevalent in the population.

Region	Coordinates (GRCh38)	Population allele frequency (x 10^{-6})	Allele frequency in diagnostic testing (x 10^{-6})	Known recessive genes	Number of coding genes
2q13 *NPHP1*	chr2:109930242-110228182	5873	2616	*NPHP1*	3
15q11.2	chr15:21311962-23261294	2764	2287	-	14
16p12.1	chr16:21754781-22502804	592.3	627.8	*OTOA, UQCRC2*	11
17p12 HNPP	chr17:14170711-15567588	314.8	388.6	*COX10, PMP22*	9
16p13.11	chr16:14772948-16330433	312.6	433.5	*NDE1, MYH11, ABCC6*	15
1q21.1 BP3-BP4	chr1:146380249-148811725	267.2	672.6	-	15
16p11.2 proximal	chr16:29416551-30202090	259.6	1674	*PRRT2, ALDOA, TBX6, CORO1A*	35
13q12.12	chr13:22911590-24323812	211.7	104.6	*SGCG, SACS, MIPEP*	7
22q11.2 LCRA-D	chr22:18530098-21214537	177.1	4499	*PRODH, SLC25A1, CDC45, GP1BB, TXNRD2, TANGO2, SCARF2, PI4KA, SNAP29, LZTR1*	49
1q21 TAR	chr1:144904297-146209950	173.9	269.0	*PEX11B, RBM8A, POLR3GL, HJV*	22

10q11.21q11.23	chr10:45765081-49954967	141.2	74.73	*RBP3, ERCC6, SLC18A3, CHAT*	38
16p11.2 distal	chr16:28706949-29049993	136.1	254.1	*TUFM, ATP2A1, CD19, LAT*	11
2q13	chr2:110494056-112385043	128.6	149.5	*ANAPC1, MERTK*	11
2q21.1	chr2:130623447-131386379	100.8	119.6	-	9
15q13.3 BP4-BP5	chr15:30246847-32496522	93.26	896.8	*FAN1, TRPM1, OTUD7A*	13
2q11.2	chr2:95759114-97430329	73.10	74.73	*ADRA2B, NCAPH, LMAN2L, CNNM4*	24
17q12	chr17:36300613-38034442	68.86	463.4	*ZNHIT3, PIGW*	21
7q11.23	chr7:73089294-74862006	39.35	1375	*NCF1*	27
15q11q13 BP3-BP4	chr15:28580349-30417865	35.29	134.5	*NSMCE3*	10
3q29	chr3:195963652-197626678	22.69	194.3	*TFRC, PCYT1A, TCTEX1D2, RNF168, NRROS, CEP19*	23
17q11.2	chr17:30621877-32037969	22.66	149.5	-	14
22q11.2 LCRD-H	chr22:21206521-24255497	12.60	59.79	*IGLL1*	45
8p23.1	chr8:7596999-12344083	10.08	134.5	*RP1L1, FDFT1*	50
10q23	chr10:79733715-87254783	7.562	59.79	*MAT1A, CDHR1*	31
17p11.2 Smith Magenis Syndrome	chr17:16777950-20450859	5.041	687.6	*TNFRSF13B, ATPAF2, MYO15A, MEIF2, TOP3A, GRAP, B9D1, ALDH3A2*	48
15q11 Prader-Willi/	chr15:21976318-	2.521	687.6	*OCA2, HERC2*	26

Angelman syndromes BP1-BP3	28537425				
15q24 BPA-BPC	chr15:72628218-75278711	2.521	74.73	*BBS4, STRA6, EDC3, MPI, COX5A*	40
15q11q13 BP3-BP5	chr15:28569118-32447357	2.521	59.79	*NSMCE3, FAN1, TRPM1*	21
7q11.23 distal	chr7:75456184-76629927	2.521	29.89	*POR, MDH2*	19

Regions are listed in descending order by population prevalence. Genes in the "Known recessive gene" column are ordered by coordinate map positions. Even though it is the third highest NAHR-mediated deletions, the Xp22.31-*STS* deletion is not included in this table because the current list focuses on autosomal recessive conditions. The gene *OTUD7A* in 15q13.3 BP4-BP5 is not reported as a recessive disease at the time of this study; however, patient analysis results from this study support this gene as a candidate recessive disease.

Table 2. Recessive genes with NAHR-mediated recurrent genomic deletions contributing to more than 20% of the overall disease burden.

Gene	Associated recessive disease	Region name	NAHR deletion frequency	NAHR-deletion Impact to Recessive Disease (NIRD)	Top 3 allele contribution to disease* (NAHR allele in red)	Fraction of NAHR allele frequency in all alleles	Aggregate carrier frequency
NPHP1	Nephronophthisis 1, juvenile, MIM# 256100	2q13 *NPHP1*	5.87E-03	7.7	***98***, 3, 2	85	6.89E-03
ADRA2B	Autosomal recessive mental retardation (from DECIPHER)	2q11.2	7.31E-05	4.1	***95***, 10, 10	90	8.12E-05
ALDOA	Glycogen storage disease XII, MIM# 611881	16p11.2 proximal	2.60E-04	4.1	***95***, 10, 10	90	2.88E-04
CEP19	Morbid obesity and spermatogenic failure, MIM# 615703	3q29	2.27E-05	4.1	***95***, 10, 10	90	2.52E-05
CORO1A	Immunodeficiency 8, MIM# 615401	16p11.2 proximal	2.60E-04	4.1	***95***, 10, 10	90	2.88E-04
COX5A	Mitochondrial complex IV deficiency, nuclear type 20, MIM# 619064	15q24 BPA-BPC	2.52E-06	4.1	***95***, 10, 10	90	2.80E-06
EDC3	Mental retardation, autosomal recessive 50, MIM# 616460	15q24 BPA-BPC	2.52E-06	4.1	***95***, 10, 10	90	2.80E-06
LAT	Immunodeficiency 52, MIM# 617514	16p11.2 distal	1.36E-04	4.1	***95***, 10, 10	90	1.51E-04
MIEF2	Combined oxidative phosphorylation	Smith Magenis	5.04E-06	4.1	***95***, 10, 10	90	5.60E-06

	deficiency 49, MIM# 619024	Syndrome					
SLC18A3	Myasthenic syndrome, congenital, 21, presynaptic, MIM# 617239	10q11.21q11.23	1.41E-04	4.1	***95***, 10, 10	90	1.57E-04
PMP22	Dejerine-Sottas disease, MIM# 145900	17p12 HNPP	3.15E-04	4.2	***94***, 11, 9	89	3.54E-04
OTUD7A	Neurodevelopmental disorder (this study)	15q13.3 BP4-BP5	9.33E-05	5.3	***93***, 10, 10	73	1.28E-04
RBM8A	Thrombocytopenia-absent radius syndrome, MIM# 274000	1q21.1 TAR	1.74E-04	9.8	***90***, 1, 1	90	1.93E-04
PRODH	Hyperprolinemia, type I, MIM# 239500	22q11.2 LCRA-D	1.77E-04	5	***87***, 36, 25	77	2.29E-04
UQCRC2	Mitochondrial complex III deficiency, nuclear type 5, MIM# 615160	16p12.1	5.92E-04	5.1	***86***, 49, 7	75	7.92E-04
NDE1	Lissencephaly 4 (with microcephaly), MIM# 614019	16p13.11	3.13E-04	5.1	***84***, 54, 15	72	4.31E-04
PEX11B	Peroxisome biogenesis disorder 14B, MIM# 614920	1q21 TAR	1.74E-04	4.7	***83***, 21, 18	71	2.44E-04
TUFM	Combined oxidative phosphorylation deficiency 4, MIM# 610678	16p11.2 distal	1.36E-04	3.9	***78***, 10, 10	64	2.11E-04
COX10	Mitochondrial complex IV deficiency, nuclear type 3, MIM# 619046	17p12 HNPP	3.15E-04	5.1	***77***, 48, 25	62	5.05E-04
SCARF2	Van den Ende-Gupta syndrome, MIM# 600920	22q11.2 LCRA-D	1.77E-04	4	***77***, 36, 7	62	2.84E-04
POLR3GL	Endosteal Hyperostosis and Oligodontia (from DECIPHER)	1q21 TAR	1.74E-04	4.1	***74***, 20, 13	59	2.95E-04
MYH11	Megacystis-microcolon-intestinal	16p13.11	3.13E-04	5	***72***, 24, 7	57	5.50E-04

	hypoperistalsis syndrome (from DECIPHER)						
NCAPH	Microcephaly 23, primary, autosomal recessive, MIM# 617985	2q11.2	7.31E-05	2.9	***71***, 27, 27	55	1.32E-04
PRRT2	Autosomal recessive mental retardation (from DECIPHER)	16p11.2 proximal	2.60E-04	5.3	***68***, 68, 10	52	5.02E-04
ANAPC1	Rothmund-Thomson syndrome, type 1, MIM# 618625	2q13	1.29E-04	3.9	***67***, 13, 11	50	2.57E-04
GRAP	Deafness, autosomal recessive 114, MIM# 618456	Smith Magenis Syndrome	5.04E-06	5.3	***67***, 85, 3	50	1.01E-05
CD19	Immunodeficiency, common variable, 3, MIM# 613493	16p11.2 distal	1.36E-04	4	***65***, 39, 28	48	2.84E-04
HJV	Hemochromatosis, type 2A, MIM# 602390	1q21 TAR	1.74E-04	5.2	***61***, 83, 5	44	3.92E-04
IGLL1	Agammaglobulinemia 2, MIM# 613500	22q11.2 LCRD-H	1.26E-05	5.2	***61***, 51, 50	44	2.87E-05
PIGW	Glycosylphosphatidylinositol biosynthesis defect 11, MIM# 616025	17q12	6.89E-05	5.2	***61***, 64, 35	44	1.56E-04
SLC25A1	Myasthenic syndrome, congenital, 23, presynaptic, MIM# 618197	22q11.2 LCRA-D	1.77E-04	4.5	***61***, 38, 15	43	4.08E-04
GP1BB	Giant platelet disorder, isolated, MIM# 231200	22q11.2 LCRA-D	1.77E-04	4.8	84, ***60***, 4	43	4.14E-04
RBP3	Retinitis pigmentosa 66, MIM# 615233	10q11.21q11.23	1.41E-04	4.1	***58***, 30, 9	41	3.44E-04
CNNM4	Jalili syndrome, MIM# 217080	2q11.2	7.31E-05	3.2	***57***, 16, 9	40	1.84E-04
FDFT1	Squalene synthase deficiency, MIM#	8p23.1	1.01E-05	5.1	***56***, 52, 52	39	2.58E-05

	618156						
SNAP29	Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome, MIM# 609528	22q11.2 LCRA-D	1.77E-04	4.5	***55***, 44, 28	38	4.62E-04
CDC45	Meier-Gorlin syndrome 7, MIM# 617063	22q11.2_LCRA-D	1.77E-04	4.5	***54***, 30, 19	37	4.73E-04
SGCG	Muscular dystrophy, limb-girdle, autosomal recessive 5, MIM# 253700	13q12.12	2.12E-04	4.7	***54***, 25, 15	37	5.73E-04
CHAT	Myasthenic syndrome, congenital, 6, presynaptic, MIM# 254210	10q11.21q11.23	1.41E-04	4.2	***52***, 25, 7	35	3.98E-04
NRROS	Seizures, early-onset, with neurodegeneration and brain calcification, MIM# 618875	3q29	2.27E-05	4.9	92, ***46***, 2	30	7.63E-05
NSMCE3	Lung disease, immunodeficiency, and chromosome breakage syndrome, MIM# 617241	15q11q13 BP3-BP4	3.53E-05	4.8	91, ***44***, 4	28	1.26E-04
TFRC	Immunodeficiency 46, MIM# 616740	3q29	2.27E-05	1.6	***40***, 41, 29	25	9.09E-05
PCYT1A	Spondylometaphyseal dysplasia with cone-rod dystrophy, MIM# 608940	3q29	2.27E-05	1.9	***38***, 39, 17	24	9.64E-05
SACS	Spastic ataxia, Charlevoix-Saguenay type, MIM# 270550	13q12.12	2.12E-04	4.4	***37***, 14, 13	23	9.20E-04
OTOA	Deafness, autosomal recessive 22, MIM# 607039	16p12.1	5.92E-04	4.4	80, ***35***, 5	21	2.78E-03
TXNRD2	Glucocorticoid deficiency 5, MIM#	22q11.2 LCRA-D	1.77E-04	3.7	48, ***35***, 36	21	8.25E-04

	617825						
MIPEP	Combined oxidative phosphorylation deficiency 31, MIM# 617228	13q12.12	2.12E-04	3.9	***33***, 15, 14	20	1.06E-03
PI4KA	Polymicrogyria, perisylvian, with cerebellar hypoplasia and arthrogryposis, MIM# 616531	22q11.2 LCRA-D	1.77E-04	4.7	***33***, 16, 13	20	9.07E-04
LMAN2L	Mental retardation, autosomal recessive, 52, MIM# 616887	2q11.2	7.31E-05	3.4	***32***, 32, 26	19	3.88E-04
TBX6	Spondylocostal dysostosis 5, MIM# 122600	16p11.2 proximal	2.60E-04	4.9	***32***, 27, 6	32	8.02E-04
ATP2A1	Brody myopathy, MIM# 601003	16p11.2 distal	1.36E-04	4.1	73, ***30***, 7	18	7.77E-04
DYNLT2B	Short-rib thoracic dysplasia 17 with or without polydactyly, MIM# 617405	3q29	2.27E-05	1.1	59, 43, ***28***	17	1.37E-04
TANGO2	Metabolic encephalomyopathic crises, recurrent, with rhabdomyolysis, cardiac arrhythmias, and neurodegeneration, MIM# 616878	22q11.2 LCRA-D	1.77E-04	2.9	78, ***27***, 23	15	1.14E-03
B9D1	Joubert syndrome 27, MIM# 617120	Smith Magenis Syndrome	5.04E-06	3.7	91, ***22***, 19	12	4.10E-05
ZNHIT3	PEHO syndrome, MIM# 260565	17q12	6.89E-05	3.7	96, ***22***, 2	13	5.47E-04
ERCC6	Cockayne syndrome, type B, MIM# 133540	10q11.21q11.23	1.41E-04	3.5	***21***, 15, 14	12	1.18E-03
FAN1	Interstitial nephritis, karyomegalic, MIM# 614817	15q13.3 BP4-BP5	9.33E-05	2.9	24, ***20***, 20	11	8.87E-04

LZTR1	Noonan syndrome 2, MIM# 605275	22q11.2 LCRA-D	1.77E-04	3.5	***20***, 17, 14	11	1.61E-03

*, the rows of the table are arranged in a descending order based on the NAHR deletion allele's contribution to the recessive disease, i.e. the percent of all affected individuals carrying at least one NAHR deletion allele, which are represented as the red numbers in the column "Top 3 allele contribution to disease". Note that the numbers from the column "Top 3 allele" are not disjunct events. For example, a patient can be compound heterozygous for the top 1 allele and the 2nd allele, and thus contribute to both numbers. Genes with an NAHR allele contribution to disease lower than 20% are not tabulated here, but can be found in Table S5.

Table 3. Clinically significant sequence variants uncovered by the deletions.

ID	Deletion/ Allele frequency	Gene (RefSeq transcript)	Genic variant	Genomic coordinate (GRCh38)	MAF in gnomAD v3.1	Classification	Category
1	10q11.21q11.23 deletion/ 1.412 $x10^{-4}$	*ERCC6* (NM_000124.3)	c.1490T>C (p.F497S)	chr10:49505920A>G	0	VUS	NDAC
2	17p12 HNPP/ 3.148$x10^{-4}$	*COX10* (NM_001303.3)	c.1277_1282dup (p.M426_L427dup)	chr17:14207158_ 14207163dup	0	VUS	NDAC
3[a]	17p12 HNPP/ 3.148$x10^{-4}$	*COX10* (NM_001303.3)	c.858G>T (p.W286C)	chr17:14192151G>T	6.567 $x10^{-6}$	VUS	NDAC
4	15q13.3 BP4-BP5 deletion/ 9.326$x10^{-5}$	*OTUD7A* (NM_130901.2)	c.2023_2066del (p.D675Hfs*188)	chr15:31484009_ 31484052del	6.58$x10^{-5}$ [b]	VUS	NDGMC
5	16p11.2 proximal / 2.596$x10^{-4}$	*PRRT2* (NM_145239.2)	c.649dup (p.R217fs)	chr16: 29813703dup	1.472$x10^{-4}$ [c]	Pathogenic	NDGMC

a. This subject is not identified from the same exome sequencing cohort as the other four subjects, but from a separate cohort for *COX10* targeted analysis.

b. This variant is marked with “low complexity region” label in gnomAD, suggesting ambiguous variant call quality. It has a variant count of 2 in gnomAD v3.1. However, manual review of the alignment data from gnomAD suggest only 1 is of higher quality. The allele frequency is adjusted in half accordingly.

c. This variant is marked with “low complexity region” label in gnomAD, suggesting ambiguous variant call quality. It is located in a homopolymer region that is susceptible to false positive variant calling. The variant allele frequency quoted here may be overestimated.

Abbreviations: MAF, minor allele frequency; VUS, variant of unknow significance; NDAC, new disease allele characterization; NDGMC, new disease gene/mechanism characterization.

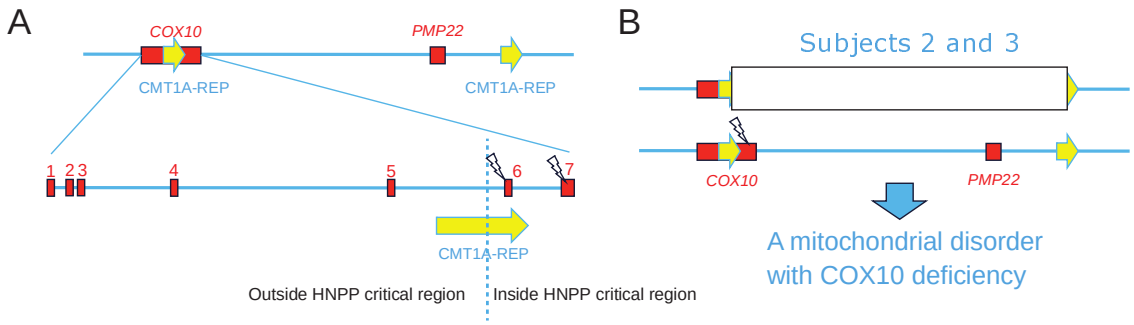


Figure 1. Compound heterozygous HNPP deletion and *COX10* variant leading to recessive COX10 deficiency in Subjects #2 and #3. **A.** The *COX10* gene spans the repeat sequence that mediate the recurrent HNPP deletion at chromosome 17p12. The *COX10* variant in the Subject #2 is located at the 3' end of the *COX10* gene on exon 7, which is inside the HNPP deletion interval. The *COX10* variant in the Subject #3 is located at the *COX10* gene exon 7, which is embedded in a CMT1A-REP. Red segments, exons of the *COX10* gene; yellow arrows, CMT1A-REPs; thunderbolts, *COX10* variants observed in Subject #2 and #3, respectively. **B.** Diagram illustrating the scheme of the relationship between the SNV/indel and recurrent deletion identified in Subject #2 and #3.

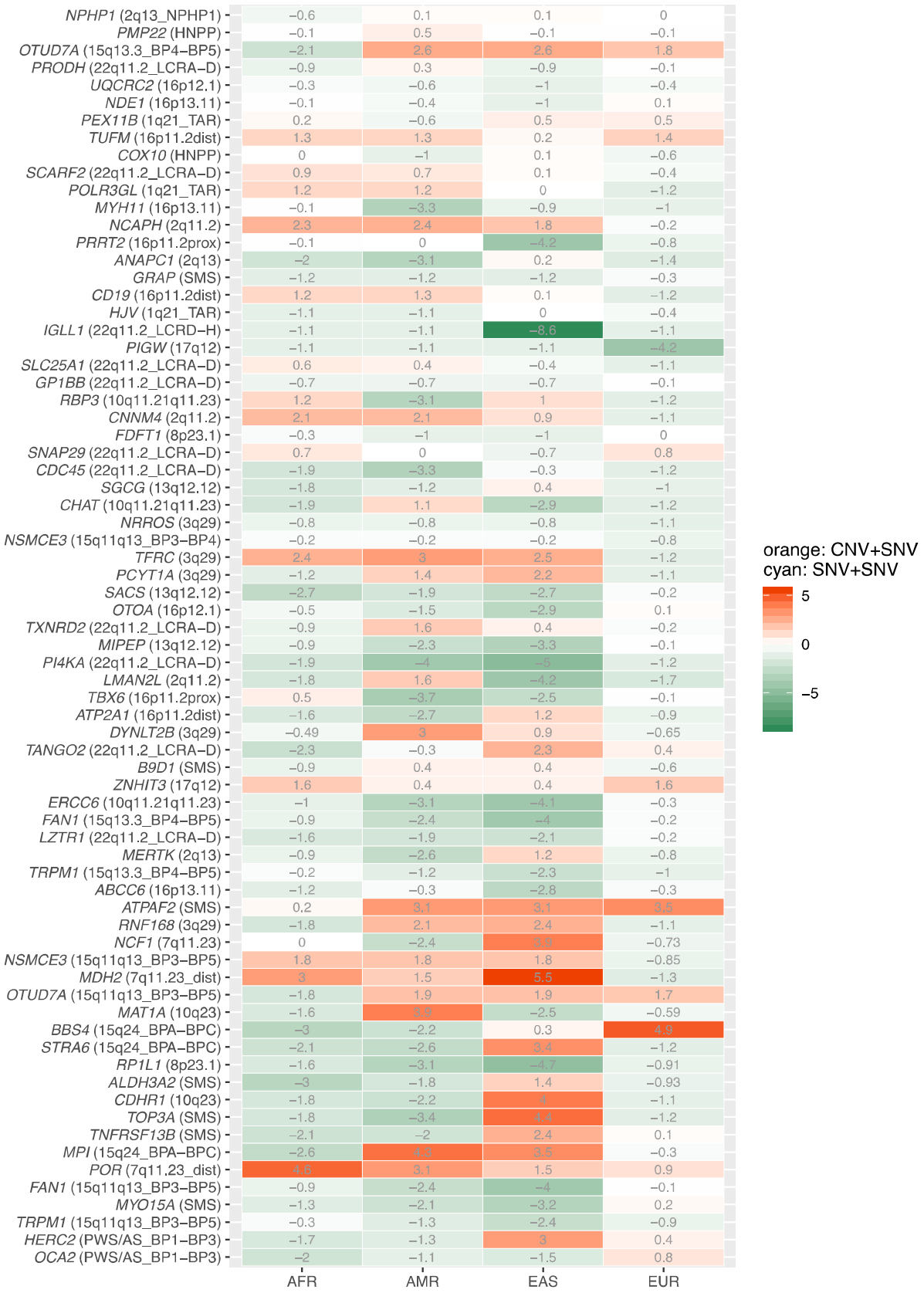

NPHP1 (2q13_NPHP1) −0.6 0.1 0.1 0
PMP22 (HNPP) −0.1 0.5 −0.1 −0.1
OTUD7A (15q13.3_BP4−BP5) −2.1 2.6 2.6 1.8
PRODH (22q11.2_LCRA−D) −0.9 0.3 −0.9 −0.1
UQCRC2 (16p12.1) −0.3 −0.6 −1 −0.4
NDE1 (16p13.11) −0.1 −0.4 −1 0.1
PEX11B (1q21_TAR) 0.2 −0.6 0.5 0.5
TUFM (16p11.2dist) 1.3 1.3 0.2 1.4
COX10 (HNPP) 0 −1 0.1 −0.6
SCARF2 (22q11.2_LCRA−D) 0.9 0.7 0.1 −0.4
POLR3GL (1q21_TAR) 1.2 1.2 0 −1.2
MYH11 (16p13.11) −0.1 −3.3 −0.9 −1
NCAPH (2q11.2) 2.3 2.4 1.8 −0.2
PRRT2 (16p11.2prox) −0.1 0 −4.2 −0.8
ANAPC1 (2q13) −2 −3.1 0.2 −1.4
GRAP (SMS) −1.2 −1.2 −1.2 −0.3
CD19 (16p11.2dist) 1.2 1.3 0.1 −1.2
HJV (1q21_TAR) −1.1 −1.1 0 −0.4
IGLL1 (22q11.2_LCRD−H) −1.1 −1.1 −8.6 −1.1
PIGW (17q12) −1.1 −1.1 −1.1 −4.2
SLC25A1 (22q11.2_LCRA−D) 0.6 0.4 −0.4 −1.1
GP1BB (22q11.2_LCRA−D) −0.7 −0.7 −0.7 −0.1
RBP3 (10q11.21q11.23) 1.2 −3.1 1 −1.2
CNNM4 (2q11.2) 2.1 2.1 0.9 −1.1
FDFT1 (8p23.1) −0.3 −1 −1 0
SNAP29 (22q11.2_LCRA−D) 0.7 0 −0.7 0.8
CDC45 (22q11.2_LCRA−D) −1.9 −3.3 −0.3 −1.2
SGCG (13q12.12) −1.8 −1.2 0.4 −1
CHAT (10q11.21q11.23) −1.9 1.1 −2.9 −1.2
NRROS (3q29) −0.8 −0.8 −0.8 −1.1
NSMCE3 (15q11q13_BP3−BP4) −0.2 −0.2 −0.2 −0.8
TFRC (3q29) 2.4 3 2.5 −1.2
PCYT1A (3q29) −1.2 1.4 2.2 −1.1
SACS (13q12.12) −2.7 −1.9 −2.7 −0.2
OTOA (16p12.1) −0.5 −1.5 −2.9 0.1
TXNRD2 (22q11.2_LCRA−D) −0.9 1.6 0.4 −0.2
MIPEP (13q12.12) −0.9 −2.3 −3.3 −0.1
PI4KA (22q11.2_LCRA−D) −1.9 −4 −5 −1.2
LMAN2L (2q11.2) −1.8 1.6 −4.2 −1.7
TBX6 (16p11.2prox) 0.5 −3.7 −2.5 −0.1
ATP2A1 (16p11.2dist) −1.6 −2.7 1.2 −0.9
DYNLT2B (3q29) −0.49 3 0.9 −0.65
TANGO2 (22q11.2_LCRA−D) −2.3 −0.3 2.3 0.4
B9D1 (SMS) −0.9 0.4 0.4 −0.6
ZNHIT3 (17q12) 1.6 0.4 0.4 1.6
ERCC6 (10q11.21q11.23) −1 −3.1 −4.1 −0.3
FAN1 (15q13.3_BP4−BP5) −0.9 −2.4 −4 −0.2
LZTR1 (22q11.2_LCRA−D) −1.6 −1.9 −2.1 −0.2
MERTK (2q13) −0.9 −2.6 1.2 −0.8
TRPM1 (15q13.3_BP4−BP5) −0.2 −1.2 −2.3 −1
ABCC6 (16p13.11) −1.2 −0.3 −2.8 −0.3
ATPAF2 (SMS) 0.2 3.1 3.1 3.5
RNF168 (3q29) −1.8 2.1 2.4 −1.1
NCF1 (7q11.23) 0 −2.4 3.9 −0.73
NSMCE3 (15q11q13_BP3−BP5) 1.8 1.8 1.8 −0.85
MDH2 (7q11.23_dist) 3 1.5 5.5 −1.3
OTUD7A (15q11q13_BP3−BP5) −1.8 1.9 1.9 1.7
MAT1A (10q23) −1.6 3.9 −2.5 −0.59
BBS4 (15q24_BPA−BPC) −3 −2.2 0.3 4.9
STRA6 (15q24_BPA−BPC) −2.1 −2.6 3.4 −1.2
RP1L1 (8p23.1) −1.6 −3.1 −4.7 −0.91
ALDH3A2 (SMS) −3 −1.8 1.4 −0.93
CDHR1 (10q23) −1.8 −2.2 4 −1.1
TOP3A (SMS) −1.8 −3.4 4.4 −1.2
TNFRSF13B (SMS) −2.1 −2 2.4 0.1
MPI (15q24_BPA−BPC) −2.6 4.3 3.5 −0.3
POR (7q11.23_dist) 4.6 3.1 1.5 0.9
FAN1 (15q11q13_BP3−BP5) −0.9 −2.4 −4 −0.1
MYO15A (SMS) −1.3 −2.1 −3.2 0.2
TRPM1 (15q11q13_BP3−BP5) −0.3 −1.3 −2.4 −0.9
HERC2 (PWS/AS_BP1−BP3) −1.7 −1.3 3 0.4
OCA2 (PWS/AS_BP1−BP3) −2 −1.1 −1.5 0.8
AFR
AMR
EAS
EUR
orange: CNV+SNV
cyan: SNV+SNV
5
0
−5

Figure 2. ΔNIRD scores for a population group relative to the general population. ΔNIRD is calculated by subtracting the general population NIRD from the specific population group NIRD. A positive ΔNIRD score suggests an even higher recessive disease contribution of the NAHR-deletion from that population group. A negative ΔNIRD score indicates a relatively higher probability for an affected patient to carry biallelic small variants in that population group. Genes that have zero ΔNIRD scores for all four population groups are not depicted in this heatmap.

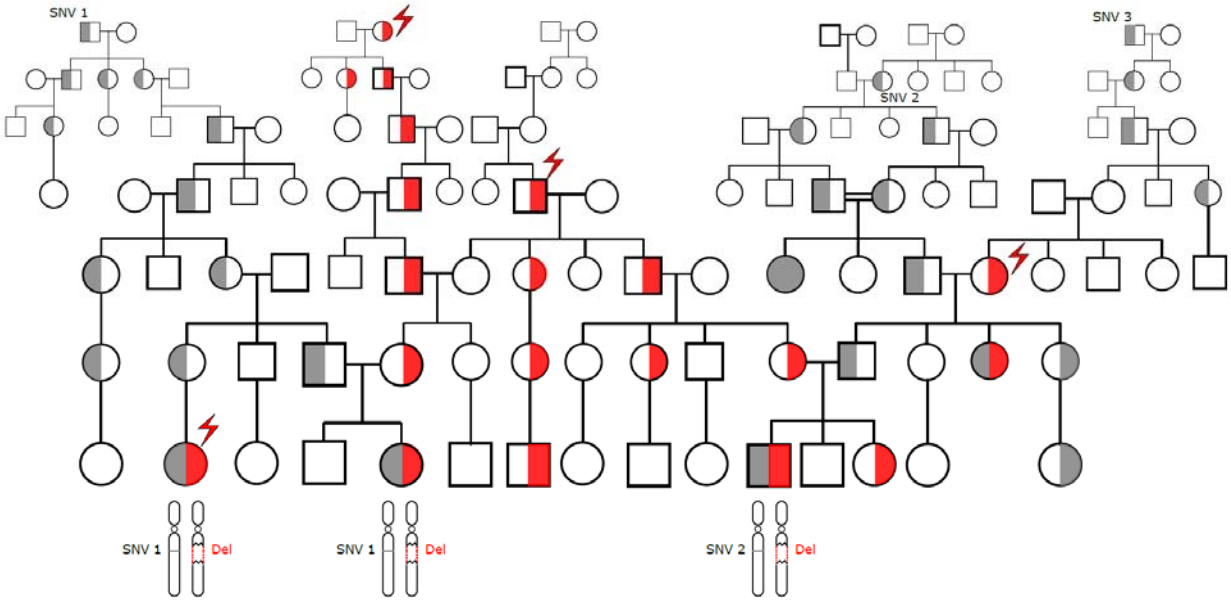


Figure 3. A 'segmental haploid genomics' approach for characterization of new disease genes or alleles for autosomal recessive conditions. The illustration above depicts dynamics of disease alleles in an autosomal recessive condition whose collective carrier burden is contributed by a recurrent genomic deletion (red area) and a few single nucleotide variants (SNVs, grey area). Individuals affected with biallelic pathogenic changes frequently carry the deletion as one of the two alleles, because the deletion arises recurrently in multiple lineages (indicated by the red lightning bolt arrow). SNV disease alleles for recessive genes tend to be passed on from ancestral generations, and may drift away without being noticed if they do not converge with another disease allele (SNV3). However, they may emerge to medical attention frequently in families with high degree of autozygosity, as illustrated in generations 3 and 4 for SNV2.

Supplemental Figure and Table legends

Figure S1. Genome-wide map for all predicted NAHR recurrent genomic deletions. Each predicted deletion event is marked as a green horizontal bar below the chromosome ideograms. The vertical bars above the chromosome ideograms illustrates the density for segmental duplications in a 1000-bp moving window.

Table S1. All 717 recurrent genomic deletions predicted based on the repeat structure in the human reference genome GRCh38.

Table S2. All clinically reported recurrent deletions and their prevalance estimates.

Table S3. Carrier disease allele frequencies by allele.

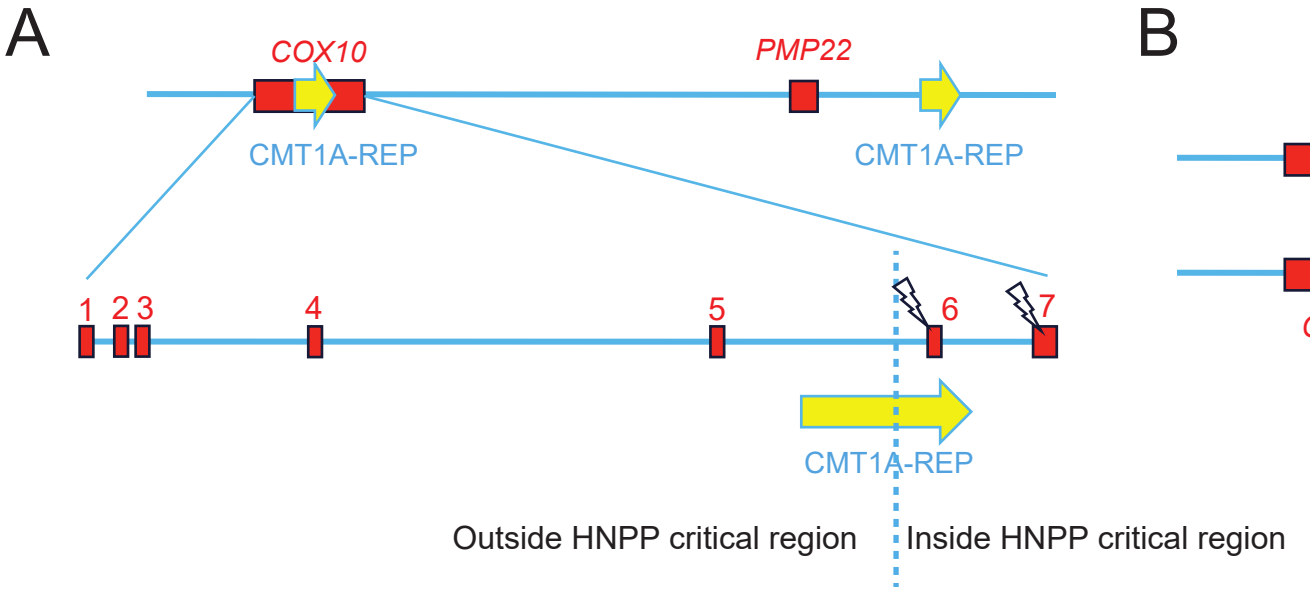
Table S4. Carrier allele frequency burden by gene. The list is ranked by genes from the highest burden to the lowest burden. The frequency burden in this list only includes the actual observed variants; the 10% extra hypothetical uncharacterized alleles as described in the Methods section are not included.

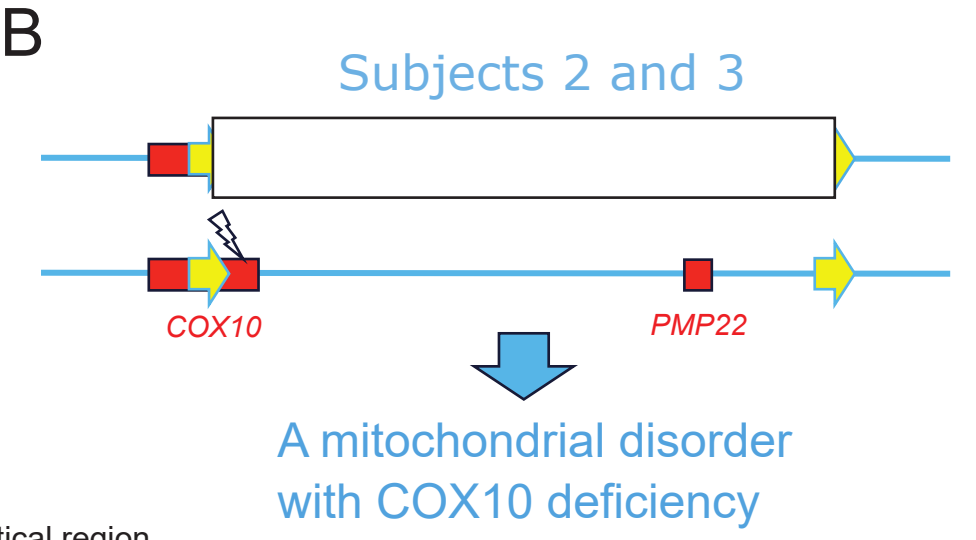
Table S5. Gene-level NAHR contribution to carrier allele and recessive disease burden as well as NAHR Deletion Impact to Recessive Traits (DIRT). This table is comprised of five panels, representing results generated using data from the general population and four specific ethnic groups, including African (AFR), Latino (AMR), East Asian (EAS), and European (EUR).

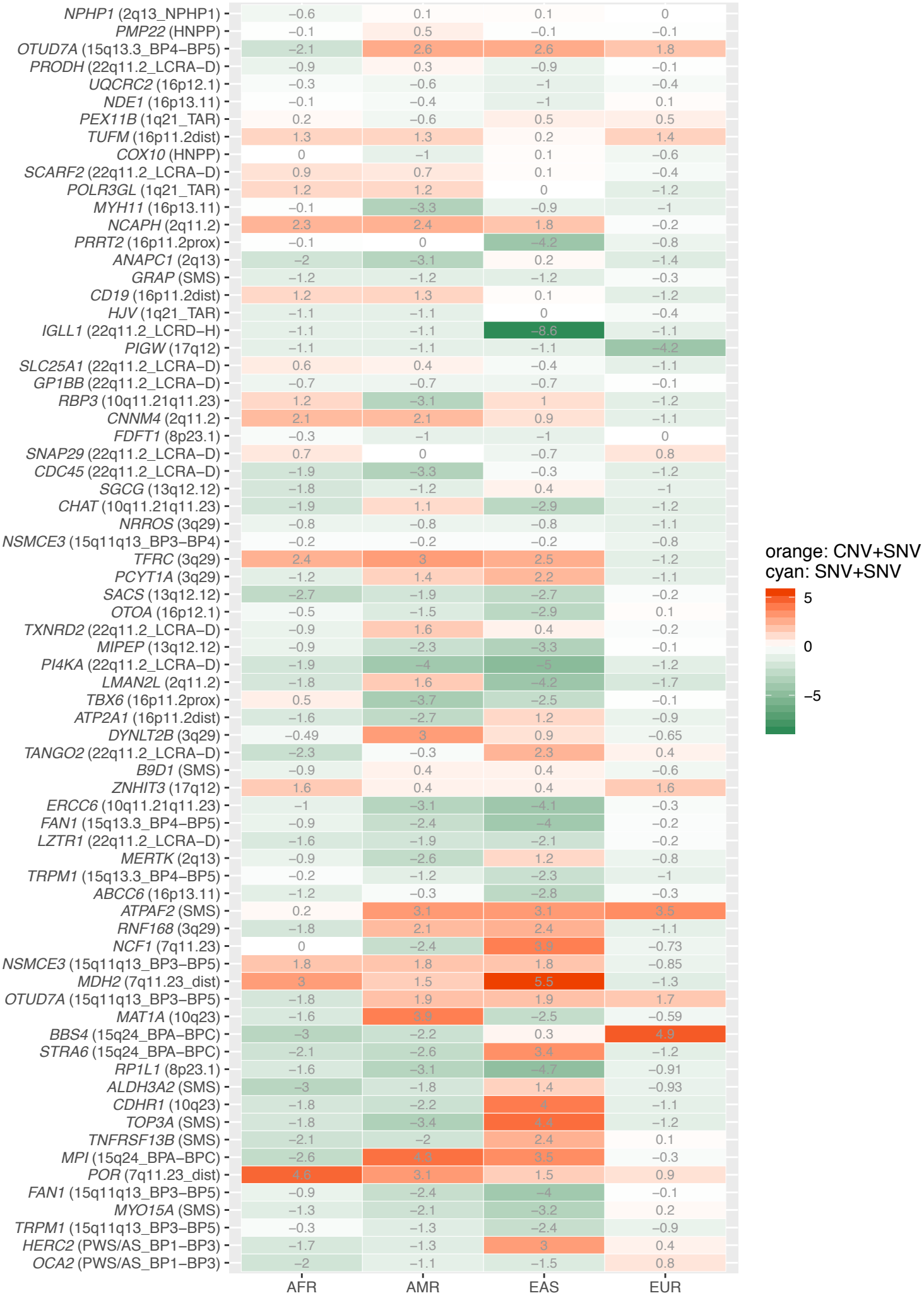
Table S6. Meta-analyses for literature reported patients affected with the 13 recessive disorders contributed by significant NAHR-mediated deletion burden.

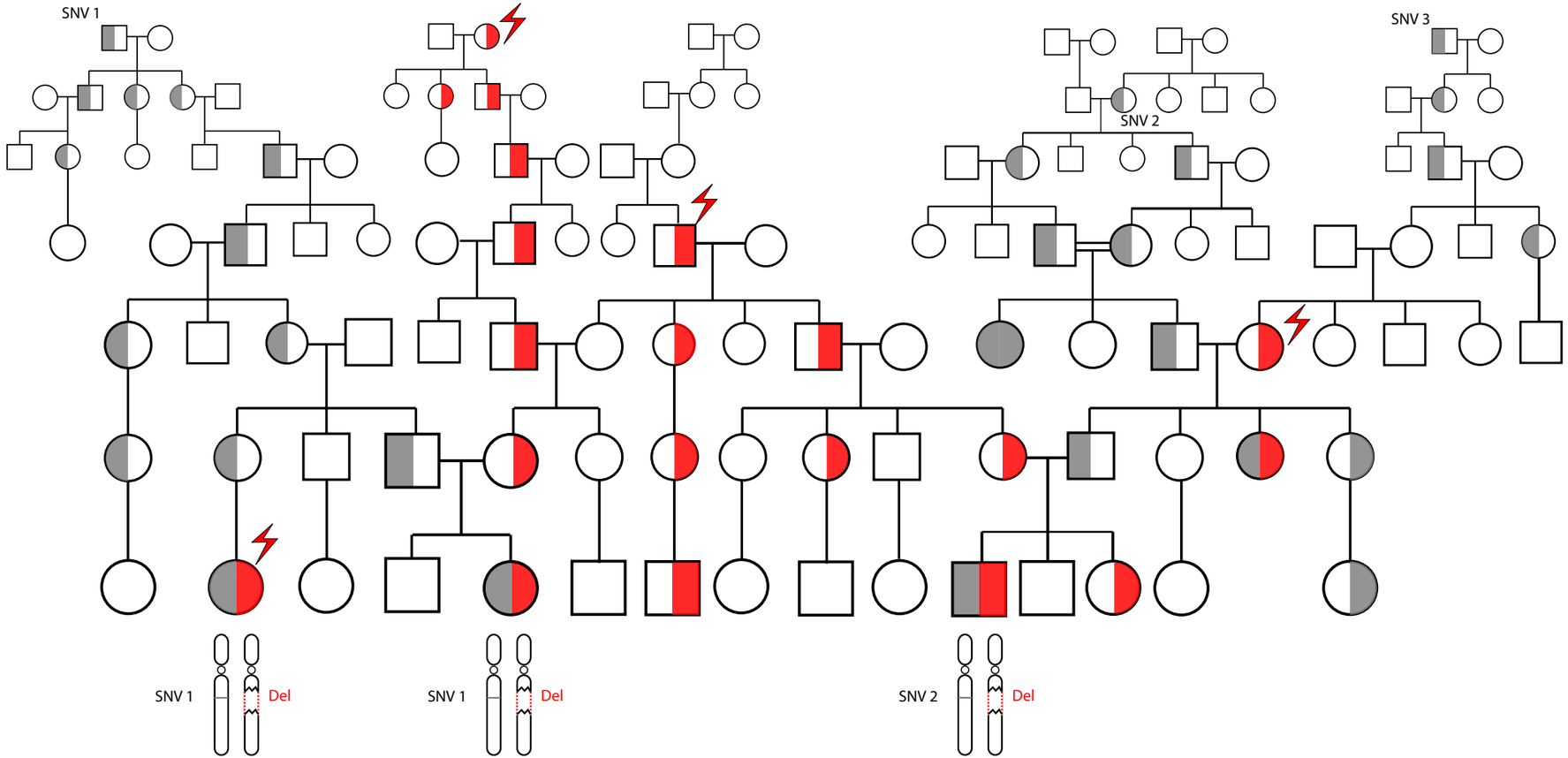
Table S7. Molecular findings of recurrent deletions identified from clinical exome sequencing.

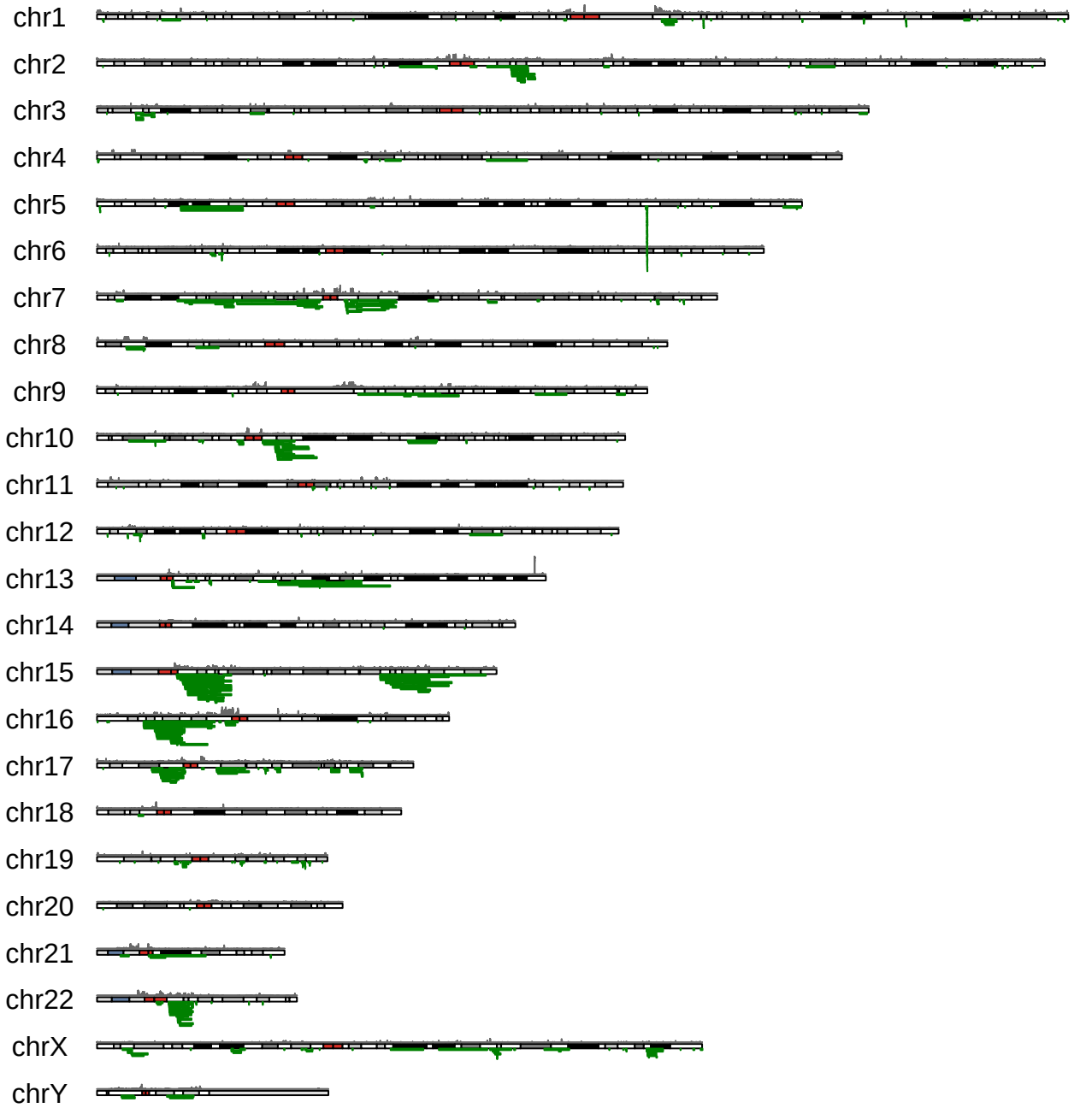
Table S8. Literature review for 15q13.3 recurrent deletions.


A
COX10
PMP22
CMT1A-REP
CMT1A-REP
1
2
3
4
5
6
7
CMT1A-REP
Outside HNPP critical region
Inside HNPP critical region


B
Subjects 2 and 3
COX10
PMP22
A mitochondrial disorder
with COX10 deficiency


NPHP1 (2q13_NPHP1) −0.6 0.1 0.1 0
PMP22 (HNPP) −0.1 0.5 −0.1 −0.1
OTUD7A (15q13.3_BP4−BP5) −2.1 2.6 2.6 1.8
PRODH (22q11.2_LCRA−D) −0.9 0.3 −0.9 −0.1
UQCRC2 (16p12.1) −0.3 −0.6 −1 −0.4
NDE1 (16p13.11) −0.1 −0.4 −1 0.1
PEX11B (1q21_TAR) 0.2 −0.6 0.5 0.5
TUFM (16p11.2dist) 1.3 1.3 0.2 1.4
COX10 (HNPP) 0 −1 0.1 −0.6
SCARF2 (22q11.2_LCRA−D) 0.9 0.7 0.1 −0.4
POLR3GL (1q21_TAR) 1.2 1.2 0 −1.2
MYH11 (16p13.11) −0.1 −3.3 −0.9 −1
NCAPH (2q11.2) 2.3 2.4 1.8 −0.2
PRRT2 (16p11.2prox) −0.1 0 −4.2 −0.8
ANAPC1 (2q13) −2 −3.1 0.2 −1.4
GRAP (SMS) −1.2 −1.2 −1.2 −0.3
CD19 (16p11.2dist) 1.2 1.3 0.1 −1.2
HJV (1q21_TAR) −1.1 −1.1 0 −0.4
IGLL1 (22q11.2_LCRD−H) −1.1 −1.1 −8.6 −1.1
PIGW (17q12) −1.1 −1.1 −1.1 −4.2
SLC25A1 (22q11.2_LCRA−D) 0.6 0.4 −0.4 −1.1
GP1BB (22q11.2_LCRA−D) −0.7 −0.7 −0.7 −0.1
RBP3 (10q11.21q11.23) 1.2 −3.1 1 −1.2
CNNM4 (2q11.2) 2.1 2.1 0.9 −1.1
FDFT1 (8p23.1) −0.3 −1 −1 0
SNAP29 (22q11.2_LCRA−D) 0.7 0 −0.7 0.8
CDC45 (22q11.2_LCRA−D) −1.9 −3.3 −0.3 −1.2
SGCG (13q12.12) −1.8 −1.2 0.4 −1
CHAT (10q11.21q11.23) −1.9 1.1 −2.9 −1.2
NRROS (3q29) −0.8 −0.8 −0.8 −1.1
NSMCE3 (15q11q13_BP3−BP4) −0.2 −0.2 −0.2 −0.8
TFRC (3q29) 2.4 3 2.5 −1.2
PCYT1A (3q29) −1.2 1.4 2.2 −1.1
SACS (13q12.12) −2.7 −1.9 −2.7 −0.2
OTOA (16p12.1) −0.5 −1.5 −2.9 0.1
TXNRD2 (22q11.2_LCRA−D) −0.9 1.6 0.4 −0.2
MIPEP (13q12.12) −0.9 −2.3 −3.3 −0.1
PI4KA (22q11.2_LCRA−D) −1.9 −4 −5 −1.2
LMAN2L (2q11.2) −1.8 1.6 −4.2 −1.7
TBX6 (16p11.2prox) 0.5 −3.7 −2.5 −0.1
ATP2A1 (16p11.2dist) −1.6 −2.7 1.2 −0.9
DYNLT2B (3q29) −0.49 3 0.9 −0.65
TANGO2 (22q11.2_LCRA−D) −2.3 −0.3 2.3 0.4
B9D1 (SMS) −0.9 0.4 0.4 −0.6
ZNHIT3 (17q12) 1.6 0.4 0.4 1.6
ERCC6 (10q11.21q11.23) −1 −3.1 −4.1 −0.3
FAN1 (15q13.3_BP4−BP5) −0.9 −2.4 −4 −0.2
LZTR1 (22q11.2_LCRA−D) −1.6 −1.9 −2.1 −0.2
MERTK (2q13) −0.9 −2.6 1.2 −0.8
TRPM1 (15q13.3_BP4−BP5) −0.2 −1.2 −2.3 −1
ABCC6 (16p13.11) −1.2 −0.3 −2.8 −0.3
ATPAF2 (SMS) 0.2 3.1 3.1 3.5
RNF168 (3q29) −1.8 2.1 2.4 −1.1
NCF1 (7q11.23) 0 −2.4 3.9 −0.73
NSMCE3 (15q11q13_BP3−BP5) 1.8 1.8 1.8 −0.85
MDH2 (7q11.23_dist) 3 1.5 5.5 −1.3
OTUD7A (15q11q13_BP3−BP5) −1.8 1.9 1.9 1.7
MAT1A (10q23) −1.6 3.9 −2.5 −0.59
BBS4 (15q24_BPA−BPC) −3 −2.2 0.3 4.9
STRA6 (15q24_BPA−BPC) −2.1 −2.6 3.4 −1.2
RP1L1 (8p23.1) −1.6 −3.1 −4.7 −0.91
ALDH3A2 (SMS) −3 −1.8 1.4 −0.93
CDHR1 (10q23) −1.8 −2.2 4 −1.1
TOP3A (SMS) −1.8 −3.4 4.4 −1.2
TNFRSF13B (SMS) −2.1 −2 2.4 0.1
MPI (15q24_BPA−BPC) −2.6 4.3 3.5 −0.3
POR (7q11.23_dist) 4.6 3.1 1.5 0.9
FAN1 (15q11q13_BP3−BP5) −0.9 −2.4 −4 −0.1
MYO15A (SMS) −1.3 −2.1 −3.2 0.2
TRPM1 (15q11q13_BP3−BP5) −0.3 −1.3 −2.4 −0.9
HERC2 (PWS/AS_BP1−BP3) −1.7 −1.3 3 0.4
OCA2 (PWS/AS_BP1−BP3) −2 −1.1 −1.5 0.8
AFR
AMR
EAS
EUR
orange: CNV+SNV
cyan: SNV+SNV
5
0
−5


SNV 1
SNV 2
SNV 3
SNV 1
Del
SNV 1
Del
SNV 2
Del


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY