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

Bo Yuan [1,2], 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 Schaaf [1,7], Yaping Yang [1,2], Weimin Bi [1,2], Christine Eng [1,2], Chad Shaw [1,2], James R. Lupski [1,8,9,10], 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, University Hospital Cologne, Cologne, Germany

[8] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

[9] Texas Children's Hospital, Houston, TX

[10] Department of Pediatrics, Baylor College of Medicine, Houston, TX

[*]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 new "disease genes" and alleles depend on patient ascertainment strategies to enrich previously uncharacterized alleles. Here, we present a novel strategy of new allele and gene discovery for recessive/biallelic disease traits. In this approach, patients with large recurrent genomic deletions mediated by nonallelic homologous recombination (NAHR) are sequenced, and new discoveries are revealed in the hemizygous chromosomal regions in *trans* to the large deletion, essentially enabling haploid genomic segment genetics. We demonstrate through computational analyses that a collection of 30 large recurrent genomic deletions scattered in the human genome contribute to more than 10% of individual disease load for 2.13% of all known "recessive disease genes". We performed meta-analyses for all literature reported patients affected with the 13 genes whose carrier burden are predicted to be almost exclusively from large recurrent genomic deletions. The results suggest that current sequencing efforts for personal genomes with large recurrent deletions is under-appreciated. By retrospective analyses of previously undiagnostic exome sequencing (ES) data on 69 subjects harboring 26 types of recurrent deletions, probable diagnostic variants were uncovered in genes including *COX10*, *ERCC6, PRRT2* and *OTUD7A*, demonstrating new disease allele/gene/mechanism characterization. Findings from this study support the contention that more whole genome sequencing (WGS) may further resolve molecular diagnoses and provide evidence for multi-locus pathogenic variation (MPV). Such analyses benefit all stakeholders in both research development and patient clinical care.

Introduction

Over the last few decades, efforts to decipher molecular and genetic mechanisms underlying Mendelian disorders have repeatedly demonstrated that mutations lead to human diseases in a continuum of a spectrum of allele number requirements, ranging from monoallelic (dominant), biallelic (recessive), to triallelic and other multiallelic 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 the traditional Mendelian inheritance mode, i.e., monoallelic or biallelic inheritance.[2] Although optimism has been increasing towards achieving disease annotation for a substantial portion of the haploinsufficient part of the human genome by dominant disease gene discoveries, statistical analysis from rare disease cohort studies suggested that the trajectory to understand the *trans* configuration biallelic-disease-causing portion of the human genome is uncertain.[3]

This apparent discrepancy in discovery of dominant versus recessive genes can be explained by the Clan Genomics Model.[1] The model predicts that dominant diseases are largely caused by emergence of new mutations in recent generations, whereas recessive diseases arise when two existing disease alleles from the population are aggregated in a personal genome in the *trans* configuration by transmission genetics and products of a mating. Assuming a dominant disease allele emerges at a rate of 10^{-8} per locus per generation, the *de novo* mutation rate at a human locus, a recessive disease with comparable incidence would require at least one of the two alleles of greater than ~10^{-4} populational allele frequency. Disease causing variants with such high allele frequencies tend to be found in established recessive disorders, often attributed to founder mutation alleles. In contrast, if a yet-to-be-defined recessive disease gene does not have

pathogenic alleles represented at a sufficient population frequency, disease annotation of the gene would be greatly hampered due to the difficulty in ascertaining individuals affected with the disease of extremely low incidence. An exception to this rule is that when disease gene discovery research is conducted in populations of elevated autozygosity, the allele pool shrinks to the Clan of the patient's extended family, which dramatically escalates the disease-allele frequency.[4] Thus, the probability of ascertaining patients with the recessive disorder increases exponentially, even when all existing alleles are ultra-rare in the general population.[5]

Here, we present a previously under-recognized source of disease alleles for recessive disorders, i.e. loss-of-function alleles in genes caused by large recurrent genomic deletions. Recurrent genomic deletions are a subset of contiguous gene deletions that are characterized by a specialized type of mutational mechanism called nonallelic homologous recombination (NAHR). NAHR is mediated by ectopic recombination between highly similar repeat sequences termed low-copy repeats (LCRs) or segmental duplications (SDs).[6] 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.[7; 8] 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 from here on).[9] The high new mutation rate ensures that genomic deletions keep arising *de novo* in the human population in unrelated individuals. This 'recurring' nature of genomic deletions distinguishes themselves from the other recessive disease alleles that are more 'stationary' in our current snapshot of the human evolutionary history. Additionally, 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,[10-15] i.e., the fitness of the deletion allele can be reasonably high, at least in certain genomic backgrounds.[16; 17] These findings all support the contention that recurrent genomic deletions are highly prevalent alleles in comparison to other small variant recessive disease alleles, potentially contributing to a considerable burden of recessive diseases.

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 genes. 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 and the concept of 'human haploid genetics' by investigating genomic intervals of recurrent segmental aneusomy.

Results

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[18] - based on the GRCh38 human reference genome sequence (Figure S1, Table S1). The collapsed NAHR map contains 721 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 population prevalence over 1/1,000,000 based on estimates from the UK Biobank, the Icelandic, and the gnomAD SV studies [11-13] (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 75 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 rarer deletions included, the span of genomic coverage increases to 82 Mb; the number of genes involved becomes 1828, with 97 representing recessive disease genes.

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

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	-	12
16p12.1	chr16:21754781-22502804	592.3	627.8	*OTOA*, *UQCRC2*	11
17p12 HNPP	chr17:14170711-15567588	314.8	388.6	*COX10*	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*, *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*	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
Smith Magenis Syndrome	chr17:16777950-20450859	5.041	687.6	*TNFRSF13B, ATPAF2, MYO15A, MEIF2, TOP3A, GRAP, B9D1,*	49

				ALDH3A2	
Prader-Willi/ Angelman syndromes BP1-BP3	chr15:21976318-28537425	2.521	687.6	*OCA2, HERC2*	26
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

We then calculated carrier allele burden for each of the known recessive genes in the human genome, in preparation for dissecting and estimating 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 2585 recessive genes were assembled. The carrier allele burden for each recessive 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 loss-of-function from gnomAD SV v2.1, all high-confidence loss-of-function small variants identified in gnomAD v3.1, and the NAHR-mediated recurrent genomic deletions, if applicable (Table S2). An aggregate of 79,158 small variant and large deletion carrier alleles were identified for the 2585 genes (Table S3). A drawback of this calculation is that SNV pathogenic missense, in-frame indel, or intronic variants not 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 carrier burden estimate, as the alleles with higher frequencies are more likely to be ascertained in screening tests of clinical diagnostic laboratories, and thus receive an entry and curation in ClinVar. Nevertheless, we assumed that on average gene-level disease variants from the above collection cover at least 90% of the total allele burden. To account for the unrepresented alleles, in the subsequent analyses, we conservatively supplemented the disease allele pool for each gene with a 10% extra variant load, which comprises of ten hypothetical variants each accounting for 1% of the overall carrier burden.

Based on the gene-level disease allele architecture, we calculated for each recessive gene the fraction of affected individuals carrying one recurrent genomic deletion. 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 observed in greater than 10% of all patients with this recessive disorder. By this arbitrary definition, 55 recessive genes, which account for 2.13% of all known recessive genes, are under significant NAHR-deletion burden for recessive disease prevalence (Table 2)! We can extrapolate based on this finding that (1) for ~2% of the known human recessive genes, one of the most effective 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, (2) perhaps a similar percentage of uncharacterized recessive disease genes may be most effectively identified by sequencing affected individuals with one of the prevalent recurrent genomic deletions- any of the 1442 genes (or 557 coding genes) within these deletion regions that have yet to be assigned an autosomal recessive inheritance could become a candidate as a novel recessive gene.

Table 2. Recessive genes with NAHR-mediated recurrent genomic deletions contributing to more than 10% of the overall disease burden.

Gene symbol	Associated recessive disease	Region name	Aggregate carrier frequency	NAHR deletion frequency	Percent NAHR contribution
NPHP1	Nephronophthisis 1, juvenile, MIM# 256100	2q13 *NPHP1*	6.53E-03	5.87E-03	99
ALDOA	Glycogen storage disease XII, MIM# 611881	16p11.2 proximal	2.89E-04	2.60E-04	94
CORO1A	Immunodeficiency 8, MIM# 615401	16p11.2 proximal	2.89E-04	2.60E-04	94
SLC18A3	Myasthenic syndrome, congenital, 21, presynaptic, MIM# 617239	10q11.21q11.23	1.57E-04	1.41E-04	94
LAT	Immunodeficiency 52, MIM# 617514	16p11.2 distal	1.51E-04	1.36E-04	94
ADRA2B	Autosomal recessive mental retardation (from DECIPHER)	2q11.2	8.12E-05	7.31E-05	94
CEP19	Morbid obesity and spermatogenic failure, MIM# 615703	3q29	2.52E-05	2.27E-05	94
NRROS	Seizures, early-onset, with neurodegeneration and brain calcification, MIM# 618875	3q29	2.52E-05	2.27E-05	94
MIEF2	Combined oxidative phosphorylation	Smith Magenis	5.60E-06	5.04E-06	94

	deficiency 49, MIM# 619024	Syndrome			
RBM8A	Thrombocytopenia-absent radius syndrome, MIM# 274000	1q21.1 TAR	1.93E-04	1.74E-04	94
COX5A	Mitochondrial complex IV deficiency, nuclear type 20, MIM# 619064	15q24 BPA-BPC	2.80E-06	2.52E-06	94
EDC3	Mental retardation, autosomal recessive 50, MIM# 616460	15q24 BPA-BPC	2.80E-06	2.52E-06	94
PRODH	Hyperprolinemia, type I, MIM# 239500	22q11.2 LCRA-D	2.29E-04	1.77E-04	85
UQCRC2	Mitochondrial complex III deficiency, nuclear type 5, MIM# 615160	16p12.1	7.92E-04	5.92E-04	83
PEX11B	Peroxisome biogenesis disorder 14B, MIM# 614920	1q21 TAR	2.43E-04	1.74E-04	82
MYH11	Aortic aneurysm, familial thoracic 4, MIM# 132900	16p13.11	4.48E-04	3.13E-04	82
TUFM	Combined oxidative phosphorylation deficiency 4, MIM# 610678	16p11.2 distal	2.11E-04	1.36E-04	77
NDE1	Lissencephaly 4 (with microcephaly), MIM# 614019	16p13.11	4.82E-04	3.13E-04	76
SCARF2	Van den Ende-Gupta syndrome, MIM# 600920	22q11.2 LCRA-D	1.97E-04	1.77E-04	75
COX10	Mitochondrial complex IV deficiency, nuclear type 3, MIM# 619046	17q12 HNPP	4.83E-04	3.15E-04	73

POLR3GL	Endosteal Hyperostosis and Oligodontia (from DECIPHER)	1q21 TAR	2.96E-04	1.74E-04	73
NCAPH	Microcephaly 23, primary, autosomal recessive, MIM# 617985	2q11.2	1.32E-04	7.31E-05	68
SLC25A1	Myasthenic syndrome, congenital, 23, presynaptic, MIM# 618197	22q11.2 LCRA-D	2.77E-04	1.77E-04	67
OTOA	Deafness, autosomal recessive 22, MIM# 607039	16p12.1	1.30E-03	5.92E-04	65
CD19	Immunodeficiency, common variable, 3, MIM# 613493	16p11.2 distal	2.84E-04	1.36E-04	61
PRRT2	Autosomal recessive mental retardation (from DECIPHER)	16p11.2 proximal	5.01E-04	2.60E-04	61
ANAPC1	Rothmund-Thomson syndrome, type 1, MIM# 618625	2q13	2.87E-04	1.29E-04	61
RBP3	Retinitis pigmentosa 66, MIM# 615233	10q11.21q11.23	3.21E-04	1.41E-04	59
CNNM4	Jalili syndrome, MIM# 217080	2q11.2	1.83E-04	7.31E-05	55
GRAP	Deafness, autosomal recessive 114, MIM# 618456	Smith Magenis Syndrome	1.01E-05	5.04E-06	55
IGLL1	Agammaglobulinemia 2, MIM# 613500	22q11.2 LCRD-H	1.40E-05	1.26E-05	54
SNAP29	Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome, MIM# 609528	22q11.2 LCRA-D	3.44E-04	1.77E-04	54

PIGW	Glycosylphosphatidylinositol biosynthesis defect 11, MIM# 616025	17q12	1.57E-04	6.89E-05	53
SGCG	Muscular dystrophy, limb-girdle, autosomal recessive 5, MIM# 253700	13q12.12	5.64E-04	2.12E-04	53
CDC45	Meier-Gorlin syndrome 7, MIM# 617063	22q11.2_LCRA-D	3.63E-04	1.77E-04	53
CHAT	Myasthenic syndrome, congenital, 6, presynaptic, MIM# 254210	10q11.21q11.23	3.94E-04	1.41E-04	51
HJV	Hemochromatosis, type 2A, MIM# 602390	1q21 TAR	3.92E-04	1.74E-04	50
FDFT1	Squalene synthase deficiency, MIM# 618156	8p23.1	2.58E-05	1.01E-05	49
GP1BB	Giant platelet disorder, isolated, MIM# 231200	22q11.2 LCRA-D	4.07E-04	1.77E-04	48
PCYT1A	Spondylometaphyseal dysplasia with cone-rod dystrophy, MIM# 608940	3q29	8.91E-05	2.27E-05	37
SACS	Spastic ataxia, Charlevoix-Saguenay type, MIM# 270550	13q12.12	9.20E-04	2.12E-04	37
TFRC	Immunodeficiency 46, MIM# 616740	3q29	9.09E-05	2.27E-05	36
MIPEP	Combined oxidative phosphorylation deficiency 31, MIM# 617228	13q12.12	1.06E-03	2.12E-04	32
PI4KA	Polymicrogyria, perisylvian, with cerebellar hypoplasia and arthrogryposis,	22q11.2 LCRA-D	1.97E-04	1.77E-04	32

	MIM# 616531				
TXNRD2	Glucocorticoid deficiency 5, MIM# 617825	22q11.2 LCRA-D	2.48E-04	1.77E-04	32
LMAN2L	Mental retardation, autosomal recessive, 52, MIM# 616887	2q11.2	3.88E-04	7.31E-05	30
TCTEX1D2	Short-rib thoracic dysplasia 17 with or without polydactyly, MIM# 617405	3q29	1.37E-04	2.27E-05	24
ATP2A1	Brody myopathy, MIM# 601003	16p11.2 distal	8.81E-04	1.36E-04	23
ERCC6	Cockayne syndrome, type B, MIM# 133540	10q11.21q11.23	1.12E-03	1.41E-04	22
TANGO2	Metabolic encephalomyopathic crises, recurrent, with rhabdomyolysis, cardiac arrhythmias, and neurodegeneration, MIM# 616878	22q11.2 LCRA-D	1.14E-03	1.77E-04	21
LZTR1	Noonan syndrome 2, MIM# 605275	22q11.2 LCRA-D	1.16E-03	1.77E-04	20
B9D1	Joubert syndrome 27, MIM# 617120	Smith Magenis Syndrome	4.10E-05	5.04E-06	15
MERTK	Retinitis pigmentosa 38, MIM# 613862	2q13	1.28E-03	1.29E-04	14
ZNHIT3	PEHO syndrome, MIM# 260565	17q12	5.48E-04	6.89E-05	14
ABCC6	Lung disease, immunodeficiency, and chromosome breakage syndrome, MIM# 617241	16p13.11	4.28E-03	3.13E-04	12

To this end, we analyzed the distributions of a near-complete catalogue of disease alleles in patients affected with one of the 13 recessive disorders whose carrier burden are predicted to be almost exclusively from NAHR-mediated large deletions. 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) those affected with homozygous small variants possibly from a close- or distant- consanguineous relationship, (2) those affected with compound heterozygous small variants, and (3) those affected with a large deletion in *trans* with a small variant. We anticipate that category #1 accounts for a substantial proportion, demonstrating the robustness of the autozygosity mapping method in new recessive gene discovery (as populational rare alleles can be escalated to much higher clan allele frequency).[5] In outbred populations corresponding to categories #2 and #3, the working hypothesis is that #3 should account for a much higher fraction. Otherwise, the opposing trend may 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.

We identified presumably unrelated patients from 40 families reported in the literature affected with one of these 13 recessive genes (Table S7). 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. As expected, 32 are in homozygous states in patients, with 30 of them being ultra-rare (not observed in

gnomAD v3.1). Among the remaining cases, six patients were affected with compound heterozygous small variants, and four families with large deletion + small variant. The recurrent deletions involved are 10q11.21q11.23 (n=2), DiGeorge (n=1), and proximal 16p11.2 (n=1). The observation of patients with compound heterozygous variants not skewing towards category #3 deviates from the expectation above, which provides preliminary evidence that current research and clinical practice in medical genetics have not devoted sufficient efforts to sequencing of individuals with recurrent genomic deletions in order to maximize disease gene/allele discovery.

To test the hypotheses delineated above about disease gene/allele discovery in individuals with segmental haploid genomes, 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 various 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,[19] and the critical gene at 2q13, *NPHP1*, has been already extensively studied.[20] We also excluded six subjects harboring a hemizygous deletion in the Xp22.31 *STS* locus. After excluding the 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 S5).

Unexpectedly, more than a quarter (26/95) of these subjects were found to have probable small variant 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. From the remaining 69 subjects with an apparent undiagnostic exome sequencing result, we identified 4 subjects with hemizygous rare variants in coding regions 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).

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/ 3.148x10^{-4}	*COX10* (NM_001303.3)	c.1277_1282dup (p.M426_L427dup)	chr17:14207158_ 14207163dup	0	VUS	NDAC
3[a]	17p12 HNPP/ 3.148x10^{-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/ 9.326x10^{-5}	*OTUD7A* (NM_130901.2)	c.2023_2066del (p.D675Hfs*188)	chr15:31484009_ 31484052del	6.58x10^{-5} [b]	VUS	NDGMC
5	16p11.2 proximal / 2.596x10^{-4}	*PRRT2* (NM_145239.2)	c.649dup (p.R217fs)	chr16: 29813703dup	1.472x10^{-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.

Subject #1 has clinical features including ataxia, developmental delay, microcephaly, and short stature. A recurrent 10q11.21q11.23 deletion[21] 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. 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 2). The *COX10* gene, defects of which cause mitochondrial complex IV deficiency (OMIM# 220110) inherited in an autosomal recessive manner, is the only disease-associated gene embedded in the deletion interval other than *PMP22*. The gene body of *COX10* spans the CMT1A-REP, the repeat sequence mediating the recurrent 17p12 deletion (Figure 1).[22; 23] 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. Based on our calculation, ~73% of all patients affected with biallelic *COX10* defects in an outbred population carry one HNPP deletion (Table 2). To test this hypothesis, we retrospectively investigated results from

596 patients suspected with a mitochondria 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. This finding, though under-powered due to the small sample size, is consistent with our prediction that most patients with COX10 deficiency carry one HNPP deletion allele.

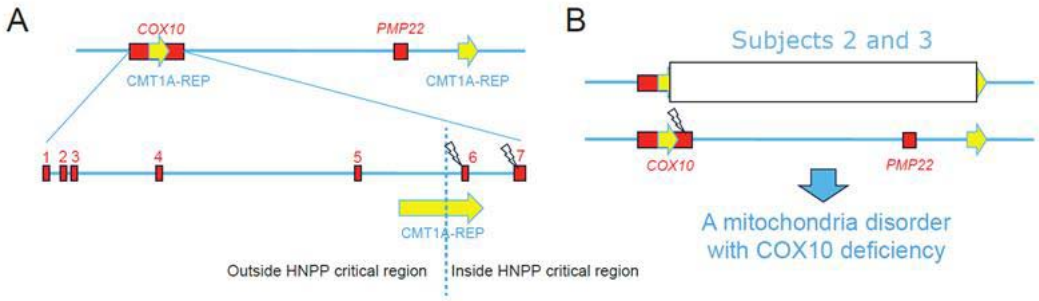

Figure 1. Compound heterozygous HNPP deletion and *COX10* variant leading to recessive COX10 deficiency in Subjects #2 and #3. **A.** The *COX10* gene spans the repeat sequence that mediate the recurrent HNPP deletion at chromosome 17p12. The *COX10* variant in the Subject #2 is located at the 3' end 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.

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 ID, epilepsy, behavioral problems, and variable dysmorphic features.[24; 25] 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 S6).[26-30] The critical gene responsible for the clinical presentations of the 15q13.3 BP4-BP5 deletion has been debated, but evidence suggest that *OTUD7A*, encoding a member of a family of deubiquitinating enzymes, may be a plausible candidate.[31; 32] Studies using syntenic heterozygous deletion mouse models suggested a critical role of Otud7a in neuronal development and brain function.[31; 32] Otud7a-null mouse models manifest many cardinal features of the 15q13.3 deletion syndrome.[31] 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 new 187 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[33] or compound heterozygous

15q13.3 deletion in *trans* with a frameshift *OTUD7A* variant,[34] corroborates that *OTUD7A* may play the critical 'driver gene' role in the 15q13.3 deletion syndrome through a biallelic autosomal recessive disease mechanism. Interestingly, the populational allele pool for *OTUD7A* is depleted for loss-of-function alleles (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.[35; 36]. The *PRRT2* gene is highly expressed in mouse brain and spinal cord during early embryonic development.[37] Heterozygous loss-of-function variants 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.[38] 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.[39] 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

biallelic pathogenic mechanism and an autosomal recessive inheritance model.[40] 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.[41]

Discussion

Considering the current clinical practice and research investigations, we postulate that the role of recurrent genomic deletions in contributing to recessive diseases is under-appreciated, therefore impeding discovery of new disease genes and alleles. Traditionally, large genomic deletions have been considered as dominant disorders, because they were almost all identified as heterozygotes by screening in symptomatic cohorts and frequently as *de novo* genomic CNV. Even though many recurrent deletions contribute to considerable carrier allele burden to individual recessive Mendelian disease genes encompassed or mapping within the genomic deletion, as discussed above, 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.[42]

Clinically, there has not been a consensus in 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 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. 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 envision 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 2). The autozygosity mapping approach has been the classical method for new disease gene discovery in medical genetics. This approach allows a straightforward strategy to target patient populations (by assessing their degree of autozygosty 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 nondiscriminatory nature of the high NAHR mutation rate. Nevertheless, this approach requires prior knowledge and screening of individuals for the recurrent deletions. It may be 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.[19; 35; 36; 43] 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.[44] While researchers are starting to sequence cohorts of individuals with large deletions,[45] it is imperative for clinical and diagnostic genomicists to foster guidelines that facilitate routine sequencing 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 of disease gene discoveries within intervals of genomic deletions, these patients will benefit from routine or perhaps even more prioritized reanalysis of sequencing data.[46]

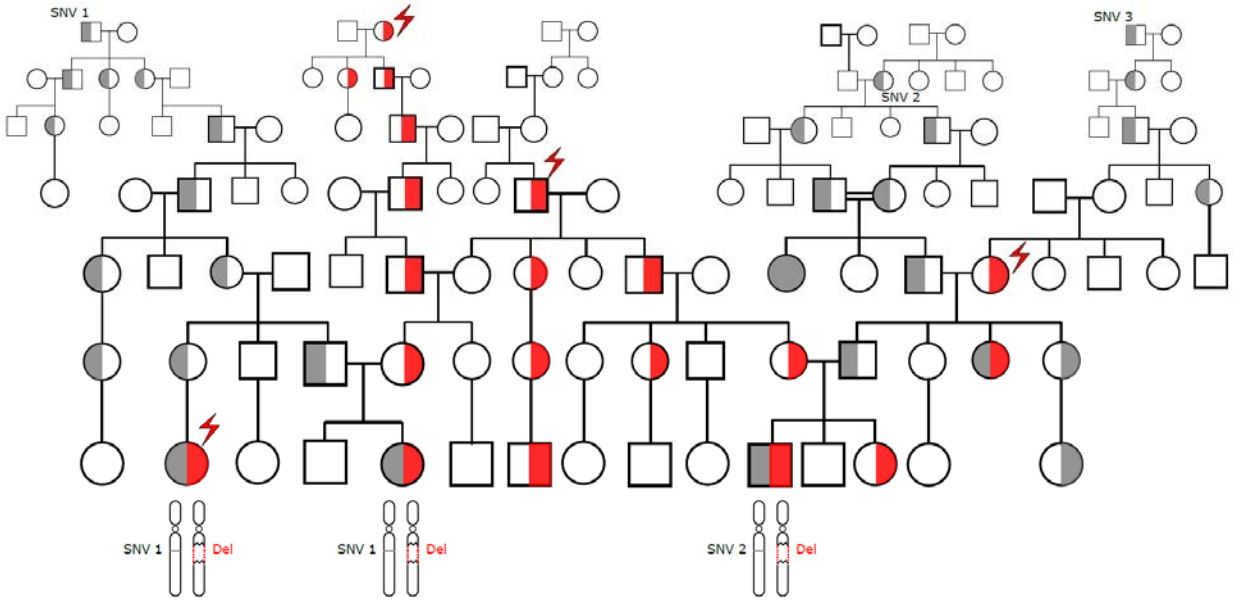

Figure 2. A 'segmental haploid genomics' approach for characterization of new disease genes or alleles for autosomal recessive conditions. The illustration above depicts dynamics of

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 thunder 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 of high degree of autozygoity, as illustrated in generations 3 and 4 for SNV2.

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. High prevalence recurrent genomic deletions are often associated with incomplete penetrance of disease phenotype of a high degree, ranging from 10-90%.[10] 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 not likely to be explained by the recessive model alone, because the observed disease penetrance is likely to dwarf the aggregation of the recessive disease allele prevalence (individually rare) on the non-deleted chromosome. It is plausible that an alternative disease model co-exists, in which the critical gene triggers disease presentation when one rare loss-of-function 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 thrombocytopenia with absent radii (TAR) syndrome [47], the *TBX6*-16p11.2 locus in association with congenital scoliosis[14], and the *F12*-5q35 Sotos deletion locus in association with blood clotting.[48] Our analysis in this study was focused on the 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 model is that large recurrent genomic deletions provide a unique perspective in characterization of new disease loci and alleles and performing human haploid genetics.

Methods

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

Hotspots of recurrent deletions in this raw map are 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, are collapsed. Metrics that can inform estimation of the new mutation rates for each genomic disorder are kept for each SD elements from the merged cluster of repeats, including repeat lengths, distance in between, and sequence similarity.[49] 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. Detailed code for generating the NAHR-deletion map is available at https://github.com/liu-lab/cnvNAHR/.

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[13] and the Icelandic cohort[12]. If the prevalence estimate is not significantly different between these cohorts (Fisher's exact test, $p>0.05$), the estimate based on the UK Biobank data is taken. Otherwise, prevalence from a third cohort (gnomAD SV[11] or other region specific literature) is used to compare with estimates from the UK Biobank and the Icelandic cohorts, and the group with a closer match is 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.[50] Of note, deletion prevalence estimates from the CMA cohort do not represent actual prevalence in the general population, but can inform relative prevalence comparison among rare events in the population.

Recessive disease carrier burden calculation

High-quality ClinVar variant is defined by having pathogenic or likely pathogenic label with at least one-star review status (accessed 01/21/2021). Loss-of-function variant from gnomAD SV is defined by variants meeting all the following criteria (1) having PASS filter in the VCF file with quality score over 500, (2) having PROTEIN_CODING__LOF flag in the VCF file, (3) POPMAX allele frequency lower than 1% and no homozygote counts, (4) is a deletion CNV, (5) have less than 80% of the span overlapping with segmental duplications, and (6) the loss-of-

function consequence affects all RefSeq transcripts of a recessive gene. High-confidence loss-of-function small variants from gnomAD v3.1 is defined by variants that fulfill all the following criteria: (1) having a 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 loss-of-function variant by LOFTEE[51].

Calculation of NAHR-deletion contribution to disease burden for a specific recessive disorder

For a recessive gene, given a collection of all disease alleles (n=*k*) and their allele frequencies ($p_1, p_2, p_3, \ldots, p_k$), if the *k*th allele is the NAHR-deletion allele, and the others are the small variant alleles, the probability of an individual carrying the NAHR-deletion allele is

$$P(A) = p_k$$

The probability for an individual to be affected with the recessive disorder is

$$P(B) = \sum_{i \leq j} p_i p_j$$

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

$$P(A \cap B) = \sum_{j} p_k p_j$$

For most large recurrent deletions, since homozygous loss of the deletion is incompatible with live birth, the probability of having NAHR-deletion alleles (p_k^2) need to be subtracted.

Finally, the NAHR-deletion contribution to recessive disease burden is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

All the above calculations are based on the concept of randomly forming individuals by sampling from a pool of alleles.

Retrospective analysis on exome sequencing data 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.[52; 53] The identification of deletions was based on a SNP array platform, which was performed concurrently with the exome assay.[54] Sanger sequencing was performed as a validation method for candidate diagnostic small variants.

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.

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.

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.

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

5. 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.

6. 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.

7. Dennis, M.Y., and Eichler, E.E. (2016). Human adaptation and evolution by segmental duplication. Current opinion in genetics & development 41, 44-52.

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

9. 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. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. Lupski, J.R. (2015). Cognitive phenotypes and genomic copy number variations. Jama 313, 2029-2030.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.
24. 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.
25. van Bon, B.W.M., Mefford, H.C., and de Vries, B.B.A. (1993). 15q13.3 Microdeletion. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, K. Stephens, andA. Amemiya, eds. (Seattle (WA).
26. 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.
27. 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.
28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

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.

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 721 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 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 S4. Carrier disease allele frequencies by allele.

Table S5. Molecular findings of recurrent deletions identified from clinical exome sequencing.

Table S6. Literature review for 15q13.3 recurrent deletions.

Table S7. Meta-analyses for literature reported patients affected with the 13 recessive disorders contributed by significant NAHR-mediated deletion burden.