Ohnologs are overrepresented in pathogenic copy number mutations

Aoife McLysaght[a,1], Takashi Makino[b], Hannah M. Grayton[c], Maria Tropeano[c], Kevin J. Mitchell[a], Evangelos Vassos[c,1,2], and David A. Collier[c,d,2]

[a]Smurfit Institute of Genetics, Trinity College, University of Dublin, Dublin 2, Ireland; [b]Department of Ecology and Evolutionary Biology, Graduate School of Life Sciences, Tohoku University, Sendai 980-8577, Japan; [c]Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London SE5 8AF, United Kingdom; and [d]Discovery Neuroscience Research, Eli Lilly and Company Ltd., Surrey GU20 6PH, United Kingdom

Edited by Arthur L. Beaudet, Baylor College of Medicine, Houston, TX, and approved December 2, 2013 (received for review May 16, 2013)

A number of rare copy number variants (CNVs), including both deletions and duplications, have been associated with developmental disorders, including schizophrenia, autism, intellectual disability, and epilepsy. Pathogenicity may derive from dosage sensitivity of one or more genes contained within the CNV locus. To understand pathophysiology, the specific disease-causing gene(s) within each CNV need to be identified. In the present study, we test the hypothesis that ohnologs (genes retained after ancestral whole-genome duplication events, which are frequently dosage sensitive) are overrepresented in pathogenic CNVs. We selected three sets of genes implicated in copy number pathogenicity: (*i*) genes mapping within rare disease-associated CNVs, (*ii*) genes within de novo CNVs under negative genetic selection, and (*iii*) genes identified by clinical array comparative genome hybridization studies as potentially pathogenic. We compared the proportion of ohnologs between these gene sets and control genes, mapping to CNVs not known to be disease associated. We found that ohnologs are significantly overrepresented in genes mapping to pathogenic CNVs, irrespective of how CNVs were identified, with over 90% containing an ohnolog, compared with control CNVs >100 kb, where only about 30% contained an ohnolog. In some CNVs, such as del15p11.2 (*CYFIP1*) and dup/del16p13.11 (*NDE1*), the most plausible prior candidate gene was also an ohnolog, as were the genes *VIPR2* and *NRXN1*, each found in short CNVs containing no other genes. Our results support the hypothesis that ohnologs represent critical dosage-sensitive elements of the genome, possibly responsible for some of the deleterious phenotypes observed for pathogenic CNVs and as such are readily identifiable candidate genes for further study.

microdeletion | microduplication | neurodevelopmental | evolution

Copy number variants (CNVs) are microdeletions or microduplications of segments of the genome, ranging from a few hundred base pairs to several megabases (1). CNVs include common polymorphic variants segregating in the population and rare mutations which can either be inherited or occur de novo. In human disease, the discovery and ability to characterize CNVs has expanded the range of known pathogenic genome variants, especially for common, complex neurodevelopmental disorders—autism spectrum disorders (ASD), intellectual disability, schizophrenia, epilepsy, attention deficit hyperactivity disorder (ADHD), and speech and language delay (2–6).

Deleterious mutations causing neurodevelopmental disorders are under strong negative genetic selection pressure (7). The majority of pathogenic CNV mutations are heterozygous deletions or single-copy duplications, i.e., increase or decrease gene dosage, and can be considered distinct from pathogenic loss of function mutations, which are not dosage effects but complete loss of function. For example, there are recessive homozygous deletions, such as nephronophthisis, in which gene deletion may be benign in heterozygous form but deleterious when homozygous (8). There are also allelic series, such as *NRXN1* (neurexin 1) gene deletions, where a heterozygous deletion is a risk factor for neurodevelopmental disorders (9) and homozygous deletion causes Pitt–Hopkins-like syndrome (Online Mendelian Inheritance in Man database, www.omim.org), a severe, syndromic form of neurodisability (10). They also tend to be pleiotropic, i.e., show association across neurodevelopmental phenotypes, and also have non-CNS phenotypes, including obesity and cardiac anomalies (11).

It is possible that genomic deletions are more likely to cause dosage sensitivity compared with duplications because the fold change is greater for deletions, i.e., there is a bigger proportional effect on dosage. There is evidence for this from nonallelic homologous recombination-mediated CNVs, such as 15q11.2 and 15q13.3, where the reciprocal mutational event should generate an equal number of deletions and duplications. The 15q11.2 duplication appears to have a benign or at least milder phenotype compared with the reciprocal deletion, and the population frequency of the duplication (0.39%) is greater than the deletion (0.18%), indicating the latter is under stronger negative genetic selection (5). On the other hand, overexpression in some developmental systems could have a bigger phenotypic effect than reduction in expression, and in some cases, both deletion and duplication are pathogenic, such as distal 16p11.2 (12).

Pathogenic CNVs provide an opportunity for stratified medicine, because as high or moderate risk factors they may allow the selection of patients for specific interventions, such as targeted pharmacotherapy, and they may clarify disease pathophysiology and the relationship between different diagnostic categories. However, it is not straightforward to identify the specific genes that give rise to the observed phenotypes. Thus, although a few

Significance

Copy number variants (CNVs) have recently emerged as an important cause of human disorders. Most pathogenic CNVs are large and contain many genes, and identification of the specific disease-causing genes has proven to be challenging. Using an evolutionary genetic approach, we tried to identify such genes through the mapping of ohnologs, genes retained after ancestral whole-genome duplication events. Comparing the proportion of ohnologs between potentially pathogenic and nonpathogenic sets of CNVs, we found that ohnologs are significantly overrepresented in the pathogenic sets. Our results support the hypothesis that ohnologs are ancestral dosage-sensitive elements that may be responsible for some of the deleterious phenotypes observed for CNVs.

Author contributions: A.M., T.M., H.M.G., K.J.M., E.V., and D.A.C. designed research; H.M.G., E.V., and D.A.C. performed research. A.M. and T.M. contributed new reagents/analytic tools; A.M., T.M., and E.V. analyzed data; and A.M., T.M., M.T., E.V., and D.A.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom corresponding may be addressed. E-mail: evangelos.vassos@kcl.ac.uk or aoife.mclysaght@tcd.ie.

[2]E.V. and D.A.C. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1309324111/-/DCSupplemental.

pathogenic CNVs can be mapped to a single gene, such as *NRXN1* or *VIPR2* (vasoactive intestinal peptide receptor 2), providing a direct route to the pathophysiology of disease, most pathogenic CNVs are large, covering several megabases, and typically delete or duplicate multiple genes (13, 14).

Phenotypes are also complex. For example, 22q11.2 deletion syndrome consists of variable combinations of over 190 phenotypic characteristics, including conotruncal heart and cleft palate defects in about 70% of cases (15). Identifying the deleted genes involved in neurodevelopmental phenotypes such as psychosis has proven difficult as there are a number of plausible candidates involved in neurotransmission, neuronal development, myelination, microRNA processing, and posttranslational protein modifications (16), including catechol-(*O*)-methyl transferase (17) and proline dehydrogenase (18).

Various strategies are possible to try and understand the relationship between specific genes and phenotypes within pathogenic CNVs. These include deletion mapping, i.e., searching for critical disease-associated intervals within larger CNV boundaries to narrow the number of genes involved; searching for inactivating mutations in single genes which carry the same phenotype as the pathogenic CNV (19); the use of bioinformatic information such as gene function and known disease associations; and the analysis of deleterious mouse knockout phenotypes to identify the most likely candidate gene(s) for a given CNV phenotype (20). Here we explore a unique evolutionary approach that identifies genes within pathogenic CNVs that have patterns of gene duplication characteristic of dosage-balanced genes (21).

Dosage balance may exist between two or more genes due to stoichiometric constraints of the protein complexes or the biochemical pathways in which the gene products participate (22, 23). Changes in gene copy number through gene duplication or loss lead to increases or decreases of gene product in the cell (24, 25). In the case of dosage-balanced genes such changes are deleterious and are removed by natural selection. Thus, a general feature of dosage-balanced genes is their low duplicability (26). An exception to this is duplication through whole-genome duplication (WGD) which increases all genes equally, maintaining the relative frequencies. Thus, WGD creates a unique opportunity for the duplication of dosage-balanced genes because it guarantees the simultaneous duplication of all components of a balanced set (26–28). Furthermore, once the genes have been duplicated, subsequent loss of individual genes would result in a dosage imbalance, thus leading to biased retention of dosage-balanced genes.

Two WGD events occurred early in the vertebrate lineage, generating initially tetraploid genomes. The genome duplication events were followed by extensive chromosomal rearrangement and loss of many duplicate copies (29–32). The 20–30% of the protein-coding genes in the human genome that can be traced back to these events (21, 32) have been postulated to have accelerated the evolution of vertebrate complexity, as they are enriched for developmental genes (33–35) and members of protein complexes (36).

Duplicated genes derived from WGD are known as "ohnologs" (37). Consistent with expectations, ohnologs have many characteristics of dosage-balanced genes. Mammalian ohnologs are more essential (i.e., knockout of one copy is more likely to result in sterility or inviability) than paralogs generated by small-scale duplication (SSD) and have comparable essentiality to singleton genes (36), i.e., copy number change is under negative selection. The overrepresentation of ohnologs in protein complexes is also consistent with the gene dosage balance hypothesis, which states that the stoichiometry of members of multisubunit complexes affects the function of the whole because of the kinetics and mode of assembly. Changing the dosage of a structural gene (for example by deletion or duplication) changes its expression in proportion to copy number (28, 38). If the product of a gene that changes in copy number belongs to a large protein complex, this may generate stoichiometric imbalance among members of the complex, resulting in a deleterious dominant negative phenotype. Finally, ohnologs were observed to be resistant to fluctuations in relative quantities by SSD and the CNV (21, 39), and, consistent with this, they are strongly associated with disease. In particular, Down syndrome caused by trisomy 21 appears to be caused in large part by the deleterious effects of the 1.5-fold increase in dosage of ohnologs on that chromosome (21).

In the present study we created three lists of genes from known or potentially pathogenic CNVs, and tested these for inclusion of ohnologs. These comprised CNVs associated with schizophrenia and other neurodevelopmental disorders, de novo CNVs under negative genetic selection identified in an Icelandic sample, which are possibly associated with disorders causing reduced fecundity (5), and CNVs from a comprehensive CNV morbidity map of neurodevelopmental delay (40). The proportion of ohnologs in these pathogenic CNV sets was compared with control copy number polymorphisms of a similar size range, which should not be under negative genetic selection, identified in the general population from two studies (41, 42). In addition, we tested whether the ohnologs mapping to pathogenic CNVs were more likely to encode proteins from multisubunit complexes, or showed enrichment for biological pathways, being thus more likely to represent dosage-balanced genes.

Results

Table 1 shows the list of CNVs associated with complex neurodevelopmental disorders (schizophrenia, autism, intellectual disability, neurodevelopmental delay, seizures, and epilepsy) that were used in the analysis. We included a total of 15 CNVs, with interval sizes ranging from 140 kb to 6.8 megabases (Mb). These covered total genome regions of 21.1 Mb containing 208 genes (minimal regions) and 37.9 Mb containing 369 genes (maximal regions). The list of CNVs is not exhaustive but includes CNVs with previous evidence of disease association from individual studies largely overlapping with a list of pathogenic CNVs proposed by Kirov et al. (43). In addition, we analyzed 440 genes that mapped to the Icelandic de novo CNV gene list (5), and 1,137 genes from the CNVs in the morbidity map list (40) (including additional genes not listed in the Cooper et al. paper, but which map within the CNV intervals).

Examination of these genes lists, plus the set of Icelandic CNVs under negative genetic selection (5) and morbidity map gene list (40), produced a list of 63 ohnolog genes which mapped to minimal CNV intervals, 93 ohnologs in the broad CNV intervals, 127 ohnologs in the Icelandic de novo CNV gene list, and 359 ohnologs from the morbidity map list (Table 2). Tables S1 and S2 show the list of genes and the ohnologs mapping to the narrow and broad intervals for the 15 established pathogenic CNV loci analyzed, and Tables S3 and S4 show the ohnologs mapping to the Icelandic de novo and the morbidity map lists, respectively.

More than 90% of the CNVs from the three pathogenic sets of genes contained an ohnolog, which is much greater than in the control CNV set, where about 30–48% of CNVs contained an ohnolog (Table 2). However, control CNVs were significantly smaller than pathogenic CNVs and thus less likely to contain an ohnolog because of encompassing fewer genes. To correct for CNV size, in the statistical analysis we also tested the proportion of genes that were ohnologs in each set of CNVs. We found a significant enrichment of ohnologs in all groups of pathogenic CNVs examined, with 63 of 208 (30%) in the narrow group of pathogenic CNVs, 127 of 440 (29%) in the Icelandic de novo CNVs, and 359 of 1,137 (32%) in the morbidity map set. The lowest proportion of ohnologs was observed in the broad definition of pathogenic CNVs (91 of 369, 25%), as expected. The

Table 1. Chromosomal bands and genome coordinates of the CNVs associated with schizophrenia and neurodevelopmental disorders

Locus	CNV region	Hg19 coordinates	N	Gene boundaries	Oh	Size (kb)	Associated disorders	Refs
1q21.1 del	Min	146,6-147,6	12	PRKAB2-NBPF24	4	1,050	Sz, ID, Epil, ASD	4, 5, 52, 53, 55
	Max	146,5-148,7	20	PRKAB2- LOC645166	4	2,220		
2p16.3 del	Min	49,8-51,4	1	NRXN1	1	1,600	Sz, ID, Epil, ASD	3, 9, 52, 55, 56
	Max	49,8-51,4	1	NRXN1	1	1,600		
3q29 del	Min	195,8-197,3	21	TFRC-BDH1	8	1,540	Sz, ID, ASD	55, 57, 58
	Max	195,8-197,3	21	TFRC-BDH1	8	1,540		
7q11.23 dup	Min	72,4-74,2	37	NSUN5P2-GTF2I	9	1,770	Sz, ASD	59, 60
	Max	72,4-74,2	37	NSUN5P2-GTF2I	9	1,770		
7q36.1 del	Min	145,3-148,1	1	CNTNAP2	1	2,860	Sz, ID	56, 61
	Max	145,3-148,1	1	CNTNAP2	1	2,860		
7q36.3 dup	Min	158,8-159,0	1	VIPR2	1	140	Sz	54, 55
	Max	157,3-159,0	7	PTPRN2-VIPR2	3	1,649		
10q11.22-23 dup	Min	49,3-51,6	28	FRMPD2-TIMM23	6	2,300	DD	62
	Max	46,8-52,4	60	FAM35BP-SGMS1	15	5,600		
15q11.2 del	Min	22,7-23,2	5	TUBGCP5-WHAMMP3	2	500	Sz, ID, ASD	5, 52, 63
	Max	22,7-23,2	5	TUBGCP5-WHAMMP3	2	500		
15q11-q13 dup	Min	23,8-28,6	14	MIR4508-HERC2	6	4,770	ASD	64
	Max	22,4-29,1	34	OR4M2- LOC100289656	8	6,709		
15q13.3 del	Min	32,3-32,5	1	CHRNA7	0	160	Sz, ID, Epil.	5, 6, 49, 52, 53, 55, 63, 65, 66
	Max	29,2-33,0	28	APBA2-SCG5	6	3,793		
16p13.11 dup/del	Min	15,5-16,3	9	MPV17L-ABCC6	4	850	Sz, ID, ASD, ADHD	46, 52, 63, 67
	Max	14,9-18,6	26	NOMO1-NOMO2	5	3,671		
16p11.2 dup	Min	29,7-30,1	27	SPN-GDPD3	9	452	Sz, ID, Epil., ASD, ADHD, BD	12, 55, 64, 68-70
	Max	29,6-30,2	32	SLC7A5P1-CORO1A	12	610		
16p12.1 del	Min	21,9-22,4	8	UQCRC2-CDR2	1	440	DD, Epil.	63, 64, 71
	Max	21,8-22,4	13	RRN3P1-CDR2	2	600		
17q12 del	Min	34,8-36,1	15	ZNHIT3-HNF1B	5	1,290	Sz, HNPP, DD	72, 73
	Max	34,4-36,1	24	CCL18-HNF1B	5	1,750		
22q11 del	Min	18,9-20,3	28	DGCR6-RTN4R	6	1,400	Sz, ID, ASD, ADHD, OCD, Mood	52, 53, 55, 64
	Max	18,9-21,8	65	DGCR6- RIMBP3C	12	2,900		

Minimal CNV region is defined as the smallest size typically associated, which should approximate to the "critical" deletion or duplication sufficient to cause the disease phenotype and maximal CNV region as the largest extent of deletion or duplication typically seen at the locus. ADHD, attention deficit hyperactivity disorder; ASD, autism spectrum disorder; BD, bipolar disorder; DD, developmental delay; Epil, epilepsy; HNPP, hereditary neuropathic pressure palsies; ID, intellectual disability; Mood, mood disorders (including depression); N, number of genes; OCD, obsessive compulsive disorder; Oh, number of ohnologs; Sz, schizophrenia.

density of ohnologs among genes mapping to pathogenic CNVs was significantly higher than that observed in the Wellcome Trust Case Control Consortium (WTCCC) control set (26 of 232; Table S5) or in the Conrad discovery set >100 kb (125 of 775; Table S6). These differences were highly statistically significant (Table 3). However, the ohnolog density in the Conrad HapMap control set (80 of 336; Table S7) was similar to the one observed in the broad group of pathogenic CNVs.

Because we previously showed that dosage-balanced ohnologs are significantly enriched in protein complex members, developmental genes, and transcription factors (21), we tested this finding in our list of CNVs associated with neurodevelopmental disorders (Table 1), and observed that a larger proportion of ohnologs mapping to these CNVs were members of protein complexes compared with ohnologs within control CNVs (20.9% vs. 10.4%), but this difference was not statistically significant (χ^2 test). Finally, we performed Gene Ontology (GO) analysis on the same set of pathogenic CNVs (Table 1) but we found no significant association with any biological process for ohnologs mapping to these CNVs.

Discussion

In the present study, we found that ohnologs were overrepresented in all three sets of genes mapping to pathogenic CNVs, with over 90% containing an ohnolog compared with about 30% of control CNVs over 100 kb in size, and the density of ohnologs was also higher. We also found that ohnologs in pathogenic CNVs are more frequently members of protein complexes, although this was not statistically significant. Our findings support the hypothesis that pathogenicity may derive from dosage sensitivity of one or more genes contained within the CNV locus.

Although we attempted to match CNVs for size, in our case CNVs were larger in size than our control CNVs. The reason control CNVs are smaller is likely to relate to the increased likelihood of larger CNVs being pathogenic (3, 12, 40). Therefore, for the statistical analysis we tested the proportion of genes that are ohnologs in each set of CNVs to control for CNV size, and this should be regarded as the main statistic of the analysis.

A limitation of the study is the differential CNV detection sensitivities of the genotyping methods (platforms and algorithms) used in different studies. Because our aim was not the direct comparison of the frequency of CNVs between cases and controls, but the identification of nonexhaustive lists of pathogenic and nonpathogenic CNVs, we were able to overcome this problem by making some assumptions: (*i*) irrespective of the method used for the studies identifying pathogenic CNVs, we accepted the findings when the same genotyping method was used in both cases and controls. The inclusion of false-positives probably reduces the power of our analysis and does not abate our findings that pathogenic CNVs contain more ohnologs. (*ii*) The nonpathogenic CNVs were identified through CNV discovery methods, which preserve a low false-positive rate. For our analysis the specificity of the control CNV list was more important than the sensitivity, as we do not require exhaustive but validated lists of control CNVs.

Table 2. Number and density of ohnologs within each class of CNV, and proportion of genes within CNVs which are ohnologs

	Cooper	Pathogenic (narrow)	Pathogenic (broad)	Iceland (de novo)	WTCCC controls	Conrad >100 kb	Conrad HapMap
CNVs (containing genes)	51	15	15	22	66	194	88
CNVs with an ohnolog	47	14	15	20	20	69	42
Ohnologs/CNV*	0.92	0.93	1.00	0.91	0.30	0.36	0.48
No. of genes	1,137	208	369	440	232	775	336
No. of ohnologs	359	63	91	127	26	125	80
Ohnolog/gene†	0.32	0.30	0.25	0.29	0.11	0.16	0.24
CNV total size (Mb)‡	100	21	38	44	19	65	33
Genes/1 Mb (CNV size)¶	11.35	9.81	9.87	10.00	12.25	11.84	10.32
Ohnologs/1 Mb (CNV size)¶	3.58	2.97	2.45	2.89	1.37	1.91	2.46

*Ohnolog/CNV is the proportion of CNVs containing at least one ohnolog.
†Ohnolog/gene is the density of ohnologs within a CNV, i.e., the number of ohnologs divided by the total number of known genes within the CNV intervals.
‡CNV total size is the total size of CNVs included in each list.
¶These last two rows in the table are the density of genes and ohnologs, respectively.

Our results are important for two reasons. First, from an evolutionary genetic perspective, they confirm the previous observation that ohnologs are more likely to be refractory to copy number change and represent a class of dosage sensitive genes. Second, from a clinical genetic perspective, they provide a tool to help determine which gene(s) from within a pathogenic CNV are more likely to be causative for the deleterious phenotypes observed at that locus. The ability to identify these is important for understanding the pathophysiology of neurodevelopmental disorders, necessary for the development of new treatments and preventative measures.

However, analysis of ohnologs alone will not provide sufficient information to determine which gene(s) within a CNV interval are pathogenic; not all ohnologs are dosage sensitive and many dosage-sensitive genes will not be ohnologs. It is interesting to consider the possibility that the pathogenicity of some CNVs may be due to the combined effect of several mildly dosage-sensitive genes which perhaps are not sufficiently dosage sensitive to show a phenotype when duplicated or deleted individually. Alternatively there may be a digenic phenomenon. One of these is likely to occur in 22q11.2 deletion syndrome, where there are many ohnologs, and a number of genes involved in neurotransmission, neuronal development, myelination, microRNA processing, and posttranslational protein modifications which could plausibly contribute to the neurobehavioral phenotypes (16). Most of the CNVs we examined contain multiple ohnologs, such as 16p13.11, with four ohnologs from nine genes in the critical region (Fig. S1). One ohnolog, *MYH11*, is thought to carry the risk of developing aortic dissection in patients with 16p13.11 duplications (44), whereas *NDE1*, also an ohnolog, is the strongest candidate for neurodevelopmental phenotypes associated with CNVs at this locus (45, 46). Two other ohnologs within the critical region, *ABCC1* and *ABCC6*, encode the multidrug resistance-associated proteins 1 and 6.

Table 3. Comparison of ohnolog presence and ohnolog density in pathogenic and control CNVs (Fisher's exact test *P* values)

	Cooper	Pathogenic (narrow)	Iceland (de novo)
Ohnolog/CNV*			
WTCCC controls	5.89E-12	8.26E-06	6.24E-07
Conrad >100 kb	9.98E-14	1.19E-05	4.98E-07
Conrad HapMap	4.62E-08	0.001	2.08E-04
Ohnolog/gene†			
WTCCC controls	2.55E-11	7.78E-07	8.09E-08
Conrad >100 kb	8.73E-15	1.04E-05	2.26E-07
Conrad HapMap	0.007	0.11	0.12

*The number of ohnologs per CNV.
†The density of ohnologs within a CNV, i.e., the number of ohnologs divided by the total number of known genes within the CNV intervals.

If we examine single genes that are pathogenic when deleted (*CHRNA7*, *VIPR2*, *CNTNAP2*, and *NRXN1*) (Table 1), three are ohnologs and one (*CHRNA7*) is not. *CHRNA7* is part of a larger family of 17 closely related mammalian ligand-gated ion channel genes that arose from SSD events (47). Their gene products form stochastic multiprotein complexes which can be heteromeric or homomeric; in their case, changes in gene dosage may affect the stoichiometry and thus function of these ion channels in the cell. In addition there is a human-specific chimeric gene to *CHRNA7* (*CHRFAM7A*), a partial duplication which arose through segmental duplication. The presence of *CHRFAM7A* could also contribute to *CHRNA7* dosage sensitivity, as it may be a dominant negative regulator of α7nAChR function (48).

The ohnolog gene list we have is unlikely to be perfect—there are likely to be false negatives (failure to identify a true ohnolog) and false positives (some genes identified may not in fact be ohnologs)—but we are confident that the errors are minimized. The fact that a gene is an ohnolog is not sufficient to determine that a particular gene within a genome imbalance locus is the source of deleterious phenotypes. Additional strategies are needed for the mapping of pathogenicity, including deletion mapping, i.e., searching for the smallest disease associated intervals within larger CNV boundaries to narrow the number of candidate genes involved, and screening for mutations within individual genes (19, 49).

In this regard, we observed that the proportion of ohnologs in our own definition of minimal regions at CNV loci is greater than in the broadly defined region. However, this difference was not statistically significant, and indeed for some CNV loci (for example 10q11.22-23) the proportion of ohnologs increased in the larger CNV. For the critical regions definition, we simply took the smallest region consistently deleted or duplicated; however, deeper phenotyping combined with case-control population association analysis will be needed to properly define critical CNV regions for specific phenotypes. Furthermore, as we know from analysis of *NRXN1*, where nonpathogenic intronic deletions are seen within disease population (9), pathogenic CNV loci may harbor nonpathogenic CNVs that could confound deletion-mapping studies.

In conclusion, we found that ohnologs are enriched in pathogenic CNVs, tend to map to multiprotein complexes, and that specific ohnolog genes within the interval have a plausible

connection with the observed phenotypes. This supports the hypothesis that these genes are critical dosage-sensitive elements of the genome and may be responsible for some of the deleterious phenotypes observed for pathogenic CNVs. This observation will also provide another tool to help determine the pathogenic gene(s) from within CNV intervals, although the ultimate certainty will be from identifying inactivating mutations in single genes within the interval that carry the same phenotype as the pathogenic CNVs (19).

Materials and Methods

Genes Mapping to Pathogenic CNVs. Three lists of pathogenic or potentially pathogenic CNVs were prepared. The first included a list of CNVs associated with schizophrenia and other neurodevelopmental disorders (Table 1). This was composed from published case-control studies that identified CNVs associated with schizophrenia, autism, neurodevelopmental delay, seizures, and intellectual disability. It is not an exhaustive list, and does not include many rare CNVs associated with severe or syndromic disorders. For each CNV we extracted the coordinates of the endpoints as defined by CNVs from the literature and from the Brain and Body Genetic Resource Exchange genome imbalance database (http://bbgre-dev.iop.kcl.ac.uk) or Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (http://decipher.sanger.ac.uk), which contain multiple incidences of pathogenic CNVs. These loci were defined by gene boundaries (inclusive distal and proximal gene) and then converted to HG19 (NCBI37/hg19). The genome-wide screen for CNVs in most disease-association studies was performed with SNP arrays usually validated by array comparative genomic hybridization (CGH). The CNV breakpoint accuracy depends on the platform used and the CNV detection algorithm. In general SNP arrays tend to underestimate the size of CNVs but high-resolution platforms are highly consistent in the assignment of start and end coordinates (50).

Because CNVs at individual loci can vary in size, we made two definitions at each locus, a minimal CNV region (the smallest size typically associated, which should approximate to the critical deletion or duplication sufficient to cause the disease phenotype) and a maximal CNV region (the largest extent of deletion or duplication typically seen at the locus). For example, because case-control association indicates that the pathogenic CNVs at 16p13.11 always include interval II (chr16: 15,380,000–16,230,000), we defined this as the critical region, whereas the maximal CNV locus was defined as including the broadest regions I, II, and III (chr16: 14,828,500–18,500,000). For the 22q11.2 deletion (velocardiofacial syndrome/DiGeorge syndrome) there are two typical-sized CNVs, 1.5 and 3 Mb (51). We defined the minimal CNV region (corresponding to the 1.5-Mb deletion), as the region chr22: 18,870,000–20,270,000 (1.4 Mb) containing the genes *DGCR6* (uc002zoh.2) at the proximal end and *RTN4R* (uc002zru.1) at the distal end, whereas the maximal CNV region chr22: 18,870,000–21,770,000 (2.9 Mb) (corresponding to the 3-Mb deletion), was defined referring to the size of the broadest 22q11.2 deletions identified by Kirov et al. (52) and by the International Schizophrenia Consortium (53), and included the genes *DGCR6* (uc002zoh.2) at the proximal end and *RIMBP3B* (uc002zuq.2) at the distal end. This method was used to test the hypothesis that minimal or critical regions of pathogenic CNVs would show enrichment over maximal regions, however, some CNV loci also contained nonoverlapping deletions and duplications, such as those in *NRXN1* (2p16.3) (9, 52) and *VIPR2* (7q36.3) (54, 55). All definitions of CNV boundaries were performed blind to the identification of ohnologs in the region.

The second list consisted of 66 de novo CNVs identified in an Icelandic sample of trios and parent–offspring pairs (9,878 transmissions) enriched for those regions that mutate most often. These CNVs under negative selection pressure may confer risk of disorders that reduce the fecundity of affected individuals (5). The third list was taken from a large study presenting a CNV morbidity map of developmental delay (40).

Genes Mapping to Control CNVs. To test for enrichment of ohnologs in pathogenic CNVs we compared genes in these regions with genes in CNVs identified in control populations without disease. Nonpathogenic CNVs were taken from two studies by the Wellcome Trust CNV Project (www.sanger.ac.uk/research/areas/humangenetics/cnv/). The study by Conrad et al. (42) used a CNV discovery sample which comprised 20 HapMap (http://hapmap.ncbi.nlm.nih.gov/) individuals of European and 20 of West African ancestry. CNVs were identified with CGH using a set of 20 NimbleGen arrays with stringent calling criteria. In collaboration with the WTCCC, a CNV-typing array was designed which comprised 105,000 long oligonucleotide probes targeted to ~11,000 candidate CNV loci. This array was used to genotype 450 HapMap samples (180 CEU [Utah Residents (CEPH) with Northern and Western European ancestry], 180 YRI [Yoruba in Ibadan, Nigeria], 45 JPT [Japanese in Tokyo, Japan], and 45 CHB [Han Chinese in Bejing, China]) in this study, and ~19,000 individuals (including 3,000 controls) in a parallel study by the WTCCC (41). Validated CNVs larger than 100 kb from the discovery experiment and the two additional control samples (450 HapMap and 3,000 WTCCC controls) were used as three separate reference lists for our study.

Ohnologs. A total of 7,294 human ohnolog genes (duplicated genes derived from WGD) were defined as described in Makino and McLysaght (21). Ohnologs are genes duplicated at the base of the vertebrate tree and located on paralogous chromosomal regions.

Protein Complexes and GO. We obtained a list of members of human protein complex from the Human Protein Reference Database (HPRD) Release 9 (www.hprd.org). GO slim annotation for biological processes of human was downloaded from ftp://ftp.geneontology.org/pub/go/GO_slims. We excluded the GO ID no. GO:0008150 (biological process unknown). The number of each GO term assigned into ohnologs in pathogenic CNVs was counted. We calculated the *P* value for each GO term by comparison of the number of observed GO term with that of expected GO term based on hypergeometric distribution using all genes. The estimated *P* values were adjusted by Bonferroni correction.

ACKNOWLEDGMENTS. D.A.C. is funded by European Commission Seventh Framework Project PsychCNVs (www.psych-cnv.eu) Grant Agreement HEALTH-2007-2.2.1-10-223423. A.M. is funded by the Science Foundation Ireland and the European Research Council (ERC). The research leading to these results has received funding from the ERC under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC Grant Agreement 309834. E.V. is funded by Guy's and St. Thomas' Charity Grant R080529.

1. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. *Nat Rev Genet* 7(2):85–97.
2. Sebat J, et al. (2007) Strong association of de novo copy number mutations with autism. *Science* 316(5823):445–449.
3. Walsh T, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320(5875):539–543.
4. Mefford HC, et al. (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* 359(16):1685–1699.
5. Stefansson H, et al.; GROUP (2008) Large recurrent microdeletions associated with schizophrenia. *Nature* 455(7210):232–236.
6. Helbig I, et al. (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nat Genet* 41(2):160–162.
7. Rees E, Moskvina V, Owen MJ, O'Donovan MC, Kirov G (2011) De novo rates and selection of schizophrenia-associated copy number variants. *Biol Psychiatry* 70(12):1109–1114.
8. Bollee G, et al. (2006) Nephronophthisis related to homozygous NPHP1 gene deletion as a cause of chronic renal failure in adults. *Nephrol Dial Transplant* 21(9):2660–2663.
9. Rujescu D, et al.; GROUP Investigators (2009) Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* 18(5):988–996.
10. Zweier C, et al. (2009) CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. *Am J Hum Genet* 85(5):655–666.
11. Malhotra D, Sebat J (2012) CNVs: Harbingers of a rare variant revolution in psychiatric genetics. *Cell* 148(6):1223–1241.
12. Weiss LA, et al.; Autism Consortium (2008) Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 358(7):667–675.
13. Sasaki M, Lange J, Keeney S (2010) Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Biol* 11(3):182–195.
14. Vissers LE, Stankiewicz P (2012) Microdeletion and microduplication syndromes. *Methods Mol Biol* 838:29–75.
15. Friedman MA, et al. (2011) Cleft palate, retrognathia and congenital heart disease in velo-cardio-facial syndrome: A phenotype correlation study. *Int J Pediatr Otorhinolaryngol* 75(9):1167–1172.
16. Schreiner MJ, Lazaro MT, Jalbrzikowski M, Bearden CE (2012) Converging levels of analysis on a genomic hotspot for psychosis: Insights from 22q11.2 deletion syndrome. *Neuropharmacology* 68:157–173.
17. van Amelsvoort T, et al. (2008) Effects of a functional COMT polymorphism on brain anatomy and cognitive function in adults with velo-cardio-facial syndrome. *Psychol Med* 38(1):89–100.
18. Raux G, et al. (2007) Involvement of hyperprolinemia in cognitive and psychiatric features of the 22q11 deletion syndrome. *Hum Mol Genet* 16(1):83–91.
19. Koolen DA, et al. (2012) Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nat Genet* 44(6):639–641.
20. Boulding H, Webber C (2012) Large-scale objective association of mouse phenotypes with human symptoms through structural variation identified in patients with developmental disorders. *Hum Mutat* 33(5):874–883.

21. Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci USA* 107(20):9270–9274.
22. Veitia RA (2002) Exploring the etiology of haploinsufficiency. *BioEssays* 24(2):175–184.
23. Veitia RA (2003) Nonlinear effects in macromolecular assembly and dosage sensitivity. *J Theor Biol* 220(1):19–25.
24. Henrichsen CN, et al. (2009) Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* 41(4):424–429.
25. Torres EM, et al. (2010) Identification of aneuploidy-tolerating mutations. *Cell* 143(1): 71–83.
26. Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424(6945):194–197.
27. Veitia RA (2004) Gene dosage balance in cellular pathways: Implications for dominance and gene duplicability. *Genetics* 168(1):569–574.
28. Veitia RA (2005) Gene dosage balance: Deletions, duplications and dominance. *Trends Genet* 21(1):33–35.
29. Ohno S (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol* 10(5):517–522.
30. McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet* 31(2):200–204.
31. Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3(10):e314.
32. Nakatani Y, Takeda H, Kohara Y, Morishita S (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* 17(9):1254–1265.
33. Blomme T, et al. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* 7(5):R43.
34. Brunet FG, et al. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23(9):1808–1816.
35. Hufton AL, et al. (2008) Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res* 18(10):1582–1591.
36. Makino T, Hokamp K, McLysaght A (2009) The complex relationship of gene duplication and essentiality. *Trends Genet* 25(4):152–155.
37. Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2(5):333–341.
38. FitzPatrick DR (2005) Transcriptional consequences of autosomal trisomy: Primary gene dosage with complex downstream effects. *Trends Genet* 21(5):249–253.
39. Pessia E, Makino T, Bailly-Bechet M, McLysaght A, Marais GA (2012) Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proc Natl Acad Sci USA* 109(14):5346–5351.
40. Cooper GM, et al. (2011) A copy number variation morbidity map of developmental delay. *Nat Genet* 43(9):838–846.
41. Craddock N, et al.; Wellcome Trust Case Control Consortium (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464(7289):713–720.
42. Conrad DF, et al.; Wellcome Trust Case Control Consortium (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464(7289): 704–712.
43. Kirov G, et al. (2013) The penetrance of copy number variations for schizophrenia and developmental delay. *Biol Psychiatry*, 10.1016/j.biopsych.2013.07.022.
44. Kuang SQ, et al.; GenTAC Investigators (2011) Recurrent chromosome 16p13.1 duplications are a risk factor for aortic dissections. *PLoS Genet* 7(6):e1002118.
45. de Kovel CG, et al. (2010) Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* 133(Pt 1):23–32.
46. Ingason A, et al.; GROUP Investigators (2011) Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry* 16(1):17–25.
47. Tsunoyama K, Gojobori T (1998) Evolution of nicotinic acetylcholine receptor subunits. *Mol Biol Evol* 15(5):518–527.
48. Araud T, et al. (2011) The chimeric gene CHRFAM7A, a partial duplication of the CHRNA7 gene, is a dominant negative regulator of α7*nAChR function. *Biochem Pharmacol* 82(8):904–914.
49. Hoppman-Chaney N, Wain K, Seger PR, Superneau DW, Hodge JC (2013) Identification of single gene deletions at 15q13.3: Further evidence that CHRNA7 causes the 15q13.3 microdeletion syndrome phenotype. *Clin Genet* 83(4):345–351.
50. Pinto D, et al. (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 29(6):512–520.
51. McDermid HE, Morrow BE (2002) Genomic disorders on 22q11. *Am J Hum Genet* 70(5):1077–1088.
52. Kirov G, et al.; International Schizophrenia Consortium; Wellcome Trust Case Control Consortium (2009) Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet* 18(8):1497–1503.
53. International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455(7210):237–241.
54. Vacic V, et al. (2011) Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* 471(7339):499–503.
55. Levinson DF, et al. (2011) Copy number variants in schizophrenia: Confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* 168(3):302–316.
56. Gregor A, et al. (2011) Expanding the clinical spectrum associated with defects in CNTNAP2 and NRXN1. *BMC Med Genet* 12:106.
57. Willatt L, et al. (2005) 3q29 microdeletion syndrome: Clinical and molecular characterization of a new syndrome. *Am J Hum Genet* 77(1):154–160.
58. Mulle JG, et al. (2010) Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet* 87(2):229–236.
59. Sanders SJ, et al. (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 70(5):863–885.
60. Van der Aa N, et al. (2009) Fourteen new cases contribute to the characterization of the 7q11.23 microduplication syndrome. *Eur J Med Genet* 52(2-3):94–100.
61. Friedman JI, et al. (2008) CNTNAP2 gene dosage variation is associated with schizophrenia and epilepsy. *Mol Psychiatry* 13(3):261–266.
62. Stankiewicz P, 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(1):165–179.
63. Mefford HC, et al. (2010) Genome-wide copy number variation in epilepsy: Novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS Genet* 6(5): e1000962.
64. Marshall CR, et al. (2008) Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 82(2):477–488.
65. Sharp AJ, et al. (2008) A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* 40(3):322–328.
66. Miller DT, et al. (2009) Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *J Med Genet* 46(4): 242–248.
67. Hannes FD, et al. (2009) Recurrent reciprocal deletions and duplications of 16p13.11: The deletion is a risk factor for MR/MCA while the duplication may be a rare benign variant. *J Med Genet* 46(4):223–232.
68. Kumar RA, et al. (2008) Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* 17(4):628–638.
69. McCarthy SE, et al.; Wellcome Trust Case Control Consortium (2009) Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 41(11):1223–1227.
70. Jacquemont S, et al. (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* 478(7367):97–102.
71. Girirajan S, et al. (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* 42(3):203–209.
72. Moreno-De-Luca D, et al.; SGENE Consortium; Simons Simplex Collection Genetics Consortium; GeneSTAR (2010) Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet* 87(5):618–630.
73. Sharp AJ, et al. (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* 38(9):1038–1042.