# DUF1220 Domains, Cognitive Disease, and Human Brain Evolution

### L. Dumas and J.M. Sikela

University of Colorado Denver School of Medicine, Aurora, Colorado 80045 Correspondence: james.sikela@ucdenver.edu

We have established that human genome sequences encoding a novel protein domain, DUF1220, show a dramatically elevated copy number in the human lineage (>200 copies in humans vs. 1 in mouse/rat) and may be important to human evolutionary adaptation. Copy-number variations (CNVs) in the 1q21.1 region, where most DUF1220 sequences map, have now been implicated in numerous diseases associated with cognitive dysfunction, including autism, autism spectrum disorder, mental retardation, schizophrenia, microcephaly, and macrocephaly.

We report here that these disease-related 1q21.1 CNVs either encompass or are directly flanked by DUF1220 sequences and exhibit a dosage-related correlation with human brain size. Microcephaly-producing 1q21.1 CNVs are deletions, whereas macrocephaly-producing 1q21.1 CNVs are duplications. Similarly, 1q21.1 deletions and smaller brain size are linked with schizophrenia, whereas 1q21.1 duplications and larger brain size are associated with autism. Interestingly, these two diseases are thought to be phenotypic opposites. These data suggest a model which proposes that (1) DUF1220 domain copy number may be involved in influencing human brain size and (2) the evolutionary advantage of rapidly increasing DUF1220 copy number in the human lineage has resulted in favoring retention of the high genomic instability of the 1q21.1 region, which, in turn, has precipitated a spectrum of recurrent human brain and developmental disorders.

### GENOME-WIDE SURVEY OF LINEAGE-SPECIFIC GENE COPY-NUMBER GAIN AND LOSS IN HUMAN AND GREAT APE EVOLUTION

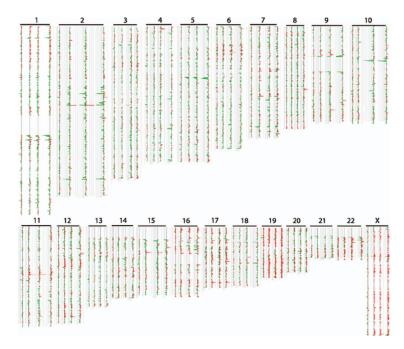
At a fundamental level, evolution has been characterized as a change in the allele frequency of a gene. More precisely, it is an alteration in the frequency of a genome sequence and may or may not involve a gene. It has been proposed that the primary types of genome alterations that underlie evolutionary change are single-nucleotide substitution, chromosomal rearrangement, and gene duplication. In 1970, Ohno (1970) put forth the argument that gene duplication is a primary mechanism of evolutionary change, due to the relaxation of selection and increase in variation afforded by its built-in redundancy. This view has also been expressed by W.H. Li (1997): "There is now ample evidence that gene duplication is the most important mechanism for generating new genes and new biochemical processes that have facilitated the evolution of complex organisms from primitive ones." More recently, a similar view has been expressed, albeit more succinctly, by E.E. Eichler (2001): "Exceptional duplicated regions underlie exceptional biology."

Given the importance of gene duplication to evolutionary change, we initiated a collaboration with Jonathan Pollack at Stanford to generate the first genome-wide and first gene-based application of array comparative genomic hybridization (arrayCGH) across human and nonhuman primate lineages (Fortna et al. 2004). The approach surveyed humans and four great ape species (bonobo, chimpanzee, gorilla, and orangutan) using multiple individuals from each. The arrays were generated from >41,000

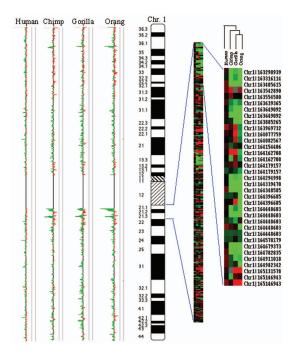
human cDNA clones (~24,000 genes), from which full inserts were amplified by polymerase chain reaction (PCR) and spotted. Each arrayCGH experiment represented a pairwise comparison of a reference DNA sample (always human) labeled with a green fluorescent dye and a test DNA sample (either human or a sample from one of the four great apes) labeled with a red fluorescent dye. Comparison of resulting array signals indicated that at least 1004 genes could be identified that gave changes in hybridization intensity consistent with lineage-specific gains or losses in gene copy number (Fig. 1). Several sets of control experiments were used to verify that the data were lineage-specific the result of copy-number changes, rather than sequence divergence.

Analysis of resulting arrayCGH data not only identified numerous candidate genes for a wide range of lineage-specific traits, but, by parsimony, also allowed one to determine when, in recent primate evolution, each of these events was likely to have occurred. In this manner, a genome-wide evolutionary history of lineage-specific gene-copy-number gain (duplication) and loss could be reconstructed covering the past 16 million years, extending from the last common ancestor (LCA) of the human and great ape lineages to the present (Fortna et al. 2004).

In addition, the data revealed that gene-copy-number changes showed a strong positional bias with respect to genome location and, in many cases, corresponded to regions known to be evolutionarily active. Among the most prominent of these were 1q21.1 (where the majority of DUF1220 domain sequences map; Fig. 2), the pericentromeric region of chromosome 9, and the fusion site on chromosome 2, which is the site at which two ancestral ape



**Figure 1.** Caryoscope view of genome-wide interhominoid arrayCGH analysis. cDNA arrays were used to perform pairwise comparisons of a reference genomic DNA (always human) labeled green and a test genomic DNA (either human, chimp, gorilla, or orangutan) labeled red. Caryoscope data are plotted as results of a five cDNA-sliding window with results from human, chimp, gorilla, and orangutan (reading right to left) for each human chromosome (Fortna et al. 2004).



**Figure 2.** Enhanced image of interhominoid arrayCGH data for chromosome 1 and the 1q21.1 region. (*Left*) Caryoscope image of arrayCGH data from human, chimp, gorilla, and orangutan, as described in Fig. 1. (*Right*) TreeView enlargement of the same data for the 1q21.1 region. In this view, each vertical column shows data from an arrayCGH experiment, whereas each horizontal row represents data for a single cDNA, organized in the order in which the genes occur in the genome (Fortna et al. 2004).

chromosomes fused to generate human Chr 2. Whereas 1q21.1 and 9p13/9q13 showed the highest concentration of genes giving human lineage-specific duplications, the fusion region on chromosome 2 contained a number of virtually contiguous genes that exhibited dramatic copy-number changes specific for several different lineages, all within a narrow (400 kb) genomic interval (Fortna et al. 2004).

There was also a pronounced enrichment of genes that showed increases in copy number specifically in the African Great Apes, compared to humans and orangutans. This finding has been reinforced by a recent report (Marques-Bonet et al. 2009) indicating that a burst of gene duplication occurred in the African Great Ape lineages. Interestingly, at least some of these African Great Ape—specific gene-copy-number increases may represent independent expansions within the African Great Ape clade; e.g., the gene was duplicated in the gorilla independent of its duplication in the *Pan* (chimp/bonobo) lineages (Fortna et al. 2004; Marques-Bonet et al. 2009).

### LINEAGE-SPECIFIC GENE-COPY-NUMBER VARIATION SPANNING 60 MILLION YEARS OF HUMAN AND PRIMATE EVOLUTION

Because of the highly informative results that we obtained from a comparison of the human and great ape genomes, we extended the application of arrayCGH to five evolutionarily more distant primate lineages (Dumas et al. 2007). By using human cDNA arrays and retaining

humans as the reference DNA in all arrayCGH experiments, we were able to interrelate data from all 10 primate species tested: human, bonobo, chimp, gorilla, orangutan, gibbon, macaque, baboon, marmoset, and lemur. This series of experiments further extended the period of primate evolutionary history that we could survey. The most distant comparison was between lemur and human, the LCA of which is estimated to have occurred ~60 million years ago. Although we were aware that sequence divergence may contribute to arrayCGH signals and that this would be more pronounced as more and more distantly related species were compared, these data indicated that valid copy-number changes could be detected even among the most distant primate species. For example, the many genes giving strong red signals in the lemur to human comparison (indicative of copy-number gains in lemur relative to humans) indicated that the method could reliably pick up copy-number gains between these highly divergent species and that the gene-copy-number expansions which were giving hybridization signals indicative of lemur-specific copy-number increases must be dramatic enough to override any contribution that sequence divergence was making to the arrayCGH signals.

This study identified 4159 genes that showed lineage-specific copy-number changes among these 10 primate species, including many that were specific to multiple lineages (Fig. 3). Among these were genes that were (1) increased in human and ape lineages relative to monkey lineages, (2) increased only in Old World monkeys relative to the other primates tested, and (3) increased in African Great Apes relative to the other species tested. If one applies parsimony to the results, the copy-number changes can be positioned at specific points in primate evolutionary time.

The number of genes showing lineage-specific copynumber changes for each lineage showed a general correlation with the age of the lineage (the age of the LCA with humans). Although the gibbon was an exception, showing an increased number of lineage-specific changes relative to its evolutionary age, this was consistent with the unusually high degree of chromosomal rearrangement that has been found within this species. In agreement with the overall trend, application of a tree-building program to this extended arrayCGH data set recapitulated the established evolutionary relationship of the species to a very high approximation (Fig. 4) (Dumas et al. 2007).

# HUMAN LINEAGE-SPECIFIC AMPLIFICATION OF DUF1220 DOMAINS

Among the genes identified in the human and great ape arrayCGH study were 134 that exhibited hybridization signals consistent with human lineage-specific increases in copy number. To characterize these more completely, the cDNA inserts of each clone were sequenced and used as BLAT queries against available primate genome sequences (Popesco et al. 2006). The most striking finding was that one gene, *MGC8902*, encoded by cDNA IMAGE:843276, gave more than 49 hits in the human genome but only 10 and 4 in chimp and macaque genomes, respectively. On the

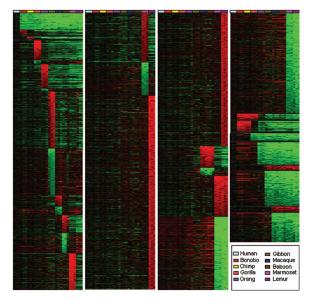
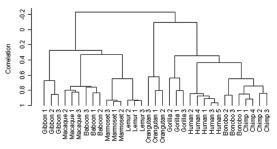


Figure 3. TreeView image of genes showing lineage-specific (LS) copy-number changes among 10 primate species. Tree-View image of 7318 genes giving LS arrayCGH signatures are shown for each of 10 lineages, including human (blue-gray), bonobo (rust), chimp (yellow), gorilla (orange), orangutan (purple), gibbon (green), macaque (blue), baboon (brown), marmoset (magenta), and lemur (light purple), as well as Old World monkeys (OWMs), marmoset and lemur, African Great Apes, Pan lineage (bonobo and chimp together), and combined arrayCGH-predicted changes relative to the remaining extended primates for the following groups: human and Pan lineage, human and African Great Apes, human and great apes, and human and all apes (great and lesser). The LS signals are grouped according to lineage and within each lineage are ordered, highest to lowest, according to the log, fluorescence ratio of the signal intensity of test sample to reference sample. Colors are displayed using a pseudocolor scale. Green and red signals correspond to LS decreases and increases, respectively, with respect to humans. (Modified, with permission, from Dumas et al. 2007.)



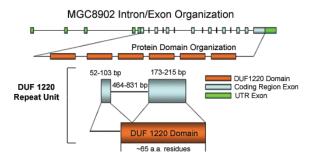
**Figure 4.** Correlation of arrayCGH data with established primate phylogeny. An unsupervised clustering of the raw data (arrayCGH log<sub>2</sub> ratios) was conducted to determine whether the individuals within each lineage cluster together as expected from the established phylogeny of these species. The complete raw data set, with each sample numbered as indicated, was used for the unbiased cluster analysis using BRB Array Tools (http://linus.nci.nih.gov/BRB-ArrayTools.html). Hierarchical clustering, using default settings for filtering, was conducted by array and included all 33 arrays from the study. The scale on the left represents the centered correlation (Pearson's coefficient) using average linkage analysis.

basis of BLAT hit span size, all human hits were shown to contain both exonic and intronic sequences and thus were not the products of retrotransposition. Further investigation of the IMAGE:843276 cDNA insert sequence indicated that it encoded six closely spaced copies of a predicted PFAM protein domain of unknown function (DUF1220). Follow-up analysis of DUF1220 domains indicated that they are ~65 amino acids in length and composed of a two-exon doublet (Fig. 5).

Sequences encoding DUF1220 domains continued to show a human lineage-specific increase in copy number when the five additional primate lineages were compared. PFAM has divided DUF1220 sequences into 11 seed domains, only one of which (O75042) is found outside primates, and there only as a single copy. All of the remaining 10 seed domains are primate specific, and, of these, O8IX162 gave the highest number of BLAT hits in the human genome. Remarkably, 37 of the 90 hits produced by Q8IX162 in the human genome gave a perfect sequence match. In humans, the DUF1220 family is found on 20-34 genes (also called the NBPF family; Vandepoele et al. 2005; Popesco et al. 2006) where they are present in from 1 to 50+ copies. Analysis of predicted DUF1220-encoding proteins indicates that the genes appear to be virtually devoid of other domains or functional signatures (Fig. 6), suggesting that the primary focus of selection was on increasing the number of DUF1220 domain copies in the genome, rather than integration of DUF1220 copies into known genes.

When all 11 DUF1220 seed domains are used for BLAT analysis and redundant hits are removed, we estimate that the human genome encodes 212 DUF1220 domains, whereas chimp, macaque, and mouse genomes encode 34, 30, and 1, respectively (Popesco et al. 2006). It should be noted that, with the exception of humans, these estimates are based on draft genome sequences, and these numbers are likely to vary when more recent genome assemblies are surveyed.

DUF1220 domains are thought to have undergone recent positive selection based on elevated  $K_a/K_s$  ratios (Popesco et al. 2006). Western analysis with antibodies directed against a human DUF1220 peptide show DUF1220-positive signals in several human tissues,



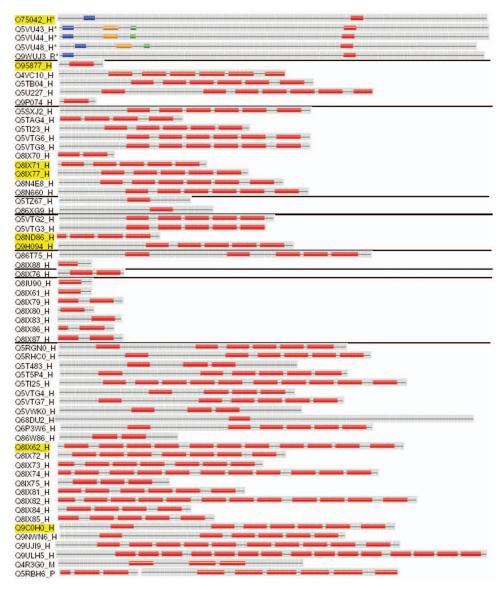
**Figure 5.** Genomic intron/exon organization of the DUF1220-encoding *MGC8902* gene and the predicted domain structure of the translated protein. A representative DUF1220 repeat unit is also shown. (Modified, with permission, from Popesco et al. 2006 [© AAAS].)

including the brain. Immunocytochemistry of postmortem brain indicates that DUF1220 proteins are expressed exclusively in neurons, and they seem to be enriched in cell bodies and dendrites (Popesco et al. 2006). Although DUF1220 sequences are found at three cytogenetic locations on chromosome 1, the great majority map to 1q21.1, which, along with the pericentromeric region of chromosome 9, are the two genomic regions that show the highest concentration of genes that exhibit human lineage-specific increases in copy number (Fig. 7).

# LINKING DUF1220 DOMAINS, BRAIN SIZE, AND COGNITIVE DISEASE

During the last few years, CNVs in the 1q21.1 region have been implicated in an increasingly higher number of human diseases, including idiopathic mental retardation (de Vries et al. 2005; Sharp et al. 2006), autism/autism spectrum disorder (Sharp et al. 2006; Autism Genome Project Consortium 2007; Mefford et al. 2008), congenital heart disease (Christiansen et al. 2004), schizophrenia (ISC 2008; Stefansson et al. 2008; Walsh et al. 2008; Need et al. 2009), microcephaly/macrocephaly (Brunetti-Pierri et al. 2008; Mefford et al. 2008), and neuroblastoma (Vandepoele et al. 2005; Diskin et al. 2009). These results have provided a number of provocative hints as to the possible role DUF1220 amplification may have in human evolution and human disease. For example, the evolutionary gene-copy-number studies that we have reported show that DUF1220 copy-number increases roughly parallel the increase in relative brain size, and/or neocortex expansion, that has occurred over recent primate and especially human evolution (Popesco et al. 2006; Dumas et al. 2007). Given this observation, it is intriguing that two reports that studied 1g21.1 CNVs and brain size found that deletions were associated with microcephaly, and duplications were associated with macrocephaly (Brunetti-Pierri et al. 2008; Mefford et al. 2008). Although no mention was made in these reports that DUF1220 copy number might underlie or be related to these phenotypes, we note here that these CNVs either encompassed or flanked DUF1220 domains (which, because of their highly duplicated nature, were not directly interrogated in these studies), raising the prospect that DUF1220 domain copy number (i.e., DUF1220 dosage) may be causally related to the observed differences in human brain size.

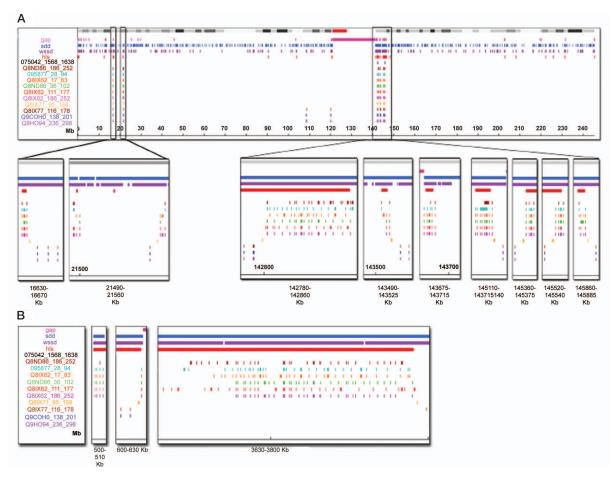
We point out here that a similar argument can be made regarding the 1q21.1 CNVs that have been reported to be involved in autism and schizophrenia. These diseases, as well as a number of pairs of other disorders, have been labeled genomic "sister disorders" by virtue of their tendency to exhibit diametrically opposite phenotypes and be caused by duplications versus deletions of the same genomic sequences (Crespi et al. 2009). Among the opposing phenotypes that distinguish autism and schizophrenia are a larger and smaller brain size, respectively (Crespi and Badcock 2008). It is therefore intriguing that the 1q21.1 CNVs that have been reported to underlie autism and schizophrenia tend to be duplications and deletions, respectively. In addition, these 1q21.1 CNVs, as with those



**Figure 6.** All InterPro proteins with DUF1220 domains. Schematic representations of all proteins in the InterPro database (www.ebi.ac.uk/interpro) that are predicted to have DUF1220 domains are shown. Protein identifications are shown on the right with accession numbers and letters. \_H, \_R, \_M, and \_P refer to proteins derived from human, rat, macaque, and orangutan, respectively. (Yellow) Accession numbers of proteins from which the 11 seed domains were derived. All protein schematics are to scale, but those marked with an asterisk are half-scale compared to the others. Horizontal lines separate groups of proteins derived from the same gene with the exception of the last group, which includes human proteins that have not been assigned to a gene. Red, green, yellow, and blue boxes are indicative of predicted DUF1220, ribosomal protein L29, prefoldin, and spindle-associated domains, respectively. (Modified, with permission, from Popesco et al. 2006 [© AAAS].)

underlying microcephaly and macrocephaly, encompass or are flanked by DUF1220 domains and span the same genomic interval. Taken together, these observations support the involvement of DUF1220 domains in the brain size differences that are known to exist between autistic and schizophrenic populations and between microcephalic and macrocephalic populations, and they suggest that a mechanistic link exists among 1q21.1 instability, the evolutionarily rapid DUF1220 copy-number amplification that has occurred in the human lineage, and the high prevalence of these diseases in human populations.

During the past several years, a number of microcephaly disease genes have been identified and these provide another link between DUF1220 and brain size. Specifically, it has been shown that of the half-dozen or so genes identified that cause microcephaly, a majority appear to encode proteins that are associated with the centrosome and control of the cell cycle (Bond and Woods 2006). Cell cycle control and the timing of when and where cells switch from symmetric to asymmetric cell division have been postulated to be key factors in changes in neuron number and brain size that have characterized mammalian and primate brain evo-



**Figure 7.** Human genomic locations of DUF1220-encoding sequences. The positions of genomic BLAT alignments to the 11 DUF1220 PFAM-predicted seed domains on chromosome 1 (*A*) and unordered chromosome 1 fragments (*B*) are noted along with the positions of gaps in the genome sequence. Notably, DUF1220 sequences are clustered around the pericentromeric C-band region on chromosome 1 and also on the unordered chromosome 1 sequences. Regions of particular interest are enlarged to show greater detail. The G-banding pattern is also shown, as well as positions of the centromere, known segmental duplications (sdd, wssd), and the BLAT hit positions of genes showing human lineage-specific increases in copy number.

lution (Rakic 1995). Given these observations, it is note-worthy that the ancestral DUF1220 domain (found once in mouse/rat and other nonprimate mammals) is found on myomegalin (PDE4DIP), a centrosomal protein whose gene is a homolog of *CDK5RAP2*, a non-DUF1220 microcephaly disease gene that also encodes a centrosomal protein (Bond et al. 2005; Dumas et al. 2007).

Recently, it has been reported that DUF1220 domains are one of only a small number of core duplicons that exist in the human genome (Jiang et al. 2007). These highly duplicated sequences have been the focal point for large-copy-number expansions in the human lineage, and in many cases, each core duplicon has been instrumental in recent chromosome-specific copy-number expansions. Typically, core duplicons are flanked by a mosaic of other duplicated sequences that appear to have been disseminated by being carried along during core duplicon transpositions. The DUF1220 domain duplicon appears to be responsible for much of the duplicated sequences found in the pericentromeric region of chromosome 1 (Jiang et

al. 2007). Core duplicon sequences are often interspersed and separated by single- or low-copy-number genes and often promote nonallelic homologous recombination (NAHR) events. Such a duplicon-rich genome architecture can often lead to NAHR between distantly spaced highly similar sequences (e.g., those of a core duplicon) that simultaneously produce duplications or deletions of the intervening sequences (segmental aneusomy). These intervening sequences often contain dosage-sensitive genes that are carried along with core duplicon duplications and deletions, resulting in disease-causing CNVs. The fact that DUF1220 sequences are associated with the numerous disease-related 1q21.1 CNVs mentioned above is fully consistent with such predictions of core duplicon behavior.

On the basis of these observations, we propose a model that links DUF1220 sequences to both human disease and human brain evolution. The model proposes the following specific testable hypotheses: (1) Increasing DUF1220 copy number may be related to increases in brain size; this

is based both on a general correlation of DUF1220 copy number with brain size between species (Fortna et al. 2004; Popesco et al. 2006; Dumas et al. 2007), as well as within the human population (Brunetti-Pierri et al. 2008; Mefford et al. 2008), (2) 1q21.1 instability promoted the rapid evolutionary copy-number increase of DUF1220 sequences, and (3) the evolutionary advantage that increased DUF1220 copy number conferred favored retention of high 1q21.1 instability, which, in turn, has resulted in many recurrent deleterious duplications (macrocephaly, autism) and deletions (microcephaly, schizophrenia) of dosage-sensitive 1q21.1 genes.

It is also noteworthy that the extreme instability of the 1q21.1 region increases the chances that somatic or de novo germ-line CNVs will occur, an observation that provides a potential explanation for the frequent examples of non-Mendelian inheritance that have been found to be associated with schizophrenia and autism. Finally, this mechanistic linking of 1q21.1 genomic instability with an evolutionarily favored process (i.e., DUF1220 copy-number increase) provides a means of reconciling the central paradox associated with schizophrenia and autism. Namely, why do diseases that are clearly maladaptive and have a significant genetic component persist at unusually high frequency throughout human populations?

# FINISHING THE 1q21.1 GENOME SEQUENCE

The 1q21.1 region is highly complex and shows a disproportionately high number of segmental duplications and sequence gaps (18), making precise assessment of its sequence content and organization difficult. As a result, caution should be exercised when relying on the current human and nonhuman primate genome sequences for 1q21.1. Before the complete evolutionary ontogeny of DUF1220 sequences can be reconstructed, and before direct comprehensive comparisons of the 1q21.1 region can be made among different individuals and among different species, a much more complete genome assembly of the region will be required. One potentially useful strategy will be the use of a haploid genomic resource: a human hydatiform mole bacterial artificial chromosome (BAC) library that has been constructed with this type of application in mind. Because of the repeat-rich nature of 1q21.1 and other genomic regions, the development of longer read sequencing technology would also provide a valuable tool for accurate genome finishing (Eid et al. 2009). Current available sequencing platforms have relatively short read lengths, a limitation that prevents accurate assembly of many of these repeat-rich regions. Such regions are increasingly being implicated in human disease, yet, due to this sequencing limitation, they continue to remain largely unexamined. Finishing 1q21.1 will provide a framework both for correctly annotating the DUF1220 domain content and organization in the human genome, and for precisely defining breakpoints and content of the many disease-related CNVs that are being identified in this evolutionarily important genomic region.

#### **CONCLUSIONS**

Presently available data suggest that the striking increase in DUF1220 domain copy number in recent evolutionary time has occurred because there is a clear adaptive advantage to greater numbers of DUF1220 domains and that this advantage and selection process has persisted throughout primate, and especially human, evolution. Indeed, as we have previously reported, sequences encoding DUF1220 domains are virtually all primate specific, show signs of strong positive selection, and are increasingly amplified generally as a function of a species' evolutionary proximity to humans, where the greatest number of copies (212) is found (Popesco et al. 2006). On the basis of the genome organization of DUF1220 sequences, the large number of human copies is not likely to have arisen by a single event, but rather by a series of small and large incremental increases, each of which conferred an adaptive advantage. This series of increases (involving multiple rounds of both gene and domain duplications) is likely to have been promoted by the high degree of genomic instability associated with the human 1q21.1 region. Such instability would have more frequently produced duplications and deletions in the region, with the DUF1220 duplications conferring a selective advantage, which in turn would result in retention of the instability in those who had more DUF1220 copies. This process could be viewed as a recurring cycle in which the increased 1g21.1 instability facilitated DUF1220 copynumber increases in certain individuals, followed by selection of individuals who exhibited the increased DUF1220 copy number, resulting in retention of the 1q21.1 instability in these individuals. In this manner, high 1q21.1 instability may have been selected for and retained in the human lineage because it more rapidly produced increases in DUF1220 copy number.

The 1q21.1 instability is a product of the genome architecture of the region and, because it works by promoting rearrangements in sequence organization and copy number, many deleterious events can be expected to occur along with the beneficial DUF1220 duplications. In this regard, the large number of recent reports implicating 1q21.1 CNVs in multiple human diseases should not be unexpected and should be viewed as a natural outcome of the genomic instability that is being favored because of the adaptive value that more DUF1220 copies may be conferring. In summary, the high number of 1q21.1 CNVs that are disease-causing may be the price that our species paid, and continues to pay, for the adaptive benefit of large numbers of DUF1220 domains.

#### **ACKNOWLEDGMENTS**

We thank the individuals who have contributed to the data presented here, including J. Pollack, A. Fortna, M. Popesco, E. MacLaren, Y. Kim, M. O'Bleness, J. Keeney, J. Hopkins, A. Karimpour-Fard, M. Cox, R. Berry, L. Meltesen, L. McGavran, G. Wyckoff, and L. Jorde. J.M.S. is supported by National Institute of Mental Health grant R01 MH81203, NIAAA grant 2 R01 AA11853, and a Butcher Foundation grant.

#### REFERENCES

- Autism Genome Project Consortium. 2007. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Genet 39: 319–328.
- Bond J, Woods CG. 2006. Cytoskeletal genes regulating brain size. Curr Opin Cell Biol 18: 95–101.
- Bond J, Roberts E, Springell K, Lizarraga SB, Scott S, Higgins J, Hampshire DJ, Morrison EE, Leal GF, Silva EO, et al. 2005. A centrosomal mechanism involving CDK5RAP2 and CENPJ controls brain size. *Nat Genet* **37**: 353–355.
- Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T, Lalani SR, Graham B, Lee B, Shinawi M, et al. 2008. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nat Genet* 40: 1466–1471.
- Christiansen J, Dyck JD., Elyas BG, Lilley M, Bamforth JS, Hicks M, Sprysak KA, Tomaszewski R, Haase SM, Vicen-Wyhony LM, Somerville MJ. 2004. Chromosome 1q21.1 contiguous gene deletion is associated with congenital heart disease. *Circ Res* 94: 1429–1435.
- Crespi B, Badcock C. 2008. Psychosis and autism as diametric disorders of the social brain. *Behav Brain Sci* 31: 241–320.
- Crespi B, Summers K, Dorus S. 2009. Genomic sister-disorders of neurodevelopment: An evolutionary approach. *Evol Appl* 2: 81–100.
- de Vries BB, Pfundt R, Leisink M, Koolen DA, Vissers LE, Janssen IM, Reijmersdal S, Nillesen WM, Huys EH, Leeuw N, et al. 2005. Diagnostic genome profiling in mental retardation. *Am J Hum Genet* 77: 606–616.
- Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mosse YP, Wood A, Lynch JE, et al. 2009. Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459: 987–991.
- Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM. 2007. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res* 17: 1266–1277.
- Eichler EE. 2001. Segmental duplications: What's missing, misassigned, and misassembled—And should we care? *Genome Res* 11: 653–656.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133– 138.
- Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T, et al. 2004. Lineage-specific gene duplication and loss in human

- and great ape evolution. PLoS Biol 2: 937-954.
- International Schizophrenia Consortium (ISC). 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455: 237–241.
- Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* 39: 1361–1368.
- Li WH. 1997. Molecular evolution. Sinauer, Sunderland, MA.
- Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* **457**: 877–881.
- Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, Huang S, Maloney VK, Crolla JA, Baralle D, et al. 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* **359**: 1685–1699.
- Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, Shianna KV, Yoon W, Kasperaviciute D, Gennarelli M, et al. 2009. A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS. Genet* **5:** e1000373.
- Ohno S. 1970. Evolution by gene and genome duplication. Springer, Berlin.
- Popesco MC, Maclaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, McGavran L, Wyckoff GJ, Sikela JM. 2006. Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science* 313: 1304–1307.
- Rakic P. 1995. A small step for the cell, a giant leap for mankind: A hypothesis of neocortical expansion during evolution. *Trends Neurosci* **18:** 383–388.
- Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, et al. 2006. Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* **38**: 1038–1042.
- Stefansson H, Rujescu D, Cichon S, Pietilainen OPH, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* **455**: 232–236.
- Vandepoele K, Van Roy N, Staes K, Speleman F, van Roy F. 2005. A novel gene family NBPF: Intricate structure generated by gene duplications during primate evolution. *Mol Biol Evol* **22**: 2265–2274.
- Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320: 539–543.