De Novo Rates and Selection of Schizophrenia-Associated Copy Number Variants

Elliott Rees, Valentina Moskvina, Michael J. Owen, Michael C. O'Donovan, and George Kirov

Background: At least 10 large and rare recurrent DNA copy number variants (CNVs) have been identified as risk factors for schizophrenia and other neurodevelopmental disorders. Because such conditions are associated with reduced fecundity, these pathogenic CNVs should be filtered out from the population by selection and must be replenished by de novo events.

Methods: To estimate the mutation rate (μ) for these CNVs and the selection pressure (s) against them, we first conducted a literature review on the rate of each of these CNVs in the population and the rate of their de novo occurrence. In each generation, the number of CNVs lost because of reduced fertility must be replenished by the same number of de novo CNVs. Therefore, the observed ratio of de novo versus all (inherited + de novo) CNVs approximates the selection coefficient (s) of that CNV. The mutation rate approximates to $\mu = s \times q$, where q is the frequency of the CNV in the population.

Results: High selection pressure operates at all these loci (s = .12 – .88), suggesting that following de novo occurrence, each of these CNVs persists in the population in only a few generations. The mutation rate for each CNV is high, affecting between 1:3500 and 1:30,000 individuals. The rarest CNVs have the highest selection coefficients.

Conclusions: The CNVs that increase risk to develop schizophrenia are caused by recent de novo mutations and are under strong selection pressure. They persist in the population because of high mutation rates.

Key Words: Autism, CNV, de novo, mutation, persistence, schizophrenia, selection

Recent research has implicated the role of large, rare DNA copy number variants (CNVs) as risk factors for schizophrenia. The best supported loci are deletions at 1q21.1, 3q29, 15q11.2, 15q13.3, 17q12, 22q11.2 (1–5); *Neurexin 1* (*NRXN1*) (2,6); and duplications at 15q11–13, 16p11.2, and 16p13.1 (7–9). The odds ratios (ORs) for carriers of these CNVs to develop schizophrenia are relatively high, ranging between 2 and greater than 30 (2,10). CNVs at each of these loci have also been implicated in additional neurodevelopmental phenotypes (11) including autism (9,12–14), attention-deficit/hyperactivity disorder (3,9,12,15), mental retardation (MR) and developmental delay (DD) (5,9,14,16–18), and epilepsy (19–21).

Patients with schizophrenia have fewer children compared with the general population (fertility ratio = .39, 95% confidence interval = .35–.44) (22) as do individuals affected with mental retardation (MR) (23) and epilepsy (24). Fecundity in those with other early-onset developmental disorders, including congenital abnormalities and autism, has received less research attention, but is also likely to be low. Given the strong reproductive disadvantages in those with these phenotypes, genetic variants that substantially increase risk of the disorders should be excluded from the population in a few generations and must be replenished through independent de novo mutation events (4,23,25,26). These predictions have been confirmed by empirical research that found de novo formation rates of CNVs to be higher in probands suffering from schizophrenia (27) and autism (25,28–30) compared with control subjects.

From the Department of Psychological Medicine and Neurology, Medical Research Council (MRC) Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, United Kingdom.

Address correspondence to George Kirov, Ph.D., MRCPsych, Department of Psychological Medicine and Neurology, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Henry Wellcome Building, Heath Park, Cardiff, United Kingdom; E-mail: Kirov@cardiff.ac.uk.

Received Jun 1, 2011; revised Jul 12, 2011; accepted Jul 14, 2011.

Itsara *et al.* (25) estimated the mutation rate (μ) and the strength of the purifying selection (s) that operates on large CNVs greater than 500 kb in size. The authors found a strong purifying selection for this class of CNVs (s = .16), that is, carriers have a 16% reduction in reproductive fitness. The mutation rate (μ) in the general population of CNVs greater than 500 kb was estimated at .0065 per genome per transmission. Until recently, information on de novo mutation rates for any specific CNV locus implicated in schizophrenia was available only for the 2.7-Mb deletion on 22q11.2, which results in the velo-cardio-facial/DiGeorge syndrome (31). There is consensus that the prevalence of the deletion is 1:4,000 live births (32), and the deletion is de novo in 72% to 92% of carriers (33,34).

The mutation rates and selection coefficients for the remaining individual CNV loci that increase risk of schizophrenia (listed earlier) have not been previously estimated. With the accumulating data on the prevalence of these CNVs, and de novo rates at these loci, we are now able to estimate these parameters. This should add to our understanding of their role in schizophrenia and in human disease in general.

Methods and Materials

To estimate the mutation rate of CNVs, we use the mutation-selection balance model proposed by Haldane (35) in the 1930s. According to the model, for a disease with a stable prevalence, deleterious mutations are excluded from the population at the same rate with which they are replaced by new mutations, thereby preserving the frequency of the mutations in the population at equilibrium. If the mutation rate (μ) and the equilibrium frequency of mutation-bearing genomes (q) are small, and assuming a dominant effect of the CNV, their relationship converges to $s = \mu/q$, where s is the selection coefficient (the selection pressure operating on the CNV) (23,25).

Here we estimated s and q based on the literature on the frequency of the 10 chosen schizophrenia-associated CNVs and from the rates of observed de novo and transmitted CNV at these loci (discussed subsequently). These two values allow us to derive μ from the foregoing formula. The estimation of confidence intervals

0006-3223/$36.00
doi:10.1016/j.biopsych.2011.07.011

(95% CI) for these values is based on the work by Itsara *et al.* (25) and presented in Supplement 1.

Estimating the Strength of the Selection Pressure (the Selection Coefficient, *s*)

To estimate the de novo ratio (the proportion of a given CNV in the population that has occurred as a de novo in the current generation), we only considered studies that reported in a systematic way the rate of de novo and inherited CNVs. To reduce publication bias in favor of overreporting de novos, we did not consider case reports, because de novo occurrences are likely to be considered of greater publication interest. Studies that only reported on de novo events but did not provide information on transmitted events also were not considered, because they would introduce a similar bias. To determine the proportions of the CNVs at a locus that are de novo in that population, we considered studies on any phenotype. The de novo ratio was then calculated by dividing the overall number of de novo events by the sum of the de novo and inherited events found in these studies. This is an estimate of the proportion of de novo CNVs out of the overall number of CNVs found in a population, at any particular locus. Assuming a simple mutation-selection balance model (as discussed earlier), this proportion equates to the selection coefficient (*s*) in the population. Only one event per multiply affected family was counted, where this was reported. Counting several individuals from a family would introduce a bias toward transmitted events, because researchers might have actively tried to contact other affected family members, who otherwise would not have been probands in the study. In the cases in which samples overlapped between studies, we took only data from the larger study, the one that used a denser microarray, or that used independent methods for validating the CNVs. The numbers of de novo and inherited CNVs found in each study that we used for our calculations, are listed in Table S1 in Supplement 1.

Estimating the Population Frequency of CNVs (*q*)

These have been established for "control" subjects in large data sets (e.g., Levinson *et al.* [2], Kirov *et al.* [10]). However, if a given CNV has a high penetrance for disorders that are underrepresented among populations recruited as healthy controls for a study (e.g., mental retardation, schizophrenia, congenital heart defects), then the frequency of that CNV among controls will be an underestimate of that in the population as a whole (*q*). To minimize this effect, we attempted to correct the estimates of *q* in the general population by taking into account the rate of these CNVs in disorders in which they are known to play a role and that are likely to be excluded from "control" populations (Table S3 in Supplement 1). We performed this correction for schizophrenia, MR/DD, autism, and congenital heart defects, if reliable estimates for the rates of CNVs in these disorders were available. Although higher rates for some of the CNVs are also reported for other disorders such as epilepsy, we did not take these into account, because we do not know whether such phenotypes are underrepresented among control groups.

To estimate the frequency of the CNVs in the general population and in any relevant disease groups, we considered only large studies, because they are more likely to reflect the real prevalence. As a rule, we considered publications based on 1000 subjects or more. Odds ratios for risk among carriers to develop schizophrenia were taken from the estimates made in meta-analyses or by combining only the largest available studies.

To extrapolate from the "control data" frequency of CNVs to that in the general population (*q*), we used the following approximate rates of these disorders in the general population, with the numbers rounded up to simplify the presentation:

1. Schizophrenia: 1%, which is the rate most widely cited by researchers (36).
2. Congenital heart disease (CHD): .5%; this is based on several estimates of the frequency of moderate to severe forms of CHD (37,38) and www.heartstats.org.uk.
3. MR/DD: 2%; the exact prevalence is not known and depends on the definition, but 2% is a widely accepted figure (39).
4. Autism spectrum disorder (ASD): 1%; this is based on reports of the prevalence of autism spectrum (40–42).

Details on the frequencies of CNVs in these disorders are presented in Table S2 in the Supplement 1.

General population frequencies (*q*) of each CNV were estimated as follows: the general population consists of individuals who are unaffected by the above disorders ("healthy controls") and groups of patients with the above phenotypes who we assume to have been excluded from those control groups. Each of these subpopulations contributes to the general population frequency of the CNV in the following proportions: schizophrenia ~1%, MR/DD ~2%, ASD ~1%, CHD ~.5%, controls ~95.5%. The size of the population is calculated from the size of the "control" group (comprising ~30,000–50,000 people, Table 1 and Table S3 in the Supplement 1) assuming that it contributes 95.5% of the size of the population (or more, if the CNV is not known to increase risk for some of the disease groups). Once the total number of CNVs and sample size had been calculated for the general population, *q* was estimated as (observed number of CNVs) / (sample size 2×2). For the 22q11.2 deletion, we used the accepted population prevalence estimates from the literature (Supplement 1).

The persistence (43) of a CNV in the general population (the average number of generations that a mutation remains in the population before it is eliminated) and the pervasiveness of a CNV (the average number of individuals carrying a mutation before it is eliminated) were estimated with the formulas derived by Garcia-Dorado *et al.* (44), which are in turn based on the work by Kimura and Ohta (45): pervasiveness $= 1/hs$, and persistence $= 2[\log_e(1/hs) + 1 - \gamma]$, where γ is Eulers's constant (~.5772) and *hs* is the deleterious effect of the heterozygous state, where *hs* (the heterozygous disadvantage) simplifies to *s* in the case of these CNVs (Supplement 1).

Results

The rate of CNVs in controls and the estimated *s*, *q*, and μ are presented in Table 1. Note that *q* is the rate in the general population divided by 2, as it refers to the frequency of the CNVs per haploid genome, and the "Rate in controls" refers to the percentage of control individuals who carry the CNV. The table also gives the estimates of the ORs for each CNV for schizophrenia. Details of how the de novo ratios and *q* were derived and the 95% CIs of *s*, *q*, and μ are presented in Supplement 1.

We found high selection pressure operating against all large schizophrenia-related CNVs ($s = .12–.88$). Using the formulas presented in Methods and Materials (44), we estimate that the most pathogenic of these CNVs (3q29, 17q12, and 22q11.2) are removed within less than two generations on average (persistence) and that none of the remaining CNVs are likely to persist in more than five generations (Table 2). Similarly, following the de novo event, each CNV is carried by no more than eight or nine individuals on average (pervasiveness), and this number is much smaller for the CNVs with the highest selection coefficients (*s*).

The schizophrenia CNVs show high and fairly uniform mutation rates (μ) ranging from 1.7×10^{-5} to 1.4×10^{-4}, with overlapping 95% CI across all loci (Figure 1A). These rates are within the range