Pre-registered Protocol. Copy number variation (CNV) and language phenotype severity in children with and without a sex chromosome trisomy

Hayley S Mountford, Dorothy V M Bishop, Paul A Thompson, Nuala H Simpson & Dianne F Newbury

16th June 2018

## Background

Children who carry an extra X or Y chromosome do not have any gross physical or mental abnormalities; however, there is a pronounced increase in the risk of language disorders and autistic features (Bishop et al, 2011). Investigation of the genetic correlates of language and communication in children with sex chromosome trisomies (SCT) may help understand the phenotypic variation seen in affected individuals, and could also lead to improved understanding of common neurodevelopmental disorders in children with a typical karyotype.

Variability of the neurocognitive phenotype is a notable feature of sex chromosome trisomies is the variability of the phenotype. Some children have severe difficulties including intellectual disability, behavioural problems or autism spectrum disorder (ASD), while others have little or no evidence of neurodevelopmental problems. A range of explanations for this phenotypic variability have been proposed. In a companion paper, Newbury et al (2018) considered the ‘double hit’ hypothesis, which maintains that the presence of an extra dose of neuroligin associated with overexpression of NLGN4 on X and Y chromosomes could amplify the impact of genetic variants that normally create only a minor risk for neurodevelopmental abnormalities. We did not, however, find any support for that hypothesis from investigation of common variants in two candidate genes.

In the current paper, we consider another way in which an extra sex chromosome could amplify genetic risk, by interacting with copy number variants (CNVs). The history of recognition of CNVs is documented by Beckmann et al (2007): these are deletions or insertions affecting chunks of DNA 1 kb in length or larger, which were first described in the 1960s and 1970s. It was a few decades later before it was recognised that this kind of large-scale submicroscopic variation is common and not necessarily pathological. Nevertheless, where CNVs are large and/or affect the function of key genes, they are likely to be associated with neurodevelopmental disorders, notably intellectual disability or autism spectrum disorder. One study also found an association with severe developmental language disorder (DLD) (Kalnak et al, 2018).

Large, disruptive CNVs tend to be relatively rare in populations. However, smaller CNVs are more common, and Simpson et al (2015) found that there was a slight increase in these in cases of DLD and their relatives, compared to a control sample. The fact that unaffected relatives showed the increase as well as affected individuals suggested that an increased burden of CNVs may play a cumulative role in mediating an increased risk of language disorder, but the precise impact may depend on the location and extent of the CNV, and whether it disrupts gene function. The specific combination of inherited events may also be important. As more information becomes available regarding the variability across the Human Genome, it has become apparent that some genes are more sensitive to change than others. Such “dosage-sensitive” genes will be more affected by copy number events. Accordingly, large databases such as the Exome Aggregation Consortium (ExAC: <http://exac.broadinstitute.org/>) have attempted to qualify the dosage sensitivity of genes by scores of sequence constraint, such as the pLI score (probability of being loss-of-function intolerant: Lek et al., 2016).

Here we consider an epistasis hypothesis that predicts that the risk of neurodevelopmental disorder associated with a CNV will be increased when there is a trisomy, because of interactions between CNVs and the overexpression of genes on the sex chromosomes. This is the first hypothesis we plan to test, and relates to the idea of a two-hit model (Veltman & Brunner, 2010), whereby the effect of a microdeletion is not deterministic, but rather acts as a risk factor that can increase the impact of deletions or duplications elsewhere on the genome. This kind of mechanism was supported by the discovery by Girirajan et al (2010) that individuals with severe neurodevelopmental disorders who had a deletion at the 16p12.1 locus often had a second CNV. A genomic alteration that may have little or no effect in an unaffected relative appeared to have a particularly detrimental effect in combination with a second ‘hit’. We extend this idea to encompass the notion that the impact of a third copy of a sex chromosome may be amplified by a CNV that might have little effect in a child with a normal complement of chromosomes.

Our second hypothesis is that there is an increased burden of CNVs in individuals with a sex chromosome trisomy across the entire genome. According to this hypothesis, the high rate of neurodevelopmental problems could be a direct consequence of an increased CNV burden - perhaps because whatever mechanism leads to a trisomy also disrupts CNV checkpoints.

The best source of evidence for an increased CNV burden comes from Rocca et al (2016), who presented evidence that men with Klinefelter syndrome (47,XXY) had an unusually high number of X-chromosome CNVs. They compared CNV burden on the X-chromosome in 94 men with Klinefelter (47, XXY) syndrome to that in 85 controls (43 males and 42 females), and reported a higher number of CNVs, especially duplications, in the Klinefelter group. Thirty-nine of them (41.5%) carried CNVs, compared to 12/42, (28.6%) of females, and 8/43, (18.6%) of males. As the authors noted, presence of additional CNVs in some individuals could provide an explanation for the variable phenotype, but they did not test for associations with phenotype in their sample. Their study raised the further question of whether an increased rate of CNVs might be seen in other sex chromosome trisomies - XXX and XYY - and whether these might be found on the autosomes as well as the X-chromosome.

Further circumstantial evidence for an impact of CNV burden in cases of sex chromosome trisomy comes from Le Gall et al (2017), who focused on a group of 14 patients with sex chromosome trisomies in whom an additional causative event was suspected because of an unusually severe phenotype involving intellectual disability or other severe developmental disorder. They found seven patients carried a pathogenic CNV (one with Williams-Beuren syndrome, one with 7q11.23 duplication, one with 17q12 duplication, three with 16p11.2 duplication and one with a 15q11.3 deletion), as well as two with a likely pathogenic CNV and five with a variant of uncertain significance. Because their report focused only on cases with a known additional micro-deletion or -duplication, the authors were not, however, able to estimate the prevalence of additional pathogenic CNVs in cases of sex chromosome trisomy, or to show that a CNV was specifically related to the severity of the phenotype.

We plan to study two groups of children: a) a sample of children with a sex chromosome trisomy (karyotypes 47,XXX, 47,XXY or 47,XYY, henceforth referred to as XXX, XXY and XYY), identified by prenatal or postnatal screening, or via support groups; b) a comparison sample of twin children selected to include a high proportion of cases with evidence of language or literacy difficulties. Both groups have undergone the same battery of language and cognitive assessments, and have had a genome-wide SNP array performed on their DNA sample.

## Research Questions *(required)*

*Please list each research question included in this study.*

1. Is the total size, number or predicted intolerance score of autosomal CNVs associated with greater phenotypic severity in individuals with sex chromosome trisomies compared to a comparison sample of normal karyotype? This would provide evidence of an interaction between copy number events in autosomal genes and genes on the sex chromosomes
2. Are individuals with a sex chromosome trisomy more likely than those of normal karyotype to carry other secondary large (>1 Mb) CNVs on the autosomes, which may have gone undiagnosed?

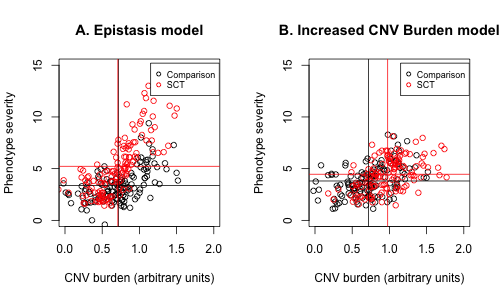
## Hypotheses *(required)*

*For each of the research questions listed in the previous section, provide one or multiple specific and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here.*

We will test two contrasting hypotheses, both of which link language and neurodevelopmental problems in SCTs to CNVs:

1. Epistasis model: SCT cases will carry a similar number, size and predicted intolerance scores of secondary CNVs as the comparison group, but these will correlate with phenotypic severity more than in the comparison group. The overall burden of secondary CNVs will predict language impairment in both comparison and SCT samples, but the regression line will have a steeper slope in the SCT group (Figure 1 A).
2. Increased burden model: SCT cases are more likely to carry a secondary pathogenic or likely pathogenic CNV that may contribute to their phenotype, compared to the comparison group (Figure 1 B).

As shown in Figure 1, in model A, the comparison and SCT groups do not differ in CNV burden (x-axis), but they differ in the slope of the regression linking CNV burden to phenotype. In model B, the SCT group has a greater burden of CNVs that is sufficient to account for the more severe phenotype in this group, without needing to postulate any difference between groups in the function linking CNV burden to phenotype. Both hypotheses are directional, and they are not mutually exclusive.

Figure 1: Schematic of predicted relationships between genotype and phenotype 

## Explanation of existing data *(required)*

*If you indicate that you will be using some data that already exist in this study, please describe the steps you have taken to assure that you are unaware of any patterns or summary statistics in the data. This may include an explanation of how access to the data has been limited, who has observed the data, or how you have avoided observing any analysis of the specific data you will use in your study. The purpose of this question is to assure that the line between confirmatory and exploratory analysis is clear.*

Both the phenotype and genome-wide SNP array data were generated for the companion study by Newbury et al (2018).

Data on language phenotypes is kept separately from data from CNV analysis and will not be linked until the protocol has been pre-registered. It is not possible to test either of the hypotheses without this information, as both involve predictions about relationships between genotype and phenotype.

Access to phenotypic data is restricted to the OSSCI team (DVMB and PAT) and has not been made available to the genetics team. For our previous study by Newbury et al (2018), we linked the phenotypic data to results from 47 specific single-nucleotide polymorphisms from two candidate genes. This level of genetic data does not provide information about the CNVs that are the focus of the current study.

The CNV analysis starts from raw data in the form of genome-wide SNP arrays generated by the Newbury et al (2018) study, which is in the form of a Genome Studio (Illumina) project file (.bsc). This contains the genotype calls, signal ratios and minor allele frequency data for 500 participants across 600,000 genome wide SNP markers. High quality calling of CNVs is being performed by HSM using software packages outside of Genome Studio which interpret signal intensity and ratios to identify copy number changes. Results of her analysis will not be made available to DVMB and PAT until after this protocol is preregistered.

## Data collection procedures *(required)*

*Please describe the process by which you will collect your data. If you are using human subjects, this should include the population from which you obtain subjects, recruitment efforts, payment for participation, how subjects will be selected for eligibility from the initial pool (e.g. inclusion and exclusion rules), and your study timeline. For studies that don’t include human subjects, include information about how you will collect samples, duration of data gathering efforts, source or location of samples, or batch numbers you will use.*

### SCT Group Recruitment

SCT cases aged from 5 to 16 years were recruited through re-contacting participants from a previous study (Bishop et al, 2011) who had agreed to be re-contacted. Additional participants were recruited via support groups (Unique: the Rare Chromosome Support Group, and the Klinefelter Syndrome Association), National Health Service Clinical Genetics Centres, and self-referred through the research project Facebook page or website. In order to be eligible for the study, SCT group participants had to have a genetic diagnosis of either XXX, XXY or XYY, and be fully aware of their genetic status.

Figure 2 shows the numbers of children with sex chromosome trisomies in relation to the type of trisomy and the reason for diagnosis. We distinguish here between those diagnosed in the course of investigations for neurodevelopmental disorder and the remainder; only the latter group can be used to estimate whether the CNV burden is unduly high in those with a sex chromosome trisomy, because their diagnosis was independent of neurodevelopmental status. However, both groups combined can be used to test predictions about CNV state/phenotype associations.

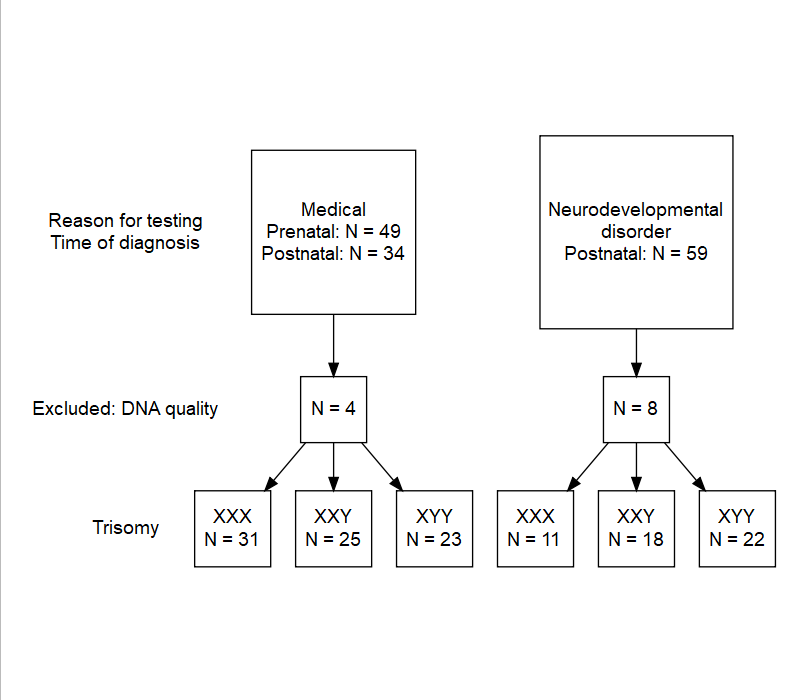


Figure 2. *N children with each type of trisomy, divided according to whether diagnosis was made when investigating neurodevelopmental disorder. From Newbury et al (2018)*

### Twin Comparison Group Recruitment

Twins in the comparison group had been recruited for a study of language and laterality (Wilson and Bishop, 2018) and had undergone the same test battery as the SCT group. Twins were recruited via information sheets sent to state schools, and also via the research project Facebook page. In this sample, we aimed to recruit a sample where around 75% pairs would include at least twin with developmental language disorder (DLD). This was achieved by selecting cases for inclusion on the basis of parental response on a telephone interview: any mention of language delay, history of speech and language therapy, current language problems or dyslexia was coded as ‘parental concern’.

A flowchart showing the numbers of twins with specific characteristics is shown in Figure 3. Because twins are not independent, the final sample was divided into two subgroups of 184 and 186 children respectively, each containing one member from each pair, selected at random. (Ns are not equal because some twins had missing DNA from just one member of the pair). Only the first twin sample is used in our main analysis, with the second sample being reserved for replication.

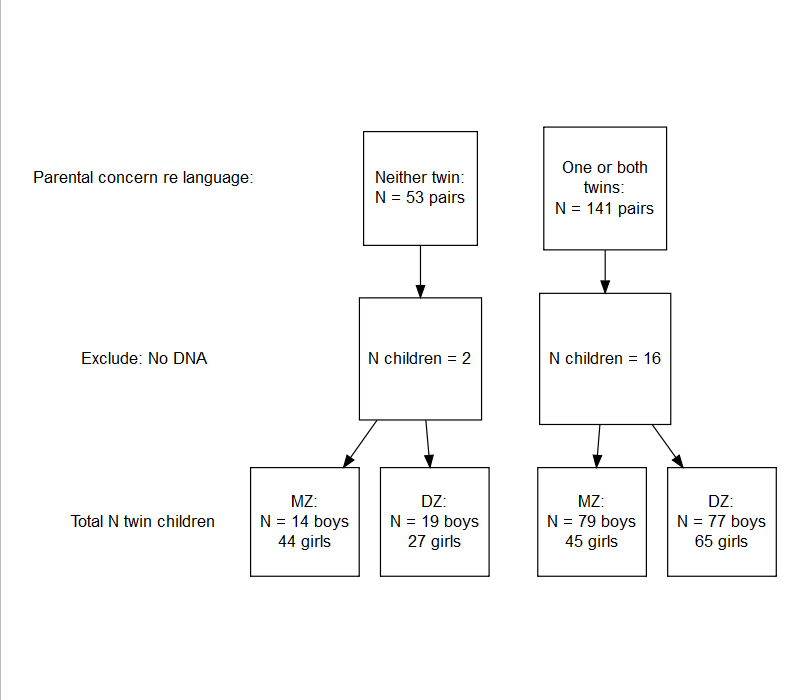


Figure 3. *Flowchart showing characteristics of children recruited to comparison groups. Information about zygosity, gender and parental concern is shown for information, but was not used in the analysis (From Newbury et al, 2018).*

### Test battery

The test battery is described by Newbury et al (2018) and specific details are given below. The battery was designed to provide a quantitative estimate of language, literacy and communication ability in children aged 5 to 16 years. In addition, parents completed a telephone interview, and were invited to complete two questionnaires and an online diagnostic interview.

### Genetic Data

DNA was collected and extracted from saliva samples (OG-500, DNA Genotek) using the manufacturer recommended protocol. DNA, but not phenotypic data, was collected from the parents of the participants where possible, and they were included on the genotyping array to assist with the confirmation of inheritance patterns.

Samples were genotyped using the genome-wide SNP array Infinium Global Screening Array-24(v1) targeting 692,824 common and rare SNPs suitable for both genotyping and CNV analysis. SNP array data were called using Illumina GenomeStudio v.2.03. Individual SNPs with a GenTrain (quality) score of < 0.5 and samples with < 0.9 of SNPs called were excluded.

## Sample size (required)

As shown in Figure 2, the SCT group consisted of 42 girls with XXX, 43 boys with XXY, and 45 boys with XYY karyotype combined into a group of 130 children with a sex chromosome trisomy. The comparison group were selected from 194 twin pairs (Figure 3), with DNA samples available for 370/388 (95%) of cases.

Genotyping data also includes 215 samples from first degree relatives of participants. This makes it possible to establish which parent the CNV was inherited from, or if it was de novo; this information is particularly useful if a potentially pathogenic CNV is discovered (see below).

All available data from genotyping of children and parents will undergo CNV analysis, as this improves the specificity of calling.

## Sample size rationale *(required)*

Our study is constrained by the numbers we were able to recruit, with further loss of cases which did not pass genotyping quality control. We had originally planned the study to detect an association between specific SNPs and language outcomes in children with SCTs. The power to detect given effect sizes in the current study is computed in the final script below: with this sample we are well-powered to detect a correlation around .35 between CNV burden and phenotype in the SCT group, assuming a correlation of around .11 in the comparison sample.

## Stopping rule *(required)*

Not applicable as data has already been obtained.

## Manipulated variables (required)

Not applicable

## Measured variables (required)

*Describe each variable that you will measure. This will include outcome measures, as well as any predictors or covariates that you will measure. You do not need to include any variables that you plan on collecting if they are not going to be included in the confirmatory analyses of this study.*

### Phenotypic measures

*Language and cognitive tests*

Verbal comprehension was assessed using the Woodcock Johnson III Tests of Cognitive Abilities (Woodcock et al 2007). Nonword repetition, oromotor sequencing and sentence repetition were tested using the NEPSY: A Developmental Neuropsychological Assessment tool (Korkman et al 1998). Phonological ability was assessed using the Phonological Assessment Battery (PhAB) (Frederickson, Frith and Reason, 1997), and speeded reading of single words and nonwords using the Test of Word Reading Efficiency (TOWRE) (Torgesen, Wagner and Rashotte, 1999). Accuracy, rate and comprehension of reading text passages were tested using the Neale Analysis of Reading Ability (2nd British Edition) (Neale, 1997). Communicative competence was assessed from parental report the Children’s Communication Checklist-2 (CCC-2) (Bishop, 2003). Nonverbal ability was estimated using the Block Design and Matrices subtests of the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999); age-normed scores on these were converted into an overall Performance IQ score.

*Psychiatric evaluation*

In an initial telephone interview, parents were asked about the child’s medical and educational history, including a question about whether anyone had diagnosed the child with a neurodevelopmental disorder such as ASD, developmental language disorder (DLD) or specific language impairment, dyslexia or dyspraxia. In addition, one or both parents were asked to complete the online Development and Wellbeing Assessment (DAWBA) (Goodman et al., 2000) in their own time. 84 parents of SCT cases and 133 parents of twins complied with this request. The DAWBA gives information on likelihood of the child meeting criteria for a range of psychiatric diagnoses; a final diagnosis is made by a trained rater who assimilates all the information and evaluates it against DSM5 criteria (American Psychiatric Association, 2013).

## Genetic measures

The CNV analysis will test for the number of CNVs, total base pairs covered by CNVs (Kbp) and a sum of the predicted likelihood of intolerance scores for each of the SCT and comparison group individuals. Individual large and pathogenic CNVs will be identified separately as these may be of clinical significance.

Genotypes have been previously called (Newbury et al 2018) using Genome Studio (Illumina), and beta allele frequencies and log 2 ratio for each of the 600,000 SNPs exported to a .csv file. CNV calling will be performed using two separate methods; PennCNV (Wang et al 2007), a HMM based method written in Perl, and QuantiSNP (Colella et al 2007) written in Matlab. CNVs will be considered high confidence when they include more than five consecutive SNPs, are called by both PennCNV and QuantiSNP at an acceptable confidence level, and overlap by at least 50% between the two methods. CNV regions will be annotated for gene names using Annovar (Wang, Li, & Hakonarson, 2010). Individual totals of CNV number, size and predicted likelihood of intolerance (pLi) score will be calculated using a bespoke R script (see below).

## Indices *(required)*

*If any measurements are going to be combined into an index (or even a mean), what measures will you use and how will they be combined? Include either a formula or a precise description of your method. If you are using a more complicated statistical method to combine measures (e.g. a factor analysis), you can note that here but describe the exact method in the analysis plan section.*

Our focus will be on one of the phenotypic measures used by Newbury et al (2018), the global index of neurodevelopmental difficulties, as CNVs have previously been associated most strongly with severe problems affecting behaviour, which are not indexed by the other two measures (nonword repetition and overall language ability). The global index is an ad hoc measure that combines all available information about a range of neurodevelopmental disorders affecting language, attention, social communication and overall functioning. The script for deriving this measure was contained in an Appendix to Newbury et al (2018), and is available on Open Science Framework (<https://github.com/oscci/Genetic_analysis_SCT_twin/blob/master/App3_global%20score%20create.Rmd>).

Three measures of CNV burden will be used: A) Mean of number of CNVs per group B) Mean of total CNV size per group C) Mean of sum of pLi scores from each CNV in each individual, and compared per group. Parent CNV events will be called but not included in the burden analyses.

## Study design (required)

*Describe your study design. Examples include two-group, factorial, randomized block, and repeated measures. Is it a between (unpaired), within-subject (paired), or mixed design? Describe any counterbalancing required. Typical study designs for observation studies include cohort, cross sectional, and case-control studies.*

This study adopts an observational design comparing cases (children with SCTs) and a comparison group (twin children with normal karyotype).

## Randomization (required)

*If you are doing a randomized study, how will you randomize, and at what level?*

The only role of randomisation in this study is the division of the twin comparison group into two subgroups. This has been done by selecting one twin from each pair at random. In effect, the second subsamples can be treated as a replication sample, though it should be noted they it is not independent, because of genetic relatedness.

### Statistical models (required)

*What statistical model will you use to test each hypothesis? Please include the type of model (e.g. ANOVA, multiple regression, SEM, etc) and the specification of the model (this includes each variable that will be included as predictors, outcomes, or covariates). Please specify any interactions that will be tested and remember that any test not included here must be noted as an exploratory test in your final article.*

The ‘increased CNV burden’ model can be simply tested by comparing the SCT group with the comparison group in terms of the CNV burden, as indexed by the three measures. A nonparametric test (Wilcoxon sum rank test) will be used given the highly skewed distribution of CNV scores. Because the three CNV measures are intecorrelated, Bonferroni correction for the number of comparisons would be overly conservative, and so we will use the Benjamini-Yekutieli method to evaluate the probability of an increase on any of the measures in the SCT group (Benjamini & Yekutieli, 2001).

The ‘double hit’ model will be tested using a regression approach, as the prediction is for an interaction, i.e. a stronger association between CNV burden and phenotype in the SCT group relative to the comparison group. Because the dependent measure is essentially a count, Poisson regression will be used (see Scripts section for more details of analysis).

## Transformations (required)

CNV burden scores will be censored to avoid a few extreme values (see draft script, below)

## Follow-up analyses (required)

*If not specified previously, will you be conducting any confirmatory analyses to follow up on effects in your statistical model, such as subgroup analyses, pairwise or complex contrasts, or follow-up tests from interactions? Remember that any analyses not specified in this research plan must be noted as exploratory.*

Any association of CNV burden and phenotype in the comparison group is expected to be weak at best, in line with previous studies of CNV burden in children with language disorders. If there is an evidence of association in this group, we will test whether this replicates in the co-twin sample.

In line with previous literature, it is possible that we will identify a few large, potentially damaging CNVs that may be contributing to an individual’s neurodevelopmental disorder. CNVs, particularly de novo CNVs, have been shown to play a major role in autism spectrum disorder (ASD) (Sebat et al., 2009, Poultney et al., 2013). De novo variants have also been implicated in language disorders (Simpson et al,. 2015, Chen et al., 2017, Kalnak et al., 2018) and therefore may provide a potential mechanism for the phenotypic variation seen the SCT cases. Large CNVs will be checked to see if they have been previously reported in the online databases (ClinVar, Decipher), and annotated in line with current American College of Medical Genetics (ACMG) guidelines (Kearney et al., 2011; South et al., 2013; Nowakowska, 2017). Parental genetic data, where available, will be used to verify the inheritance mechanism and consequent potential for pathogenicity.

## Inference criteria (required)

*What criteria will you use to make inferences? Please describe the information you’ll use (e.g. specify the p-values, Bayes factors, specific model fit indices), as well as cut-off criterion, where appropriate. Will you be using one or two tailed tests for each of your analyses? If you are comparing multiple conditions or testing multiple hypotheses, will you account for this?*

Because we have directional predictions, we will use one-tailed statistical tests with alpha set to .05. As noted above, we will use the Benjamini-Yekutieli method to control the false discovery rate by adjusting for multiple comparisons.

## Data exclusion (required)

*How will you determine which data points or samples (if any) to exclude from your analyses? How will outliers be handled?*

Poor quality genetic data will be excluded from the analysis at a cut off of <0.5 GenTrain score for SNPs and <90% SNP data per individual. As high quality data is key to achieving accurate CNV calls, any low quality or poorly performing samples will be removed from the analysis. This is identified by quality scores, but also excessive CNV calls or excessive heterozygosity levels.

Only high confidence CNVs which overlap by at least 50% between the two detection algorithms will be considered in the analysis; those that do not meet this bench mark will be excluded.

## Missing data (required)

*How will you deal with incomplete or missing data?*

For phenotypic analysis, the global neurodevelopmental index uses all available data: thus if the DAWBA (online interview) was not completed, reliance is placed on report from parental interview, but where DAWBA is available, that information is incorporated in the index. With this approach, phenotypic scores are available for all cases.

Regarding the genetic analysis, no data is missing from the study, though individuals may have data missing from some SNPs on the array. This is not a problem for downstream analyses, and these are coded as missing by Genome Studio. Individual with more than 10% SNP calls missing have been excluded, and individual SNPs were excluded if they were missing more than 5% of calls.

## Exploratory analysis (optional)

*If you plan to explore your data set to look for unexpected differences or relationships, you may describe those tests here. An exploratory test is any test where a prediction is not made up front, or there are multiple possible tests that you are going to use. A statistically significant finding in an exploratory test is a great way to form a new confirmatory hypothesis, which could be registered at a later time.*

Genes covered by high confidence CNVs will be analysed using the gene ontology analysis package WebGestalt (Wang et al., 2017) to identify common molecular pathways. This pathway analysis may identify potential mechanisms that underlie the high levels of language disorder seen in children with an sex chromosome trisomy.

## Ethical considerations

In line with our ethics committee approval, we will describe results of this analysis in a newsletter to participant families, and encourage those who wish to do so to ask for information specific to their child. For cases who have a pathogenic or likely pathogenic CNV, and whose parents request feedback, we will liaise with their local clinical geneticist, who will be able to discuss the implications of the finding with the family.

## Scripts

Scripts are contained within a R Markdown document that was used to generate this protocol, and can be found on <https://github.com/oscci/CNV_hayley>.

### 1. Predictions from models A and B

Our first script was used to generate Figure 1. It uses simulated data to illustrate models A and B, contrasting the predictions from the Epistasis model and Increased CNV burden model for sex chromosome trisomies. Distributions of CNVs are simulated for both models, with a probabilistic rule linking CNV size to presence of phenotype. Ten binary phenotypes are summed with added error to give continuous phenotype score.

#NB CNV size currently modeled as normally distributed; would be better if modelled to   
#reflect fact that the freq goes down with size, as discussed here:  
# Conrad et al: 2006 Nature genetics 38(1):75-81.: A high-resolution survey of deletion polymorphism in the human genome.  
  
#See also this review: by Zhang et al: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4472309/  
#Database where structural variation is cataloged (the Database of Genomic Variants or DGV,  
#http://projects.tcag.ca/variation/).  
  
#Define some variables to be simulated  
myvars<-c('sct','sizecnvA','sizecnvB','phenoA1','phenoA2','phenoA3','phenoA4','phenoA5',  
 'phenoA6','phenoA7','phenoA8','phenoA9','phenoA10',  
 'phenoAsum','phenoB1','phenoB2','phenoB3','phenoB4','phenoB5',  
 'phenoB6','phenoB7','phenoB8','phenoB9','phenoB10','phenoBsum')  
  
#set up dataframe to save simulation results  
ncases<-130 #number of cases each for sct and control (roughly correct, though more controls in practice)  
mydata<-data.frame(matrix(NA,nrow=ncases\*2,ncol=length(myvars)))  
colnames(mydata)<-myvars  
range1<-1:ncases  
range2<-(ncases+1):(ncases\*2)  
mydata[range1,1]<-0  
mydata[range2,1]<-1 #half control and half with trisomy  
mymean<-0; mysd<-1; #mean and SD for size of cnv - arbitrary  
mydata[,2]<-rnorm(ncases\*2) #random number ZSCORE represents size cnv  
#model B - higher rate of cnv in trisomy cf control - add .8 sd to cnv size  
mydata[range1,3]<-mydata[range1,2]  
mydata[range2,3]<-mydata[range2,2]+mysd\*.8  
  
#model A: same rate of CNV but bigger effect with SCT - Epistasis  
#relation between sizecnv and neurodev is probabilistic; use lower cutoff for disorder for sct  
  
 pvect<-pnorm(mydata$sizecnvA) #sizecnvA at this point is a zscore, so can just convert to pvalue  
 for (i in 1:10){ #phenotypes model A (have 10, each independently determined)  
 rvect<-runif(ncases) #random number from 0 to 1  
 dvect<-pvect-rvect #if pvect value is high, then v likely to be +ve and vice versa  
 mydata[,(i+3)]<-0 #whether or not phenotype present for this variable  
 mycutoff<-.7 #the higher this value, the lower the correl between cnv/pheno  
 w<-which(dvect>mycutoff) #if dvect is above cutoff, phenotype present  
 mydata[w,(i+3)]<-1  
 sctcutoff<-.3 #can try changing this and mycutoff to see how relationship affected  
 w<-which(dvect[range2]> sctcutoff) #different cutoff for sct cases  
 mydata[(w+ncases),(i+3)]<-1 #nb add ncases, because reference is just to part of vector in formula  
 }  
 mydata$phenoAsum<-rowSums(mydata[,4:13])+rnorm(ncases)+3 #rnorm + 3 adds random error  
  
#add 3 to the cnv variables, so they could plausibly be Mb (ie megabase units, million base pairs)  
mydata[,2:3] <-(3+mydata[,2:3])/3-.3  
  
  
png(filename = "CNV\_models.png", width = 500, height = 300)#uncomment to create png file  
par(mfrow=c(1,2))  
   
plot(mydata$sizecnvA,mydata$phenoAsum,col=(mydata$sct+1),  
 xlab='CNV burden (arbitrary units)',ylab='Phenotype severity',main='A. Epistasis model',  
 ylim=c(0,15),xlim=c(0,2))  
legend(1.15, 15.2, legend=c("Comparison", "SCT"),  
 col=c("black", "red"), pch=1,cex=0.75)  
  
conphenomean<-mean(mydata$phenoAsum[range1])  
sctphenomean<-mean(mydata$phenoAsum[range2])  
abline(h=conphenomean, col='black') #add lines to plot to show means for SCT and control on x and y axis  
abline(h=sctphenomean, col='red')  
concnvmeana<-mean(mydata$sizecnvA[range1])  
sctcnvmeana<-mean(mydata$sizecnvA[range2])  
abline(v=concnvmeana, col='black')  
abline(v=sctcnvmeana, col='red')  
  
  
#model B: higher rate of CNV but same effect with SCT - added burden model  
#otherwise identical : relation between sizecnv and neurodev is probabilistic; use lower cutoff for disorder for sct  
  
pvect<-pnorm(mydata$sizecnvB)  
for (i in 1:10){ #phenotypes model B  
 rvect<-runif(ncases)  
 dvect<-pvect-rvect  
 mydata[,(i+14)]<-0  
 w<-which(dvect>mycutoff)  
 mydata[w,(i+14)]<-1  
}  
mydata$phenoBsum<-rowSums(mydata[,15:24])+rnorm(ncases)+3  
  
  
plot(mydata$sizecnvB,mydata$phenoBsum,col=(mydata$sct+1),pch=1,  
 xlab='CNV burden (arbitrary units)',ylab='Phenotype severity',  
 main='B. Increased CNV Burden model', ylim=c(0,15),xlim=c(0,2))  
legend(1.15, 15.2, legend=c("Comparison", "SCT"),  
 col=c("black", "red"), pch=1,cex=0.75)  
conphenomean<-mean(mydata$phenoBsum[range1])  
sctphenomean<-mean(mydata$phenoBsum[range2])  
abline(h=conphenomean, col='black')  
abline(h=sctphenomean, col='red')  
concnvmeanb<-mean(mydata$sizecnvB[range1])  
sctcnvmeanb<-mean(mydata$sizecnvB[range2])  
abline(v=concnvmeanb, col='black')  
abline(v=sctcnvmeanb, col='red')  
dev.off() #uncomment this if printing to png file

## quartz\_off\_screen   
## 2

### 2. CNV counting script

This script by Hayley Mountford processes the CNV data. N. B. At time of pre-registration, only the genetics team (Newbury/Mountford) have access to the raw genetic data file, and only the psychology/stats team (Bishop/Thompson) have access to the phenotype data files. The CNV data file has not yet been generated.

# R script for counting total CNV size, number of genes covered, counts of dosage sensitive and insensitive  
# and total pLI score per individual  
# Written by Hayley Mountford: 8th June 2018  
runhm <- 0  
if(runhm==1){  
# needs to be repeated for both cases and controls.   
  
# import package dependencies  
#install.packages('tidyr', dependencies=T)  
#install.packages('sqldf', dependencies=T)  
#install.packages('dplyr', dependencies=T)  
  
library(tidyr)  
library(sqldf)  
library(dplyr)  
  
# set dose senstivity threshold for pLI score (>0.9 is considered dosage intolerent)  
DOSE\_THRESH <- 0.9  
  
hayleydir<-'/Users/hayle/Dropbox/\_SLI/Trisomy/SLIC data CNV/'  
#Will need to set directory for reading data  
  
# load annoted CNV regions from Annovar output into a dataframe  
df <- read.table(paste0(hayleydir,"cases.hg19\_multianno.txt"), header=T, sep='\t', strip.white=T,  
 blank.lines.skip=T, stringsAsFactors=F)  
  
# original file from -   
# ftp://ftp.broadinstitute.org/pub/ExAC\_release/release0.3.1/functional\_gene\_constraint/fordist\_cleaned\_nonpsych\_z\_pli\_rec\_null\_data.txt  
# working with a subset of this file which contains the Gene name (Gene.refGene), and pLI to make it run quicker  
pLi <- read.table(paste0(hayleydir,"pLi\_full.txt"), col.names=c('Gene.refGene', 'pLi'), header=T,  
 sep='\t', strip.white=T, blank.lines.skip=T, stringsAsFactors=F)  
  
# full official HGNC gene names (March 2018 version) downloaded from https://www.genenames.org/cgi-bin/statistics  
# read in hgnc data file hgnc\_complete\_set for name matching  
hgnc <- read.table(paste0(hayleydir,"hgnc\_complete\_set.txt"), header=T, sep='\t', strip.white=T,  
 blank.lines.skip=T, stringsAsFactors=F, quote='"', comment.char="")  
  
  
#-----------------------------------------------------------------------------------#  
  
# fixing up Annovar output issues  
# some fields have multiple genes but no distances, lets split the DF into two  
df.hasdist <- df[df$GeneDetail.refGene != '',]  
df.nodist <- df[df$GeneDetail.refGene == '',]  
  
# df has badly formatted data! Annovar has put the gene function details in the wrong column  
# ExonicFunc data is in the GeneDetail column so extract all these rows  
df.exonfunc <- df.hasdist[!startsWith(df.hasdist$GeneDetail.refGene, 'dist='),]  
  
# remove the bad ExonicFunc rows from df.hasdist  
df.hasdist <- df.hasdist[startsWith(df.hasdist$GeneDetail.refGene, 'dist='),]  
  
# move the ExonicFunc data into the correct column  
df.exonfunc$ExonicFunc.refGene <- df.exonfunc$GeneDetail.refGene  
df.exonfunc$GeneDetail.refGene <- ''  
  
# concatenate df.exonfunc and df.nodist  
df.nodist <- rbind(df.nodist, df.exonfunc)  
  
# split the delimited fields  
df.split <- rbind(separate\_rows(df.hasdist, c('Gene.refGene', 'GeneDetail.refGene'), sep=';'),  
 separate\_rows(df.nodist, 'Gene.refGene', sep=';'))  
  
# this dealt with a couple of rogue rows   
tmp <- df.hasdist[1250:1270,]  
separate\_rows(tmp, c('Gene.refGene', 'GeneDetail.refGene'), sep=';')  
  
  
#-----------------------------------------------------------------------------------#  
  
  
# adding the pLI score for each of the genes in the correctly formatted dataframe  
# join the pLi data onto df.split  
df.pli <- merge(df.split, pLi, by='Gene.refGene', all.x=T)  
  
# split the df into two: those with a pLi score and those without  
df.nopli <- df.pli[is.na(df.pli$pLi),]  
df.pli <- df.pli[!is.na(df.pli$pLi),]  
  
# drop pLi column, as it's all NA and we're going to join again via an alias  
df.nopli <- subset(df.nopli, select=-c(pLi))  
  
#-----------------------------------------------------------------------------------#  
  
# now join using the gene aliases from hgnc...  
# get all the gene symbols and aliases  
symbol.alias <- hgnc[c('symbol', 'alias\_symbol')]  
symbol.alias <- symbol.alias[symbol.alias$alias != '',]  
  
# split delimitered aliases  
hgnc.split <- separate\_rows(symbol.alias, 'alias\_symbol', sep='[|]')  
  
# add the symbol to symbol pairs  
hgnc.symbol <- hgnc[c('symbol')]  
hgnc.symbol$alias\_symbol <- hgnc.symbol$symbol  
  
# concatenate the aliases and symbol pairs  
hgnc.all <- rbind(hgnc.split, hgnc.symbol)  
  
# do a self join to give us all possible alias pairs  
hgnc.pairs <- merge(hgnc.all, hgnc.all, by='symbol', suffixes = c('','2'))  
  
# join pLi data onto the aliases  
df.pli.alias <- merge(hgnc.pairs, pLi, by.x='alias\_symbol', by.y='Gene.refGene')  
  
# tidy up the df  
df.pli.alias <- df.pli.alias[,c('alias\_symbol', 'alias\_symbol2', 'pLi')]  
  
# join pLi data using the alias name  
df.nopli <- merge(df.nopli, df.pli.alias, by.x='Gene.refGene', by.y='alias\_symbol2', all.x=T)  
  
# merge the two datasets back together  
df.pli <- bind\_rows(df.pli, df.nopli)  
  
#-----------------------------------------------------------------------------------#  
  
# This is what we want in the 5 columns for the text output  
  
# Col1: individual sample identifier = Otherinfo  
# Col2: Count of overlapping dose sensitive genes  
# Col3: Count of overlapping non-dose sensitive genes  
# Col4: Count of overlapping genes not on the list  
# Col5: Count of CNVs containing overlapping genes (includes non-coding RNAs, as these have potential function)  
# Col6: Count of total CNVs per individual  
# Col7: total Kbp of cnvs (end - start/1000) per individual  
  
  
# using sql to filter for each criteria for col1 to col5  
df.query1 <- sqldf(  
 sprintf(  
 "SELECT `Otherinfo` AS `id`,  
 IFNULL(SUM(pLi >= %s), 0) AS `dose\_sense`,  
 IFNULL(SUM(pLi < %s), 0) AS `dose\_insense`,  
 SUM(pLi IS NULL) AS `no\_pLi`,  
 COUNT(DISTINCT `Chr` || '-' || `Start` || '-' || `End`) AS `overlap\_cnvs`,  
 SUM(pLi) AS `sum\_pLi`  
 FROM `df.pli`  
 WHERE `Func.refGene` IN ('exonic', 'intronic', 'UTR5', 'UTR3',  
 'ncRNA\_exonic', 'ncRNA\_intronic')  
 GROUP BY `Otherinfo`",  
 DOSE\_THRESH,  
 DOSE\_THRESH  
 )  
)  
  
# Col6: Count of total CNVs per individual  
# Col7: total Mbp of cnvs (end - start/1000) per individual  
  
df.query2 <- sqldf(  
 "SELECT `Otherinfo` AS `id`,  
 COUNT(\*) AS `total\_cnvs`,  
 SUM((`End` - `Start`)/1000) AS `total\_Kbp`  
 FROM `df`  
 GROUP BY `Otherinfo`"  
)  
  
  
# join the two results together (old version which drops all intronic because of the merge function)  
df.out <- merge(df.query1, df.query2, by='id', all = TRUE)  
df.out[is.na(df.out)] <- 0  
  
  
# Write data to a tsv file which is easier to import into Excel, linux, R, etc  
write.table(df.out,paste0(hayleydir,"cases\_out.tsv"), sep="\t", row.names = F)  
}

### 3. Using a previous dataset to work out how to quantify CNV burden

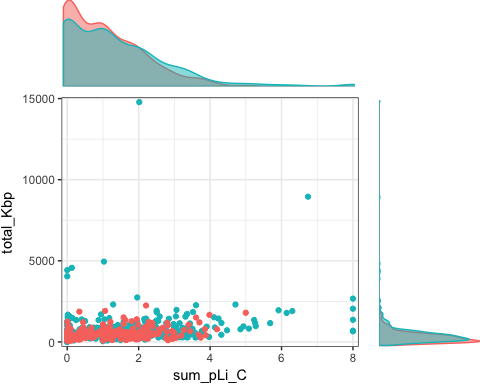
We will have data on two groups, as specified in previous paper on the Epistasis hypothesis by Newbury et al (2018). One group are children with sex chromosome trisomies (SCTs), and the comparison group is twins, oversampled for children with language problems. Only one randomly-selected member of each twin pair features in the main analysis.

To anticipate properties of the CNV data, we use an existing CNV dataset from a sample from the SLI Consortium, previously reported by Simpson et al (2015). This is SLIC\_df.txt. We do not have phenotypic data, so aim is just to establish the distributions of scores on different measures of CNV burden.

First we upload the data and apply censoring to recode extreme values.

##############################################################################  
#Using test\_script\_cnv.R developed by Paul Thompson  
#Using version of file with IDs anonymised = see renameSLICs.R  
CNV\_dat<-read.table('SLIC\_df\_anon.txt',header=TRUE,stringsAsFactors=FALSE)  
#Function borrowed and adapted from the 'scales' package. This applies the censoring to the two potentially interesting variables.  
#str(CNV\_dat) #uncomment to inspect variables  
censor\_PT<-function (x, limit = 7, only.finite = TRUE)   
{  
 force(range)  
 finite <- if (only.finite)   
 is.finite(x)  
 else TRUE  
 x[finite & x > limit] <- limit+1  
 x  
}  
  
CNV\_dat$sum\_pLi\_C<-censor\_PT(CNV\_dat$sum\_pLi,limit=7,only.finite = F)  
#CNV\_dat$total\_Kbp\_C<-censor\_PT(CNV\_dat$total\_Kbp,limit=3000,only.finite = F)  
  
#plot(CNV\_dat$sum\_pLi,CNV\_dat$sum\_pLi\_C) #show effect of censoring  
  
#hist(CNV\_dat$sum\_pLi\_C) #NB still v skewed  
  
#-----------------------------------------------------------------------------#  
p1 <- ggplot(CNV\_dat,aes(x=sum\_pLi,y=sum\_pLi\_C)) + geom\_point() +  
 scale\_x\_continuous(expand=c(0.02,0)) +  
 scale\_y\_continuous(expand=c(0.02,0)) +  
 theme\_bw() +  
 theme(legend.position="none",plot.margin=unit(c(1,1,1,1),"points"))  
  
theme0 <- function(...) theme( legend.position = "none",  
 panel.background = element\_blank(),  
 panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),  
 panel.margin = unit(0,"null"),  
 axis.ticks = element\_blank(),  
 axis.text.x = element\_blank(),  
 axis.text.y = element\_blank(),  
 axis.title.x = element\_blank(),  
 axis.title.y = element\_blank(),  
 axis.ticks.length = unit(0,"null"),  
 axis.ticks.margin = unit(0,"null"),  
 panel.border=element\_rect(color=NA),...)  
  
p2 <- ggplot(CNV\_dat,aes(x=sum\_pLi)) +   
 geom\_density(alpha=0.5) +   
 scale\_x\_continuous(breaks=NULL,expand=c(0.02,0)) +  
 scale\_y\_continuous(breaks=NULL,expand=c(0.00,0)) +  
 theme\_bw() +  
 theme0(plot.margin = unit(c(0,-1,0.5,1.5),"lines"))   
  
p3 <- ggplot(CNV\_dat,aes(x=sum\_pLi\_C)) +   
 geom\_density(alpha=0.5) +   
 coord\_flip() +   
 scale\_x\_continuous(labels = NULL,breaks=NULL,expand=c(0.02,0)) +  
 scale\_y\_continuous(labels = NULL,breaks=NULL,expand=c(0.00,0)) +  
 theme\_bw() +  
 theme0(plot.margin = unit(c(0,0,1.75,1),"lines"))  
  
#mylegend<-g\_legend(p1) #This throws error, even after installing lemon  
  
#grid.arrange(arrangeGrob(p2,ncol=2,widths=c(3,1)),  
 # arrangeGrob(p1,p3,ncol=2,widths=c(3,1)),  
 # heights=c(1,3))  
  
  
CNV\_dat$Group<-as.factor(CNV\_dat$Group)  
#-----------------------------------------------------------------------------#

Scatterplots with marginal histograms show how the different measures of CNVs are related.

 This dataset has a group variable indicating source of sample: SLIC (includes affected and unaffected relatives) and controls. We see how the different CNV burden indicators compare. The first table shows that the different measures of CNV burden are intercorrelated.

mycor <- rcorr(as.matrix(cbind(CNV\_dat$total\_cnvs,CNV\_dat$total\_Kbp,CNV\_dat$sum\_pLi\_C)),type='pearson')  
rtab<-round(mycor$r,3)  
colnames(rtab) <- c('total cnv','total Kbp','sum pLi')  
rownames(rtab) <- c('total cnv','total Kbp','sum pLi')  
  
knitr::kable(rtab,Caption='Correlations between CNV measures')

|  |  |  |  |
| --- | --- | --- | --- |
|  | total cnv | total Kbp | sum pLi |
| total cnv | 1.000 | 0.464 | 0.476 |
| total Kbp | 0.464 | 1.000 | 0.259 |
| sum pLi | 0.476 | 0.259 | 1.000 |

Next we consider how well the different measures differentiate between the groups. Because the data is non-normal, a Wilcoxon rank sum test is used. In a preliminary analysis, we analyse each dependent measure separately.

### Corrections for multiple testing

Given correlations between the CNV measures, the Benjamini-Yekutieli (BY) correction is applied. We see that the pattern of significance among the tests remains unchanged.

|  |  |  |
| --- | --- | --- |
|  | Uncorrected | Benjamini-Y |
| Total CNVs | 0.10 | 0.183 |
| Total Kbp | 0.00 | 0.000 |
| sum pLi | 0.03 | 0.082 |

## 4. Testing Epistasis hypothesis: Phenotype analysis

We have three possible phenotypes that were used by Newbury et al (2018). We decided *a priori* that the measure of global neurodevelopmental impairment is optimal, because previous research on CNVs implicates more severe conditions such as ASD and intellectual disability, which are captured by this measure. We do not have this measure for the SLIC cohort, so we simulated some to test analyses.

## Poisson regression

Poisson regression is the optimal method, as the dependent variable is a count. This will naturally allow for the lower bound of zero.

Reference for analysis: <https://stats.idre.ucla.edu/r/dae/poisson-regression/>

#-----------------------------------------------------------------------------#  
summary(m1 <- glm(phenord ~ sum\_pLi\_C + Group + sum\_pLi\_C:Group, family="poisson", data=CNV\_dat))

##   
## Call:  
## glm(formula = phenord ~ sum\_pLi\_C + Group + sum\_pLi\_C:Group,   
## family = "poisson", data = CNV\_dat)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1369 -1.0542 -0.3332 0.9532 2.3205   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.38601 0.04548 30.475 < 2e-16 \*\*\*  
## sum\_pLi\_C -0.02213 0.03023 -0.732 0.46415   
## GroupSLIC 0.13895 0.05381 2.582 0.00981 \*\*   
## sum\_pLi\_C:GroupSLIC 0.03862 0.03330 1.160 0.24607   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 1099.0 on 776 degrees of freedom  
## Residual deviance: 1071.1 on 773 degrees of freedom  
## (4 observations deleted due to missingness)  
## AIC: 3565.1  
##   
## Number of Fisher Scoring iterations: 5

#check the assumption that mean and variance ar roughly equal. If not, we can use robust estimates. I think this will make little difference in our case but included for completeness.  
  
cov.m1 <- vcovHC(m1, type="HC0")  
std.err <- sqrt(diag(cov.m1))  
r.est <- cbind(Estimate= coef(m1), "Robust SE" = std.err,  
 "Pr(>|z|)" = 2 \* pnorm(abs(coef(m1)/std.err), lower.tail=FALSE),  
 LL = coef(m1) - 1.96 \* std.err,  
 UL = coef(m1) + 1.96 \* std.err)  
  
r.est

## Estimate Robust SE Pr(>|z|) LL  
## (Intercept) 1.38601474 0.05441780 4.252391e-143 1.27935585  
## sum\_pLi\_C -0.02212557 0.03525869 5.303178e-01 -0.09123260  
## GroupSLIC 0.13895160 0.06360255 2.891187e-02 0.01429061  
## sum\_pLi\_C:GroupSLIC 0.03862436 0.03843621 3.149474e-01 -0.03671062  
## UL  
## (Intercept) 1.49267362  
## sum\_pLi\_C 0.04698146  
## GroupSLIC 0.26361259  
## sum\_pLi\_C:GroupSLIC 0.11395933

#test the effect of dropping group. We anticipate that this will be impair model fit.  
m2 <- update(m1, . ~ . - Group)  
## test model differences with chi square test  
anova(m2, m1, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: phenord ~ sum\_pLi\_C + sum\_pLi\_C:Group  
## Model 2: phenord ~ sum\_pLi\_C + Group + sum\_pLi\_C:Group  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 774 1077.9   
## 2 773 1071.1 1 6.7655 0.009294 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#-----------------------------------------------------------------------------#

### 5. Statistical power analysis for Poisson regression interaction via Monte Carlo simulation

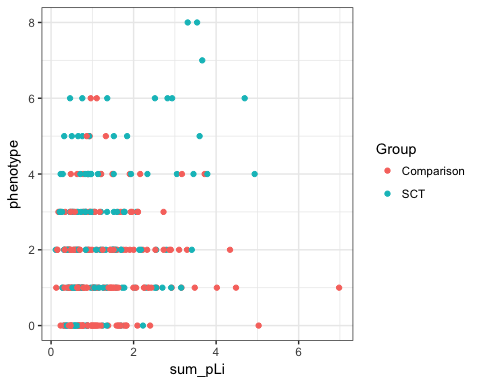
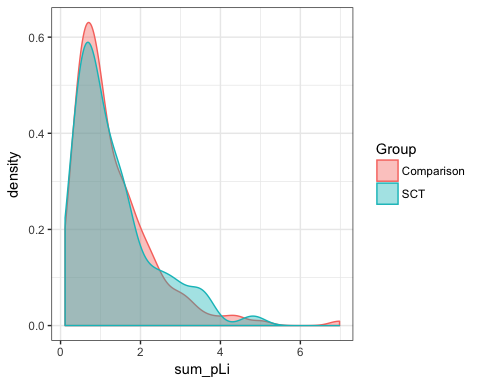
Analyses so far have used an existing dataset (SLIC dataset) to work out optimal methods of analysis for our new data.

For the planned analysis, our sample size is fixed (130 SCT cases and 184 comparison cases), as the data are already collected. We conduct a power analysis using some new simulated data to give an idea of the effect size that is detectable with this sample.

Our main goal is to see whether the association between the CNV burden and phenotype is stronger for the SCT vs comparison group. For Poisson regression we need to specify coefficients corresponding to main effects of CNV burden and group status, as well as the interaction between these. We know in advance that the effect of CNV burden tends to be small and not always detectable, whereas the effect of group status is large - with SCT cases being notably more impaired on average than comparison cases. If we fix the beta for the CNV effect at .10, and the beta for the group effect at .18, we get plausible-looking data. (Though this does vary from run to run with our smallish sample size. Note also that because we use Poisson regression, it is the exponentials of the beta values that enter into the regression).

#generate a population of data to sample and refit model.  
#The first run is just illustrative to show what the simulated data looks like relative to the Epistasis model.  
  
n <- 100000 #size of population from which we will sample  
 myn<-314 #size of sample (based on true size in our SCT/comparison sample)  
   
   
 #regression coefficients for simulation (to see the change in "real" terms use exp(beta).)  
 #These are guestimates - selected to give distributions that look plausible  
 beta0 <- 0.4 #intercept  
 beta1 <- 0.10 #main effect of pLI score on phenotype  
 beta2 <- 0.18 #main effect of Group on phenoptype  
 beta3 <- 0.18 #effect of Group x pLI on phenotype (epistasis)  
   
#generate covariate values (assumption that no correlation between predictors)  
 sum\_pLi <- rlnorm(n=n, 0, 0.7) #approximate distribution of sum\_pLi (looking at real sum\_pLi distribution)  
 #sum\_pLi <- rnorm(n=n, 0, 3) #approximate distribution of sum\_pLi  
 Group <- rbinom(n,1,prob=130/314) #simulate binary variable (0=comparison, 1=SCT)  
   
#compute mu's  
 mu <- exp(beta0 + beta1 \* sum\_pLi + beta2\*Group + beta3\*(sum\_pLi\*Group)) #the rhs is exponentiated as we are using Poisson regression which uses a Log link function.  
   
#generate Y-values  
 phenotype <- rpois(n=n, lambda=mu)  
 w<-which(phenotype>10) #truncate high values  
 phenotype[w]<-10  
   
#data set  
 simdata <- data.frame(phenotype=phenotype, sum\_pLi=sum\_pLi,Group=Group)  
 simdata$Group<-as.factor(simdata$Group)  
 levels(simdata$Group)<-c("Comparison","SCT")

## Distribution plots of sample of simulated data.



## [1] "Correlation of pLI/phenotype: comparison = 0.043; SCT = 0.219"

## Power analysis

In the simulation, we keep the beta values specified for the main effects, and compute the power as we vary the value of beta for the interaction. In addition, we compute the Spearman correlation between the CNV burden and phenotype for the comparison and SCT groups on each run, so we can get a sense of how big a difference we are able to detect.

|  |  |  |  |
| --- | --- | --- | --- |
| beta3 | Pow.CNVxGrp | r.avg.Comp | r.avg.SCT |
| 0.05 | 0.104 | 0.11 | 0.19 |
| 0.06 | 0.120 | 0.11 | 0.19 |
| 0.07 | 0.178 | 0.11 | 0.21 |
| 0.08 | 0.188 | 0.11 | 0.22 |
| 0.09 | 0.314 | 0.10 | 0.23 |
| 0.10 | 0.276 | 0.10 | 0.24 |
| 0.11 | 0.360 | 0.12 | 0.26 |
| 0.12 | 0.482 | 0.11 | 0.27 |
| 0.13 | 0.488 | 0.10 | 0.28 |
| 0.14 | 0.534 | 0.11 | 0.29 |
| 0.15 | 0.640 | 0.11 | 0.31 |
| 0.16 | 0.720 | 0.10 | 0.32 |
| 0.17 | 0.732 | 0.11 | 0.34 |
| 0.18 | 0.764 | 0.10 | 0.33 |
| 0.19 | 0.834 | 0.11 | 0.35 |
| 0.20 | 0.864 | 0.10 | 0.36 |

The table shows the power to detect the interaction term with beta1 and beta2 fixed to .10 and .18. The right-most columns show the average Spearman correlations for the Comparison group and the SCT group. The correlation for the Comparison group does not change: with these parameters, it is fixed around .11. As beta3 increases, we see the correlation increases for the SCT group: power is above 80% to detect a differential effect of CNV on SCTs (Epistasis model) if the correlation for the SCT group is around .35 or above.

## Session information

## R version 3.3.3 (2017-03-06)  
## Platform: x86\_64-apple-darwin13.4.0 (64-bit)  
## Running under: macOS 10.13.5  
##   
## locale:  
## [1] en\_GB.UTF-8/en\_GB.UTF-8/en\_GB.UTF-8/C/en\_GB.UTF-8/en\_GB.UTF-8  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] conflicted\_0.1.0 simglm\_0.7.0 MASS\_7.3-50 lemon\_0.3.3   
## [5] msm\_1.6.5 sandwich\_2.4-0 scales\_0.5.0 ggpubr\_0.1.6   
## [9] magrittr\_1.5 gridExtra\_2.3 Hmisc\_4.1-1 ggplot2\_2.2.1   
## [13] Formula\_1.2-3 survival\_2.40-1 lattice\_0.20-35 knitr\_1.18   
##   
## loaded via a namespace (and not attached):  
## [1] zoo\_1.8-1 reshape2\_1.4.3 purrr\_0.2.4   
## [4] splines\_3.3.3 colorspace\_1.3-2 expm\_0.999-2   
## [7] htmltools\_0.3.6 yaml\_2.1.16 base64enc\_0.1-3   
## [10] rlang\_0.2.0 foreign\_0.8-70 glue\_1.2.0   
## [13] RColorBrewer\_1.1-2 bindrcpp\_0.2 plyr\_1.8.4   
## [16] bindr\_0.1.1 stringr\_1.2.0 munsell\_0.4.3   
## [19] gtable\_0.2.0 htmlwidgets\_0.9 mvtnorm\_1.0-5   
## [22] psych\_1.7.3.21 evaluate\_0.10.1 labeling\_0.3   
## [25] latticeExtra\_0.6-28 parallel\_3.3.3 highr\_0.6   
## [28] htmlTable\_1.11.1 broom\_0.4.3 Rcpp\_0.12.17   
## [31] acepack\_1.4.1 backports\_1.1.2 checkmate\_1.8.5   
## [34] mnormt\_1.5-5 digest\_0.6.15 stringi\_1.2.2   
## [37] dplyr\_0.7.4 grid\_3.3.3 rprojroot\_1.3-2   
## [40] tools\_3.3.3 lazyeval\_0.2.1 tibble\_1.3.4   
## [43] cluster\_2.0.5 crayon\_1.3.4 tidyr\_0.7.2   
## [46] pkgconfig\_2.0.1 Matrix\_1.2-8 data.table\_1.11.2   
## [49] assertthat\_0.2.0 rmarkdown\_1.9 rstudioapi\_0.7   
## [52] R6\_2.2.2 rpart\_4.1-13 nlme\_3.1-131   
## [55] nnet\_7.3-12

## References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

Beckmann, J. S., Estivill, X., & Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nature Reviews Genetics, 8(8), 639-646.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29, 1165–1188.

Bishop, D. V. M. (2003). The Children’s Communication Checklist, version 2 (CCC-2). London: Pearson.

Bishop, D. V. M., Jacobs, P. A., Lachlan, K., Wellesley, D., Barnicoat, A., Boyd, P. A., . . . Scerif, G. (2011). Autism, language and communication in children with sex chromosome trisomies. Archives of Disease in Childhood, 96, 954-959. <doi:10.1136/adc.2009.179747>

Chen, X. S., Reader, R. H., Hoischen, A., Gilissen, C., Veltman, J. A., Simpson, N. H., . . . Fisher, S. E. (2017). Next-generation sequencing identifies novel gene variants and pathways involved in Specific Language Impairment. Scientific Reports, 7(46105). <doi:10.1038/srep46105>

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., . . . Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Research, 35(6), 2013-2025. <doi:10.1093/nar/gkm076>

Frederickson, N., Frith, U., & Reason, R. (1997). Phonological assessment battery (PhAB). Windsor: NFER-Nelson.

Girirajan, S., Rosenfeld, J. A., Cooper, G. M., Antonacci, F., Siswara, P., Itsara, A., . . . Eichler, E. E. (2010). A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nature Genetics, 42, 203-209. <doi:10.1038/ng.534>

Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The Development and Well-Being Assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. Journal of Child Psychology and Psychiatry and Allied Disciplines, 41, 645-655.

Kalnak, N., Stamouli, S., Peyrard-Janvid, M., Rabkina, I., Becker, M., Klingberg, T., . . . Tammimies, K. (2018, in press). Enrichment of rare Copy Number Variation in children with Developmental Language Disorder. Clinical Genetics.

Kearney, H. M., South, S. T., Wolff, D. J., Lamb, A., Hamosh, A., & Rao, K. W. (2011). American College of Medical Genetics recommendations for the design and performance expectations for clinical genomic copy number microarrays intended for use in the postnatal setting for detection of constitutional abnormalities. Genetics in Medicine, 13, 676-679. doi: 10.1097/GIM.0b013e31822272ac

Korkman, M., Kirk, U., & Kemp, S. I. (1998). NEPSY: A developmental neuropsychological assessment. San Antonio: Psychological Corporation.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536, 285-292. <doi:10.1038/nature19057>

Le Gall, J., Nizon, M., Pichon, O., Andrieux, J., Audebert-Bellanger, S., Baron, S., . . . Isidor, B. (2017). Sex chromosome aneuploidies and copy-number variants: a further explanation for neurodevelopmental prognosis variability? European Journal of Human Genetics, 25(8), 930-934. <doi:10.1038/ejhg.2017.93>

Neale, M. D. (1999). Neale Analysis of Reading Ability Second Revised British Edition. London: GL Assessment.

Newbury, D. F., Simpson, N. H., Thompson, P. A., & Bishop, D. V. M. (2018). Stage 1 Registered Report: Variation in neurodevelopmental outcomes in children with sex chromosome trisomies: protocol for a test of the double hit hypothesis. Wellcome Open Research.

Nowakowska, B. (2017). Clinical interpretation of copy number variants in the human genome. Journal of Applied Genetics, 58(4), 449-457. doi: 10.1007/s13353-017-0407-4

Poultney, C. S., Goldberg, A. P., Drapeau, E., Kou, Y., Harony-Nicolas, H., Kajiwara, Y., . . . Buxbaum, J. D. (2013). Identification of small exonic CNV from whole-exome sequence data and application to Autism Spectrum Disorder . American Journal of Human Genetics, 93(4), 607-619. <doi:10.1016/j.ajhg.2013.09.001>

Rocca, M. S., Pecile, V., Cleva, L., Speltra, E., Selice, R., Di Mambro, A., . . . Ferlin, A. (2016). The Klinefelter syndrome is associated with high recurrence of copy number variations on the X chromosome with a potential role in the clinical phenotype. Andrology, 4(2), 328-334. <doi:10.1111/andr.12146>

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., . . . Wigler, M. (2007). Strong association of de novo copy number mutations with autism. Science, 316(5823), 445-449. <doi:10.1126/science.1138659>

Simpson, N. H., Ceroni, F., Reader, R. H., Covill, L. E., Knight, J. C., Hennessy, E. R., . . . Consortium, S. L. I. (2015). Genome-wide analysis identifies a role for common copy number variants in specific language impairment. European Journal of Human Genetics, 23(10), 1370-1377. <doi:10.1038/ejhg.2014.296>

South, S. T., Lee, C., Lamb, A. N., Higgins, A. W., & Kearney, H. M. (2013). ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. Genetics in Medicine, 15(11), 901-909. <doi:10.1038/gim.2013.129>

Torgesen, J. K., Wagner, R., & Rashotte, C. (1999). Test of Word Reading Efficiency (TOWRE). New York: Psychological Corporation.

Veltman, J. A., & Brunner, H. G. (2010). Understanding variable expressivity in microdeletion syndromes. Nature Genetics, 42, 192-193. <doi:10.1038/ng0310-192>

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S., . . . Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genone Research, 17(11), 1665-1674. <doi:10.1101/gr.6861907>

Wang, K., Li , M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research, 38(16), e164. <doi:0.1093/nar/gkq603>

Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. (2017). WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Research, 45 (Web Server Issue), W130–W137. <doi:10.1093/nar/gkx356>

Wechsler, D. (1999). Wechsler Abbreviated Scale of Intelligence. San Antonio: Psychological Corporation.

Wilson, A. C., & Bishop, D. V. M. (2018). Resounding failure to replicate links between developmental language disorder and cerebral lateralisation. PeerJ, 6, e4217. <doi:10.7717/peerj.4217>

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). Woodcock-Johnson III Tests of Cognitive Abilities. Rolling Meadows, IL: Riverside Publishing.