Identifying repeat expansions in individuals with autism spectrum disorder

Beyza Duymayan

UCLA BIG Summer

Introduction

Repeat expansions are a highly polymorphic class of genetic variation in the human genome. The expansion contributes to neurological genetic disorders when reaching beyond a pathogenic threshold.[1]

A resurgence of repeat expansion discovery (Fig. 1) fueled by long-read sequencing era led to an extensive loci catalog.

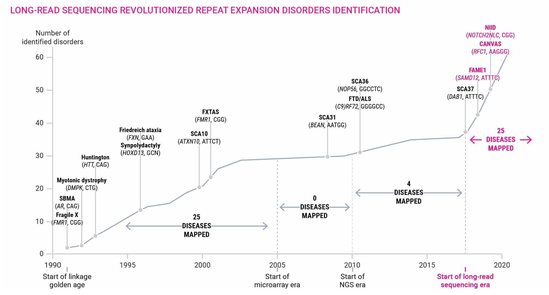

Figure 1 – History of repeat expansions (Depienne et al., 2021)

Family data from the Simons Simplex Collection (SSC) and SPARK aimed at one child with Autism Spectrum Disorder (ASD) and unaffected parents and siblings.

GWAS for sporadic and familial ALS revealed a strong association with SNPs in a 170-kb region at chr9p21.2. Recent findings have identified a GGGGCC hexanucleotide repeat expansion in intron 1 of *C9ORF72* as the pathogenic mutation linked to both familial and sporadic ALS and FTD.[2]

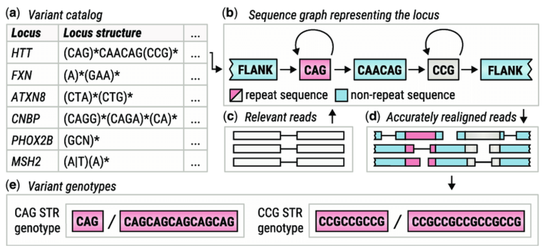

Figure 2 – ExpansionHunter (Dolzhenko et al., 2019)

This study aimed to identify known repeat expansions in individuals with ASD using ExpansionHunter (EH) (Fig. 2), the computational tool that uses short-read sequencing data, which estimates the sizes of the repeats defined in the catalog.

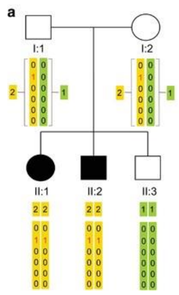

Figure 3– Haplotypes from phasing (Wang et al., 2015)

Tools such as SHAPEIT5, phase whole genotypic data for estimating haplotypes. This study also utilizes this tool to identify the most common haplotypes associated with the *C9ORF72* locus, as well as specific haplotypes linked to pathogenic GGGGCC repeat expansions within the same locus.

Methods

Part 1

~30 loci

~1989 samples (SSC)

Reference human genome 38

Expansion Hunter

.fam file ~1989 samples

Contains phenotypic information, which is the number of repeat units on the long allele for all samples

Filtering using PLINK

.bim file SNP data

.fam file ~1944 samples

.bed file binary genotype data

For consistency, only family of 4 was kept. Removed samples without phenotypic data. The .bim file was filtered against non-SNPs and multiallelic SNPs. The .bed contains genotype data.

ClassicMendel

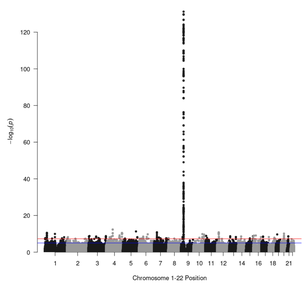

Figure 4 – Manhattan plot of all autosomal chromosomes of *C9ORF72* locus (previous study SSC data) – (MAF > 5%)

ClassicMendel performs pedigree-GWAS on all autosomal chromosomes for statistical analysis of quantitative trait, number of repeats in long allele.

Part 2

~12519 samples (SPARK)

Filtered with PLINK and ran ClassicMendel

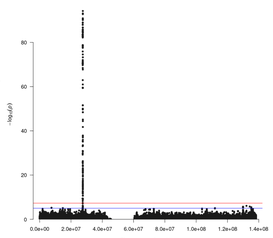

Figure 5– Manhattan Plot of chromosome 9 of *C9ORF72* locus (previous study SPARK data) – (MAF > 5%)

~11419 samples

Figure 5 has a peak of significant SNPs which are strongly associated with the repeat in the *C9ORF72* locus. With this range (Fig. 6), through phasing by the tool SHAPEIT5, alleles are assigned to the paternal and maternal chromosomes, creating common haplotypes. Haplotypes were assessed through parent samples only (~6099 samples)

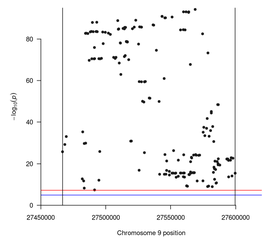

Figure 6 – Zoomed in peak of manhattan plot of chromosome 9 of *C9ORF72* locus (previous study SPARK data)

Normal	Intermediate	Pathogenic
1-19	20-23	≥24

Table 1– Ranges of repeat units in *C9ORF72* for pathogenicity of ALS/FTD

Figure 7 shows the samples, represented as black dots, that hold the pathogenic repeat expansion of GGGGCC of *C9ORF72*, considering them as carriers of the disease.

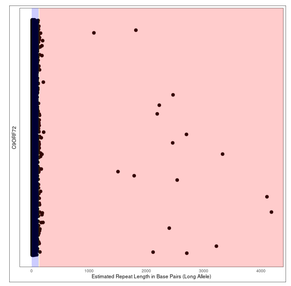

Figure 7 – Graph of normal, intermediate, and pathogenic ranges for samples' repeat expansions of GGGGCC in *C9ORF72* locus

Haplotypes specific to these pathogenic carriers are identified by comparing haplotypes of the pathogenic parent and child, similar to Figure 3. Originally, 21 families contain carriers, but only 9 families show parent to child transmission. A 5 kb range of chr9: 27570000-27575000 is used for the haplotype.

Family ID	Sample ID	Father ID	Mother ID	Sex	Repeat Unit Length
SF098106	SP098167	0	0	1	6
SF098106	SP097910	0	0	2	411
SF098106	SP098106	SP098167	SP097910	2	353
SF098106	SP098107	SP098167	SP097910	1	6

Table 2 – A family with a pathogenic repeat carrier mother and daughter

Results

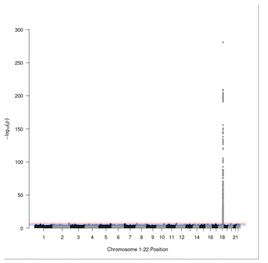

Figure 8 – Manhattan plot for CAG repeat in *TCF4* (pedigree format) – (MAF > 5%)

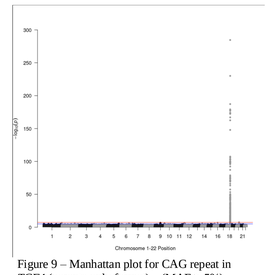

Figure 9 – Manhattan plot for CAG repeat in *TCF4* (parents only format) – (MAF > 5%)

Multiple loci followed the pattern of a strong peak, similar to Figure 4. In Figure 8, there are a group of SNPs transmitted together, indicating linkage disequilibrium (LD). These SNPs demonstrate both linkage and association with the repeat when analyzed using a pedigree-formatted .fam file. In Figure 9, when plotting with only parent information in the .fam file, a strong peak still prevailed, although with fewer significant SNPs.

The SNPs between Figure 8 and 9 are positively correlated with each other, indicating association. Similar SNPs show higher significance, by p-value, in pedigree format than parents only.

Correlation Coefficient: 0.9807

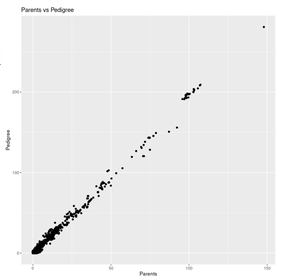

Figure 10 – Correlation graph between Figure 8 and Figure 9 SNPs

When identifying the common haplotypes with samples of parents only, 4 major haplotypes existing at a frequency of about 67% of the total haplotypes in *C9ORF72* were found in Figure 11.

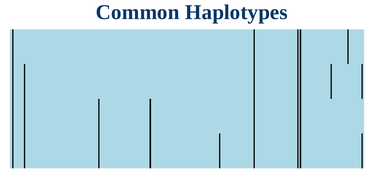

Figure 11 – Four common haplotypes of parents only in *C9ORF72*

Total Haplotypes: 12198 (6099 samples of parents only)
Each haplotype is 276 SNPs long
Total frequency of common haplotypes: ~67%

When identifying the common haplotypes with samples of carriers of the pathogenic GGGGCC repeat in *C9ORF72*, family data was used to identify a total of 3 haplotypes (Fig 12).

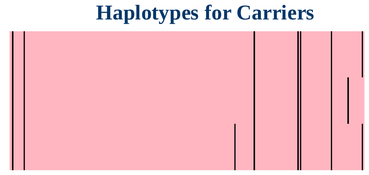

Figure 12 – Three haplotypes of carriers of pathogenic GGGGCC repeat in *C9ORF72*

Total Haplotypes: 19 (9 families)
Each haplotype is 276 SNPs long

Between Figure 11 and 12, we see SNPs that are present in these haplotypes as the following positions: 27570156, 27573523, 27574017, and 27574089.

Discussion

Repeat expansions are known to contribute towards neurological disorders at pathogenic thresholds. Therefore, it is crucial to characterize the distribution of associated repeat expansions at the population level, through significant SNPs. In Figures 10 and 11, SNPs in the locus, defined by the peak, are associated with the repeat length of the *C9ORF72* repeat expansion, GGGGCC.

By analyzing haplotypes, researchers can identify SNPs of the genome that are in LD. The utility of identifying haplotypes lies in the ability to efficiently pinpoint carriers of pathogenic repeat expansions by analyzing associated haplotypes and SNPs. EH can be used for a preliminary check, serving as a proxy to identify potential carriers before proceeding with more expansive whole-genome sequencing techniques

Future Directions

Hopefully, common haplotypes of other loci besides *C9ORF72* can be identified with the new-found chromosomal range of significant SNPs. In addition, larger families of carriers of pathogenic repeats is needed as the current study of 9 families was too small of a sample size. Future studies should look to analyze the association of the repeat to ASD individuals.

Acknowledgements

I would like to extend my thanks to the BIG Summer Program, Roel Ophoff, and Lingyu Zhan for contributions towards this research.

References

1. Leitão, E., Schröder, C., & Depienne, C. (2024). Identification and characterization of repeat expansions in neurological disorders: Methodologies, tools, and strategies. *Revue Neurologique*.
2. Smith, B. N., Newhouse, S., Shatunov, A., Vance, C., Topp, S., Johnson, L., Miller, J., Lee, Y., Troakes, C., Scott, K. M., Jones, A., Gray, I., Wright, J., Hortobágyi, T., Al-Sarraj, S., Rogelj, B., Powell, J., Lupton, M., Lovestone, S., Sapp, P. C., … Shaw, C. E. (2013). The C9ORF72 expansion mutation is a common cause of ALS+/-FTD in Europe and has a single founder. *European journal of human genetics : EJHG*, *21*(1), 102–108. https://doi.org/10.1038/ejhg.2012.98