Characterisation of the C9orf72 repeat expansion in the UK Biobank: populations, phenotypes and modifiers.

Stuart Lee[1,2], Liam Fearnley[1,2], Haloom Rafehi[1,2], Mark Bennett[1,2], Melanie Bahlo[1,2]

1. Genetics and Gene Regulation Division, WEHI, Parkville, Victoria
2. Department of Medical Biology, University of Melbourne, Parkville Victoria

Background: The hexamer short tandem repeat GGGGCC in the promoter region of *C9orf72* locus is known to cause amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) when a person carries at least 30 copies of the pathogenic motif on a single allele. It is also the underlying genetic risk factor for the GWAS peaks identified for both ALS and FTD (Reus et al. 2021; van Rheenen et al. 2021), highlighting its importance in these two disorders. Surprisingly, recent work (Ibañez et al. 2024; Gagliardi et al. 2025) has demonstrated that the prevalence of the pathogenic allele is more common than expected in healthy volunteer biobank cohorts suggesting that penetrance is currently overestimated and/or that disease is underdiagnosed.

Methods: Using short-read whole genome sequencing of unrelated individuals in UKB (n = 377,776), we have estimated the copy number of the repeat unit using ExpansionHunter5 (Dolzhenko et al. 2019) and estimate the pathogenic carrier allele frequency across PANUKB genetic ancestries and biological sex over all participants and excluding those with no prior neurological condition. We also performed PheWAS between C9orf72 GGGGCC longer allele length with PhecodeX phenotypes and other multi-omic modalities in UKB. We leverage known familial relationships in the UKB to form monozygotic twins (n=126 pairs) for quality control, and parent-child trios (n = 989 trios) and duos (n = 3781) to perform quality control on our repeat calls and to assess genetic anticipation.

Results: We confirm that the prevalence of the C9orf72 GGGGCC pathogenic allele length, defined as having >29 copies of the repeat motif, is more common in the UKB participants with 1 in every 449 persons of European descent and 1 in every 278 persons without prior neurological diseases. PheWAS analysis recapitulates known disease risk Phecode associations like *Systemic atrophies primarily affecting the central nervous system* (OR: 1.127 95% CI [1.006, 1.113]), *dementias and cerebral degeneration* (OR: 1.087 95% CI [1.005, 1.076]), as well as *symptoms related to mental disorders* (OR: 1.029 95% CI [1.007, 1.015]). The family-based analysis demonstrates that inheritance of the repeat is relatively stable when the repeat copy number is below the pathogenic threshold with there being on average no change in allele length in a child's repeat size, however when a parent is carrying an intermediate (between 23 and 29 copies) or above pathogenic repeat size the child is more likely to have a decreased number of copies. Further analysis is required to confirm these results. These insights will be incorporated into the modifier analyses.