**Getting SNP RSIDs for Parkinson and CJD GWAS data**

Firstly, the Parkinson GWAS summary-statistics are available here <http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST009001-GCST010000/GCST009325/> , and the CJD GWAS summary-statistics are available here <http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90001001-GCST90002000/GCST90001389/>

Unfortunately, the CJD and Parkinson GWAS summary statistics came without SNP RSIDs, but these can be inferred using the Chromosome number, base position, and the ref and alt alleles. According to the authors of both papers, the genome build used is GRCh37. Apart from this information found in the GWAS summary statistics table, we also need a comprehensive list of variants from this build (<https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/00-All.vcf.gz>). The file name for the latter is “00-All.vcf.gz”. What I will be explaining is an adapted version of what was written by “cpad0112” in this thread <https://www.biostars.org/p/160302/>

The first step was extracting Chromosome number, base position, and the ref and alt alleles of the CJD and Parkinson GWAS summary stats. Then I copied the base position again so as to copy the bed format (structure: Chromosome number, base position, base position, ref allele, alt allele). I did this for every chromosome so that 1) there wasn’t overlap in the same position on different chromosomes and 2) it made it so my computer was able to compute things, otherwise it wouldn’t have had the memory to do it. This is all documented in “Get per chrom file.R”. I made sure that these were tab separated and deleted the header line on the command line using the following line:

tail -n +2 CJD\_chr21.tsv > CJD\_chr21\_1.tsv

As for “00-All.vcf.gz”, I firstly unzipped it using the “gunzip” command on bash 3.2 and later also separated it by chromosomes using the awk command on bash. The following is an example of how I extracted all the chromosome 2 variants:

awk '$1=="2"' 00-All.vcf > chr2\_grch37.vcf

I then added the header section of the original “00-All.vcf.gz” to all the chrX\_grch37.vcf by creating a copying the header, accessing each “chrX\_grch37.vcf” manually with “nano” and pasting the header.

To get SNP RSIDs, I used bedtools 2.30.0 using the following command (CJD chr 21 as an example):

bedtools intersect -a CJD\_chr21\_1.tsv chr21\_grch37.vcf -b .vcf -wb > intersect\_chr21\_CJD.tsv

Problems with this approach: Too many variants, more than one per base pair location. Making sure that there is only one variant per base pair location: There is one R script per chromosome called “Get\_final\_SNPs\_chrX\_disease.R” which explains how that was done. Next step was to see which base pair locations didn’t get any SNP RSIDs, and then aggregate all the final SNPs into one table and add them to the GWAS summary statistics. This is shown in “SNPs\_in\_GWAS\_disease.R”, where disease is either CJD or Parkinson