Additional file 1: The Avon Longitudinal Study of Parents and Children (ALSPAC)

ALSPAC: Cohort Description

ALSPAC recruited 14,541 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992. 14,541 is the *initial* number of pregnancies for which the mother enrolled in the ALSPAC study and had either returned at least one questionnaire or attended a "Children in Focus" clinic by 19/07/99. Of these *initial* pregnancies, there was a total of 14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age.

When the oldest children were approximately 7 years of age, an attempt was made to bolster the initial sample with eligible cases who had failed to join the study originally. The number of new pregnancies not in the initial sample (known as Phase I enrolment) that are currently represented and reflecting enrolment status at the age of 18 is 706 (452 and 254 recruited during Phases II and III respectively), resulting in an additional 713 children being enrolled.

The total sample size for analyses using any data collected after the age of seven is therefore 15,247 pregnancies, resulting in 15,458 foetuses. Of this total sample of 15,458 foetuses, 14,775 were live births and 14,701 were alive at 1 year of age.

A 10% sample of the ALSPAC cohort, known as the Children in Focus (CiF) group, attended clinics at the University of Bristol at various time intervals between 4 to 61 months of age. The CiF group were chosen at random from the last 6 months of ALSPAC births (1432 families attended at least one clinic). Excluded were those mothers who had moved out of the area or were lost to follow-up, and those partaking in another study of infant development in Avon.

ALSPAC: Genotyping and Imputation Description

ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms by 23andme subcontracting the Wellcome Trust Sanger Institute, Cambridge, UK and the Laboratory Corporation of America, Burlington, NC, US. The resulting raw genome-wide data were subjected to standard quality control methods. Individuals were excluded on the basis of gender mismatches; minimal or excessive heterozygosity; disproportionate levels of individual missingness (>3%) and insufficient sample replication (IBD < 0.8). Population stratification was assessed by multidimensional scaling analysis and compared with Hapmap II (release 22) European descent (CEU), Han Chinese, Japanese and Yoruba reference populations; all individuals with non-European ancestry were removed. SNPs with a minor allele frequency of < 1%, a call rate of < 95% or evidence for violations of Hardy-Weinberg equilibrium (P < 5E-7) were removed. Cryptic relatedness was measured as proportion of identity by descent (IBD > 0.1). Related subjects that passed all other quality control thresholds were retained during subsequent phasing and imputation. 9,115 subjects and 500,527 SNPs passed these quality control filters.

ALSPAC mothers were genotyped using the Illumina human660W-quad array at Centre National de Génotypage (CNG) and genotypes were called with Illumina GenomeStudio. PLINK[1,2] (v1.07) was used to carry out quality control measures on an initial set of 10,015 subjects and 557,124 directly genotyped SNPs. SNPs were removed if they displayed more than 5% missingness or a Hardy-Weinberg equilibrium P-value of less than 1.0e-06. Additionally SNPs with a minor allele frequency of less than 1% were removed. Samples were excluded if they displayed more than 5% missingness, had indeterminate X chromosome heterozygosity or extreme autosomal heterozygosity. Samples showing evidence of population stratification were identified by multidimensional scaling of genome-wide identity by state pairwise distances using the four HapMap populations as a reference, and then excluded. Cryptic relatedness was assessed using a IBD estimate of more than 0.125 which is expected to correspond to roughly 12.5% alleles shared IBD or a relatedness at the first cousin level. Related subjects that passed all other quality control thresholds were retained

during subsequent phasing and imputation. 9,048 subjects and 526,688 SNPs passed these quality control filters.

477,482 SNP genotypes in common between the sample of mothers and sample of children were combined. SNPs with genotype missingness above 1% due to poor quality were removed (11,396 SNPs removed). 321 subjects were removed due to potential ID mismatches. This resulted in a dataset of 17,842 subjects containing 6,305 duos and 465,740 SNPs (112 were removed during liftover and 234 were out of HWE after combination). Haplotypes were estimated using ShapeIT (v2.r644) which utilises relatedness during phasing. A phased version of the 1000 genomes reference panel (Phase 1, Version 3) was obtained from the Impute2 reference data repository (phased using ShapeIt v2.r644, haplotype release date Dec 2013). Imputation of the target data was performed using IMPUTE[3] V2.2.2 against the reference panel (all polymorphic SNPs excluding singletons), using all 2,186 reference haplotypes (including non-Europeans).

This gave 17,842 mothers and children eligible for study with available genotype data.

References

1. *Purcell S et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81(3): 559-575.*
2. *PLINK v1.07 by S.Purcell http://pngu.mgh.harvard.edu/purcell/plink/*
3. *Howie et al. (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics 5(6): e1000529.*