Fragile Families
Genotype Component/DNA
Restricted Use Data Appendage

January 2019

Prepared by the staff at the Bendheim-Thoman Center for Research on Child Wellbeing (CRCW), the Department of Molecular Biology, and the Department of Computer Science, Princeton University. For more information about Fragile Families, please visit our web site at http://www.fragilefamilies.princeton.edu/ or email ffdata@princeton.edu

DATA APPENDAGE OVERVIEW

The Fragile Families Genotype Component/DNA Restricted Use Appendage (ff_snp_pub1) contains genetic information on focal children from the Fragile Families and Child Wellbeing Study (FFCWS). In order to obtain and process genetic information, saliva samples were provided by focal children during in-home visit assessments at the Year 9 follow-up wave and by mail at the Year 15 follow-up wave. The goal of collecting genetic information was to allow researchers to test hypotheses about the relationships among genes, family and community environments, and child development.

DATA COLLECTION AND PROCESSING PROCEDURES

Overview

As part of the Years 9 and 15 follow-up waves, we attempted to collect saliva samples for genetic analysis from all focal children completing the in-home visit activities. Ultimately, 3,221 samples from unique focal children were collected. Samples were received, processed, and genotyped under the supervision of Dr. Daniel Notterman, Department of Molecular Biology at Princeton University, Principal Investigator of the NICHD award that supports genetic research with the Fragile Families and Child Wellbeing Study.

Collection kit

Interviewers used the Oragene® DNA Self-Collection Kit (DNA Genotek Inc., Ontario, Canada) to collect saliva samples from focal children. The Self-Collection Kit is a repository for the collection, preservation, transportation, and purification of DNA from saliva.

Process for administering

The respondents were instructed to spit into the container until the liquid portion reached a line on the interior of the container (the ideal volume of saliva to be collected was 2 ml). The container was then capped. In the process of screwing the cap onto the container, a liquid preservative was released. The container was then put into a small plastic biohazard bag that contained absorbent material if the container were to leak. The plastic bag was then put into a mailer.

In cases where the child had developmental or physical limitations prohibiting the interviewer from collecting the full sample by having the child spit into the collection kit, the child accessory kit was used. The child accessory kit contained a set of five saliva sponges used with the Oragene® self-collection kits. The saliva sponges were inserted into the child's mouth and moved around the upper and lower cheek pouches on both sides of the mouth to collect saliva. The sponges were stored inside the containers and then sealed as described above. Respondents were instructed to rinse their mouth out 5 minutes prior to the saliva sample collection. They were also provided with a packet of sugar and instructed to use 1⁄4 tsp. if they were having difficulty stimulating saliva.

Lab receipt and identification of kits

Specimen containers (placed in the bubble wrap mailers) were mailed back to Westat. As Westat received specimen containers from the field, they were receipted and placed in a shipment box with other received containers. Until they were mailed, these boxes were secured in the locked field room and maintained at room temperature. On an approximately monthly basis during the field period, Westat shipped boxes of specimen containers to the laboratory at Princeton University. A transmittal form containing the IDs of the enclosed containers was emailed to lab staff. The lab confirmed receipt of the boxes with Westat. In a few cases, samples were collected by other investigators at Columbia University or the University of Michigan, and then transferred to the Notterman lab.

Extraction and storage

From October 2007 through May 2010, the lab at Princeton received monthly FedEx shipments from October 2007 through May 2010 and again from April 2014 through January 2017 containing DNA saliva sample collection kits from Westat. In total, 163 shipments were received. Upon receipt of the shipments, lab technicians used a barcode reader to inventory the individual samples. These data were imported into a Microsoft Access database where a full inventory of receipted samples was kept. In addition, smaller shipments were received from Colter Mitchell, PhD at the University of Michigan between February 2017 and March 2018 for additional Fragile Families participants located in that area.

Extraction was completed 1 to 2 weeks after receipt of samples from Westat. DNA was extracted from the entire sample using the Oragene® Laboratory Protocol for Manual Purification of DNA (DNA Genotek) and divided into three aliquots: one working (4°C), and two long-term storage (-80°C). Samples processed during the Year 9 wave were divided into three aliquots: one working (4°C) and two long-term storage (-80°C). Samples processed during the Year 15 wave were divided into two aliquots and stored at -80°C.

When samples were ready to be processed, they were incubated at 50°C in a water incubator for a minimum of 1 hour or in an air incubator for a minimum of 2 hours. The mixed Oragene®-DNA/saliva sample was transferred to a 15 ml centrifuge tube. A 1/25 ul volume portion of Oragene®-DNA Purifier was added to the microcentrifuge tube and mixed by vortexing for a few seconds. The sample was incubated on ice for 10 minutes, then centrifuged at room temperature for 10 minutes at 4,000 rpm. The clear supernatant was carefully transferred with a pipet into a fresh centrifuge tube. The same volume of room- temperature 95-100% ethanol was added to an equal volume of supernatant and gently mixed by inversion 10 times. The sample was allowed to stand for 10 minutes at room air to allow the DNA to fully precipitate. The tube was then placed in the centrifuge for 10 minutes at room air at 4,000 rpm. Supernatant was then removed with a pipet and discarded, taking care to avoid disturbing the DNA pellet. An ethanol wash consisting of 1 ml of 70% ethanol was added to the tube without disturbing the pellet. After standing at room temperature for 1 minute, the tube was gently swirled to completely remove the ethanol, taking care not to disturb the pellet. The DNA was rehydrated by adding 0.5 to 1.0 ml of

TE solution, vortexing the sample for 30 seconds, incubating it at room temperature, vortexing it again and transferring the rehydrated DNA to 1.5 ml micro centrifuge tubes. DNA concentration was determined using the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher) and stored as described above.

GENOTYPING

Genomic DNA (200 ng) was genotyped using the Illumina PsychChip (Psychiatric Genomics Consortium, Illumina) at the Pennsylvania State University College of Medicine Genome Sciences Core (Hershey, PA) according to the manufacturer's protocol (Illumina, Inc., San Diego, CA). Two different bead pools were used during the course of the project (PsychChip_15048346_B (v 1.0) and PsychChip_v1-1_15073391 (v 1.1)). The original data was clustered in three batches in GenomeStudio v2.0 (Illumina Inc., San Diego, CA) and TOP allele data (GRCh37) exported in plink format. Plink v1.07 was used to remove samples with >0.02 missing SNPs (n=362) and SNPs with >0.05 missing. SNPs were then filtered again to remove those with a missing rate > 0.02. The genotyping rate of the remaining samples was >0.99. Variants that were duplicated on the PsychChip (different identifiers with the same position) were merged. If non-biallelic variation was introduced during this merge, the least frequent allele was set to missing. We then corrected the strandedness and reference position of the genotyped SNPs by running the WTCCC strand correction workflow (Rayner and McCarthy pipeline: (http://www.well.ox.ac.uk/~wrayner/strand/WR-ASHG2011posterPP-portrait.pdf). In short, this workflow uses BLAT alignments of probe sequences against a recent human genome version to correct position and orientation and removes poorly mapped variants. As a final cleaning step, the last several variants were corrected to match the 1000 Genome Reference (The 1000 Genomes Project Consortium, 2015).

FILE DESCRIPTIONS

This data appendage includes results from genotyping 3119 Fragile Families focal children. All coordinates correspond to the February 2009 human reference sequence (GRCh37) produced by the Genome Reference Consortium. There are four files. Three of these files, ff_snp_pub1.bim, ff_snp_pub1.bed, and ff_snp_pub1.fam, contain the genotype data in binary plink format (Purcell et al., 2007). The fourth file, ff_snp_pub1.pheno contains limited phenotype data from the Fragile Families and Child Wellbeing Study. All of these files are text files except ff_snp_pub1.bed.

The ff_snp_pub1.bim file is a six-column white-spaced file containing one line per variant with the following fields:

1. Chromosome code
2. Variant ID
3. Position in centimorgans
4. Base-pair coordinate (1-based)
5. ALT allele code
6. REF allele code

The ff_snp_pub1.fam file is a six-column white-spaced file containing one line per sample with the following fields:

1. Family ID ('FID')
2. Within-family ID ('IID')

3. Within-family ID of father (0 since father is not in dataset)
4. Within-family ID of mother (0 since mother is not in dataset)
5. Sex code (‘1’= male, ‘2’= female, ‘0’=unknown)
6. Phenotype value (‘-9’/’0’ = missing data)

The ff_snp_pub1.bed file is a binary biallelic genotype table. For a complete explanation of the file format, please see: http://zzz.bwh.harvard.edu/plink/binary.shtml.

The ff_snp_pub1.pheno file contains child’s sex, parent and self-reported race/ethnicity, and self-reported or directly-measured height, weight, and body mass index (at Year 15).

REFERENCES

1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. Nature. 526:68-74.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., Sham, P.C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559-75 Epub 2007 Jul 25.