# Next Generation Genotyping (NGG) Applied to the MindCrowd Study Cohort

Ignazio Piras,[1] Joshua Talboom,[1] Azeem Siddique,[2,3] Phillip Ordoukhanian,[2,3] Steve Head,[2,3] Keith Brown[3] (kbrown@igenomx.com), Matt Huentelman[1] (mhuentelman@tgen.org) | 1. TGen, Phoenix, AZ; 2. The Scripps Research Institute, La Jolla, CA; 3. iGenomX, Carlsbad, CA

## Impact

The future of population scale genotyping will feature:

- Internet based recruitment and phenotyping
- Self-reported health status and family history
- At home sample collection (through the mail)
- Low input requirements: 1ng
- Speed to results: < 1 week
- High content genotyping: > 80M genotypes/ sample
- High call rate: 99% average
- Low failure rate: <2%
- Consistent performance across ethnicities (no ascertainment bias)
- Low cost: $50/sample

## Introduction

It is estimated that over 5 million people suffer from Alzheimers Disease (AD) in the United States, and that number will grow to over 13 million by 2050. Having a first degree relative with AD is a well-documented risk factor, but the influence of family history on cognition across the lifespan is poorly understood. To address this issue, the Mindcrowd study (mindcrowd.org) was initiated. Mindcrowd is an internet based paired-associates learning (PAL) task program that has recruited over 150K participants across the globe and has shown that family history is associated with a reduced PAL scores across both sexes approximately four decades prior to the typical onset of AD. To explore the genetic risk associated with family history and the effect on PAL and AD onset or modified effect, we applied the Riptide High Throughput DNA library preparation with low pass sequencing and imputation to genotype over 80 million bi-allelic variant positions annotated from the 1000genomes project on a subset of samples (n=960). This proof of concept study represents the future of population scale genotyping as it includes internet-based phenotyping and recruitment, self-reporting of relevant health information and family history, at home sample collection (blood), low input sample processing (1-2ng), high capacity genotyping by sequencing (>80M SNPs) and imputation (1kgenome project phase III). Data was generated in one week at a cost of approximately $50 per sample.

## Methods:

Phenotyping was performed by testing three replicate paired word associations. During the learning phase, each participant was presented with each word pair at 2 second intervals. During the recall phase, each participant was presented the first word in the pair and used their keyboard to type in the second word in the pair from memory. The test was repeated two more times and a cumulative score was determined. Samples for consenting participants were obtained by self-submission of dried blood spot on a 903-protein saver card (Whatman, UK) and shipped via the mail. DNA extraction: Perkin/Elmer Chemagic 360 CMG 1030 cDNA from DBS Kit. Approximately 1ng of DNA was input into the Riptide library prep using the low input protocol (iGenomX, CA, USA) and sequenced on an Illumina Novaseq s4 flow cell with 96 samples per lane (Psomagen, MD, USA). FASTQ files were demultiplexed using the opensource demux tool(Fulcrum Genomics, AZ, USA) and individual sample FASTQs were processed through the Gencove pipeline to produce individual sample Variant Call Files (VCF). All the following analysis were conducted with bcftools, vcftools and PLINK. We excluded low quality imputed and low called variants (Genotype Probability < 0.9 and genotype rate < 95%, respectively), as well as low frequency variants (minimum allele frequency < 5%). Finally, we excluded sex-mismatched, outliers and related samples. Association analysis with PAL score was conducted by linear regression, adjusting for age and education attainment. Population stratification was assessed by λ inflation factor.

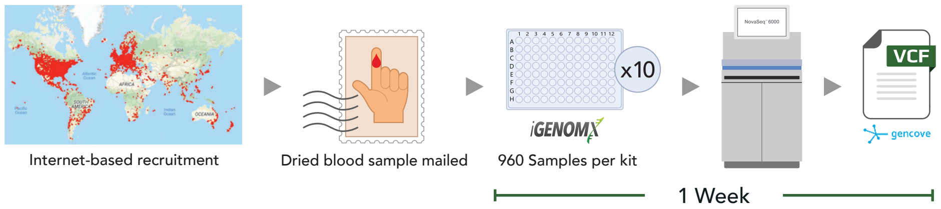

Figure 1: Sample acquisition and genotyping workflow.

| | |
|---|---|
| Sample Success Rate<br>Sample number = 960 | 98.2% |
| Genotyping Call Rate<br>85M variant positions | >99.6%<br>Average: 99.6 ± 0.003<br>Range: 96.5-99.8% |
| Library Prep Cost Per Sample | $5<br>Sequencing cost:<br>~ $35 per sample |

Table 1: Success rate percentage of samples per 96 well plate that pass quality control filters from NGG pipeline and percent of the 80M genotype positions with a call made from each 96 well plate.

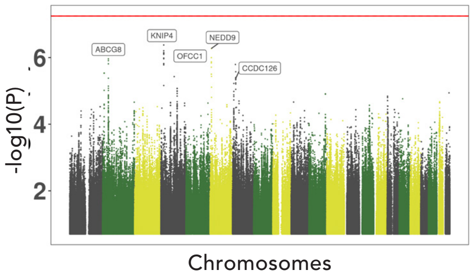

Figure 2: Manhattan plot showing the results of the genome-wide quantitative trait loci (QTL) association analysis with PAL score. The red line indicates genome-wide significance (p < 5.0E-08).

## Results:

After imputation was performed, 98.2% of samples passed quality control filters based on minimum read coverage of the 80M+ variant positions from the phase III release of the 1000genomes project (Table 1). Of the over 80M variant positions, an average call rate of 99.6% was achieved with a range of 96.5% to 99.8%. (Table 1). After selecting the family history subsample and conducting quality controls, we obtained 5,046,139 variants in 581 participants. After association analysis, we did not detect SNPs significant at the genome-wide level (p < 5.0E-08), and p-value distribution showed no significant inflation (λ= 1.003). The manhattan plot (Figure 2) shows the top hits including KNIP4, and NEDD9. NEDD9 is a neural precursor cell expressed developmentally downregulated protein. The protein is an intermediate in a number of important signaling pathways relevant for proliferation, survival, and migration and has been previously linked to Late Onset Alzheimer's Disease. Four SNPs in the NEDD9 gene correlate with increased PAL scores across all age deciles (Figure 3). KNIP4 encodes Kv channel interacting protein 4 and belongs to a family of voltage gated potassium channel-interacting proteins which may regulate neuronal excitability in response to changes in intracellular calcium. KNIP4 also interacts with Presenilin, a known Familial Alzheimer's associated protein. Six SNPs in the KNIP4 gene show correlation with decreased PAL scores across all age deciles (Figure 4).

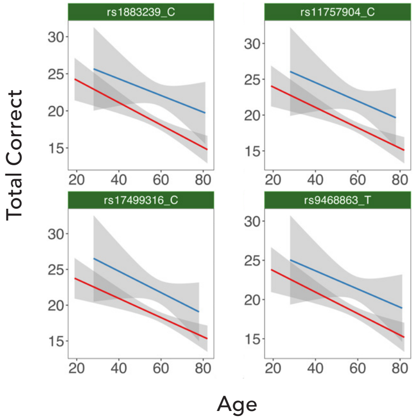

Figure 3: Correlation of PAL score (y-axis) with age (x-axis) by genotype for the top SNPs located in NEDD9.

– Carrier of the minor frequency allele
– Non carriers of the minor frequency allele

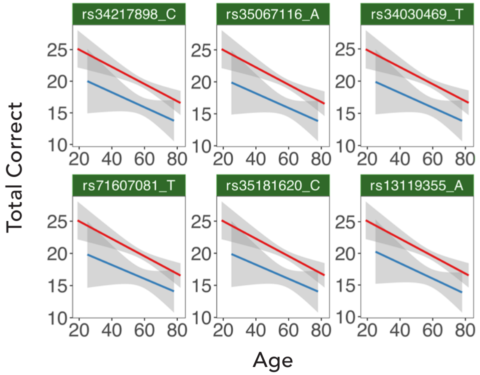

Figure 4: Correlation of PAL score (y-axis) with age (x-axis) by genotype for the top SNPs located in KNIP4.

– Carrier of the minor frequency allele
– Non carriers of the minor frequency allele

## Conclusions:

Next generation genotyping is enabled by the iGenomX Riptide High Throughput Library preparation and is shown here to maximize unbiased genotyping across ethnicities with high call rates, low failure rates, and at a cost and scale previously unobtainable with older-generation genotyping technologies. The combination of internet-based recruitment and phenotyping, at home sample collection, low input requirements, easy workflows and significantly reduced cost represents the future of population scale genomic research. This research highlights both known and novel genetic associations with Alzheimers disease. Increased statistical power is required to confirm these associations, and further biological studies are required to determine the role and mechanism by which these genes contribute to familial Alzheimer's risk.

**To learn more, visit igenomx.com and mindcrowd.org**