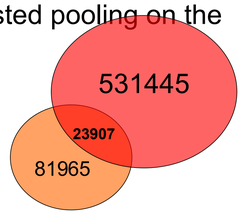
DNA Pooling for Whole-Genome Association Studies with Illumina Infinium Assays

Amber E. Baum[1], Nirmala Akula[1], Ben Klemens[3], Imer Cardona[1], Winston Corona[1], Andrew Singleton[2], John Hardy[2], Sevilla Detera-Wadleigh[1], Francis J. McMahon[1]

[1]Genetic Basis of Mood and Anxiety Disorders Unit, National Institute of Mental Health, and
[2]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, [3]Brookings Institution

INTRODUCTION

Genome-wide association studies are now feasible. Measuring allele frequencies of pools of cases and controls, instead of between individuals, would greatly decrease costs and facilitate discovery. DNA pooling has successfully been used on the 10K Affymetrix microarray platform, but no groups have shown that it works on the Illumina platform. We tested pooling on the Infinium chips Human-1 (~109K SNPs) and HumanHap550 (~555K SNPs), which have ~23K overlap.



METHODS

Individuals and Pool Construction

- Coriell panel NDPT008 (neurologically normal Caucasian males and females ages 55-84, n=88)
- Individual genotyping in Hardy lab, call rates >95%
- Six equimolar pools manually prepared and genotyped by McMahon lab
- Concentration assayed by PicoGreen (serial dilutions from 200 to 50 to 10ng and reconcentration to 50ng, variance: 10% of mean) and verified by nanodrop

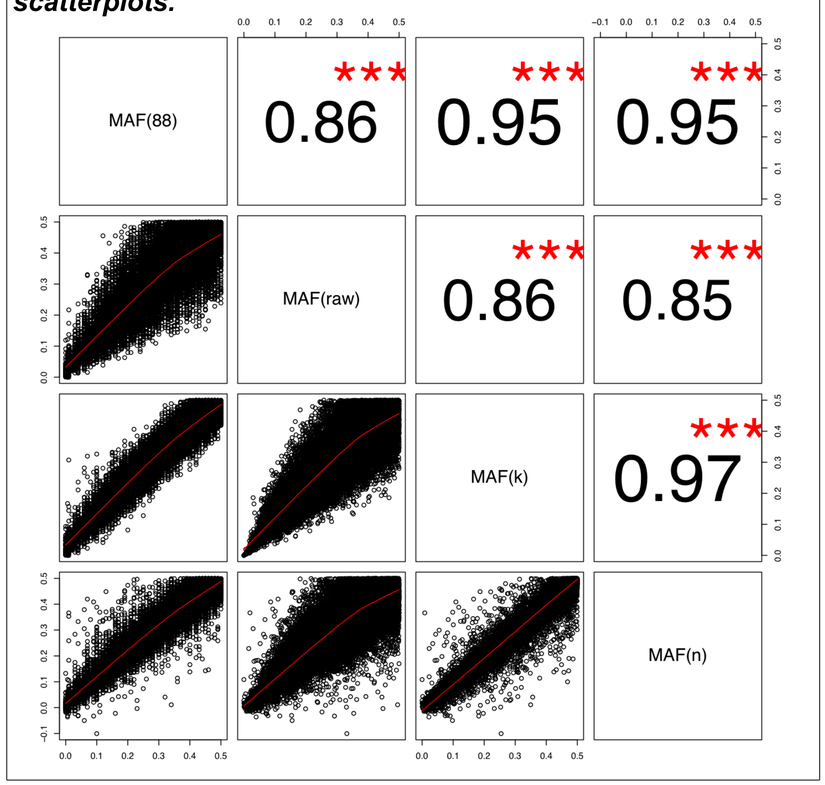
Allele Frequency Calculations, Statistical Analysis

- MAF_{88} calculated from 88 individual Hardy lab chips
- RAF_{raw} calculated by Illumina's BeadStudio
- SNP-specific correction factors *k* and AA_{avg}/BB_{avg} derived separately for each platform:
  - 109K: 242 individuals genotyped in McMahon lab
  - 550K: 120 HapMap individuals genotyped by Illumina

RESULTS

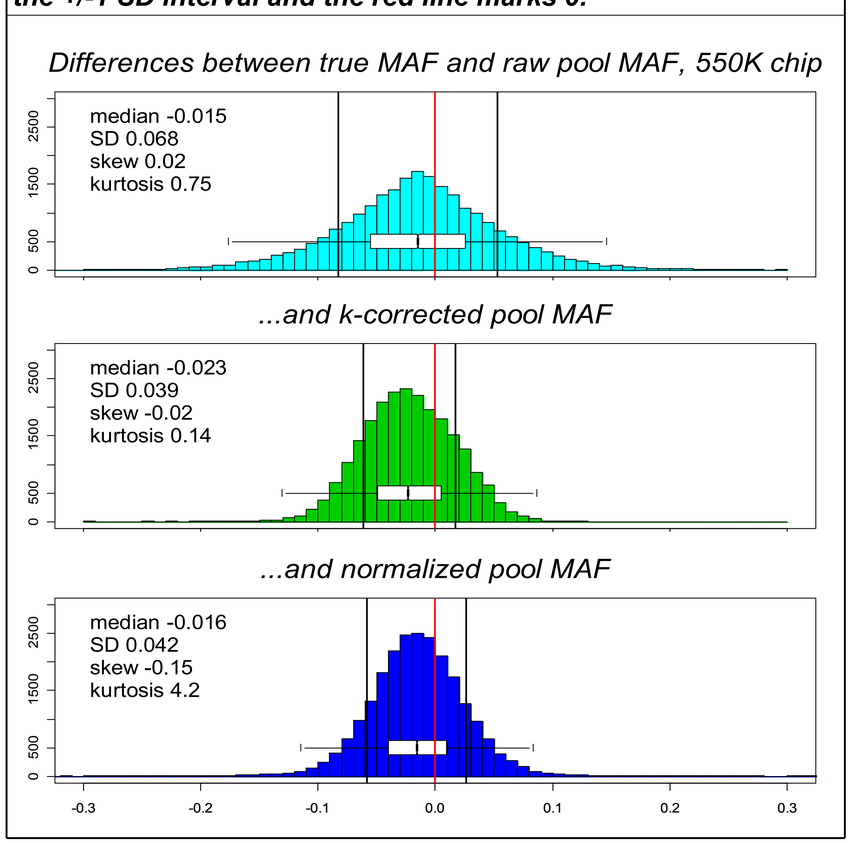
Correlation Matrices

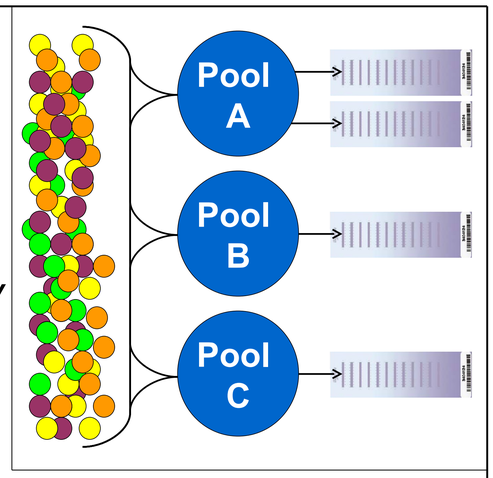
Scatter plots comparing raw and normalized pool frequencies to known population allele frequency. Upper panels: font-scaled Pearson's r, with red stars indicating significance (all $p < 10^{-16}$). Lower panels: loess-fitted scatterplots.



Difference Distributions

Histograms of (MAF_{88} - MAF_{pool}) for raw and normalized frequencies, overlaid with quartile plots. Black lines illustrate the +/-1 SD interval and the red line marks 0.





Pooling schematic. Three separate pools were made to assay the variance introduced by the pooling process; one pool was sampled with two chips to assay chip variance. Correlations between all pools were highly significant with r=0.997.

Formulae for raw and normalized relative allele signals. Parameters are calculated individually for each SNP.

RAF_{raw}	*raw image data, no correction*	$\frac{X_{raw}}{X_{raw} + Y_{raw}}$
RAF_k	k_{SNP}= avg(X_{raw}/Y_{raw}) over AB loci *Corrects for deviation of heterozygote from 50% A*	$\frac{X_{raw}}{X_{raw} + kY_{raw}}$
RAF_n	AA_{avg} = avg(RAF_{raw}) over AA loci *Normalizes homozygotes: AA to 1, BB to 0*	$\frac{RAF_k - BB_{avg}}{AA_{avg}}$

***k*-correction:** Hoogendoorn et al, *Hum Genet* (2000) 107:488-493
normalization: Craig et al, *BMC Genomics* (2005) 6:138

SUMMARY AND RECOMMENDATIONS

- ❑Simple normalization methods applied to the HumanHap550 chip produce data **highly correlated with data from individual genotyping**.
- ❑The HumanHap550 showed **greater precision and accuracy than Human-1** (population vs normalized r=0.90, median differences -0.036 (raw), -0.045 (k-corrected), and -0.026(normalized).
- ❑Additional improvements may be possible with **chip-specific normalization methods** proposed by Illumina (personal communication).

We recommend:

9. Use of **more than one DNA quantification method**. The greatest source of variance in the pooling process is DNA quality and pool construction.
10. Use of **test datasets** for verification of accuracy of pooling technique.
11. Conceptualization of pooling/whole-genome association as **one of several tools to prioritize SNPs** for individual genotyping

Research supported by the Intramural Research Programs of the