NGS Pretesting and QC Using Illumina Infinium Arrays

J. Romm, H. Ling, P. Zhang, I. McMullen, P. Witmer, K. Doheny

Center for Inherited Disease Research (CIDR), IGM, JHU-SOM, Baltimore, MD

Introduction

The Center for Inherited Disease Research (CIDR) provides high quality next-generation sequencing (NGS), genotyping and statistical genetics consultation to investigators working to discover genes that contribute to disease. To maintain a high-quality service at a cost effective price, CIDR routinely performs sample pretesting to evaluate DNA quality, detect contamination/sample mixtures, confirm biological gender assignments, and check for Mendelian inconsistencies and inbreeding prior to initiating laboratory processes. In addition, this process creates a genetic barcode for each samples which assures proper sample and data tracking and allows for an independent measure of NGS variant calls quality.

Methods

For NGS pretesting, CIDR uses a dense Illumina Infinium array, which permits extended analyses for 1st and 2nd degree relationships, unexpected relationships, ethnicity and large chromosomal anomalies. In addition, we calculate concordance and sensitivity to call heterozygotes, providing a critical quality control measure for NGS data. We routinely used the Illumina CoreExome array, but have recently switched to the newly available Illumina QC array, which allows for everything mentioned and provides a cost savings.

Workflow

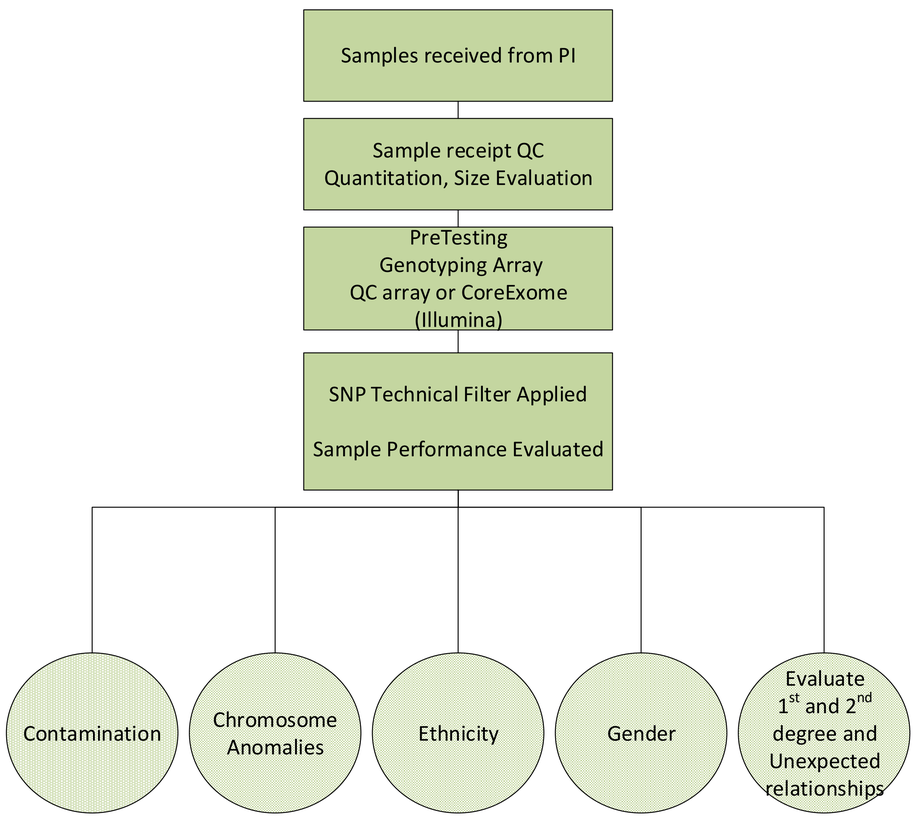

Figure 1
Contamination is detected using VerifyIDintensity[1]
Chromosome Anomalies are detected using GWASTools[2]
Ethnicity is evaluated using SmartPCA[3]
Gender is evaluated by plotting mean R Chr X x ChrY
Relationships are evaluated using KING [4]

Results

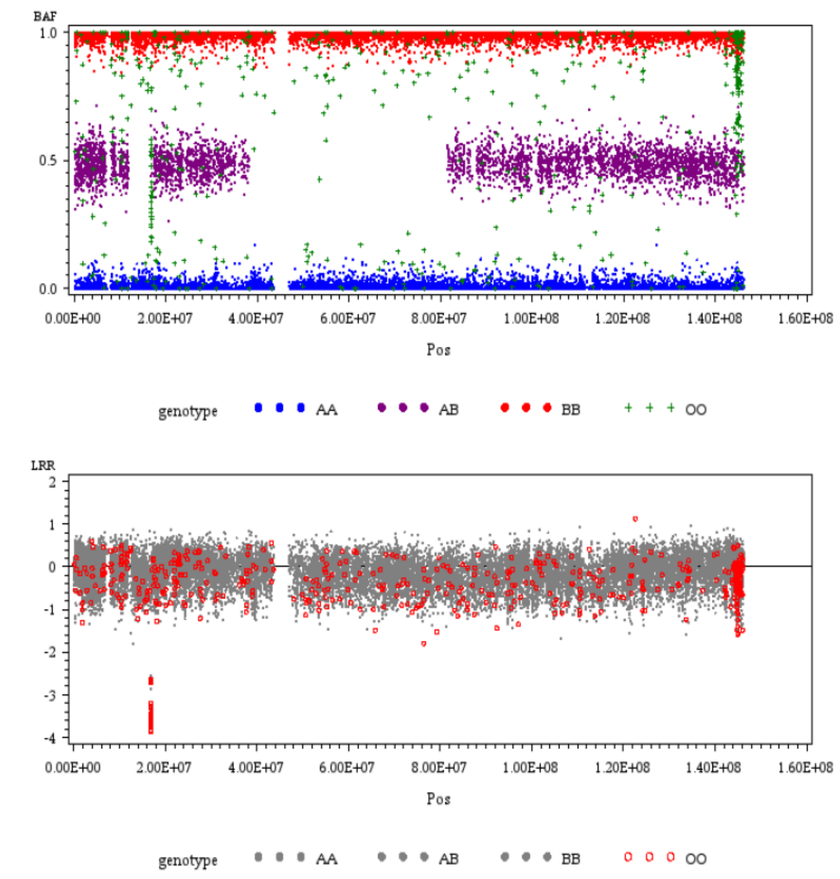

Figure 2
Chr 8 interstitial homozygous deletion detected with the Illumina CoreExome Array.

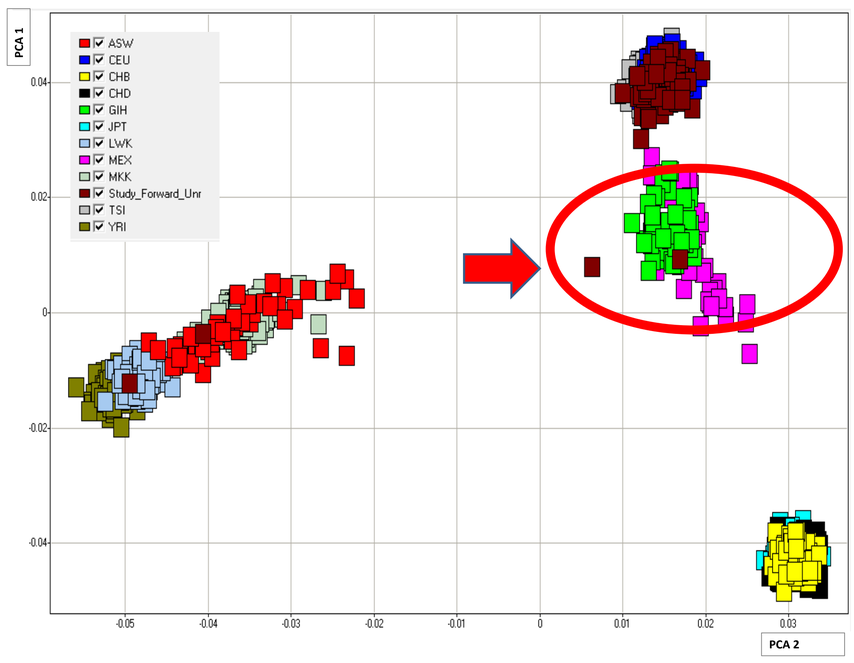

Figure 3
Samples that were annotated by the sender as one ethnicity that are genotyping as another. Detected with the Illumina QC array.

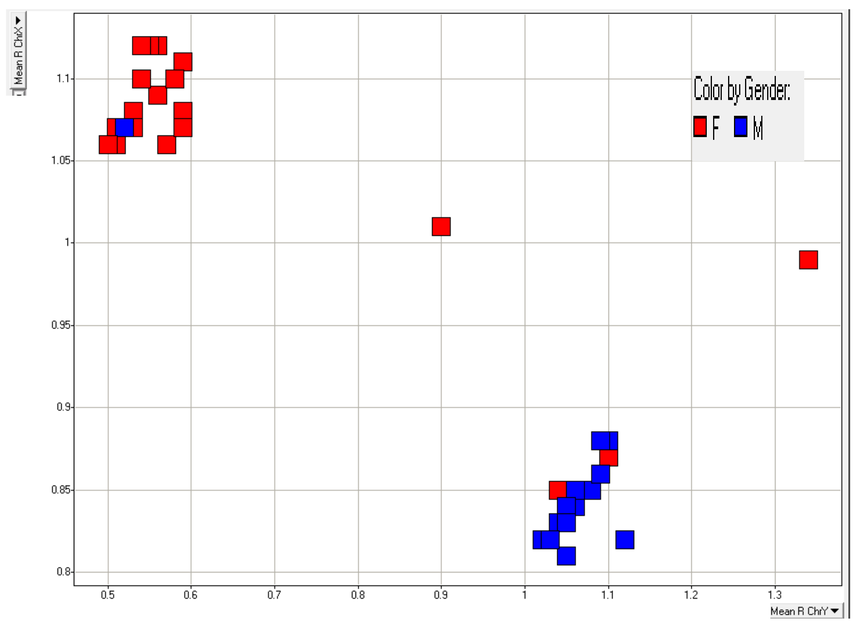

Figure 4
Several samples annotated as the incorrect gender. Detected with the Illumina QC array.

Conclusion

Using a dense Illumina Infinium array is effective for identification of various sample issues before a project begins NGS. This is a cost-saving measure for the NGS project – better to identify sample issues up front cost-effectively than the costly investment of sequencing samples with unresolvable issues. In addition, the array genotype calls allows for an measure of NGS variant call quality (concordance) and completeness (sensitivity to heterozygote array calls).

References:
[1] G. Jun, et. al. (2012) Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. American Journal of Human Genetics 91(5):839-848
[2] Gogarten SM, et. al. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics, 28(24), pp. 3329-3331.
[3] Price AL, et. al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904-9. Epub 2006 Jul 23.
[4] Manichaikul A, et. al (2010). Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867-73. doi: 10.1093/bioinformatics/btq559.