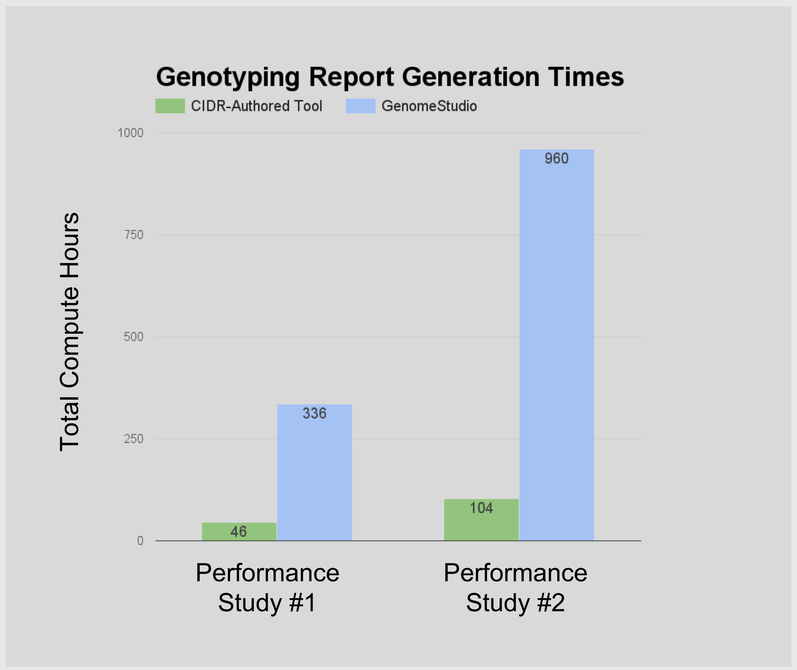
Rapid Generation of Illumina Infinium Genotyping Release Data

SML Griffith*, H Ling*, K Hetrick*, J Romm*, EW Pugh*, IA McMullen*, DR Leary**, BR Myers*, MZ Mawhinney*, ME Hurley*, AB Robinson*, L Watkins, Jr.*, KF Doheny*

*Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD.
**New York Genome Center, New York, NY.

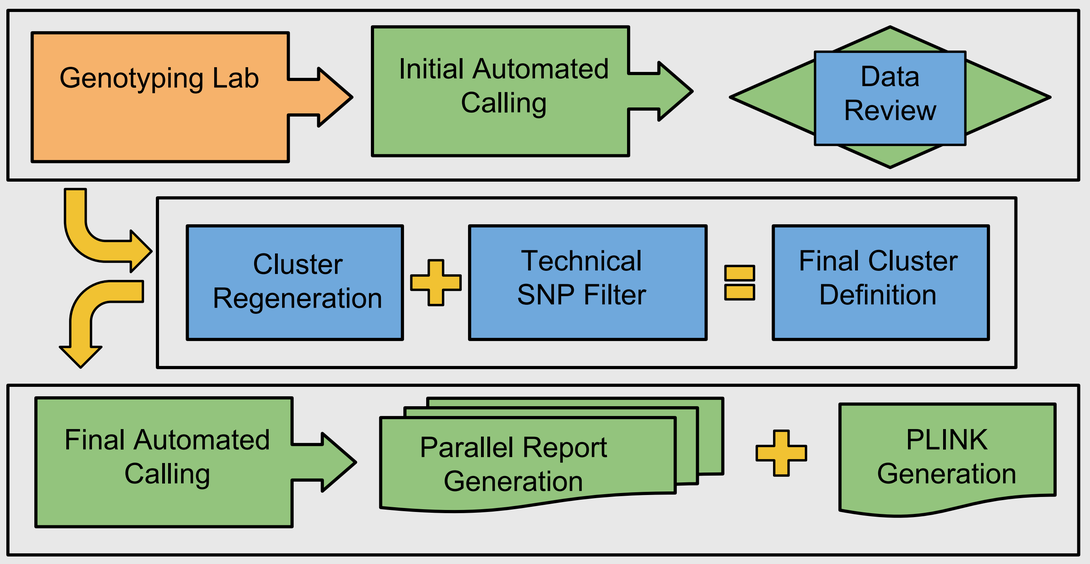
The Problem

When using Illumina's Infinium range of genotyping arrays, Illumina's GenomeStudio is the typical way to generate final genotyping data. However, for large numbers of genotyped samples, generating these data via this method is time consuming and resource intensive. We have developed software for generating this data in the form of both individual reports and project-wide PLINK files.



Data Generation Process

- An automated genotype calling pipeline is employed to both initially call genotypes and to generate final genotypes prior to release. This process results in genotype call files (.gtc). This pipeline, described in (2), has been updated with a faster .gtc file parser and greater parallelism.
- Final cluster definitions are generated via GenomeStudio (3).
- The .gtc files associated with each sample are parsed on the fly and used to generate genotype files, which also contain normalized raw intensity data and several metrics, including Log R Ratio and B Allele Frequency.
- Genotyping report generation is embarrassingly parallel, and can be done across several machines for rapid results.
- The same .gtc files are parsed to generate PLINK (1) formatted files containing identical genotypes to those in the genotyping reports.



Performance Study #1

- ~7,500 samples on the Omni5Exome+Custom array.
- Blade servers had 192 GB RAM and 32 processing cores.
- Genotyping report generation took 23 hours across two blades. PLINK generation took 7 hours on one blade.
- GenomeStudio report generation projected to take two weeks on two blades at fastest observed speed.

Performance Study #2

- ~50,000 Samples on the MEGA Array.
- Blade servers had 192 GB RAM and 32 processing cores.
- Genotyping report generation took 26 hours across four blades. PLINK generation took 22 hours on one blade.
- GenomeStudio report generation projected to take ten days on four blades at fastest observed speed.

References

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

2. The CIDR AutoCall Pipeline: An automated analysis pipeline for Illumina® Infinium Products. M. W. Barnhart, K. Hetrick, C.W. Bark, D.R. Leary, G. Lowe, E. Hsu, J.L. Goldstein, K.F. Doheny, L. Watkins, Jr. Center for Inherited Disease Research, Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD. Presented at ASHG 2007.

3. Microarray Data Analysis Workflows. Illumina Technical Note: DNA Analysis.

Future Work

- More complete GUI and command line interfaces for both the PLINK and genotyping report generators.
- Comparison between control samples in the PLINK data to HapMap and 1000 genomes data.
- Reports containing the call rates across both SNPs and individuals.

CIDR
Center for Inherited Disease Research
Johns Hopkins University

The Center for Inherited Disease Research (CIDR) provides high quality genotyping and sequencing services and statistical genetics consultation to investigators working to discover genes that contribute to disease.

JOHNS HOPKINS
UNIVERSITY & MEDICINE