Comparison of Genotyping products

Introduction
Over the past two decades, large population genotyping studies have delivered a better understanding of disease, an expanded pipeline of therapeutic targets and new diagnostic/predictive tests to improve health and wellness. These advancements are a result of innovation in genotyping technology. From real-time PCR and TaqMan assays through microarrays and sequencing, the drive for more information at a reduced cost has been consistent. Today’s most comprehensive microarrays can provide over 5 million single nucleotide variants (SNV) at a cost of approximately $0.000002 per genotype ($450 / sample). We are presenting a Next-Generation Genotyping (NGG) approach and a novel DNA library prep product that enables the sequence-based genotyping of more than 37 million genetic markers at a cost of less than $0.000002 / genotype (less than $80/sample). In addition to the dramatic increase in marker density that enables the sequence-based genotyping of more than 37 million genetic markers at a cost of less than $0.000002 / genotype (less than $80/sample). In addition to the dramatic increase in marker density throughout the genome, NGG avoids the ascertainment bias associated with static content microarrays. NGG is enabled by high-throughput sequencers with the main application criteria being low error rates and high read counts. As the widespread adoption of NGS platforms for applications like NIPT has increased, achieving high quality results at low cost is simply a matter of obtaining just enough evidence and high read counts. As the widespread adoption of NGS platforms for applications like NIPT has increased, achieving high quality results at low cost is simply a matter of obtaining just enough evidence (defined as number of reads) to make high confidence calls on known targets. Sequencer reads are mapped to a reference genome, a haplotype is identified, and genotypes are assigned. Twenty plus years of sequencing and genotyping have made this approach robust today. In this study, we evaluate the minimum performance criteria required for genotyping to high precision and sensitivity, using the iGenomX Riptide high throughput DNA library preparation product.

Method
960 genomic DNA samples from the 1,000 genomes project (1KG) were obtained from the Coriell Institute (New Jersey). 50ng of input per sample was prepared using Riptide and sequenced on a single Illumina NovaSeq 54 flow cell by Macrogen (Maryland). FastQ files were sent to Gencove (New York) for genotype calling and VCF generation. VCFs were compared to public genotypes (Illumina GSA microarray) at NIST high confidence loci to analyze precision, sensitivity and accuracy across platforms. Processing time is estimated as less than 5 hours for library prep (with automation), 40 hours for sequencing and 24 hours for genotyping and VCF generation.

Results
The first 960 sample set generated about 2.8Gbp of data. Per sample read counts averaged 20 million across 960 samples with a range of a few million to over 100 million. More than 90% of the samples achieved a minimum read count less than 1 standard deviation below the mean. The Gencove pipeline can make calls on 38 million bi-allelic variants for each sample regardless of read count or coverage. Comparing Gencove VCF calls of these samples with GSA array genotypes from public sources allows us to determine the relationship between read count and coverage with precision, sensitivity and accuracy (see tables below). Of the 934 unique samples, 840 had GSA array data available through the 1KG project (phase 3 release). Published specifications for the NovaSeq 54 flow cell shows the sequencer will generate 16-20 billion reads, meaning that 384-480 individually barcoded samples per flow cell will generate between 35 million and 52 million average reads per sample with optimal performance. Using a published price for 54 flow cell from Texas A&M University of approximately $25,000, the cost per sample for sequencing amounts to $52-$65 (384 -480 samples per flow cell).

Advantages
• >10x more variants than microarrays
• No fixed content bias
• Improved signals for GWAS and polygenic risk scores