Describe the sequencing platform:

Describe the QC platform:

Quality control of sequence data falls under two broad categories: raw sequence data QC, and analysis QC. Raw sequence data quality control begins with data provided by the sequencing center, and aims to produce the highest quality set of variants while maintaining a set that can be used in any down stream analyses. Analysis QC aims to further subset genotypes and samples to reduce spurious signals, and is tailored to a specific statistical procedure.

We have received reads aligned to a reference for all 253 samples, as well as genotypes called with GATK's Unified Genotyper. Additionally, the sequencing center provided genotypes derived from an Illumina Exome chip generated as a part of their standard sequencing pipeline.  
  
Sequence QC proceeds by recalling genotypes using the aligned PGRNseq reads using three independent algorithms. We then characterize them by genotype concordance, observed Transition-Transversion mutation ratio, QUAL scores, read depths, genotype missingess, and rates of rediscovery for previously identified variant sites. Using these metrics, we have identified the intersection of these three genotype sets as the highest quality. This result is concordant with previous experience.

Finally, we use the Exome chip data as a reference set of genotypes. We found a high level of concordance among genotypes derived from the Exome chip data, and the consensus PGRNseq data.

Given a high quality set of input data, we then move to analysis QC. We identify the statistical tests of interest, and potential confounding factors for each of these. The first round of analysis aims to find associations between genotype and phenotype, and can be split into two categories: single marker, and group-based tests.

Potential sources of confounding are similar for both tests: ancestral genetic relatedness, high rates of genotype missingness, small sample size, and clinically relevant covariates. The major difference between the two is statistical power for rare alleles. Single marker tests cannot be used for alleles with low frequency, and group-based tests are designed specifically to test the aggregate information of rare alleles.

We remove ancestral confounding in both single marker and gene-based tests. We use the Exome chip genotypes to cluster samples based on genetic similarity, and identify a set of 161 European samples that form a genetically homogenous population. This is achieved by subsetting the markers to uncorrelated sites, performing Principal Components Analysis on the resulting genotype matrix, and clustering of the samples projected onto the major axes of variantion using a mixture model. This cluster is tagged by samples with self-reported European ancestry, and is concordant with similar PCA plots found in the literature. All first pass analyses are restricted to this subset of samples.

Genotypes undergo further QC based on the statistical tests used. For single marker tests this involves removing the following: SNPs with high rates of missing genotypes, SNPs with significant departures from Hardy Weinberg Equilibrium, SNPs with minor allele frequency less than 5%. The same thresholds are used prior to conducting gene-based tests, except only singleton alleles are excluded.

We adjust each set of tests using available covariates: dosing regimen, sex, and site of ascertainment. Finally, given results of a single marker association scan, we assess whether or not genetic relatedness is still inflating results by computing the Genomic Inflation Factor.

To accurately identify variants from the sequencing data, a consensus approach will be employed. Base calling and read alignment to the human genome was performed at the sequencing center at the University of Washington.. Aligned reads will be passed to a custom platform that performs variant calling using three common methods: GATK Unified Genotyper (PMID: 20644199, 21478889), FreeBayes (Garrison, Marth submitted), and Atlas2 (PMID:22239737). The platform identifies consensus variants as those found using all three calling methods, reducing the number of false positive variant calls that may be identified by any single calling algorithm but increasing false negative calls by the conservative nature of the consensus approach.