Interim Phase 1 Variant quality score recalibrator project consensus process

A high-level view of the process used to build the project consensus is as follows:

1. First pool together all SNP calls made by any center. For the phase1 calls this list was approximately 46.3 million SNPs.
2. Re-call at all SNP sites using the GATK's Unified Genotyper with project BAM files that have been fully indel realigned at the population level. The unified genotyper adds several important statistics, calculated on a per-site basis, which will be used by the variant quality score recalibrator. These statistics (explained with more detail below) are: QualByDepth, HaplotypeScore, MappingQualityRankSum, ReadPositionRankSum, and HomopolymerRunLength.
3. Apply the Variant Quality Score Recalibrator genome-wide to train a Gaussian mixture model over the five error covariates listed above. Each input variant is assigned a VQSLOD score which is the log odds ratio between the probability that the SNP is true or false given the model. In addition to using the error covariate statistics the model incorporates a prior probability of being a true variant which is based on the number of callsets that the original variant is found in. This captures the intuition that variants called independently by multiple callers are more likely to be real. The prior used here is Q<10X>, where X is number of callsets the variants was found in.
4. Partition the list of variants into those that are PASSing and those which are filtered out by choosing the VQSLOD value which gives 99.8% sensitivity to the accessible HapMap3 variants.

http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper

http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration

Variant quality score recalibration:

Given a set of putative variants along with SNP error covariate annotations, variant quality score recalibration employs a variational Bayes Gaussian mixture model to estimate the probability that each variant is a true polymorphism in the samples rather than a sequencer, alignment or data processing artifact. The set of variants are treated as an *n*-dimensional point cloud in which each variant is positioned by its covariate annotation vector. A mixture of Gaussians is fit to a set of likely true variants, here used the variants already present in HapMap3 as well as those variants which were found to be polymorphic by the Omni chip. Following training, this mixture model is used to estimate the probability of each variant call being true, capturing the intuition that variants with similar characteristics as previously known variants are likely to be real, whereas those with unusual characteristics are more likely to be machine or data processing artifacts.

The error statistics:
The following statistics are calculated on a per-site basis and are used by the variant quality score recalibrator to model error.

- **QualByDepth**. This is the variant quality score (the confidence assigned by the unified genotyper in the site being a variant site) divided by the number of reads in the pileup. This statistic captures the intuition that as sequencing depth increases the confidence in the site should also increase if it is a real variant.
- **HaplotypeScore**. This is a measure for how well the data from a 10 base window around the SNP can be explained by at most two haplotypes. In the case of mismapped reads, the pattern of mismatches around the SNP would seem to imply many more than two haplotypes and is indicative of error.
- **MappingQualityRankSum**. This is a Wilcoxon rank sum test which tests the hypothesis that the reads carrying the alternate base have a consistently lower mapping quality than the reads with the reference base.
- **ReadPositionRankSum**. This is a Wilcoxon rank sum test which tests the hypothesis that the alternate base is consistently found more often at the beginning or ending of the read instead of randomly distributed throughout. A bias would indicate that the reads are mismapped.
- **HomopolymerRunLength**. This is the length of the longest consecutive homopolymer run made from the alternate allele and the surrounding reference bases.

Citation for the GATK methods (both Unified Genotyper and Variant Quality Score Recalibrator) used here:

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl., C., Philippakis, A., del Angel, G., Rivas, M.A, Hanna, M., McKenna, A., Fennell, T. Kernytsky, A., Sivachenko, A, Cibulskis, K., Gabriel, S., Altshuler, D. and Daly, M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011 Apr; 43(5):491-498.