Quality control of variant calling from exome sequence data in the UK10K project

Lucy Crooks[1], Olli Pietiläinen[1,2], Karola Rehnström[1], Jeffrey Barrett[1] and Aarno Palotie[1]
on behalf of the UK10K Consortium (http://www.uk10k.org/consortium.html)

[1]Wellcome Trust Sanger Institute, Cambridge, UK; [2]National Institute for Health and Welfare, Helsinki, Finland. lc8@sanger.uk.ac

UK10K

The UK10K project is a study of sequence data for 10,000 individuals from the UK and Finland. In total 6,000 individuals from three disease groups, schizophrenia and autism, obesity, and a set of eight rare diseases, will be exome sequenced. There are 4,000 control individuals that have been whole-genome sequenced. With this large sample size, we will improve our understanding of rare variants and hope to identify new genetic factors affecting the investigated diseases.

Sequencing and calling

Exons were targeted with baits that covered 52 Mb. Paired end reads of 75 bp were sequenced by Illumina HiSeq 2000. Average depth was 70x. Variants were called by SAMtools on the combined exome samples (4,060 in the current release). Calling was restricted to the baits ± 100 bp. SNV sites were filtered by VQSR from GATK at a truth sensitivity of 99.5% for HapMap 3.3 sites.

http://samtools.sourceforge.net/ http://www.broadinstitute.org/gatk/

Purpose

Establish an effective method for filtering genotype calls in each sample

Genotypes are reported for each sample at all sites that pass VQSR. Several quality control metrics are available at the sample level, and the aim is to evaluate which best indicate accurate calls and determine an appropriate filter threshold.

Methods

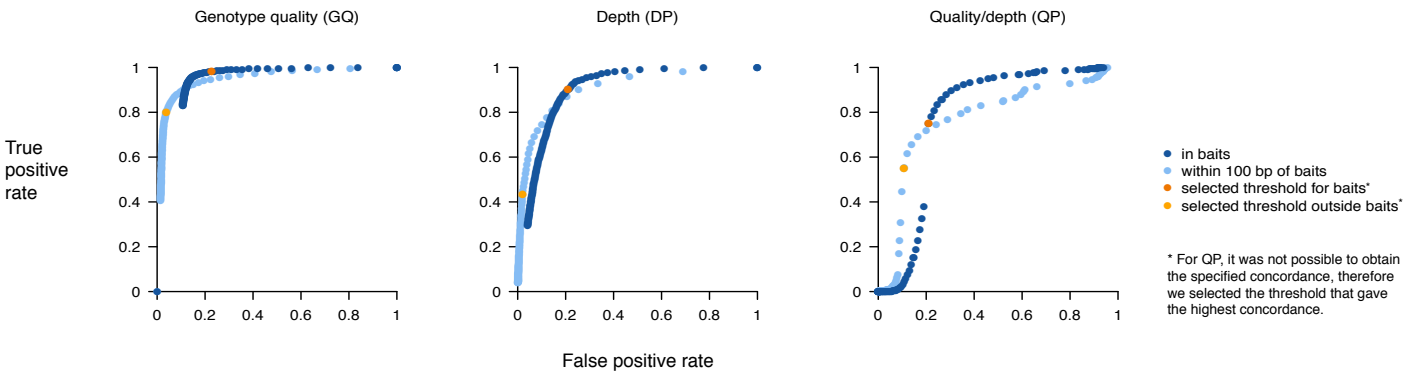
We compared genotypes from the Illumina 660W BeadChip (GWAS genotypes) with the sequencing calls for 257 Finnish individuals. There were 13,303 shared SNPs in the baits and 7,523 flanking the baits (autosomes only). True positives were defined as cases where the GWAS and sequence genotypes matched and false positives as cases where they were different. ROC curves were produced by measuring the true and false positive rates for filtering at different thresholds. We selected the threshold that increased the concordance between the sequence and GWAS genotypes to 99.9%.

Genotype quality likelihood of the called genotype divided by the sum of the likelihoods for all possible genotypes (phred scaled)

Depth number of high quality reads for the site

Quality/depth if genotype is homozygous variant, sum of PL likelihoods that genotype is homozygous variant and heterozygous divided by the PL likelihood that genotype is homozygous reference, divided by the depth. Else, sum of PL likelihoods that genotype is homozygous reference and heterozygous divided by the PL likelihood that genotype is homozygous variant, divided by the depth. (Phred scaled)

ROC curves



Genotype quality was the best discriminator

		Baits	± 100 bp of baits
	Exclusion filter	GQ < 20	GQ < 30
Before filtering	% concordance	99.6	98.0
	% non ref concordance	99.2	96.1
After filtering	% concordance	99.9	99.9
	% non ref concordance	99.8	99.8
	True positive rate	98.2	80.1
	% genotypes removed	2.3	23

Conclusions

The initial concordance of the sequence and GWAS genotypes was very high in the baits and could be further improved by filtering on GQ, whilst retaining a high proportion of accurate genotypes. Outside the baits, the initial concordance was lower and filtering to the same level of concordance discarded a substantial proportion of accurate genotypes. Depth and quality/depth were less effective filters than GQ.