**Workshop on Establishing a Central Resource of Data from Genome Sequencing Projects**

# Sequence Data Processing

Gonçalo Abecasis, Lisa Brooks

Sequence data processing involves mapping the sequence reads and making variant calls. Calling SNPs is relatively robust for most of the genome, but methods for calling indels and structural variants are still evolving rapidly.

Some questions can be addressed without re-analyzing the raw sequence data generated by each project. For example, meta-analysis of sequencing study results can be done even when the studies use different algorithms for calling variants; these analyses may lose some power compared to joint analyses of all available data, but generally they will not produce false positive signals. Other analyses will benefit strongly from joint processing of the data. For example, studies looking at associations between rare variants and disease will benefit from using very large sets of control samples, and appropriate analysis requires uniform data processing. Analyzing complex variants (for example, looking for specific breakpoints of copy number variants) or distinguishing true variants from sequencing artifacts (for example, evaluating support for the two alternate alleles in heterozygotes) would benefit greatly from examining raw sequence data across many projects.

Here, we focus on the types of central data processing, the challenges, and the benefits.

1. Using variants and genotypes called by each project

   The simplest combined analyses would rely on variants and genotypes called by each project. We should expect that the rates of false-positive variants and the ability to discover and genotype different types of variants will vary greatly by project, even when the underlying sequence data are similar in quality.

   Nevertheless, some types of analyses are relatively robust to these underlying differences. Meta-analysis of studies of the same phenotype could provide a more powerful view of the relationship between genetic variation and a trait than analysis of any single study.

   The main hurdles for these analyses are data access and consistent use of file formats for genotype data and for phenotypes. There are now standard formats for sharing genotypes and variants from sequencing projects. Although there are standards and ontologies for phenotypic data, they are not used consistently, are cumbersome, and are not amenable to analysis.

2. Each project re-calling or filtering variants in standard ways

In principle, analyses would benefit if every sequenced sample were analyzed consistently with the same data processing pipeline. Although this would be helpful, it probably would not be sufficient to allow joint analysis of samples sequenced at different sites, because, for example, if each study filters variants independently, large studies will have greater power and better ability to remove artifacts than small studies. This option would provide only an incremental improvement over the first strategy.

3. Re-analysis of all sequence read data and re-calling all variants with a standard pipeline

   This is potentially the most challenging option to implement; however, it potentially has great value. It would enable studies to combine sequence data from many sources, potentially increasing the power of rare variant studies by enabling the use of very large control sets. It would allow analysis of previously sequenced samples to improve in accuracy and completeness, as they incorporate new analysis strategies, models of individual variants refined on large numbers of individuals, and high power to identify artifacts of sequencing based on using very large numbers of samples. For example, many methods call a variant only when it has been seen twice; large combined sample sets thus allow many more rare variants to be called than could be called using each sample set alone.

   Data processing pipelines can now handle tens of thousands of sequenced samples. This approach becomes more practical if we focus on large subsets of data with common characteristics (sequenced at a relatively high depth, with widely used technologies, and with relatively long read lengths) and exclude smaller subsets of data (sequenced with unusual technologies or short read lengths).