

Exploring 1000 Genomes with Bioconductor: Dense SNP Imputation and Sequence-based Expression Genetics

VJ Carey, PhD, Channing Laboratory, Harvard Medical School

September 26, 2010

Contents

1 Introduction

The 1000 genomes (1KG) project aims to develop a “deep catalog of human variation”. By employing new techniques for determining the full sequence of individuals’ genomes and publishing results at various levels of resolution, the project promises significant data resources for use in the enrichment of theory and methods of statistical genetics. In this paper two applications of resources provided through 1KG are discussed in the context of Bioconductor’s R-based facilities for analysis of genome-scale data. First, we discuss methods and performance of imputation of genotypes from relatively sparse SNP panels to the full 1KG panels for CEPH populations. Second, we examine methods for analysis of the genetics of gene expression when next-generation sequencing is used to characterize both DNA variation and gene expression, the latter via the family of methods known as “RNA-seq”.

2 SNP imputation to the 1KG panel

2.1 Concept and prevalent approaches

It is widely accepted that genetic association analyses can be enhanced when unobserved genetic markers are suitably imputed for individuals who have only been sparsely genotyped (??). Imputation schemes have been systematically compared for use in applications, but no clearly dominant method has been identified, using metrics related to accuracy for array-based genotypes compared across array versions (e.g., imputation from the Affymetrix genomewide 5.0 SNP panel to Affymetrix genomewide 6.0 compared to 6.0 calls) or enhancement of power for association studies (?). The imputation