Overlapping Pools for High Throughput Targeted Resequencing

Snehit Prabhu* and Itsik Pe'er*

Department of Computer Science
Columbia University, New York
{snehitp,itsik}@cs.columbia.edu

Abstract. Resequencing genomic DNA from pools of individuals is an effective strategy to detect new variants in targeted regions and compare them between cases and controls. There are numerous ways to assign individuals to the pools on which they are to be sequenced. The naïve, disjoint pooling scheme (many individuals to one pool) in predominant use today, offers insight into allele frequencies, but does not offer the identity of an allele carrier. We present a framework for overlapping pool design, where each individual sample is resequenced in several pools (many individuals to many pools). Upon discovering a variant, the set of pools where this variant is observed reveals the identity of its carrier. We formalize the mathematical framework for such pool designs, and list the requirements from such designs. Next, we build on the theory of error-correcting codes to design arrangements that overcome pitfalls of pooled sequencing. Specifically, three practical concerns of low coverage sequencing are investigated: (1) False positives due to errors introduced during amplification and sequencing; (2) False negatives due to undersampling particular alleles aggravated by non-uniform coverage; and consequently (3) Ambiguous identification of individual carriers in the presence of errors. We show that in practical parameters of resequencing studies, our designs guarantee high probability of unambiguous singleton carrier identification, while maintaining the features of naïve pools in terms of sensitivity, specificity, and the ability to estimate allele frequencies. We demonstrate the ability of our designs by extracting rare variations on pooled short read data of 12 individuals from the 1000 Genome Pilot 3 project.

* S.P. was supported in part by NSF CCF 0829882 ; I.P. was supported in part by NIH 5 U54 CA121852.

S. Batzoglou (Ed.): RECOMB 2009, LNCS 5541, p. 270, 2009.