Optimization of Bait Selection for Sequence Capture Arrays

Kevin Silverstein[1], Kenneth Beckman[2], Matthew Bower[3], Kari Bunjer[2], Matthew Schomaker[4], Teresa Kemmer[4], Lisa Schimmenti[5], Sophia Yohe[6], Randolph Peterson[6], Amy Karger[6], Monika Roychowdhury[6] and Bharat Thyagarajan[6]

[1]Biostatistics and Bioinformatics, Masonic Cancer Center, University of Minnesota, Minneapolis, MN 55455; email: silve023@umn.edu

[2]Biomedical Genomics Center, University of Minnesota, Minneapolis, MN 55455

[3]Division of Genetics and Metabolism, University of Minnesota Medical Center-Fairview, Minneapolis, MN 55455

[4]Molecular Diagnostics Laboratory, University of Minnesota Medical Center-Fairview, Minneapolis, MN 55455

[5]Departments of Pediatrics, Ophthalmology, Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN 55455

[6]Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55455

Next-Generation Sequencing (NGS) technologies will fundamentally alter how clinicians test patients for disease-associated genetic variants. While sequencing costs are falling at a rapid pace, the goal of an affordable full-genome sequence is not likely to be realized in the immediate future. Sequence capture arrays, such as the Agilent SureSelect platform, provide an interim solution allowing NGS technologies to be adapted for specific clinical testing applications. Capture arrays enrich for targeted regions of the genome via hybridization to designed long oligo "baits" for subsequent sequencing by NGS technologies. This offers a cost-effective means for identifying disease-associated variants in genes of interest. To date, oligo bait design has been rather simplistic. Typically, baits are evenly tiled across the region of interest. The final cost of a capture array can be heavily influenced by the total number of bases covered in the reference genome and the number of baits in the design. For large arrays involving many small target exons, the overhang of baits at both ends of these exons can be substantial, nearly doubling the cost of design. We have sought to systematically explore the effects of a variety of layout parameters on the depth of coverage at the boundaries and throughout the targeted exons among ~600 genes. Specifically, we have matched target exons for length, GC content, and problematic bait count (defined as an oligo with GC > 65% or deltaG of folding < 2 SEM below the mean). Within these matched exons (each having N=500), we have systematically compared the default layout against an expanded, compressed, and bait-reduced layout. In another set of matched-exon experiments (each having N=50), we have systematically tested the replacement of problematic baits with nearby improved ones, or enriched their number by 2X or 3X in an attempt to recover difficult-to-capture regions. A comparison of exon mean depth of coverage and fraction of sufficiently covered bases (>10X/20X/30X) across our experimental variables helps us to maximize desired target coverage while minimizing design costs.