Accurate Detection and Genotyping of SNPs utilizing Population Sequencing Data: Supplemental Text

Vikas Bansal[1*], Olivier Harismendy[1,2], Ryan Tewhey[1], Sarah S. Murray[1], Nicholas J. Schork[1], Eric J. Topol[1], Kelly A. Frazer[1,2]

Application of SNIP-Seq to 1000 Genomes data

On the suggestion of a reviewer, we applied SNIP-Seq to population sequence data from the 1000 Genomes project. The pilot project 3 of the 1000 Genomes project aims to sequence the coding regions of 1000 genes in $\sim$ 1000 individuals at roughly 20x coverage. The targeted regions (defined in the file ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/technical/reference/P3_consensus_exonic_targets.bed) represent a total of $\sim$ 1.4 megabase of sequence across the human genome. To limit the amount of downloaded sequence data, we limited the analysis to reads that aligned to chromosome 1. The total sequence length of the targeted regions on chromosome 1 was $\sim$ 150 kilobases. Aligned sequence data in the form of .sam files was downloaded from the 1000 genomes ftp website (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/) using the Samtools package (Li et al., 2009). For evaluating SNIP-Seq, we considered 120 samples that were sequenced using the Illumina platform using 76 bp reads and for which the alignments were generated using the MAQ alignment method.

For running SNIP-Seq, we generated pileup files for each of the 120 samples (restricted to only the targeted regions in the .bed file) from the SAM files using a simple python script. SNIP-Seq was run on the pileup files with the default parameters allowing each read to have upto 5 mismatches for it be used for SNP calling. SNIP-Seq identified 626 SNPs across the $\sim$ 150 kilobases of targeted sequence in the population.

List of 120 samples that were utilized for SNP calling:

NA11830,NA11918,NA11930,NA12005,NA12154,NA12748,NA12829,NA12830,NA12843,NA12889,NA17978,NA18000,

NA18113,NA18123,NA18126,NA18145,NA18164,NA18488,NA18498,NA18519,NA18520,NA18632,NA18633,NA18635,

NA18636,NA18637,NA18641,NA18642,NA18643,NA18647,NA18669,NA18671,NA18674,NA18679,NA18683,NA18684,

NA18685,NA18687,NA18689,NA18690,NA18694,NA18695,NA18696,NA18698,NA18699,NA18701,NA18704,NA18707,

NA18708,NA18745,NA18747,NA18748,NA18749,NA18757,NA18853,NA18867,NA18868,NA18910,NA18950,NA18960,

NA18982,NA18983,NA18985,NA18988,NA18989,NA18999,NA19000,NA19003,NA19006,NA19011,NA19012,NA19054,

NA19055,NA19056,NA19057,NA19058,NA19059,NA19060,NA19062,NA19063,NA19065,NA19066,NA19067,NA19068,
NA19089,NA19090,NA19091,NA19092,NA19130,NA19213,NA19235,NA19236,NA19247,NA19248,NA19546,NA19550,
NA19551,NA19552,NA19553,NA19554,NA19555,NA19556,NA19558,NA19559,NA19560,NA19561,NA19562,NA19563,
NA19564,NA19565,NA19566,NA19568,NA19569,NA19572,NA19573,NA19574,NA20511,NA20587,NA20763,NA20764

References

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**:2078–2079.