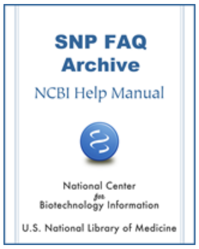
NIH U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: SNP FAQ Archive [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Double Hit SNP Computation. 2005 Jul 23 [Updated 2014 Feb 18].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



SNP FAQ Archive
NCBI Help Manual
National Center for Biotechnology Information
U.S. National Library of Medicine

Double Hit SNP Computation

Created: July 23, 2005; Updated: February 18, 2014.

What criteria does dbSNP use to determine double-hit SNPs independently of Dr. Jim Mullikin's algorithm?

I have not made the double-hit, two-allele computation for some time now. Currently, we rely exclusively on Dr. Mullikin's data AFAIK. As for my double-hit computation, I made the initial calculation of double-hit SNPs based on submitter-supplied clone accessions. If we can establish that two different submitters working with different clone libraries had independently identified each allele, we confirmed the SNP as a double hit. I believe this mined something on the order of 10 K of double-hit SNPs.

We also knew that we had a bolus of TSC SNPs mined from traces known to be from sources other than the clone libraries used in the human genome. I'm a bit fuzzy on the details now; I do, however, recall that the individuals supplying the TSC traces were pooled together, but that a statistical argument, based on the number of individuals in the pool, allowed us to consider each trace as an independent sample with high confidence. Additionally, the allele appearing on the genome itself constituted one hit of that allele. If we found at least one other trace with the genomic variant and two traces of the variant not on the genome in the TSC dataset, the SNP was classified as double-hit, two-allele. I believe we classified about 100 K double-hit SNPs by this method.

Because dbSNP contains Dr. Jim Mullikin's double-hit SNPs, can you tell me what method he uses to determine these double-hits?

In an email, Dr. Mullikin described his double-hit method (*reprinted with permission*):

First, I align the following sequences to the human reference sequence: all human traces from the trace archive; all clone sequences not used in the reference sequence; cDNA sequence; the Celera WGSA assembly; and Celera reads from non-donor B individuals. For any rsIDs, I look at the alignment, and count how many times I see each allele. If I see each allele two or more times in different DNAs, I classify it as a double-hit SNP.

I also use chimp to promote an allele from a count of one to two. For example, let's say for an A/G SNP, A is seen in human DNA seven times, and G once. Then, if chimp is a G, it becomes a double-hit SNP. Also, if the chimp sequence is polymorphic, or does not agree with either human allele, the chimp allele(s) is not used.

For human DNA, if the sequence comes from a single individual, I do not allow that individual to contribute to the allele counts more than once per allele. For the clone overlaps, I do a similar thing, but never allow the clones to increment the reference allele count, and I limit the variant allele to a single increment no matter how many clones overlap that base. This way I do not have to look up the source of the reference clone DNA or the overlapping clone DNA sources.

Do you know what computational methods Dr. Jim Mullikin uses to identify sequences/reads?

In an email Dr. Mullikin described his computational method (*reprinted with permission*):

The (computationally identified) SNPs were discovered largely from the same reads used in the double hit determination. I submit SNPs to dbSNP from the traces. Once they assign them rsIDs, I inspect the alignment data I described to count alleles.