Question: Alternate nucleotide is more frequent than reference nucleotide. OMG I'm dizzy. How do I stop the twirl?

3

1. How did it come to be that the alternate nucleotide was more frequent than the reference nucleotide?
2. How does one account for this phenomenon when designing a strategy to filter for variants of interest? Should I go through the complicated process of selecting those individuals who DO NOT have the variant and calculate that the REFERENCE frequency in the population is probably around (1 - esp6500siv_all)?

(/u/592/)
14 months ago by Farrel (/u/592/) • **160**
Pittsburgh, PA, USA

I am researching a rare disease and have whole exome sequence data with the corresponding variant calls. Each variant call has been passed to annovar and among other data, we have looked up the frequency of the variant in the esp6500siv2_all data. Clearly a variant that was observed to have a high frequency in our sample but that had low frequency in esp6500siv2_all would be of disproportionate interest.

Low and behold I was surprised to find that 13% of the all of our variants (4055 out of 32131) had an allele frequency that was greater than 0.5. How can that be? I expected that all the allele frequencies would be <0.5.

I had thought that the variants would be akin to a minor allele frequency (MAF). Clearly I was wrong. I pulled 3 random variants from among the variants that had more than 0.5 frequency, to check them against the Exome Variant Server.

```
    avsnp147 Chr      Start       End Ref Alt Gene.refGene   esp6500siv2_all
1: rs3803530  15  89632842  89632842   C   A       KIF7             0.5373
2:  rs621383   3 125118840 125118840   T   C     SLC12A8            0.9988
3:  rs633561  11  64229857  64229857   A   G      NUDT22            0.9418
```

Looking up at NHLBI Exome Sequencing Project (ESP) Exome Variant Server (http://evs.gs.washington.edu/EVS/) and using All Allele

1. rs3803530: C>A; A=6984/C=6014 which means A is 6984/(6984+6014) or 0.537
2. rs621383: T>C; C=12479/T=15 which means C is 12479/(12479+15) or 0.999
3. rs633561: A>G; G=12240/A=756 which means G is 12240/(12240+756) or 0.941

variant (/t/variant/) maf (/t/maf/) exome (/t/exome/) • 1.5k views

ADD COMMENT • link (/p/282029/) • Not following

modified 14 months ago by Emily_Ensembl (/u/7019/) ♦ **16k** • written 14 months ago by Farrel (/u/592/) • **160**

2 *1) How did it come to be that the alternate nucleotide was more frequent than the reference nucleotide?*

the reference genome carries the rare allele.

ADD REPLY • link (/p/282029/#282064) written 14 months ago by Pierre Lindenbaum (/u/30/) ♦ **116k**

1 OK and the reference genome would be just one person's at any one spot? The entire genome could be made up of many people's genome but at any one locus it would be just one person's sequence? So would I be correct that all the SNPs mentioned in the NHLBI Exome Sequencing Project (ESP) would be from an individual who is homozygous at that point? If they were heterozygous there could be no ref vs alt.

ADD REPLY • link (/p/282029/#282724) written 14 months ago by Farrel (/u/592/) • **160**

1 Hey Farrel, hg38 was released in the wake of the 1000 Genomes Project, where whole genome sequence data from ~2500 individuals became available. This information was in part used to construct hg38 where, for example, many of the rare disease risk alleles of hg19 were modified to represent more common alleles. hg38 also improves on sequence in centromeric and other repeat regions, which were previously difficult to sequence. So, yes, it's still a linear representation of the genome and has its flaws.

I'm not sure that I understand your point about the ESP. The ESP is a disease association study and includes heterozygous and homozygous variants.

> *The goal of the NHLBI GO Exome Sequencing Project (ESP) is to discover novel genes and mechanisms contributing to heart, lung and blood disorders by pioneering the application of next-generation sequencing of the protein coding regions of the human genome across diverse, richly-phenotyped populations and to share these datasets and findings with the scientific community to extend and enrich the diagnosis, management and treatment of heart, lung and blood disorders.*

ESP allele frequencies, like all others, are just counted 1 for het and 2 for hom. If my variant is observed as het in 1 individual in my cohort of 500 patients, then the allele frequency is `1 / (500 * 2) = 0.001%`

ADD REPLY • link (/p/282029/#282756) written 14 months ago by Kevin Blighe (/u/41557/) ♦ **35k**

10

This is due to the fact that the very reference genomes that we use for re-alignment are themselves based on individuals who carry rare risk alleles. Thus, when we call variants against these genomes, we are, at many loci, comparing against rare disease risk alleles.

(/u/41557/)
14 months ago by Kevin Blighe (/u/41557/) ♦ **35k**
Republic of Ireland

As the best/worst example (depending on your point of view), hg19 / GRCh37 was used for more than a decade as the primary reference genome, yet ~70% of the genomic sequence of this genome was based on a single individual from the Buffalo area, New York, USA. Amongst the many 1 000s of rare disease susceptibility alleles that this individual carried was one called *Factor V Leiden*, which statistically increases the risk of deep vein thrombosis (DVT). If you're researching DVT (I was), you have to be aware of this.

Thus, if I perform exome-seq on an individual who does not have *Factor V Leiden* and re-align the data to hg19 / GRCh37, the *Factor V Leiden* variant position will show a SNV because the reference allele in my patient sample (which doesn't increase risk of DVT) is being compared against the disease allele that's contained in the very reference genome against which I'm re-aligning my data. Without careful screening, I may assume that my patient has increased risk of DVT, erroneously so.

There was a publication on this listed in PubMed but it's very difficult to find, even by Google. It's a critical problem yet has not received the attention that it deserves.

The situation improved with hg38 / GRCh38, as this reference build was based on much more individuals, but the same problems still persist, broadly speaking.

So, you really have to get to know your target panel and all of these nuances related to whatever variants you're studying., particularly if you're dealing with live patient data.

Kevin

Update 3rd January 2018

It has come to my attention that there is automated method to search for these types of variants in your VCF:

- http://rmahunter.bioinf.me/ (http://rmahunter.bioinf.me/)
- https://github.com/bioinf/RMAhunter (https://github.com/bioinf/RMAhunter)

ADD COMMENT • link (/p/282029/#282033)

modified 12 days ago • written 14 months ago by Kevin Blighe (/u/41557/) ♦ **35k**

8

reference ≠ major ≠ ancestral ≠ wildtype

As Kevin says, the reference is just whatever is in the reference sequence, which is the sequence of whoever they happened to sequence for that region.

GRCh38 is an improvement compared to GRCh37 because the GRC sought out some loci where the reference allele was not the major allele in the 1000 Genomes project, and replaced those regions with tiny contigs which did have the major allele. Some, not all.

(/u/7019/)
14 months ago by Emily_Ensembl (/u/7019/) ♦ **16k**
EMBL-EBI

ADD COMMENT • link (/p/282029/#282093) written 14 months ago by Emily_Ensembl (/u/7019/) ♦ **16k**

1 Hi Emily, How to understand ancestral ≠ wildtype? Thanks.

ADD REPLY • link (/p/282029/#328290) written 6 months ago by Shicheng Guo (/u/19400/) • **7.3k**

3 The ancestral allele is identified by tracing back up the evolutionary tree to see what other primates have at the same location. A mutation way back in our lineage at a particular locus may be one of the small evolutionary changes that *make us human*. Individuals who have the ancestral allele may have a phenotype that makes them more like our ancestors (maybe long arms, a heavy brow or slight mental retardation), in which case we could say that the ancestral allele is associated with the phenotype.

Generally I don't like the word "wildtype" at all because it infers that one allele confers a phenotype, and the other does not, but in fact both alleles confer phenotypes, it just depends on your perspective as to which one is "normal". When we're talking about human phenotypes, that perspective is often drawn along racial lines, which is not acceptable. For example rs4988235 (http://www.ensembl.org/Homo_sapiens/Variation/Explore?r=2:135850576-135851576;v=rs4988235;vdb=variation;vf=3211518). The reference allele is G, which happens to also be the ancestral allele which is found in all our primate relatives. As a European, I consider the reference allele G to be the one associated with a phenotype: lactose intolerance. However, a non-European might say that the phenotype association is with the alternative allele, A: lactase persistence. I would, therefore be uncomfortable with assigning either of those alleles the term "wildtype".

ADD REPLY • link (/p/282029/#328494) written 5 months ago by Emily_Ensembl (/u/7019/) ♦ **16k**

Please log in (/site/login/) to add an answer.

Similar posts • Search » (/local/search/page/)

- How to fix this strand flips for Michigan Imputation Server? (/p/350238/)
  Hi Every one I am a bit stuck at this error from Michigan Imputation Server, does any one know ...
- QUAL and AF GATK (/p/173703/)
  Hi. I am a bit confused. I have my vcf files from GATK (exome data) and I want to filter out the...
- How Does Average Heterozygosity Relate To Alellic Frequency Data (/p/76418/)
  Forgive me if this is a dumb question but I assumed that the Average Heterozygosity was somehow r...
- Allele frequency ExAC database (/p/221860/)
  Hi all, I have a doubt regarding the ExAC database about the allele frequency. For some variants...
- Duplicated Allele Using Varscan (/p/90585/)
  Hi friends, Anyone have any similar experience about the duplicated allele with different allele...
- Why does shoremap calculate allele frequency this way? (/p/275956/)
  Hello everyone, I'm learning how to analyze genome sequencing data and questions answered on the...
- subseting VCF by bcftools (/p/267683/)
  I'm trying to include only single nucleotide variant, or some say SNPs, from the 1000 genomes pro...
- Bam to nucleotide frequencies (/p/109798/)
  I have whole genome sequences of several clinical isolates and I want to get the nucleotide freq...
- understanding of dbSNP info on MAF, Alleles and HGVS Names (/p/132500/)
  I have some problems with understanding of the concepts of SNP and related MAF, Alleles and NGVS ...
- Why Freebayes Allele frequency(AF) is 0.5 or 1.0, instead of reporting actual allele frequencies ? (/p/243073/)
  Hi Biostars Leaders, Freebayes(version:v1.0.1-1-g683b3cc-dirty) defines AF as Description="Estim...
- vcf-consensus with more than one alternative alleles (/p/103424/)
  Hi all, If I have a vcf file, and some variants have more than one alternative alleles, like: #...
- Deletion detected at less frequency than expected for a plasmid sample (/p/337976/)
  Hi, I was performing variant analysis for a plasmid sample sequenced on MiSeq with read length P...
- literature search in DbSNP (/p/233768/)
  hai frds, I am using the DBsnp database how to find the literature paper? Filters activ...
- LOH and CNV data from VCF files (/p/119866/)
  I was asked to find both the CNV and the LOH datas from 3 VCF files (all from the same patient bu...
- Need help understanding read depth for somaticsniper vcf's (/p/105827/)
  Hi, I'm having trouble understanding read depth in vcf's called by somatic-sniper. Here is a sa...
- How should I calculate the variant density, per kb for a gene, for two different groups in a case/control study? (/p/199403/)
  I have variant data for a group of people. Some of those people have a disease. I'm trying to cre...

Traffic: 2262 users visited in the last hour
REVOLVERMAPS