Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling

Yury A. Barbitoff, BS^{1,2,3}, Igor V. Bezdovornikh, BS¹, Dmitrii E. Polev, PhD², Elena A. Serebryakova, MD², Andrey S. Glotov, PhD², Oleg S. Glotov, PhD³ and Alexander V. Predeus, PhD¹

Purpose: We comprehensively assessed the influence of reference minor alleles (RMAs), one of the inherent problems of the human reference genome sequence.

Methods: The variant call format (VCF) files provided by the 1000 Genomes and Exome Aggregation Consortium (ExAC) consortia were used to identify RMA sites. All coding RMA sites were checked for concordance with UniProt and the presence of same codon variants. RMA-corrected predictions of functional effect were obtained with SIFT, PolyPhen-2, and PROVEAN standalone tools and compared with dbNSFP v2.9 for consistency.

Results: We systematically characterized the problem of RMAs and identified several possible ways in which RMA could interfere with accurate variant discovery and annotation. We have discovered a systematic bias in the automated variant effect

prediction at the RMA loci, as well as widespread switching of functional consequences for variants located in the same codon as the RMA. As a convenient way to address the problem of RMAs we have developed a simple bioinformatic tool that identifies variation at RMA sites and provides correct annotations for all such substitutions. The tool is available free of charge at <http://rmahunter.bioinf.me>.

Conclusion: Correction of RMA annotation enhances the accuracy of next-generation sequencing–based methods in clinical practice.

Genet Med advance online publication 26 October 2017

Key Words: allele frequency; next-generation sequencing; reference; minor allele; whole-exome sequencing

INTRODUCTION

Over the past several years, the rapid increase in availability of next-generation sequencing has led to sequencing of hundreds of thousands of human exomes and genomes. Among many other uses, sequencing is quickly becoming a new standard in diagnostics of Mendelian disease. However, variant calling and especially interpretation still remain a challenge for clinicians and bioinformaticians. A recent publication by the Exome Aggregation Consortium (ExAC) suggests that the majority of variants currently annotated as pathogenic are actually benign or insufficient to cause disease.¹ Variant interpretation depends on several major determinants, such as clinical and epidemiological information, variant effect prediction by numerous available software tools, and allelic frequency (AF) information from large sequencing consortia. Clinical experience of variant interpretation was summarized in recent American College of Medical Genetics and Genomics guidelines for interpretation of sequence variants.² However, it is rarely mentioned that the reference genome sequence itself represents a problem for discovery and interpretation of variation.

The reference human genome sequence used in virtually all of the clinical studies originates from the sequence obtained by the Human Genome Project.³ One of the inherent problems of the reference genome is the presence of minor alleles that became

occasionally included in the reference sequence. Notably, these reference minor alleles (RMAs) include several well-established rare pathogenic variants that are occasionally misinterpreted in clinical practice. For example, in a recent publication concerning Bardet–Biedl syndrome (MIM 615981), the authors stated that the N70S substitution in *BBS2* (c.209A > G [p.Asn70Ser], NM_031885.3; dbSNP accession ID rs4784677) is a false positive because it is present in 31 of 32 of their patients.⁴ It appears that the pathogenic allele could have been present in only one of the patients, as the variant is called only when the sample has a wild-type allele at the RMA site. In a different example, the discordance between the annotations in different databases has led to exclusion of the same variant from further analysis in a recent study of the Qatari population.⁵

Given these examples, it is not surprising that the existence of the RMA problem has already been described and several approaches to address the issue have been proposed.^{6–8} The first approach implies the usage of a corrected population-specific reference genome sequence.⁶ However, this method requires additional corrections to all existing variant annotation databases and complicates the comparison of results obtained in different studies. The second approach suggested introduction of the new variant call format (VCF), called eVCF, which would allow keeping information about nonvariant sites.⁷ The major drawback of such an approach

¹Bioinformatics Institute, St. Petersburg, Russia; ²Biobank of the Research Park, St. Petersburg State University, St. Petersburg, Russia; ³Department of Genetics and Biotechnology, St. Petersburg State University, St. Petersburg, Russia. Correspondence: Alexander V. Predeus (predeus@bioinf.me)

Submitted 21 March 2017; accepted 18 August 2017; advance online publication 26 October 2017. doi:10.1038/gim.2017.168

Figure 1 The scale and complexity of the reference minor allele problem. (a) Hierarchical classification of reference minor alleles (RMAs) into several distinct classes. Top bar, noncoding versus coding variants; middle bar, different effects of in-coding sequence RMA on protein sequence; bottom bar, concordance of the reference alleles and the UniProt reference protein alleles for all missense variants. Top numbers and percentages represent RMAs with allele frequency (AF) of 0.5 or greater; bottom numbers and percentages represent rare RMAs (AF > 0.99). Note that 89% of the RMAs are in agreement with the corresponding UniProt sequence. (b) Population specificity of the RMA sites (based on 1000 Genomes data). Heat map represents Jaccard indices characterizing relative similarity between RMA in five major ancestral groups and in global population (see **Supplementary Methods** for more detail). Bar plot represents total amount of sites with reference allele frequency < 0.5 for each group. (c) Number of misclassified single-nucleotide polymorphisms (SNPs) located in the same codon as the RMA. Rescued variants are initially classified as nonsense or missense, but switch their class to same sense or missense after the RMA correction. Gained variants are the opposite cases, and become missense or nonsense when the RMA is corrected from minor to major nucleotide (see Definitions in the **Supplementary Material**). (d) Comparison of variants annotated as pathogenic in dbNSFP and after correction for the three variant effect predictors: PROVEAN, PolyPhen-2 (HumVar), and SIFT. (e) Empirical distributions of the number of predicted pathogenic substitutions in 10,000 random sets of 4,408 non-RMA variants with matched minor allele frequency (MAF) distributions (see **Supplementary Methods** for more information). The left panel illustrates consensus of the three tools and right panels show distributions for individual predictors. Blue marks indicate the values obtained for RMAs set by our corrected variant prediction and orange marks indicate automated predictions by dbNSFP. AFR, African; AMR, American; EUR, European; EAS, East Asian; SAS, South Asian.

is incompatibility with all conventional variant annotation and interpretation protocols. The most recent way to counter the RMA problem is using a wrapper around the now-retired Genome Analysis Toolkit UnifiedGenotyper, which forces genotyping at RMA loci.⁸ Apart from using an outdated variant caller, such an approach limits the user's options to

choose among available software packages for data manipulations. Altogether, all of the mentioned ways to correct the RMA problem are either not compatible with the current variant-calling software, interfere with the variant annotation and interpretation protocols, or both. These approaches also require additional effort to reanalyze existing data from

scratch to gain insights into the hidden variation. These drawbacks have resulted in the fact that the above-mentioned approaches are virtually never used in published studies in medical genetics.

Perhaps more importantly, the scope of the problem and the levels on which RMAs could alter variant calling and annotation have never been comprehensively assessed. Hence, in our study we undertook a systematic effort to analyze the problem of RMAs and its influence on variant annotation and interpretation, and developed an easy-to-use tool to tackle this problem in research and clinical practice.

MATERIALS AND METHODS

Data collection and analysis

To obtain a comprehensive list of RMAs we have filtered standard b37-based VCF files from two major sequencing projects: 1000 Genomes (build 2013-05-02)⁹ and ExAC (release 0.3.1),¹ retaining only variants with nonreference AF of 0.5 or more in either database.

Human reference proteome was extracted from SwissProt/UniProtKB database (<http://www.uniprot.org/>). Local alignment of protein sequences was done using the FASTA suite (http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml). To obtain the pathogenicity predictions we used local installations of PROVEAN,^{10,11} SIFT,^{12,13} and PolyPhen-2.¹⁴ The dbNSFP v. 2.9 database¹⁵ was used to compare the predictions for consistency. To analyze the abundance of RMA in real patients' data we have selected 107 whole-exome samples from the Biobank of St. Petersburg State University that fit the ExAC quality control criteria.

More details on methods can be found in the **Supplementary Methods** online.

RESULTS

Our initial filtering of the two VCF files resulted in a compendium of 2,094,954 sites. As expected, most of the RMA sites resided in noncoding rather than coding regions. Of 12,709 sites residing in coding sequence, 6,149 (48.4%) were synonymous, 4,952 (40.0%) were missense, 200 (1.6%) were indels, and remaining 1,451 (10.0%) included nonsense and other types of variants (**Figure 1a**). Of those, 1,020 RMA variants have a corresponding ClinVar entry, with 19 in-coding sequence variants having "pathogenic" (15) or "likely pathogenic" (4) clinical significance, including the previously described variant in *BBS2*. Importantly, global population shares on average 2/3 of RMA sites with each of the major ancestral groups (**Figure 1b**, **Supplementary Figure S1**).

An important problem that has never been addressed is misannotation of nucleotide substitutions in the same codon with the RMA. Indeed, separate annotation of RMA and the neighboring variants may change the presumed amino acid substitution type from missense to nonsense or same sense, and vice versa. To evaluate this problem, we have identified all the cases where such reversal takes place. We have found 1,683 potential rescued protein-truncating variants and 1,237 gained protein-truncating variants associated with the RMA

loci, along with 9,457 cases when the presence of an RMA makes same-sense variants appear nonsynonymous, and vice versa, with gained variants being slightly more abundant (4,628/4,829, **Figure 1c**) (see the Definitions section of the **Supplementary Information**).

We were also surprised to discover that the majority of the RMAs are in concordance with the corresponding UniProt protein sequence (**Figure 1a**). Because the protein sequences resulting from translation of the reference genome usually harbor mutant amino acid, automated pathogenicity predictions, like those stored in the widely used dbNSFP database¹⁵ or generated by automated VCF-based query, would represent the effect of the incorrect amino acid substitution—that is, minor to major instead of the converse. To validate this hypothesis we have annotated all RMAs with the custom pathogenicity predictions by three commonly used variant effect predictors: PROVEAN,^{10,11} SIFT,^{12,13} and PolyPhen-2¹⁴ (see **Supplementary Methods** for more information). We found a dramatic discrepancy between corrected and precalculated predictions that is most probably explained by the swapped substitution. Of all coding missense substitutions in the RMA list, 264 were annotated as potentially pathogenic by all three tools after our correction, which is in sharp contrast to only nine in the dbNSFP database (**Figure 1e**). Most importantly, there was virtually no overlap between the corrected and the original predictions—that is, the majority of the variants predicted as pathogenic are classified as benign after the correction, and vice versa (**Figure 1d**). To statistically validate our findings we have performed a bootstrapping pseudosampling of 10,000 sets of 4,408 non-RMA variants with matched minor allele frequency distribution from the dbNSFP database (see **Supplementary Methods**). We have discovered that the corrected predictions fall much closer to the expected number of pathogenic predictions for all individual tools. The same is true for the consensus between all three predictors (**Figure 1e**).

Finally, we have developed a simple way to identify possible RMA-related problems in variant calling using a regular single- or multisample VCF file. Unlike all other approaches suggested so far, the intuitive tool we dubbed RMA Hunter allows extraction and analysis of variation at RMA sites in under a minute without the need to reanalyze raw next-generation sequencing data (See **Supplementary Methods**). The tool searches for RMAs and their neighboring variants and outputs three major groups of variants (false-negative and false-positive RMAs, and variants with switched functional types, **Figure 2a**) annotated with appropriate allele frequencies and corrected pathogenicity predictions in the form of a web report (**Supplementary Figure S3**). We utilized RMA Hunter to assess the average number of variants in each of the three groups described above using a set of 107 whole-exome sequencing experiments. Our analysis shows that a typical exome harbors on average 2.24 false-negative as well as 450 false-positive rare (minor allele frequency <0.01) nonsynonymous coding RMAs. Since our average exome contained 12,026 nonsynonymous substitutions, 2,596 of

Importantly, a significant amount of RMAs emerge as potentially clinically relevant variants that may have been overlooked in earlier studies. Several hundreds of protein-coding RMAs have a nonreference allele frequency of 1 throughout ExAC and 1000 Genomes, meaning that the

DISCUSSION

Upon careful examination, the RMA problem goes well beyond previously described issues and involves virtually every stage of variant calling and interpretation (**Figure 2c**). First, RMA variants influence read alignment efficiency,⁶ mostly attributable to RMA indels. Second, RMAs elude many existing variant-calling pipelines as those only report variants different from the reference sequence. The third and perhaps the most significant effect of the

reference allele has never been observed in any of the sequenced individuals. These variants either belong to highly evolutionary conserved regions, or reside in functionally relevant protein domains or motifs. One such example is the L545F substitution in the far C-terminus of SOWAHA (c.1635G > T [p.Leu545Phe], NM_175873.5; dbSNP accession ID rs40470). The C-terminal fragment of this protein harbors an amino acid motif conserved among the majority of mammalian species based on UniProtKB. This motif is disrupted by the F residue that corresponds to the reference genome allele (**Figure 2d**). Interestingly, the reference allele at this site was observed 23 times as a somatic mutation in cancer according to the COSMIC database.¹⁷ The combination of extremely low germ-line AF and high frequency of appearance as a somatic mutation is characteristic to known cancer driver mutations in oncogenes (**Supplementary Figure S2**).

An interesting example of an RMA-caused switch of the variant type is the rs10769028 (c.621A > G [p.Ser270Ser], NM_021926.3, TCA → TCG) variant with minor allele frequency <0.01 in *ALX4*. The nearby variant (c.620C > G [p.Ser270*], NM_021926.3, TCA → TGA) is annotated as the nonsense substitution that has known-pathogenic effect¹⁸ and an ultralow AF in ExAC, although it does not cause a premature stop codon when the major allele is present at the RMA site in the same codon (c.620CA > GG [p.Ser270Trp], NM_021926.3, TCA → TGG); in other words, 99% of individuals harboring a C allele are far less likely to develop disease.

Despite the fact that the probability of observing a clinically significant false-negative RMA is low, it becomes much more relevant with the growing number of samples analyzed. We realize that substantial alterations to existing pipelines or formats to tackle all the above-mentioned problems would require a gargantuan effort given the amount of modern clinical sequencing data. The RMA Hunter tool allows fast RMA correction, and thus, it should be useful when processing large amounts of whole-genome/whole-exome sequencing data, and is especially applicable to samples that remained without a clear diagnosis in massive next-generation sequencing-based clinical studies. Altogether, it is evident that RMAs might cause certain confusion and potential misinterpretation of sequence variants. Systematic correction of RMAs in variant calling enhances the power of next-generation sequencing techniques in clinical practice and helps avoid potentially costly mistakes in interpretation of sequencing data.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at <http://www.nature.com/gim>

ACKNOWLEDGMENTS

The study was supported by Russian Science Foundation grant 14 50 00069. Equipment from the Biobank of the Research Park of St. Petersburg State University was used for whole-exome sequencing experiments analyzed in the present study.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Lek M, Karczewski KJ, Samocha KE, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:283–291.
2. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–423.
3. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
4. Geoffroy V, Pizot C, Redin C, et al. VaRank: a simple and powerful tool for ranking genetic variants. *PeerJ* 2015;3:e796.
5. Rodriguez-Flores JL, Fakhro K, Hackett NR, et al. Exome sequencing identifies potential risk variants for Mendelian disorders at high prevalence in Qatar. *Hum Mutat* 2014;35:105–116.
6. Dewey FE, Chen R, Cordero SP, et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet* 2011;7:e1002280.
7. Ferrarini A, Xumerle L, Griggio F, et al. The use of non-variant sites to improve the clinical assessment of whole-genome sequence data. *PLoS One* 2015;10:e0132180.
8. Magi A, D'Aurizio R, Palombo F, et al. Characterization and identification of hidden rare variants in the human genome. *BMC Genomics* 2015;16:340.
9. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature* 2015;526:68–74.
10. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 2012;7:e46688.
11. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 2015;31:2745–2747.
12. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–1081.
13. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;11:863–874.
14. Adzhubey IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;7:248–249.
15. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* 2013;34:E2393–E2402.
16. Fu W, O'Connor TD, Jun G, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 2012;493:216–220.
17. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015;43:D805–D811.
18. Mavrogiannis LA, Taylor IB, Davies SJ, et al. Enlarged parietal foramina caused by mutations in the homeobox genes *ALX4* and *MSX2*: from genotype to phenotype. *Eur J Hum Genet* 2006;14:151–158.