Short Communication

A performance evaluation study: Variant annotation tools - the enigma of clinical next generation sequencing (NGS) based genetic testing

Sachleen Tuteja a, Sabah Kadri b,c, Kai Lee Yap b,c,e

a Illinois Mathematics and Science Academy, 1500 Sullivan Road, Aurora, IL 60506, USA
b Department of Pathology and Laboratory Medicine, Ann and Robert H. Lurie Children’s Hospital of Chicago, 225 E. Chicago Ave, Chicago, IL 60611, USA
c Department of Pathology, Northwestern University Feinberg School of Medicine, 420 E. Superior St, Chicago, IL 60611, USA
e Corresponding author at: Molecular Diagnostics, Department of Pathology & Laboratory Medicine, Ann & Robert H. Lurie Children’s Hospital of Chicago, Northwestern Feinberg School of Medicine, 225 E. Chicago Ave, Box 82, Chicago, IL 60611, USA.
E-mail address: klyap@luriechildrens.org (K.L. Yap).

A R T I C L E   I N F O

Keywords:
Variant annotation
Genetic testing
Alamut®
VEP
ANNOVAR
Gene panel

A B S T R A C T

Dramatically expanding our ability for clinical genetic testing for inherited conditions and complex diseases such as cancer, next generation sequencing (NGS) technologies are allowing for rapid interrogation of thousands of genes and identification of millions of variants. Variant annotation, the process of assigning functional information to DNA variants based on the standardized Human Genome Variation Society (HGVS) nomenclature, is a fundamental challenge in the analysis of NGS data that has led to the development of many bioinformatic algorithms. In this study, we evaluated the performance of 3 variant annotation tools: Alamut® Batch, Ensembl Variant Effect Predictor (VEP), and ANNOVAR, benchmarked by a manually curated ground-truth set of 298 variants from the medical exome database at the Molecular Diagnostics Laboratory at Lurie Children’s Hospital. Of the 3 tools, VEP produces the most accurate variant annotations (HGVS nomenclature for 297 of the 298 variants) due to usage of updated gene transcript versions within the algorithm. Alamut® Batch called 296 of the 298 variants correctly; strikingly, ANNOVAR exhibited the greatest number of discrepancies (20 of the 298 variants, 93.3% concordance with ground-truth set). Adoption of validated methods of variant annotation is critical in post-analytical phases of clinical testing.

Introduction

Genetic testing has witnessed a transformative revolution in the last decade with the introduction of next generation sequencing (NGS) technologies.1–3 With its unprecedented cost-effective scalability, continuously improving efficiencies, and diagnostic yield, NGS has not only allowed for an exponential increase in the elucidation of genetic causes for both rare Mendelian and complex heterogenous disorders, but it is also proving to be an essential tool for identifying therapeutic targets in neoplasms, and screening for prenatal aneuploidy and pediatric onset disorders.4–7 The rapid proliferation of diagnostic tests, from small hotspot mutation and large panels to whole exome and whole genome platforms, has allowed for rapid interrogation of tens of thousands of genes and identification of millions of variants, many of which we have never seen before.8–10

While whole genome sequencing results in the detection of about 4 million variants per individual, exome sequencing that covers only the 1% protein-coding portion of the genome, detects about 20,000 variants.11,12 Though most of these variants likely only contribute to human population diversities since a single exome harbor only about 100–200 potential disease-causing changes, identification of the 1 or 2 disease-causing variants among these many alterations is a classic needle in the haystack problem that ends being a meticulous and arduous process as part of the NGS bioinformatics pipeline.13 Thus, translating sequencing information into clinical practice in this era of genomic testing is limited by the accurate annotation and interpretation of the variants rather than detection of variants alone. This discovery of novel genetic variants at this extraordinary pace has also reiterated the need of a standardized approach for describing, documenting, and communicating genetic variants, first recognized by the scientific community in the pre-NGS era,14,15 and has resulted in the widespread adoption of the Human Genome Variation Society (HGVS) nomenclature system by research and clinical genetic testing laboratories.16–19

Variant annotation is the process of describing the nature and the effect of the genomic aberrations produced by a variant that involves adding auxiliary metadata to quality filtered raw putative variant calls.20 The most basic annotations will classify variants based on their relationship to genes, their transcripts, and other key features such as exons, introns, and splice sites, in addition to frequently providing information regarding known allele frequencies, predicted deleterious effects, and involvement in known human diseases and phenotypes.21

Variant annotation is dependent wholly on the gene models within which they reside, and has seen a prolific growth in both capability and scale in the past decades, evolving into its own research field.21,22 Conforming variant
coordinates from the transcript to the genome, and vice versa, is a complicated process that is dependent on the genomic and transcript sequence accessions and versions, and the alignment tools used, consequently resulting in significantly different locations of the variants.\textsuperscript{23,24}

Both theoretical and empirical based variant annotation approaches have been used. A series of empirically based tools including the open source SnipEff,\textsuperscript{25} Variant Effect Predictor (VEP, Ensembl),\textsuperscript{26} VarReporter,\textsuperscript{30} ANNOVAR,\textsuperscript{28} the commercial Alamut Batch,\textsuperscript{29} and tools developed by individual laboratories such as Invitae,\textsuperscript{30} for outputting HGVS syntax from next generation sequencing data have gained prominence over the years. A significant proportion of these tools enable variant annotation based on Ensembl transcripts while parallelly leveraging the rich annotations in dbNSFP,\textsuperscript{31} a database of curated annotations and functional effect predictions for all potential non-synonymous and splice site SNVs in the human reference genome.

Alamut® Batch (Sophia Genetics, formerly Interactive Biosoftware), a licensed gene annotation software, is widely used in clinical laboratories, including Lurie Children’s Molecular Diagnostics Lab, for supporting variant annotation and classification using HGVS standards. However, the efficiency of performing variant annotation by Alamut® Batch is limited by its licensing structure, and thus prompted our evaluation of other potential, well supported open-source replacement tools like VEP. In this study, we compare the concordance of variant nomenclature generated by the open source ANNOVAR and Variant Effect Predictor (VEP), and the commercial Alamut® Batch.

Materials and methods

A test set of 298 intronic and exonic variants across 191 genes, previously classified and reviewed in clinical reports by the Molecular Diagnostics Laboratory at Lurie Children’s Hospital was curated and used for this study. These variants were generated by targeted gene panel sequencing using the ~4700 gene medical exome panel on 105 patients (Fig. 1). Briefly, 150 bp paired end reads generated using the Illumina NextSeq 550 at an average depth of 250 per sample were mapped to the human genome GRCh38/hg38 using Burrows Wheeler aligner (BWA) v0.7.12, sequence alignment and map (SAM) files converted to binary alignment and map (BAM) format and sorted using SAMtools v1.9. Local realignment around the known indels was performed by GATK v4.1.4.1 on the sorted BAM files and Picard-tools v2.18.27 was used to remove Polymerase chain Reaction (PCR) duplicates. Finally, base quality score recalibration was performed using GATK again to generate the VCF (Variant Call File), a de facto standard used in reporting genetic variants.

The 298 variants represented the following 191 genes: ACTA2, ADAM17, ADAMTS10, ADAMTS17, ADAMTS14, ADS1, AICDA, AK2, ALDH5A1, ALDH7A1, ANKRD26, AP3D1, ASNS, ATM, ATRX, BGN, BLOC1S6, BTX, C1QC, C1S, C2, C3, C5, C6, C9, CACNA2D2, CARML2, CDDC151, CD257, CD46, CD79A, CD1A1, CFB, CHFR, CHD7, CIITA, CLN6, CLBP, CNTN2, CNTNAP2, COL12A1, COL1A2, COL3A1, COLS1A2, COLEC11, CR2, CSF2RB, CSF3R, CTCL1, CTSC, CCR4, DEPDC5, DGKE, DHAH1, DNAH5, DOCK8, DRC1, DYNCH1H1, ELANE, ENG, EPFG, EPHB4, ERC66L2, FANCA, FANCB, FANCD2, FANCE, FANCI, FASN, FAT4, FB11, FB22, FCRG2A, FOXP3, GABBR2, GABRD, GAMT, HRAS, HYOU1, ICAM1, IF1H, IPFGR2, IKKBG, IL1OAR, IL12RB1, IL17A, IL17R, IL7R, JAK3, IMD1C, KCND2, KCNT1, KMT2D, KRIT1, LCK, LIG1, LTR2, LTP2B, LYST, MAS1P, MAP2, MECPR2, MEVF, MCLU1, MLH1, MSH2, MSH6, MTHFD1, MVK, MYH11, MYLK, NALCN, NF1, NF2BIA, NLRC4, NLBP1, NOD2, NOTCH1, NRXN1, OFD1, PARK, PGQ, PLCP2, PLCD1, PNKD, POL2, POLG, PRKCE1, PSTP11, PTPRC, RAF1, RBCK1, RECQ4, RFX1, RFXAP, RMRP, RNASEH2B, RFN213, RFN31, RTEL1, RUNX1, RYR3, SBDS, SCN3A, SCN8A, SERPIN1, SERPING1, SETD2, SLCSA15A5, SLCS2A1, SLCS2A1, SMARCA1, SOS1, SOS2, SPINK5, STG31A3, STAT3, STAT5B, STRAD, STZB2, TBK1, TGBP1, TCA1, TINFB, TIRAP, TMG6, TMCG, TNFRSF13B, TNFRSF1A, TNXB, TPPI, TPP2, TRAF3IP2, TSC1, TSC2, TCC3, TCCA7, TYK2, UNC13D, VPS13B, VPS4S, WRAP53, ZBTB24, ZCCHC8, and ZNF341.

The 298 variants were manually reviewed using Alamut® Visual (v2.15, Sophia Genetics, formerly Interactive Biosoftware) and Integrative Genomics Viewer (IGV v2.10.2),\textsuperscript{10,10} and orthogonally confirmed with Sanger sequencing when necessary. After the validation process, the variants were distributed with respect to their classification (Single nucleotide variants (SNVs), Deletions, Duplications, Insertions, and Complex variants), thereby constituting the ground-truth set, and subsequently independently annotated using Alamut® Batch (v1.4.2, July 2015, Sophia Genetics), Ensembl Variant Effect Predictor (VEP v105.0, 2017) and ANNOVAR (v. October 24, 2019). Custom functions were written in python to compare the results of the 3 software packages.

Fig. 1. Overview of the targeted gene panel sequencing pipeline using the ~4700 gene medical exome panel on 105 patients used to generate the test set of 298 intronic and exonic variants across 191 genes evaluated in this study. Briefly, 150 bp paired end reads generated using the Illumina NextSeq 550 at an average depth of 250 per sample were mapped to the human genome GRCh38/hg38 using Burrows Wheeler aligner (BWA) v0.7.12, sequence alignment and map (SAM) files converted to binary alignment and map (BAM) format and sorted using SAMtools v1.9. Local realignment around the known indels was performed by GATK v4.1.4.1 on the sorted BAM files and Picard-tools v2.18.27 was used to remove Polymerase chain Reaction (PCR) duplicates. Finally, base quality score recalibration was performed using GATK again to generate the VCF (Variant Call File), a de facto standard used in reporting genetic variants.
Genomics Viewer,52 (IGV v2.10.2), and orthogonally confirmed with Sanger sequencing when necessary. As a HGVS nomenclature tool, Mutalyzer,53 (v2.0.35), was used to verify the outputted data with respect to the standard HGVS nomenclature guidelines. After this validation process, the variants were distributed with respect to their classification (Single nucleotide variants (SNVs), Deletions, Duplications, Insertions, and Complex variants), thereby constituting the ground-truth set.

Subsequently, these 298 variants were independently annotated using Alamut® Batch (v1.4.2, July 2015, Sophia Genetics, formerly Interactive Biosoftware), Ensembl Variant Effect Predictor (VEP v105.0, 2017) and ANNOVAR (v. October 24, 2019). For single nucleotide variations (SNVs), the substitutions were further divided into missense variants (219 variants), which causes a substitution in the amino acid residues, and is a non-synonymous change; nonsense variants (19 variants), which prematurely end the protein with a stop codon, also a non-synonymous change; and silent variants (19 variants), which had no effect on the amino acid, and is a synonymous change. The effect of 39 variants is not predicted to have an effect on the amino acid since those variants were in the intronic region. Genomic variants were recognized as duplications if the alternate nucleotides were a repeat of the sequence prior to the genomic location of the variant. Complex variants were identified when multiple base pairs were involved and can be described as having deleted nucleotide(s) as well as inserted nucleotide(s). Custom functions were written in Python to compare the results from the 3 software packages (https://github.com/sachT19/Variant-Nomenclature-Comparer).

Previous manual evaluation for each of the software was completed by Lurie’s Bioinformatics team. Alamut Batch annotations were part of clinical testing at Lurie Children’s, therefore a standardized operating procedure was followed by different bioinformaticians. VEP annotations were run by a bioinformatician and ANNOVAR was added as an additional tool subsequently and the annotation process performed by ST.

Fig. 2. Relative concordance of variant nomenclature generated by ANNOVAR, Variant Effect Predictor (VEP), and Alamut® Batch. A. Distribution of variant types in the ground-truth set. Variant types evaluated were: single nucleotide variants (SNVs), deletions (del), duplications (dup), insertions (ins), and complex (complex) variants. The majority of the variants were classified as SNVs, followed by deletions, duplications, complex, and insertions. B. Variant distribution profile of the 3 annotation packages - Alamut® Batch, Variant Effect Predictor (VEP), and ANNOVAR. The relative annotation profiles are very similar between the 3 software, although VEP was unable to identify one of the variants, and ANNOVAR was unable to annotate 4 changes. C. Exact concordance of HGVS syntax at the coding level between the ground truth and the 3 variant annotation tools - Alamut® Batch, Variant Effect Predictor (VEP), and ANNOVAR. A total 278 of the 298 variants were found to be concordant between the 3 software packages. VEP and Alamut® Batch were 99% concordant, however, ANNOVAR produced 20 discrepancies. D. Distribution profile of the 20 discrepant variants called by ANNOVAR. Variant types evaluated were: single nucleotide variants (SNVs), deletions (del), duplications (dup), insertions (ins), and undefined (Undef). The majority of the discrepancies were classified as SNVs and ANNOVAR was unable to annotate four variants.
Table 1
Detailed characterization of the functional annotation of the discrepancies between the 3 variant annotation packages - Alamut® Batch, Variant Effect Predictor (VEP), and ANNOVAR. A majority of the discrepancies were due to the various transcripts used by the 3 different software.

| Chrom | Gene     | Alamut Batch cNomen   | Alamut Batch pNomen | VEP cNomen   | VEP pNomen | ANNOVAR cNomen   | ANNOVAR pNomen | Ground Truth cNomen | Ground Truth pNomen | Ground Truth Location | Pathogenicity |
|-------|----------|-----------------------|---------------------|--------------|-------------|------------------|----------------|---------------------|---------------------|-----------------------|---------------|
| 1     | CSF3R    | NM_000760.3:c.2092C>T  | p.Arg698Cys         | NM_000760.4: c.2041–30C>T | .            | NM_000760:exon17:c.2041–30C>T | .            | NM_000760.4:c.2041–30C>T | NM_02007.7:c.1375G>T | p.? Intronic         | VUS           |
| 9     | ERCC6L2  | NM_020207.4:c.1408G>T  | p.Val470Phe         | NM_064592.3:p.Val459Phe | T            | NM_020207:exon1:c.1375G>T | p.V459F       | NM_020207.7:c.1375G>T | p.Val459Phe        | p.? Exonic            | VUS           |
| X      | RTK      | NM_000661.2:c.4456C>T  | p.?                 | .            | .            | NM_000661.2:c.4456C>T | .            | .                   | .                   | p.? SUTR             |               |
| 2     | FANCL    | NM_018602.3:c.1096_1099dup | p.Thr367            | NM_018602.4: c.1096_1099dup | .            | NM_018602:exon14: c.1099_1100insATTA | p.T367fs     | NM_018602.3:c.1096_1099dup | NM_153461.1:c.1674_1676del | p.? Exonic            | Pathogenic     |
| 3     | IL17RC   | NM_153461.3:c.1674_1676del | p.Leu559del         | NM_000898.4:c.432 + 2 T>A | T            | NM_000898:exon12:c.1672_1674del | p.?          | NM_000898.3:c.432 + 2 T>A | NM_000898.5:c.1674_1676del | p.? Exonic            | VUS           |
| 7     | COL1A2   | NM_000089.3:c.432 + 2 T > A | p.?                 | .            | .            | .                   | .            | .                   | .                   | .                    |               |
| 7     | DNAH11   | NM_00277115.1:c.10691 + 2 T > C | p.?                 | .            | .            | .                   | .            | .                   | .                   | .                    |               |
| 8     | DNAH11   | NM_00277115.1: c.15323,15345dup | p.Arg4508, Leu4514dup | NM_017890.5:c.5513_5527del | .            | NM_017890:exon23: c.5513_5527del | .            | NM_017890.4:c.5513_5527del | NM_00277115.1:c.10691 + 2 T > C | p.? Exonic            |               |
| 9     | VPS13B   | NM_000933.4:c.2153del   | p.Arg523            | NM_001081.2:c.1566_1567del | p.?          | NM_001081:exon10: c.1564_1566del | p.?          | NM_001081.2:c.1566_1567del | NM_000089.3: c.524_527del | p.? Exonic            |               |
| 10    | MICU1    | NM_000677.3:c.1A>G     | p.?                 | .            | .            | .                   | .            | .                   | .                   | .                    |               |
| 11    | WRAP53   | NM_01081.2:c.1566_1567del | p.Arg323            | NM_012452.2:c.2044del | p.?          | NM_012452:exon11: c.2043_2045del | p.?          | NM_012452.2:c.2044del | NM_01081.2:c.1566_1567del | p.? Exonic            |               |
| 11    | WRAP53   | NM_01081.2:c.1566_1564dup | p_ARG523            | NM_01081.2:c.1566_1567del | p.?          | NM_01081:exon10: c.1564_1566del | p.?          | NM_01081.2:c.1566_1567del | NM_000089.3: c.524_527del | p.? Exonic            |               |
| 16    | TSC2     | NM_000548.4:c.599 + 5,599 + 7del | p.?                 | .            | .            | .                   | .            | .                   | .                   | .                    |               |
| 17    | WRAP53   | NM_01081.2:c.1566_1564dup | p.Arg522            | NM_018014.1:c.1566_1564dup | p.?          | NM_01801:exon11: c.1559_1561dup | p.?          | NM_01801.2:c.1566_1567del | NM_000548.4:c.599 + 5,599 + 7del | p.? Exonic            |               |
| 17    | FASN     | NM_000140.4:c.1513C>T  | p.Arg1705Trp        | NM_002627.3:c.6834del | p.?          | NM_002627:exon4:c.6833del | p.?          | NM_002627.3:c.6834del | NM_000140.4:c.1513C>T | p.? Exonic            |               |
| 17    | NF1      | NM_000267.3:c.6834del  | p.Leu69             | NM_0012452.2:c.2044dup | p.?          | NM_001245:exon11: c.2043_2045del | p.?          | NM_0012452.2:c.2044dup | NM_01081.2:c.1566_1567del | p.? Exonic            |               |
| 17    | TNFRSF13B| NM_012452.2:c.2044dup  | p.?                 | .            | .            | .                   | .            | .                   | .                   | .                    |               |
| 18    | ICAMI    | NM_000201.2:c.1546C>T  | p.Glu516*           | NM_000192.2:p.Glu516* | .            | NM_000192:exon12: c.1546 + 92C>T | .            | NM_000192.1:c.1546C>T | NM_000201.2:c.1546C>T | p.? Exonic            |               |
| 19    | JAK3     | NM_000215.3:c.566 + 6,566 + 41del | p.?                 | .            | .            | .                   | .            | .                   | .                   | .                    |               |
| 20    | KMT2D    | NM_003482.3:c.2992C>A   | p.Pro998Thr         | NM_003482.4:c.2992C>A | A            | NM_003482:exon11:c.2992C>A | p.?          | NM_003482.3:c.2992C>A | NM_003482:exon11:c.2992C>A | p.? Exonic            |               |
| X      | DCK1     | NM_001363.4:c.1512_1514dup | p?.                 | .            | .            | .                   | .            | .                   | .                   | .                    |               |
| X      | MECF2    | NM_000992.3:c.806del   | p.Ile506             | NM_000992.4:c.806del | T            | NM_000992:exon2:c.806del | p.?          | NM_000992.3:c.806del | NM_000992:exon2:c.806del | p.? Exonic            |               |
Results

In this study, we evaluated the performance of 3 variant annotation software packages: the commercial Alamut® Batch, and the open source Ensembl Variant Effect Predictor (VEP), and ANNOVAR, against a manually curated ground-truth set of 298 variants from the medical exome database at the Molecular Diagnostics Laboratory at Lurie Children’s Hospital. All 3 platforms generate transcript and protein-based variant nomenclature from genomic coordinates according to the guidelines by the HGVS. Our analysis of the 298 variants revealed that a vast majority of them were single nucleotide variants (SNVs, 92.6%) in the ground-truth set; the rest comprised a smaller number of deletions, duplications, complex, and insertions (Fig. 2a). SNVs are the most prevalent form of genetic variants and have been shown to play a potential role in the predisposition of disease.34

Although the input transcript alignments for VEP, ANNOVAR, and Alamut® Batch were identical, the tools produced a different number of transcripts and annotations. As shown in Fig. 2b, individually, all 3 tools - Alamut® Batch, VEP, and ANNOVAR - exhibited similar annotation profiles with regards to distribution of variant types, although VEP was unable to identify one of the variants, and ANNOVAR was unable to annotate 4 changes.

To investigate the degree of concordance between the 3 software packages, we compared the variant annotation calls with reference to the ground-truth set. When the annotations from all 3 software tools are exactly equivalent at the coding level, it is referred to as 100% concordance. A total of 278 of the 298 variants were found to be concordant between the 3 software packages (Fig. 2c). While Alamut® Batch and VEP exhibited comparable accuracy and precision with the ground truth with >99% concordance, ANNOVAR exhibited the greatest number of discrepancies (20 of the 298 variants, 93.3% concordance), a majority of which were non-SNVs (13/20) [Fig. 2c and d].

The functional annotation of the discrepancies between the 3 software packages are further characterized in Table 1. As observed for the 2 discrepancies in the CSF3R and the ERCG6L2 gene, the NM transcript versions between Alamut® Batch and VEP are different, resulting in varying genomic coordinates, and thus, differing nomenclature based on the locations in the gene transcript versions. In the case of the CSF3R gene, while Alamut® Batch called the variant as exonic, VEP was in agreement with the ground truth with >99% concordance, ANNOVAR exhibited the greatest number of discrepancies (20 of the 298 variants, 93.3% concordance), a majority of which were non-SNVs (13/20) [Fig. 2c and d].

The functional annotation of the discrepancies between the 3 software packages are further characterized in Table 1. As observed for the 2 discrepancies in the CSF3R and the ERCG6L2 gene, the NM transcript versions between Alamut® Batch and VEP are different, resulting in varying genomic coordinates, and thus, differing nomenclature based on the locations in the gene transcript versions. In the case of the CSF3R gene, while Alamut® Batch called the variant as exonic, VEP was in agreement with the ground truth with >99% concordance, ANNOVAR exhibited the greatest number of discrepancies (20 of the 298 variants, 93.3% concordance), a majority of which were non-SNVs (13/20) [Fig. 2c and d].

The functional annotation of the discrepancies between the 3 software packages are further characterized in Table 1. As observed for the 2 discrepancies in the CSF3R and the ERCG6L2 gene, the NM transcript versions between Alamut® Batch and VEP are different, resulting in varying genomic coordinates, and thus, differing nomenclature based on the locations in the gene transcript versions. In the case of the CSF3R gene, while Alamut® Batch called the variant as exonic, VEP was in agreement with the ground truth with >99% concordance, ANNOVAR exhibited the greatest number of discrepancies (20 of the 298 variants, 93.3% concordance), a majority of which were non-SNVs (13/20) [Fig. 2c and d].

The 20 discrepancies between ANNOVAR and the ground truth, 16 were exonic, while 4 were intronic. The transcript version for ANNOVAR was not provided by the software, so the cause for these discrepancies is unknown. For the 4 variants that it was unable to call, 2 were splicing variants and 2 were intronic variants. These 20 discrepancies additionally exhibit a spectrum of pathogenicity, ranging from likely benign to pathogenic. An overwhelming majority of the variants (14/20) were classified as Variants of Unknown Significance (VUS), so their phenotypic effect has not been discovered yet. Despite substantial progress in variant detection genome-wide, a significant majority of annotated genes have yet to be assigned function in the context of human disease traits.35,36

Discussion

Based on this data, it is evident that with respect to the HGVS nomenclature standards and clinical integrity, VEP has produced most accurate variant annotations. The inconsistencies in the data were observed in SNVs, and with the exception of the upstream gene variant, VEP was able to correctly identify and produce the HGVS nomenclature for 297 out of 298 variants. On the other hand, Alamut® Batch called 296 out of the 298 variants correctly. Performing significantly worse than the former 2 softwares, ANNOVAR was only able to correctly annotate 278 out of the 298 variants.

Similar observations have been reported by other studies evaluating the performance of other variant annotation tools. McCarthy et al. (2014) recorded 551 983 concordant variants out of the 637 841 (86.5%) between ANNOVAR and VEP; the discrepancies arising due to a difference in transcript versions.34 A 92.6% concordance (100/108 variants) between SnpEff and VEP was reported by Yen et al. (2017).25

To our knowledge, our study is the first relative performance evaluation of all 3 tools - ANNOVAR, Alamut® Batch, and VEP. Due to the advantage of the VEP algorithm to default the usage of the latest transcript version and unlimited licensing requirements, the Lurie Molecular Diagnostics Laboratory has decided to incorporate VEP as the variant annotator replacing Alamut® Batch.

While the chemistries of NGS library generation have matured and costs of sequencing have greatly declined over the last decade, the primary challenges facing a clinical genetic testing laboratory are in the expeditious analysis of large amounts of genetic data and interpretation of their clinical significance, as the scope of testing has approached exome and genome scales. While extensive cataloging of the human genetic variation is happening at a rapid pace, accurate detection and annotation of genetic variants are crucial to ensuring pediatric patients are receiving an accurate molecular diagnosis for their genetic conditions. Accurate identification and annotation of genetic variants will enable the establishment of substantial literature and information on the genetics of specific disease areas. Without accurate genetic variant identification and annotation to form the basis of a strong scientific literature, clinical interpretation of Variants of Unknown Significance (VUS) will continue to be a challenge for clinical diagnostics. Therefore, adoption of appropriate and validated methods of variant annotations is critical in the post analytical phases of clinical testing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

None.

Acknowledgments

Mr. Chris McCabe and Mr. Nicholas Miller – Bioinformatics Group, Ann & Robert H. Lurie Children’s Hospital of Chicago, IL.

References

1. Hussen BM, Abdullah ST, Salibi A, et al. The emerging roles of NGS in clinical oncology and personalized medicine. Pathol Res Pract 2022;230, 153760. https://doi.org/10.1016/j.prp.2022.153760.
2. Levy SE, Myers RM. Advancements in next-generation sequencing. Annu Rev Genomics Hum Genet 2016;17:95-115. https://doi.org/10.1146/annurev-genom-083115-022413.
3. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell 2015;58(4):586-597. https://doi.org/10.1016/j.molec.2015.05.004.
4. Clark MM, Stark Z, Farah L, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genomic Med 2018;3(1):1-10. https://doi.org/10.1038/s41525-018-0053-8.
5. Retierer K, Jumula J, Chot MT, et al. Clinical application of whole-exome sequencing across clinical indications. Genet Med 2016;18(7):696-704. https://doi.org/10.1038/gim.2015.148.
6. Farwell KD, Shalminzadi L, El-Khechen D, et al. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 selected families with undiagnosed genetic conditions. Genet Med 2015;17(7):578-586. https://doi.org/10.1038/gim.2015.145.
7. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 2013;369(10):1502-1511. https://doi.org/10.1056/NEJMoa1306555.
8. Koblodt DC. Best practices for variant calling in clinical sequencing. Genome Med 2020;12(1):1-13. https://doi.org/10.1186/s13073-020-00791-w.
9. Horton RH, Luccassen AM. Recent developments in genetic/genomic medicine. Clin Sci 2019;133(5):697-708. https://doi.org/10.1042/CS20180436.
10. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. Nature 2011;470(7333):204–213. https://doi.org/10.1038/nature09764.

11. Gillissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet 2012;20(5):490–497. https://doi.org/10.1038/ejhg.2011.258.

12. Lam HY, Clark MJ, Chen R, et al. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 2012;30(1):78–82. https://doi.org/10.1038/nbt.2065.

13. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 2011;12(9):628–640. https://doi.org/10.1038/nrg3046.

14. Beaudet AL, Tsui LC. A suggested nomenclature for designating mutations. Hum Mutat 1993;2(4):245–248. https://doi.org/10.1002/humu.1380020402.

15. Beutler E. The designation of mutations. Am J Hum Genet 1993;53(3):783.

16. Callenberg KM, Santana-Santos L, Chen L, et al. Clinical implementation and validation of automated human genome variation society (HGVS) nomenclature system for next-generation sequencing–based assays for cancer. J Mol Diagn 2018;20(5):628–634. https://doi.org/10.1016/j.jmoldx.2018.05.006.

17. Li MM, Datto M, Duncavage EJ, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn 2017;19(1):4–23. https://doi.org/10.1016/j.jmoldx.2016.10.002.

18. Donnen AJ, Dalgleish R, Maglott DR, et al. HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat 2016;37(6):564–569. https://doi.org/10.1002/humu.22983.

19. 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(5):405–423. https://doi.org/10.1038/gim.2015.30.

20. Samuels DC, Yu H, Guo Y. Is it time to reassess variant annotation? Trends Genet 2016;32(9):423–428. https://doi.org/10.1016/j.tig.2016.08.002.

21. Seifid Dafti MJ, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. Biotechniques 2017;62(1):18–30. https://doi.org/10.2144/000144912.

22. Chakravorty S, Hegde M. Gene and variant annotation for Mendelian disorders in the era of advanced sequencing technologies. Annu Rev Genomics Hum Genet 2017;18:229–256. https://doi.org/10.1146/annurev-genom-083115-022545.

23. Yen JH, Garcia S, Montana A, et al. A variant by any name: quantifying annotation discordance across tools and clinical databases. Genome Med 2017;9(1):1–14. https://doi.org/10.1186/s13073-016-0396-7.

24. McCarthy DJ, Humburg P, Karapin A, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Med 2014;6(3):1–16. https://doi.org/10.1186/gm543.

25. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnPEFF: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 2012(6):80–92. https://doi.org/10.1116/fly.2016.10.19695.

26. McLaren W, Gil L, Hunt SE, et al. The ensemble variant effect predictor. Genome Biol 2016;17(1):1–14. https://doi.org/10.1186/s13059-016-0974-4.

27. Huang EJ, Lee CC, Chiu LY, et al. VAReporter: variant reporter for cancer research of massive parallel sequencing. BMC Genomics 2018;19(2):1–11. https://doi.org/10.1186/s12864-018-4468-5.

28. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(16):e164. https://doi.org/10.1093/nar/gkq603.

29. Alamut® Batch. Interactive Biosoftware, France. https://www.interactive-biosoftware.com/alamut-batch/.

30. Hart RK, Rico R, Hare E, Garcia J, Westbrook J, Funaro VA. A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. Bioinformatics 2015;31(2):268–270. https://doi.org/10.1093/bioinformatics/btu630.

31. Lu X, Li C, Mou C, Dong Y, Tu Y, dNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 2020;12(1):1–8. https://doi.org/10.18664/jim.20654.

32. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol 2011;29(1):24–26. https://doi.org/10.1038/nbt.1754.

33. Wildeman M, Van Ophuizen E, Den Dunne JT, Tsuchner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 2008;29(1):6–13. https://doi.org/10.1002/humu.20654.

34. Eichler EE, Nickerson DA, Altschuler D, et al. Completing the map of human genetic variation: a plan to identify and integrate normal structural variation into the human genome sequence. Nature 2007;447(7141):161. https://doi.org/10.1038/nature05716.

35. Varga E, Chao EC, Yeager ND. The importance of proper bioinformatics analysis and clinical interpretation of tumor genomic profiling: a case study of undifferentiated sarcoma and a constitutional pathogenic BRCA2 mutation and an MLH1 variant of uncertain significance. Familial Cancer 2015;14(3):481–485. https://doi.org/10.1007/s10439-015-9760-3.

36. Posey JE, O’Donnell-Luria AH, Chong JX, et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet Med 2019;21(4):798–812. https://doi.org/10.1038/s41436-018-0408-7.