Supplementary Information for:

BISQUE: locus- and variant-specific conversion of genomic, transcriptomic, and proteomic database identifiers

Michael J. Meyer[1,2,3,†], Philip Geske[1,2,†], and Haiyuan Yu[1,2,*]

[1]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, 14853, USA

[2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, New York, 14853, USA

[3]Tri-Institutional Training Program in Computational Biology and Medicine, New York, New York, 10065, USA

[†]The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors

[*]To whom correspondence should be addressed. Tel: 607-255-0259; Fax: 607-255-5961; Email: haiyuan.yu@cornell.edu

Contents

Supplementary Note 1: Extended Methods

1.1 Database dependencies

BISQUE is decoupled from runtime dependencies on other databases, and does not owe its allegiance to any one database convention. It must nevertheless rely on mappings provided by these databases to their peers. A useful mapping table contains two parts: the actual correspondence of identifiers in two databases and the biological sequences to which these identifiers correspond. Together, these two pieces of information provide much of the basis required to convert in both directions between the identifiers.

When both identifier correspondence tables and sequences are available, MySQL tables were constructed to map these values between databases and a template conversion algorithm was customized to convert loci and variants along this new edge in the BISQUE conversion graph (Figure 1a).

BISQUE stores parsed versions of files at the following URLs in a MySQL database:

UniProt (http://www.uniprot.org/)

- ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping_selected.tab.gz
- ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping.dat.gz
- ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/HUMAN.fasta.gz

Ensembl (http://www.ensembl.org/)

- ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh38.pep.all.fa
- ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh38.cds.all.fa
- ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/Homo_sapiens.GRCh38.79.gtf.gz

NCBI (http://www.ncbi.nlm.nih.gov/)

- ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2ensembl.gz
- ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/protein/protein.gbk.gz
- ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/RNA/rna.fa.gz
- ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/H_sapiens/GFF/ref_GRCh38.p2_top_level.gff3.gz
- ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b142_GRCh38/ASN1_flat/*
- ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b142_GRCh37p13/ASN1_flat/*

EBI (http://www.ebi.ac.uk/)

- ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/split_xml

We acknowledge that curation practices of BISQUE's source databases may sometimes lead to erroneous results. Since BISQUE is a conversion utility for existing identifiers, it does not attempt to reconcile these errors, which are rare and should be addressed by the databases themselves. However BISQUE will provide the best possible results given the data that is available, and will be updated frequently to incorporate the latest revisions to each external database.

1.2 Database updates

Updates of BISQUE can be automated through scripts available in the GitHub repository (http://github.com/hyulab/bisque) that are designed to rebuild all conversion MySQL tables by re-accessing the latest versions of all source databases. However, due to inevitable changes in file formats and FTP server architectures in the source databases, some manual maintenance needs to be performed before each update. Advanced users of BISQUE will be able to perform this maintenance themselves, and we have released all necessary source code to accomplish manual updates. We are also committed to our own routine updates of BISQUE. We intend to update the GitHub update scripts, the downloadable full and lite versions of the software (and associated MySQL tables for download), and the web server twice each year.

1.3 Identifiers

BISQUE uses identifiers as reported by their associated databases. Many databases include alternate identifier names, however for internal consistency, BISQUE only uses primary names for each. The following identifiers are currently available in BISQUE's conversion graph (command-line and API handles in parantheses):

> **Protein identifiers:** UniProt (uniprot), The Protein Data Bank (pdb), Ensembl Protein (ensp), RefSeq Protein (refp).
>
> **Transcript identifiers:** Ensembl Transcript (enst), RefSeq Transcript (reft).
>
> **Genomic identifiers:** Ensembl Gene (ensg), GRCh37 (grch37), GRCh38 (grch38), dbSNP (dbsnp).

Gene symbols: Gene symbols may be used as input to BISQUE, and will be interpreted as equivalent to the corresponding UniProt protein. This means that associated loci and variants should be made in reference to the UniProt identifier. All other inputs should use systematic database names.

Identifier versions: Some identifiers (most visibly, RefSeq) have associated version numbers appended after the identifiers themselves (i.e., RefSeq transcript NM_000346.3). These version numbers are updated in each new release of the database if the sequence associated with that transcript is updated. BISQUE maintains copies of only the most recent versions of these transcripts, and it is often the case that older transcript versions are forward compatible with the latest database build. Outdated (and therefore potentially erroneous) versions of RefSeq transcripts will be treated as the most recent versions (and notice of the correction will be displayed on the output page). RefSeq transcripts can also be inputted by the user without any version number indicated, in which case BISQUE will assume the latest available version of that transcript. Most other databases (i.e., Ensembl and UniProt) perform versioning of their identifiers, but the version numbers are hidden from the user. For these, BISQUE uses the latest

available versions of all identifiers and their associated sequences and ID mappings to other databases, and does not allow input of version numbers from the user.

1.4 BISQUE conversion architecture

BISQUE's internal database (built from the sources identified in Supplementary Material 2.1), primarily consists of two types of tables: (1) mapping identifiers to their residue sequences, and (2) mapping identifiers in one database to those of another, both within and across molecular class (i.e., gene, transcript, protein). Not all identifiers can be mapped to all other identifiers through direct mappings. Rather, BISQUE maintains a requisite number of available conversion tables between identifiers such that all identifiers can be converted to all other identifiers through one or more conversions. This architecture can be depicted as a graph, wherein nodes represent database identifiers and edges represent bidirectional conversions that may take place between identifiers (Figure 1a). By traversing this conversion graph, BISQUE can produce conversions between any identifiers contained in its database.

1.5 Determining optimal paths in the conversion graph

Although BISQUE is capable of traversing any user-defined path in the conversion graph (command-line only), optimal paths from input to output databases are chosen for users of the web server. In defining these paths, we have attempted to avoid spurious results by minimizing the number of steps in each traversal, avoiding descending and climbing the graph in a single traversal, and minimizing the number of databases used. For instance, when converting from Ensembl Gene to RefSeq Transcript, BISQUE will traverse the path through Ensembl Transcript rather than through the human genome which could yield results in overlapping genes (i.e., on the opposite genomic strand) that are unrelated to the original query. When possible, a single best path is chosen, however in cases where multiple paths are equally optimal based on the aforementioned criteria, BISQUE combines the results from traversing both paths. For example, when converting from Ensembl Transcript to RefSeq Protein, BISQUE traverses from Ensembl Transcript→***Ensembl Protein***→RefSeq Protein and from Ensembl Transcript→***RefSeq Transcript***→RefSeq Protein. Both paths share the same number of steps, number of databases traversed, and neither path traverses both upstream and downstream, making them equally optimal.

1.6 Locus- and variant-specific conversion

Conversions in BISQUE are performed stepwise along optimal paths in the conversion graph, with the output from one conversion step being fed as input into the next until an entire conversion path has been traversed from input to output. In cases where multiple outputs are produced from a single conversion step (i.e., due to codon degeneracy), all outputs from that step will be presented to the user (if it is the terminal step in a conversion), or fed as multiple inputs into the next conversion algorithm. Conversion algorithms associated with each edge vary throughout the graph, and must take into account different parameters depending on the type of conversion being made. For instance, converting from a genomic locus to transcript locus follows inherently different rules than converting an amino acid substitution to all potential transcript variants. The primary conversion algorithm types are:

Genomic Locus → Transcript

When a genomic locus is provided, it will be checked to see if it exists within an annotated transcript region. In all transcripts, by default, a locus must be specified relative to the transcript's coding sequence (CDS). However, users may also specify loci in relation to full transcript sequences (including 5' and 3' UTRs as annotated by the source database), by selecting the cDNA option through the web interface or command line tool.

To determine the ordinal position of the genomic variant within the transcript, the exons comprising the transcript are joined to create a locus mapping from genomic positions to transcript positions. Genomic positions falling within introns (even those contained between exons in the same transcript) will not map in BISQUE. Both input and output positions are 1-indexed.

Variants given at genomic loci are assumed to be in reference to the sense (+) genomic strand regardless of whether annotated transcripts in that region are derived from the sense or antisense (–) strand. Therefore, when converting variants, BISQUE will return complements of both reference and alternate nucleotides provided as input if the mapped transcript is derived from the antisense strand. If the given reference nucleotide does not match the true reference nucleotide (from the transcript database), the true reference will be assumed and the web server will alert the user that a correction has been made.

Due to performance concerns, it is not possible to convert an entire chromosome without a provided locus to transcripts on the web server (though this can be accomplished through the downloadable command-line tool). Even when a locus is provided, there are many cases in which several transcripts are derived from the same genomic region (occasionally on opposite genomic strands—see Figure 1b). BISQUE will report all transcripts when this is the case.

Transcript → Genomic Locus

The reverse procedure can be performed to convert transcripts and associated annotations to the genome. Transcript loci should be provided as 1-indexed positions within the transcript's coding sequence (or full sequence when the cDNA option is selected). Variants should be provided in reference to the raw transcript sequence (single strand mRNA containing ATG start codon), regardless of the polarity of the genomic strand it was derived from. Returned genomic variants will always be reported on the sense (+) strand.

Gene Conversions

Conversions to and from genes follow the same rules as transcript ↔ genomic locus conversions. The only genes currently in the BISQUE core conversion graph are from the Ensembl database, which maintains genomic regions in both GRCh37 and GRCh38 corresponding to their annotated genes. These regions are strict supersets of Ensembl Transcripts, which allows a simple numerical offsets to account for the differences in annotating loci within genes, transcripts, and genomic regions.

A major difference between gene conversions and transcript conversions is that genes are always referenced on the sense (+) genomic strand. Therefore, loci of genes need to be provided as 1-indexed from the first annotated base in the gene, 5'-most on the sense strand, regardless of the polarity of transcripts derived from this gene. Variants annotated in reference to genes should also match the genomic bases of the sense strand.

Please note that the vast majority of intragenic sequence is non-coding, and therefore loci referenced within these regions will not map to other identifiers besides the genome and are not subject to reference-base checking as BISQUE does not store the sequences outside coding regions as a space-saving measure.

Transcript → Protein

Basic conversion of identifiers is handled by mapping tables derived from the database associated with the protein. Loci are converted by separating the coding sequence of the transcript into codons, with codons 1–*n* in the transcript directly aligning to amino acids 1–*n* in the protein. Variants are converted by replacing the

reference nucleotide with the given alternate nucleotide and observing the effect on the amino acid encoded by the codon.

If a reference amino nucleotide provided by the user does not match the true reference nucleotide (from the transcript database), the true reference will be assumed and the web server will alert the user that a correction has been made. In rare cases where the reference codon does not match the reference amino acid in the converted protein, the protein reference amino acid will be displayed.

Protein → Transcript

Basic conversion of identifiers is handled by mapping tables derived from the database associated with the protein. Loci are converted by separating the coding sequence of the transcript into codons, and aligning amino acids 1–*n* in the protein directly to codons 1–*n* in the transcript. All possible single-nucleotide variants in the transcript codon (3 possible alternate nucleotides × 3 possible positions = 9 potential variants) are assessed for matches to the provided alternate amino acid, and all potential nucleotide variants that could produce the provided amino acid substitution are presented to the user. In cases where a single protein can be mapped to multiple transcripts, the entire procedure is completed for each matching transcript, with all results that contain the given locus and could produce the given substitution presented to the user.

If a reference amino acid provided by the user does not match the true reference amino acid (from the protein database), the true reference will be assumed and the web server will alert the user that a correction has been made. In rare cases where the true reference amino acid does not match the codon in the transcript, the protein reference amino acid is assumed true and the given alternate amino acid is matched against all possible single-nucleotide variants within the codon as given by the transcript database.

Transcript ↔ Transcript

Conversions between transcripts are performed by mapping transcript IDs between databases. Loci mapping between transcripts is performed by referring to annotated nucleotide positions within the genome, leveraging the fact that both transcripts have been pre-aligned to the genome. In rare instances where after locus mapping reference alleles at the given positions do not match, the correct reference allele will be shown for both transcripts and variants will not be possible to map.

Protein ↔ Protein

Conversions between proteins are computed using ID mappings from the protein databases (UniProt mappings for conversions involving UniProt proteins and NCBI mappings between Ensembl and RefSeq). In rare cases wherein protein sequences do not match for matched IDs in different databases, a Needleman-Wunsch-based algorithm, EMBOSS Stretcher[1], is used to align the two sequences to create a mapping between their loci. These are isolated cases of incompatible identifier versions referring to slightly different protein isoforms, or inter-database inconsistencies over which is considered the ‘canonical’ isoform of a protein.

1.7 Conversion quality filtering options

The following options are available via the BISQUE web interface, API, and command-line tool. These are specifically designed to help users choose the most relevant results in cases of many identifier mappings.

Swiss-Prot only (--swissprot)

When this option is selected, conversion output will be filtered to only include UniProt identifiers with the reviewed, Swiss-Prot designation. Additionally, when converting to identifier types other than UniProt, only those identifiers with direct mappings to SwissProt UniProt identifiers will be included. For instance, when converting a genomic locus to Ensembl transcript position, only transcripts that encode SwissProt UniProt proteins will be returned.

Canonical only (--canonical)

When this option is selected, conversion output will be filtered to only include UniProt identifiers representing canonical isoforms (i.e. with the "-1" designation). Additionally, when converting to identifier types other than UniProt, only those identifiers with direct mappings to canonical UniProt isoforms will be included. For instance, when converting a genomic locus to Ensembl transcript position, only transcripts that encode canonical UniProt protein isoforms will be returned.

Calculate alignment scores (--quality)

When this option is selected, BISQUE will calculate the alignment scores between sequences for all identifier-based conversions in the conversion path. Alignment scores are calculated as the fraction of identical residues in a sequence alignment for each step in a conversion path for which the two identifiers have directly comparable sequences (i.e. protein-protein, transcript-transcript, gene-transcript, and transcript-protein). These identity scores are then averaged across all steps in a conversion for which alignments could be performed and are reported on a 0-1 scale (1 being identical).

Alignment identity scores are calculated using EMBOSS Stretcher[1], a Needleman-Wunsch-based algorithm designed to be faster than EMBOSS Needle. All alignments are performed using Stretcher default parameters, protein alignments are performed with the default EBLOSUM62 substitution matrix, and gene and transcript alignments are performed using the default EDNAFULL substitution matrix. For conversion steps that span molecular classes, for instance transcript to protein, nucleotide sequences are converted to amino acid sequences using a standard codon mapping table, and a protein alignment is performed.

Alignment scores for mapping to and from the PDB are performed without Stretcher. Rather, since PDB's only connection to the rest of the conversion graph is through UniProt, the alignment identity score between PDB and UniProt is taken as the number of UniProt residues contained in the PDB structure based on SIFTS residue mappings[2] divided by the total number of residues in the UniProt protein. As there are sometimes many PDB structures with varying degrees of coverage of the same UniProt protein, these alignment scores are very useful to identify which PDB structures contain the largest number of the original UniProt residues.

It should be noted that sequence alignment is never used to determine identifier mappings; these are always determined using mapping tables provided by the source databases (although alignments are used to determine residue mappings for protein-protein alignments as discussed in Supplementary Note 1.6). The alignment score simply provides a metric to assess the compatibility of these identifier mappings.

Supplementary Note 2: Database & Conversion Scope

The set of databases over which BISQUE can perform conversions was selected to best represent the most popular conventions of current large datasets[3-5]. Many datasets are released with annotations to several database conventions, making it highly likely that BISQUE will be useable for the vast majority of applications. We have chosen to focus on biomolecules within these databases that can be converted completely within our conversion framework (i.e., proteins, mRNA transcripts, and coding regions of genes and genomes). Such conversions are more suited to an all-by-all conversion framework as they can begin and end at any node within the conversion network. Furthermore, this ensures a small database footprint, improving BISQUE's speed and portability.

However, we realize that research takes place outside of coding regions of our chosen core set of databases. To ensure that BISQUE is more generally useful, we have designed it to be extensible beyond this core set of identifiers. Experienced programmers will appreciate the ability to *fork* BISQUE on GitHub (http://github.com/hyulab/bisque) and the scripts and documentation we have made available to aid in the integration of more databases (nodes) within the BISQUE conversion graph.

Since BISQUE focuses on mutations with functional effects in all molecular classes, we have also included two databases that fall outside of our core conversion graph, but have high functional relevance: The Protein Databank[6] (PDB) and The Single Nucleotide Polymorphism Database[7] (dbSNP). Maintaining our curation of molecules and regions with functional effects throughout the entire conversion graph, we curate missense SNPs in dbSNP (~ 1.3 million) and regions of PDB structures that map to UniProt proteins[2, 8].

Supplementary Note 3: Comparison with other conversion utilities

Part of the reason, we believe, that there are few off-the-shelf options for locus and variant-specific conversion is that many online data analysis tools compute these conversions on a case-by-case basis. For instance, predictors of variant function, which are not designed to act as standalone conversion tools, still must maintain some *under-the-hood* conversion capability if they are to be user-friendly and accept data annotated in variety of conventions. One popular functional SNP predictor, PolyPhen-2[9], allows a variety of inputs, but ultimately converts all of these to UniProt in order to perform the multiple sequence alignments that contribute to its classifier. Other functional annotation tools such as snpEff[10] and VAT[11] also convert downstream starting from chromosome variants in VCF files to variants in transcripts and proteins (Supplementary Table 1).

However the conversion frameworks underlying these tools and others are limited to converting in one direction, from many types of inputs into a single type of output, not freely among all identifiers. For instance, popular variant effect predictor VEP[12], while it provides built-in conversion capability to provide a better user experience, is limited to a nucleotide-centric form of variant conversion. Since it assesses variant effects at the nucleotide level, this is not an issue for the functionality of VEP, however, it means that it cannot handle conversion from one protein identifier to another (i.e. Ensembl Protein to UniProt) when there is more than one possible nucleotide mutation encoding an amino acid substitution. For a non-nucleotide-centric conversion utility, such as BISQUE, these conversions are handled without passing through a nucleotide identifier at all.

BISQUE also provides the ability to perform reverse mappings of variants (i.e. amino acid substitutions to nucleotide mutations). While this functionality may be less often used than forward mapping, it is still addresses some practical scientific needs. For instance, reverse mappings are required for designing PCR primers to introduce specific amino acid substitutions in a protein expression system by altering the nucleotides of the encoding transcript. Such a task is not suitable for nucleotide-centric variant annotation software given the potential multiplicity of nucleotide mutations giving rise to the same amino acid substitution. BISQUE simply returns all potential nucleotide substitutions, which a researcher may filter based on their criteria for primer design.

Many biological resources do require external conversion for custom queries. For instance, a biologist may be interested in determining the precise 3D locations of UniProt amino acids affected by non-synonymous SNPs in X-ray crystallographic protein structures in the PDB[6] (to assess for proximity to other variants, binding domains, etc.). Alternatively, a biologist having identified a residue of interest in a PDB structure may wish to determine its coding location in the human genome. The PDB does not currently have an inbuilt method to convert between genomic variants and amino acid substitutions, nor should it as its developers cannot possibly anticipate all of the potential uses of their data. BISQUE removes the onus of developing conversion utilities from databases like the PDB, which are likely to have little interest in maintaining conversion utilities in addition to their own data.

Resource	URL	Type	Input Vehicle / Type	Outputs	Variant- or locus-specific conversion?
UniProt[8]	http://www.uniprot.org	Proteomic Database	Web form & service; UniProtKB, Ensembl, RefSeq, etc.	UniProtKB*	No
DAVID[13]	http://david.abcc.ncifcrf.gov	Bioinformatics Suite	Web form & service; UniProtKB, Ensembl, RefSeq, Microarray (Agilent, Affy), etc.	Same as inputs	No
Synergizer[14]	http://llama.mshri.on.ca/synergizer/translate	Database ID Conversion	Web form; UniProtKB, Ensembl, RefSeq, PDB, etc.	Same as inputs†	No
PICR[15]	http://www.ebi.ac.uk/Tools/picr	Protein ID Conversion	Web form; UniProtKB, Ensembl, RefSeq, etc.; amino acid sequences	Protein Identifiers: UniProtKB, PDB, etc.	No
Ensembl Biomart[16]	http://www.ensembl.org/biomart	Bulk Ensembl Data Retrieval	Web form & API; Ensembl Identifiers	UniProt, RefSeq, Genebank, etc.	No
PolyPhen-2 [9]	http://genetics.bwh.harvard.edu/pph2	Functional Annotation	Web form; Proteins and SNPs; amino acid sequences	UniProtKB, Annotation	Yes‡ to UniProtKB
SnpEff[10]	http://snpeff.sourceforge.net	Functional Annotation	Downloadable Tool; VCF files§	Annotated VCF files§	Yes‡ to Ensembl transcript/gene and amino acid subs
VAT[11]	http://vat.gersteinlab.org/index.php	Functional Annotation	Downloadable Tool; VCF files§	Annotated VCF files§	Yes‡ to Ensembl transcript/gene and amino acid subs
Condel[17]	http://bg.upf.edu/fannsdb/query/condel	Functional Annotation	Web form; UniProtKB, Ensembl, GRCh37	Text tables among all input databases	Yes, among noted databases
VEP[12]	http://www.ensembl.org/Tools/VEP	Functional Annotation	Web form and API; UniProtKB, Ensembl, RefSeq, etc.	Same as inputs	Yes, except for protein-protein conversions
Variant Annotation Integrator[18]	https://genome.ucsc.edu/cgi-bin/hgVai	Functional Annotation	Web form; Genomic variants as pgSnp or VCF	Text table of UCSC gene identifiers	Yes, only from genome to UCSC gene identifiers

Supplementary Table 1. Survey of available tools for database identifier conversion and variant or locus-specific conversion coupled with functional annotation.

*UniProt can output as many identifier types as can be input, but either the output or the input identifier must be UniProtKB.
†Varies depending on chosen input type.
‡Conversion only available with added overhead of functional annotation and only in one direction to noted outputs.
§Format for storing genetic variants (chromosomal loci and variations).

References

1. Li, W. et al. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* **43**, W580-584 (2015).
2. Velankar, S. et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research* **41**, 9 (2013).
3. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220 (2013).
4. Abecasis, G.R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
5. Forbes, S. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* **39**, 50 (2011).
6. Berman, H.M. The Protein Data Bank. *Nucleic Acids Research* **28** (2000).
7. Sherry, S. et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311 (2001).
8. UniProt-Consortium UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212 (2015).
9. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010).
10. Cingolani, P. 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 (Austin)* **6**, 80-92 (2012).
11. Habegger, L. et al. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* **28**, 2267-2269 (2012).
12. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070 (2010).
13. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).
14. Berriz, G.F. & Roth, F.P. The Synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics* **24**, 2272-2273 (2008).
15. Cote, R.G. et al. The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* **8**, 401 (2007).
16. Cunningham, F. et al. Ensembl 2015. *Nucleic Acids Res* **43**, D662-669 (2015).
17. Gonzalez-Perez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* **88**, 440-449 (2011).
18. Karolchik, D. et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**, D764-770 (2014).