

Database

Open Access

CASCAD: a database of annotated candidate single nucleotide polymorphisms associated with expressed sequences

Victor Guryev*, Eugene Berezikov and Edwin Cuppen

Address: Hubrecht Laboratory, Netherlands Institute for Developmental Biology, Uppsalalaan 8, 3584CT, Utrecht, The Netherlands

Email: Victor Guryev* - guryev@niob.knaw.nl; Eugene Berezikov - berezikov@niob.knaw.nl; Edwin Cuppen - ecuppen@niob.knaw.nl

* Corresponding author

Published: 27 January 2005

Received: 11 October 2004

BMC Genomics 2005, 6:10 doi:10.1186/1471-2164-6-10

Accepted: 27 January 2005

This article is available from: <http://www.biomedcentral.com/1471-2164/6/10>

© 2005 Guryev et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: With the recent progress made in large-scale genome sequencing projects a vast amount of novel data is becoming available. A comparative sequence analysis, exploiting sequence information from various resources, can be used to uncover hidden information, such as genetic variation. Although there are enormous amounts of SNPs for a wide variety of organisms submitted to NCBI dbSNP and annotated in most genome assembly viewers like Ensembl and the UCSC Genome Browser, these platforms do not easily allow for extensive annotation and incorporation of experimental data supporting the polymorphism. However, such information is very important for selecting the most promising and useful candidate polymorphisms for use in experimental setups.

Description: The CASCAD database is designed for presentation and query of candidate SNPs that are retrieved by *in silico* mining of high-throughput sequencing data. Currently, the database provides collections of laboratory rat (*Rattus norvegicus*) and zebrafish (*Danio rerio*) candidate SNPs. The database stores detailed information about raw data supporting the candidate, extensive annotation and links to external databases (e.g. GenBank, Ensembl, UniGene, and LocusLink), verification information, and predictions of a potential effect for non-synonymous polymorphisms in coding regions. The CASCAD website allows search based on an arbitrary combination of 27 different parameters related to characteristics like candidate SNP quality, genomic localization, and sequence data source or strain. In addition, the database can be queried with any custom nucleotide sequences of interest. The interface is crosslinked to other public databases and tightly coupled with primer design and local genome assembly interfaces in order to facilitate experimental verification of candidates.

Conclusions: The CASCAD database discloses detailed information on rat and zebrafish candidate SNPs, including the raw data underlying its discovery. An advanced web-based search interface <http://cascad.niob.knaw.nl> allows universal access to the database content and allows various queries supporting many types of research utilizing single nucleotide polymorphisms.

Background

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation within species. As a

result, SNPs are now becoming the most popular type of marker in genetic association and mapping studies. SNPs are also most likely to be the molecular basis for the

majority of phenotypic variation in (outbred) populations. In particular, SNPs in regulatory and protein-coding regions can have an effect on gene expression levels and protein activity, respectively. The phenotypic differences observed between selected (sub) strains in model organisms may be the result of specific (combinations of) natural occurring polymorphisms. Hence, a comprehensive inventory of SNPs, including extensive annotation will be extremely valuable in the search for functional polymorphisms.

There is often a vast unexplored potential in large sequence datasets that have been collected for other purposes, for example, EST and whole genome sequencing (WGS) projects. In an effort to address these two issues, we have developed an *in silico* candidate SNP mining pipeline that uses all publicly available sequence data for a specific organism, and designed a database, CASCAD (CASCAD SNP CANDIDATES Database), that allows storage of a wide variety of primary source data, cross-annotation to other databases, and analysis parameters for SNPs associated with expressed sequences.

Construction and content

We applied the SNP discovery pipeline to both rat [1] and zebrafish (unpublished results) and identified about 33,000 and 52,000 high-quality candidates, respectively, that were extensively annotated and stored in the CASCAD database (Table 1). The database includes detailed primary information on which the discovery of the candidate SNPs was based. This information, including sequence quality information (Phred score), the number of supporting reads for every nucleotide observed at the SNP position, and expected alignment lengths, was found to be very valuable for filtering for predicted variants that have the highest likelihood to be experimentally confirmed. Two verification experiments for the rat resulted in confirmation rate estimates of 59% (68 candidate SNPs in 10 different laboratory rat strains) and 50.3 % (340 candidates in 5 laboratory and 2 wild rat isolates) [1]. A set of 139 CASCAD entries tested in 7 zebrafish isolates confirmed 67.6% of them as true polymorphisms (unpublished data). The success rate values obtained are likely to be underestimates since only limited number of isolates/samples were used and we were unable to include exactly the same isolates that were used for generating the primary data (e.g. EST sequencing).

We designed a web-based interface with Perl scripts communicating to a MySQL database, and displaying HTML pages through Apache server running on SuSE Linux. The interface provides simple, advanced (Figure 1), and sequence-based search forms. Parameters that can be used in a search include strain information, different formats of sequence identifiers (e.g. GenBank, UniGene, LocusLink

Table 1: Input data (number of sequence reads) for the CASCAD pipeline and number of predicted candidate SNPs.

	<i>Rattus norvegicus</i>	<i>Danio rerio</i>
Input data		
mRNA	25, 634	3, 366
EST	244, 518	283, 572
WGS	19, 813, 313	11, 588, 394
Candidate SNPs predicted	33, 305	51, 769
synonymous	3, 842	9, 111
non-synonymous	3, 708	6, 217
nonsense	162	158

accessions or gene symbol), map positions (genetic and physical), and a wide variety of SNP characteristics, such as experimental evidence, a likelihood score for verification as deduced from extensive verification experiments [1], and information regarding restriction sites that have been affected, facilitating the design of RFLP based assays. To this end CASCAD is tightly linked to primer design [2] and local genome assembly [3] interfaces, enabling a fast, reliable, and universal primer design for a chosen SNP candidate even when there is no assembled genome sequence available. To facilitate the retrieval of candidate SNPs with higher verification rates, represented by more reliable and common SNP candidates, we implemented the possibility to restrict searches to entries characterized by location at hypervariable CpG site, by minimal base-calling quality (Phred score), sequence match size, and/or minimal number of supporting reads for either allele.

In addition to primary sequence data analysis, the effect of all SNPs on protein coding capacity was evaluated and non-synonymous SNPs were categorized in classes reflecting the severity of the polymorphism using a BLOSUM-based score. The predicted missense SNPs were analyzed by SIFT [4] and Polyphen [5] programs that utilize not only substitution information but also phylogenetic conservation and structural protein information to predict a potential effect of the polymorphism on protein function.

Query results are summarized on the SNP details page (Figure 2), listing the SNP characteristics and including active links to other databases and resources, such as dbSNP, Ensembl, UniGene, and LocusLink. More detailed information regarding raw data underlying the candidate SNP (links to the original sequence files and a full nucleotide and protein alignment) can be obtained by clicking on the observed nearly exact hits between nucleotide sequences. Moreover, statistics on the data that support the SNP (number of occurrences for every nucleotide at

CASCAD advanced search - Rattus norvegicus

Strain / data partition

AND Strain: SD

AND Comparison: BNxSD

AND Partition: EST x WGS

Ids / accessions

AND CASCAD ID

AND GenBank accession

AND UniGene ID

AND LocusLink ID

AND Gene symbol

Map / Position

NOT Chromosome: 15

OR RH panel: RH22 (cR)

AND Map position, from

AND Map position, to

AND Position in genome assembly, from

AND Position in genome assembly, to

SNP characteristics

NOT Substitution type: Transition

AND 'BLOSUM' score at least: 5

AND SNP category at least: 7

NOT SIFT prediction: Tolerated

AND PolyPhen prediction: Probably damaging

AND CpG site: not at CpG site

AND Restriction site affected

AND Sequence quality at least

AND Observed alignment length at least

AND Observed / expected alignment length ratio at least

AND Identity score at least

AND Supported by # of reads: Multiple

NOT Experimentally checked: Not confirmed

Options

Limit output entries amount to

Sort by: Alignment length

mRNA treated as high quality?: No

Display: Results

Submit

Figure 1
CASCAD advanced search form.

CASCAD SNP 001876 - Rattus norvegicus

Stats

Strain (partition): SDxSD(EST x EST), SDxBN(EST x trace)
 nt replacement: S
 SNP basecalling quality: 51
 Supporting reads: 2 (+)
 CpG dinucleotide: yes (+)
 Restriction sites affected: BsiYI; NlaIII; CdiI;
 AA replacement: D<->H (6)
 PolyPhen prediction: probably damaging (2.0250)
 SIFT prediction: intolerant (0.0000)

Links to other resources

Gene name: *Apba2*
 dbSNP ID: rs8146038
 Ensembl geneID: ENSRNOG00000016358
 UniGene ID: 62687
 LocusLink ID: 83610

Map data

Map:
 Position in genome assembly: 01:119177118, (snpview)

Status

Validated: Not assayed
 Maximal identity score: 81.56
 SNP category (0..9): 5

Observed hits (click to get complete info)

001876.1 001876.10 001876.11 001876.2 001876.3 001876.4

Figure 2
SNP details page

SNP position, range of Phred basecalling quality scores) are provided.

Utility and discussion

For many applications, it is important to be able to distinguish between SNP candidates by their characteristics, as they may be predictive for verification success rate or carry biologically relevant information. Non-confirmed candidate polymorphisms may represent variants uncommon for a given population, but also sequencing errors (all types of sequences), RNA editing events and reverse transcriptase errors (EST reads). In order to minimize the contribution of false positives, one can exclude polymorphisms based on a single read for either allele, as is common for many *in silico* discovery pipelines [6]. Although this is a valid approach when selecting SNPs for population or association genetics, one could inadvertently discard many rare variants that may be associated with phenotypic variation, for example by affecting protein structure or function. Information on such poly-

morphisms can be very useful when mapping disease or QTL alleles.

We have developed our database to fulfill the needs of any particular SNP application by providing control over every parameter we used in the polymorphism discovery step.

Applications of the CASCAD database include queries for potentially deleterious SNPs in a specific genomic region of interest, for example a QTL interval, design of SNP-based mapping panels using either RFLP or any other technology, and identification of informative SNPs for fine-mapping. Custom sequences can be provided to search for known SNPs in any sequence of interest. In addition, the CASCAD pipeline [1] can be used to build a candidate SNP database for any model organism of interest for which sufficient sequencing data is available.

Conclusions

The main purpose of CASCAD database is to provide flexible access to candidate single nucleotide polymorphisms, which were predicted using a computational approach from publicly available sequence data of the rat and zebrafish. The resulting database is crosslinked to most common public databases and can be queried for SNPs using accession numbers, sequence context, SNP characteristics, but also using parameters specific to the SNP discovery process, allowing stringent or relaxed conditions suitable for different types of applications.

Availability and requirements

The database is freely accessible through the website <http://cascad.niob.knaw.nl>. Programs, scripts, MySQL database dumps, and instructions for setting up a species-specific SNP database can be obtained from the authors upon request.

Authors' contributions

VG designed and implemented the CASCAD database. EB tested database and interface. EC provided supervision and guidance for the project.

Acknowledgements

This work was supported by the Dutch Ministry of Economic Affairs through the Innovation Oriented Research Program on Genomics, grant #IGE010017.

References

- Guryev V, Berezikov E, Malik R, Plasterk RHA, Cuppen E: **Single nucleotide polymorphisms associated with rat expressed sequences.** *Genome Res* 2004, **14**:1438-1443.
- Primer design interface** [<http://primers.niob.knaw.nl>]
- Berezikov E, Plasterk RHA, Cuppen E: **GENOTRACE: cDNA-based local GENOME assembly from TRACE archives.** *Bioinformatics* 2002, **18**:1396-1397.
- Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31**:3812-3814.
- Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: Server and survey.** *Nucleic Acids Res* 2002, **30**:3894-3900.
- Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino MB: **Mining SNPs from EST databases.** *Genome Res* 1999, **9**:167-174.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

