BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 11 2009, pages 1419–1421
doi:10.1093/bioinformatics/btp166

Genome analysis

CNVVdb: a database of copy number variations across vertebrate genomes

Feng-Chi Chen[1], Yen-Zho Chen[2] and Trees-Juen Chuang[2,*]

[1]Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli County 350 and [2]Genomics Research Center, Academia Sinica, Taipei 11529, Taiwan

Received on December 3, 2008; revised on March 2, 2009; accepted on March 17, 2009

Advance Access publication March 25, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: CNVVdb is a web interface for identification of putative copy number variations (CNVs) among 16 vertebrate species using the-same-species self-alignments and cross-species pairwise alignments. By querying genomic coordinates in the target species, all the potential paralogous/orthologous regions that overlap ≥80–100% (adjustable) of the query sequences with user-specified sequence identity (≥60%∼≥90%) are returned. Additional information is also given for the genes that are included in the returned regions, including gene description, alternatively spliced transcripts, gene ontology descriptions and other biologically important information. CNVVdb also provides information of pseudogenes and single nucleotide polymorphisms (SNPs) for the CNV-related genomic regions. Moreover, multiple sequence alignments of shared CNVs across species are also provided. With the combination of CNV, SNP, pseudogene and functional information, CNVVdb can be very useful for comparative and functional studies in vertebrates.

Availability: CNVVdb is freely accessible at http://CNVVdb.genomics.sinica.edu.tw.

Contact: trees@gate.sinica.edu.tw

Table 1. Currently available target/subject species in CNVVdb

Target	Subject
Human	Human, chimpanzee, orangutan, rhesus macaque, dog, mouse, rat, horse, cow, opossum, chicken, zebrafish, tetraodon, fugu, stickleback, medaka
Chimpanzee	Chimpanzee, human, orangutan, rhesus macaque, mouse, rat, opossum, chicken, zebrafish
Rhesus macaque	Rhesus macaque, human, chimpanzee, orangutan, mouse, rat
Mouse	Mouse, human, chimp, orangutan, rhesus macaque, rat, horse, dog, cow, opossum, chicken, tetraodon, fugu, stickleback, medaka, zebrafish
Rat	Rat, human, chimp, rhesus macaque, mouse, dog, horse, cow, opossum, chicken, zebrafish
Dog	Dog, human, mouse, rat, horse, cow
Chicken	Chicken, human, orangutan, mouse, rat, horse, opossum, zebrafish, fugu
Stickleback	human, mouse, chicken, zebrafish, tetraodon, fugu, medaka

1 INTRODUCTION

Genomic copy number variations (CNVs) have been demonstrated to be relevant to alterations of gene expression and translation levels (Johnson *et al*., 2001; McLysaght *et al*., 2003), which may result in significant phenotypic changes or functional differences both within and between species. Therefore, genome-wide identification and comparative analyses of CNVs may further our understanding of the mechanisms of genome evolution and functional divergence. Notwithstanding the existence of several web interfaces for human intra-species CNV analysis (Arner *et al.*, 2007; Bruford *et al*., 2008; Chen *et al*., 2008), as our understanding, no database is currently available for genome-wide identification of inter-species CNVs.

Here, we present a database (‘CNVVdb’) to provide potential inter-species CNV information by finding duplicated regions within a genome (paralogues) and between different genomes (orthologues). The paralogues/orthologues are inferred from pairwise sequence alignments between 16 vertebrate species (Table 1). Note that the CNVs referred here are not equal to copy number polymorphisms, which are defined as variations of copy numbers in >1% of a population (Lee *et al.*, 2008). Further experimental evidence is needed to support the polymorphism status of the inferred CNVs in this database. In CNVVdb, the compared species should include one target species (of which the genomic regions serve as query sequences) and at least one subject species (of which the genomes are searched against). The eight target species are selected into CNVVdb because the self-alignments and cross-species pairwise sequence alignments are both available for these species. Therefore, the interface can identify not only orthologous sequences between the target and the subject species but also paralogues in the target species. The differentiation of genomic copy numbers between the target and the subject species is evolutionarily important. If the genomic copies encompass genes, the gene family sizes may be different among the compared species, potentially giving rise to functional divergences. Therefore, the information of well-annotated genes (Ensembl) that overlap with the identified paralogous/orthologous regions is also provided. Additional information of these genes is also given, including gene description, alternatively spliced transcripts/isoforms, GC content, expression data (from Unigene), Gene Ontology (GO) descriptions, pathway information and protein domain descriptions. For human

[*]To whom correspondence should be addressed.

genes, the disease association information is also provided. Since duplicated gene copies may be subject to relaxed selection pressure and probably result in pseudogenes, information of pseudogenes located in the indentified paralogues/orthologues is also displayed. Moreover, information of single nucleotide polymorphisms (SNPs) that are included in the CNV-related genomic regions is also provided. For the human CNV-related genomic regions, the polymorphism data derived from two published human individual genomes [the Venter (Levy *et al*., 2007) and Watson (Wheeler *et al*., 2008) genomes] are also provided. The previously published evidence that supports genomic CNVs between human individuals is also collected. In addition, CNVVdb also provides multiple sequence alignments of all paralogues/orthologues identified.

2 MATERIALS AND METHODS

Based on the user-specified genomic regions in the genome of the target species, the system extracts paralogous/orthologous regions from the UCSC (University of California Santa Cruz) target species self-alignments and the UCSC target–subject species orthologous alignments [both alignments are chained Blastz alignments (Schwartz *et al*., 2003)], respectively. The corresponding coordinates of the paralogous/orthologous regions are calculated. Based on the coordinate pairs and the UCSC alignments, the 2 nt sequences compared are extracted from the corresponding genomes and the alignable sequences are displayed. For accuracy, we only consider the paralogues/orthologues that overlap ≥80% of the query sequence and the sequence identity of the alignable regions are >60%. As well, to reduce the noises caused by uncharacterized transposons and retroelements, we limit the query sequences to ≥1 kb in size.

The target–subject species orthologous alignments, human/dog self-alignments, UCSC-annotated transcripts and RefSeq transcripts were downloaded from the UCSC genome browser (the browser provides a wealth of data, such as the reference sequences, annotation databases, alignments and so on) at http://genome.ucsc.edu/. The self-alignments of the other six target species were provided directly by the system manager of the UCSC genome browser. The gene annotation, GO descriptions, expression information, genomic sequences and SNP data were all downloaded from the Ensembl genome browser (release 49) at http://www.ensembl.org/. The polymorphism data of the human reference genome versus the Venter and Watson genomes were downloaded from the corresponding sites at ftp://ftp.jcvi.org/pub/data/huref/ and ftp://ftp.hgsc.bcm.tmc.edu/pub/uers/wheeler/, respectively. The protein domain descriptions [Interpro (Mulder *et al*., 2007), SMART and PFAM], the KEGG pathways (Kanehisa *et al*., 2008) and the disease association information [OMIM, HIV interaction, and the Genetic Association Database (Becker *et al*., 2004)] were downloaded from the DAVID knowledgebase (Huang da *et al*., 2007) at http://david.abcc.ncifcrf.gov/. The experimental data supporting the existence of genomic CNVs between human individuals were downloaded from the Database of Genomic Variants (Iafrate *et al*., 2004) at http://projects.tcag.ca/variation/ and the CNV project (including the WGTG, 500KEA and Redon CNVs) at http://www.sanger.ac.uk/humgen/cnv/redon2006/cnv_data/. Duplications identified by the WGAC (whole-genome assembly comparison) approach (Bailey *et al*., 2001) were downloaded from http://eichlerlab.gs.washington.edu/database.html. The pseudogenes were identified using the PseudoPipe program (Zhang *et al*., 2006). The multiple sequence alignments were generated using the MUSCLE package (Edgar, 2004a, b).

3 WEB INTERFACE

The users may search for CNVs across vertebrates using genomic coordinates (single/multiple regions) in the target species. Two query parameters are adjustable: the proportion of overlapped regions (≥80–100%) and the level of sequence identity (≥60%∼≥90%) between the query sequence(s) in the target species and the matched sequence(s) in the subject species.

The output of CNVVdb includes four main parts: information summary about the query region in the target species, information summary about the matched region(s) in the subject species, detailed information for each matched region and multiple sequence alignments of the identified paralogues/orthologues. For Parts I and II, the CNVVdb system provides rich information for the identified paralogous/orthologous regions for all of the target/subject species, including information of well-annotated genes (transcripts/proteins), duplicated/processed pseudogene, CNV data (including experimental evidence and *ab initio* predictions) collected from other databases and SNP data. For well-annotated genes/transcripts, more detailed information is provided, such as protein domains, GO descriptions, pathways involved, associations with human diseases and so on. Part II also provides a genome-wide overview of the CNV distribution for each subject species. The user can click on the species name to view the corresponding graph. For Part III, the user can select one of the subject species and then select one of the identified paralogues/orthologues for more details. The pairwise sequence alignments between the query and the matched sequences are given, in which the genic regions are highlighted. All these results are downloadable. The multiple sequence alignments among the query and all matched sequences (paralogues) in the specified species are also provided here. Finally, Part IV allows the users to download all paralogous/orthologous sequences identified and perform multiple sequence alignments (‘Multiple_Alignment’) among these sequences. The interface will be updated every 3 months.

4 DISCUSSION

CNVVdb identifies potential CNVs across vertebrate genomes using paralogous/orthologous sequence alignments. CNVVdb is thus very suitable for studies of comparative CNVs and species-specific duplications in vertebrates. In addition to identification of paralogous and orthologous regions, CNVVdb also provides information of SNPs for the CNV-related regions. The information may facilitate population genetics and evolutionary studies of the vertebrate species, especially for those that currently have considerable SNP data. Moreover, the identified paralogues/orthologues can be useful for the identification of copy number polymorphisms because such polymorphisms are suggested to be strongly associated with orthologous genomic regions and intrachromosomal segmental duplications (Perry *et al*., 2008).

CNVVdb can be applied to a number of different research fields. For gene-focused CNV studies, the combination of the identified paralogues/orthologues and pseudogenes can facilitate studies of gene gain/loss events, and provide useful information regarding gene family expansions/contractions across species. Meanwhile, for gene annotation, potentially unannotated protein-coding sequences may be submitted to the CNVVdb to search for matches with gene-containing CNVs or pseudogenes in the vertebrate genomes. Moreover, since pseudogenes are genomic fossils and can be regarded as a virtual resource of ancient transcripts (Huang *et al*., 2008; Shemesh *et al*., 2006), the pseudogene-related CNVs are useful for studies in alternatively spliced variants and transcriptome divergence across species. With CNVVdb, the users can also perform comparative functional analyses by combining CNV data and the functional information, such as GO/protein domain descriptions and KEGG pathway information, to probe the possible association between CNVs and functional divergences across species. For human CNV-related genes, the disease associations are also provided. Such information is potentially useful for exploring the evolutionary origin of human diseases.

ACKNOWLEDGEMENTS

We thank the system manager of the UCSC Genome Browser for providing Chimpanzee, Macaque, Mouse, Rat, Chicken and Stickleback self-alignments.

Funding: National Health Research Institutes (NHRI), Taiwan (NHRI-EX98-9408PC); National Science Council, Taiwan (NSC 96-2628-B-001-005-MY3) (above to T.-J.C.); NHRI intramural funding (to F.-C.C.).

Conflict of Interest: none declared.

REFERENCES

Arner,E. *et al.* (2007) Database of Trypanosoma cruzi repeated genes: 20,000 additional gene variants, *BMC genomics*, **8**, 391.

Bailey,J.A. *et al.* (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.

Becker,K.G. *et al.* (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.

Bruford,E.A. *et al.* (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.*, **36**, D445–D448.

Chen,P.A. *et al.* (2008) CNVDetector: locating copy number variations using array CGH data. *Bioinformatics*, **24**, 2773–2775.

Edgar,R.C. (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar,R.C. (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Huang da,W. *et al.* (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, **35**, W169–W175.

Huang,Y.T. *et al.* (2008) Identification and analysis of ancestral hominoid transcriptome inferred from cross-species transcript and processed pseudogene comparisons. *Genome Res.*, **18**, 1163–1170.

Iafrate,A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Johnson,M.E. *et al.* (2001) Positive selection of a gene family during the emergence of humans and African apes. *Nature*, **413**, 514–519.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Lee,S. *et al.* (2008) Quantitative analysis of single nucleotide polymorphisms within copy number variation. *PLoS ONE*, **3**, e3906.

Levy,S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.

McLysaght,A. *et al.* (2003) Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Natl Acad. Sci. USA*, **100**, 15655–15660.

Mulder,N.J. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.

Perry,G.H. *et al.* (2008) Copy number variation and evolution in humans and chimpanzees. *Genome Res.*, **18**, 1698–1710.

Schwartz,S. *et al.* (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.

Shemesh,R. *et al.* (2006) Genomic fossils as a snapshot of the human transcriptome. *Proc. Natl Acad. Sci. USA*, **103**, 1364–1369.

Wheeler,D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

Zhang,Z. *et al.* (2006) PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics*, **22**, 1437–1439.