BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 23 2013, pages 3003–3006
doi:10.1093/bioinformatics/btt534

Genome analysis

Advance Access publication September 12, 2013

CruzDB: software for annotation of genomic intervals with UCSC genome-browser database

Brent S. Pedersen[1,*], Ivana V. Yang[1] and Subhajyoti De[1,2,*]

[1]Department of Medicine, University of Colorado Denver, School of Medicine, Denver, CO, USA and
[2]Molecular Oncology Program, University of Colorado Cancer Center, Aurora, CO, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: The biological significance of genomic features is often context dependent. Annotating a particular dataset with existing external data can provide insight into function.

Results: We present CruzDB, a fast and intuitive programmatic interface to the University of California, Santa Cruz (UCSC) genome browser that facilitates integrative analyses of diverse local and remotely hosted datasets. We showcase the syntax of CruzDB using microRNA binding sites as examples, and further demonstrate its utility with three biological discoveries. First, DNA replication timing is stratified in gene regions—exons tend to replicate early and introns late during S phase. Second, several non-coding variants associated with cognitive functions map to lincRNA transcripts of relevant function, suggesting potential function of these regulatory RNAs in neuronal diseases. Third, lamina-associated genomic regions are highly enriched in olfaction-related genes, indicating a role of nuclear organization in their regulation.

Availability: CruzDB is available at https://github.com/brentp/cruzdb under the MIT open-source license.

Contact: bpederse@gmail.com or subhajyoti.de@ucdenver.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 16, 2013; revised on September 4, 2013; accepted on September 6, 2013

1 INTRODUCTION

Biological significance of many genomic and epigenomic features is context dependent. Recently, large scale integrative projects such as the Encyclopedia of DNA Elements (ENCODE) project (ENCODE, 2012) have systematically analyzed the regions of active transcription, gene regulation and chromatin patterns in the genome. Even though decades of research provided insights into many individual functional elements, integrative analyses have presented a systems-level picture that could not be captured previously. Moreover, these integrative projects have highlighted that biological function of certain features can be appreciated in the context of other genomic and epigenomic features in the genomic neighborhood (Ernst and Kellis, 2013).

Systematic presentation of large-scale datasets from the ENCODE (ENCODE, 2012) and other projects in the University of California, Santa Cruz (UCSC) genome browser (Kent *et al.*, 2002) has enabled individual investigators to analyze their local data in the context of these already available features. Already, we are beginning to see the utility of such a community-wide integration of diverse datasets and their role in uncovering new facets of basic biology and clinical research. Researchers routinely use publicly available data-tables from the ENCODE project and many other large-scale projects from the UCSC genome browser, which also allow programmatic access to much of the information used on that site via its public MySQL servers (Dreszer *et al.*, 2012). Even so, there exists no user-friendly computational framework that allows integration of multiple in-house and publicly available data-tables and parallelized context-dependent analyses of the integrated datasets. Today, in the era of 'the $1000 genome, the $100 000 analysis' (Mardis, 2010), we believe that such a computational framework can increase the speed and efficiency of integrative analyses in many areas of biomedical research.

We present CruzDB, a programmatic interface to the genome data resources from UCSC (Dreszer *et al.*, 2012) that offers a simple, parallelizable and intuitive syntax to address common use-cases including annotation and spatial querying. We first describe the design features of CruzDB, flexibility of the user interface and potential utilities. We present example code from the library and then describe four diverse findings that we made using CruzDB.

2 APPROACH

CruzDB uses the python programming language and sqlalchemy (SQL-alchemy) library to access publicly available data hosted at the UCSC genome browser database (Dreszer *et al.*, 2012). By using sqlalchemy, we are able to wrap the database tables dynamically rather than requiring explicit code for each of the thousands of available tables (10 076 in the hg19 database) for each organism and version-specific database.

Although CruzDB can function using only the remote data from UCSC's MySQL instance, we show that substantial improvements in speed can be achieved from having a local mirror and using built-in parallelization (described in next section). The library contains a suite of tests to ensure correctness. CruzDB requires python 2.6 or 2.7, the MySQL client libraries and the python sqlalchemy library. Installation is available using standard python tools from http://pypi.python.org/pypi/cruzdb or from the source repository at https://github.com/brentp/cruzdb/.

*To whom correspondence should be addressed

3 METHODS

CruzDB simplifies common tasks such as those that return upstream or downstream features, exons, introns, untranslated regions (UTRs) and transcription start sites. Location-based queries can use the UCSC bin column (Kent *et al*., 2002) when available for more efficient queries. The bin column that is present in some of the database tables is used to implement an efficient k-nearest neighbor search for a given feature along with methods to find nearest upstream and downstream neighbors. The query results from each table can be customized, such that, for example, an interval within a CpG island can be annotated with 'island', whereas one that is nearby will be annotated as 'shore'. Other operations include the generation of browser URLs to view a specific feature, the extraction of coding exons and retrieval of the genomic sequence for any of those feature types from the UCSC Distributed Annotation Server. One can also obtain a list of BLAST-Like Alignment Tool (BLAT) (Kent, 2002) hits for a particular feature.

Using CruzDB, it is possible to mirror a subset of tables from UCSC to a local MySQL or SQLite database using a single line of python code. A local copy allows a user to add data that are not in UCSC and then use that new table just as one would use any other table in the database. This expands the utility of our tool to any dataset with a start, end and chromosomal designation. Though it improves the speed of otherwise network-intensive operations, having a local copy is not necessary, and all of CruzDB's features are available on the public MySQL instance, except for those that modify the database.

To further speed up large numbers of queries, we provide a memory-efficient implementation of an interval tree that can be much faster than performing repeated SQL queries. Because all features must be read into memory to create an interval tree, there is a trade-off between the time to read all features into memory versus the time spent querying. That trade-off depends on the number of intervals. Figure 1 shows the comparison between local and remote instances and whether parallelization is used when annotating about 3300 intervals (timing data are available in Supplementary File S1). Note that SQLite is quite fast, even without parallelization; however, the time for repeated queries to the remote (UCSC) MySQL instance can be greatly reduced by reading the entire table into a local interval tree to reduce network back-and-forth. As the number of intervals to annotate increases, so does the speed improvement from reading the intervals into a tree. Some speed improvement may be achieved by modifying MySQL settings; here we have used the default.

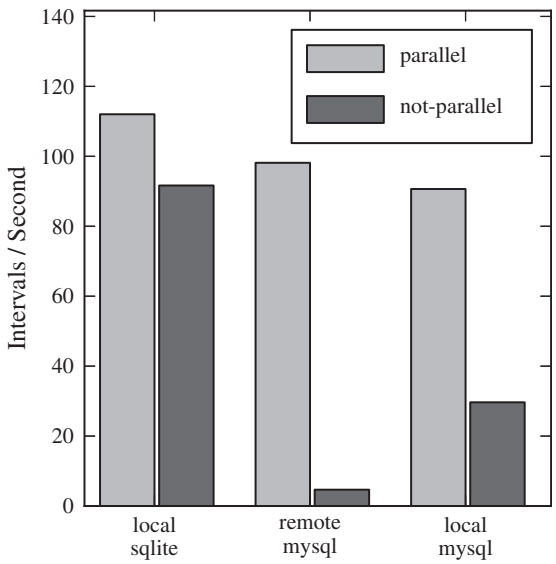

Fig. 1. Intervals annoted per second for a set of about 3300 intervals using a local SQLite, local MySQL or remote (UCSC) MySQL instances for parallel SQL queries (light) or traditional serial queries (dark)

The most common use-case has been to annotate a list of intervals with any table from the UCSC genome-browser database. We provide an interface, by which, with a single command, a user can annotate a file of intervals with a list of tables present in the database. For gene-like tables, the output lists the nearest gene, and whether the interval overlaps an exon, intron, untranslated region or other gene feature.

4 EXAMPLES

4.1 Code example: microRNA (miRNA) targets

Because CruzDB is a library, we show a short code example, using the target-scan database of predicted miRNA targets (Grimson *et al*., 2007) available in the UCSC genome browser as targetScanS. We will walk through the important parts of the code. The full code to perform the analysis is 12 lines (excluding comments) and is available as Supplementary File S2. First, we import the needed libraries:

```
from cruzdb import Genome
from cruzdb.sequence import sequence
```

Then, we mirror the refGene and targetScanS tables from UCSC (version hg19) to a local SQLite database:

```
local=Genome(''hg19'')\
    .mirror((''refGene", ''targetScanS''),
            "sqlite:///hg19.mirna.db'')
```

Now that we have mirrored these tables from the remote UCSC server, they will always be available in the local SQLite database as long as we keep the hg19.mirna.db file. We then iterate over the rows of refGene, where each row is a python object with methods such as 'is_coding'.

```
for gene in (rgene for rgene in
             local.refGene if rgene.is_coding):
```

Inside that loop, we extract the gene's 3′ UTR and search for any miRNA in targetScanS that it overlaps using the efficient bin query:

```
  utr_start, utr_end = gene.utr3
  sites = local.bin_query('targetScanS',
                          gene.chrom,
                          utr_start,
                          utr_end)
```

Still inside the gene loop, we then filter to those sites that contain at least one miR-96 binding site with a score >85 and then print those to a file along with the UTR sequence. We also save the gene name for later gene-ontology analysis:

```
  if any(''miR-96'' in s.name
         and s.score > 85 for s in sites):
      print gene, sequence('hg19', gene.chrom,
                           utr_start, utr_end)
      ref_seq_ids.append(gene.name)
```

After this loop, we will have a file of the genes that have a miR-96 binding site in their 3′ UTR. We can also send the genes to DAVID (Huang *et al*., 2009) in a single command:

```
Genome.david_go(refseq_ids)
```

This will open a genome browser window with the genes loaded into DAVID. Even with this short example, we identify relationships that are biologically plausible. We know that miR-96 is associated with hearing loss (Menca *et al*., 2009); when we look at the ontology enrichment from DAVID (Supplementary File S3), we see terms associated with synapses and cell junction which are, in turn, known to be associated with deafness and hearing loss (Martinez *et al*., 2009). Although our findings in this example are not necessarily novel, it does demonstrate the use of our approach in identifying enrichment of biologically relevant functions in the set of genes with a common miR binding site, which can be helpful in prioritizing gene lists to identify disease (or other condition) relevant regulatory elements.

4.2 Replication timing

DNA replication in the human genome is spatiotemporally segregated such that some genomic regions are replicated early and some late (Hansen *et al*., 2010). It was previously suggested that gene-rich regions replicated early. But it was not surveyed whether both exons and introns replicate early, or whether the replication timing pattern is context dependent even at a finer scale. Integrating DNA replication timing data from multiple cell types, and using the definition provided by Hansen *et al*. (2010), we marked the 'constant early' and 'constant late' replication timing regions—i.e. the regions that were replicated early and late irrespective of the cell type tested. Integrating this locally hosted dataset with CpG island, and refGene data-tables from the UCSC genome browser, we find that early-replicating regions are enriched for gene bodies and for CpG islands relative to the late-replicating regions (Supplementary Files S4 and S5), which is consistent with that reported by Hansen *et al*. (2010). In contrast, we find that introns are relatively more likely to be replicated late. For instance, among those regions that fall within a gene, there is 152% enrichment for late-replicating regions that fall entirely in an intron (without touching an exon) relative to early-replicating regions. When we restrict to coding genes with at least one intron, the enrichment goes up to 159% (Supplementary Files S6 and S7). Although it requires further investigation, this is a novel finding that suggests that even though gene-rich regions are replicated early, there are finer-scale replication timing patterns that correlate with intron–exon structures.

4.3 LincRNAs

Complex genetic diseases are usually associated with multiple common and rare genetic variants. Although a small subset of these variants overlap with known genes, many reside in non-protein coding regions. Some of these variants were shown to affect regulatory elements that affect expression of known genes. Non-coding RNAs (ncRNAs) are a class of regulatory RNAs that play important roles in development, cancer and other diseases. LincRNAs are a relatively recently identified class of ncRNA, which play key role in epigenetic regulation (Lee, 2012), and there are >20 000 predicted lincRNA genes in the human genome. So far, the genetic variants have not been systematically surveyed in the context of different classes of ncRNAs including lincRNAs.

Here, we use lincRNA transcripts available in the UCSC hg19 from Cabili *et al*. (2011) and overlap with the genome-wide association study (GWAS) catalog from NHGRI (Hindorff *et al*., 2009) as available in UCSC's gwasCatalog table. The catalog contains a list of 12 194 single nucleotide polymorphisms (SNPs) that have been associated with one of over 600 traits. After annotating with CruzDB (Supplementary File S8), we examined SNPs from the GWAS catalog that overlapped a lincRNA, and especially those that were >10 kb from the nearest gene. Using these criteria, we found 388 SNPs that overlapped a lincRNA and were also sufficiently distant from known RefSeq genes. When we enumerate the trait (disease category) with the highest proportion of SNPs that fall within a lincRNA distant to a gene and then filter to those that show at least five SNPs within a lincRNA, some traits among the highest by this metric are intelligence (5 of 57 SNPs fall in lincRNAs) and other categories related to cognitive disorders (Supplementary File S9). Although overlap does not automatically indicate causality, it is consistent with the role of these lincRNAs in development. There are several more instances where disease-associated variants overlap with lincRNAs with relevant biological functions.

Using more relaxed criteria, where an SNP was selected simply if it was closer to a lincRNA than to the nearest gene, we found 2153 SNPs (Supplementary File S10). Our findings combined with the recent study showing a lower incidence of SNPs within lincRNAs (Chen *et al*., 2013) show the importance of annotating GWAS results with lincRNAs in addition to genes.

4.4 Lamina-associated domains

Within the nucleus, different genomic regions occupy distinct nuclear territories, such that some regions are in contact with nuclear lamina–termed lamina-associated domains or LADs (Dittmer and Misteli, 2011; Guelen *et al*., 2008). These regions usually have repressive chromatin marks and lower levels of gene expression. However, it has not yet been investigated systematically whether certain classes of genes are more clustered in LADs compared with that expected by chance. Overlaying data on LADs from Guelen *et al*. (2008), and known genes, we find over 5000 genes overlap completely/partially with the LADs (Supplementary File S11). Furthermore, piping the genes that overlap a LAD with a score >0.9 (the fraction of probes with a positive smoothed log-ratio) to the DAVID gene-ontology enrichment software (Huang *et al*., 2009), we report strong enrichment for categories related to olfaction (adjusted P <1e-80), G-protein coupled receptor (adjusted P <1e-60) and other categories related to sensing (Supplementary File S12). Our findings are consistent with a recent report (Clowney *et al*., 2012) that nuclear clustering of olfactory receptor genes governs their monogenic expression. It is suspected that laminB receptor-induced changes in nuclear architecture influences singular transcription pattern of the olfactory receptor genes (Clowney *et al*., 2012).

When we create a stricter subset of genes by filtering to those with a score greater than 0.9 that fall entirely within an LAD, we

find even stronger enrichment of olfaction and related terms (adjusted P <1e-106), G-protein coupled recepter (adjusted P <1e-95, Supplementary File S13).

5 DISCUSSION

We have showcased the programmatic interface of CruzDB using miRNA binding sites as motivation, and further demonstrated its utility using three biological examples. The biological examples and their CruzDB code demonstrate the simple syntax, and the potential of this utility to facilitate hypothesis-driven studies. Although the examples we have shown are in human version *hg19*, CruzDB can be used for any organism and version available in the UCSC database by using, for example, '*dm3*' as the initializer.

Although previous studies have demonstrated higher order organization of DNA replication timing patterns in the human genome, our observation that replication timing patterns correlate with intron–exon structures reveals a finer-scale stratification of DNA replication patterns within gene regions. LincRNAs play major regulatory roles in different biological processes in neuronal and other tissue types, but their role in complex cognitive traits and diseases has not been systematically assessed yet. We report several instances where disease-associated variants overlap with lincRNAs with relevant biological functions. Our findings, combined with the recent study showing a lower incidence of SNPs within lincRNAs (Chen *et al*., 2013), highlight the importance of examining GWAS hits in this context. Finally, integrating data on LADs and protein-coding regions, we find that olfactory receptor genes are highly enriched in the LADs. Our findings are consistent with a recent report that nuclear clustering of olfactory receptor genes governs their monogenic expression, and that laminB receptor-induced changes in nuclear architecture influence singular transcription pattern of the olfactory receptor genes (Clowney *et al*., 2012).

Further work needs to be done to validate this work and to demonstrate the broader impact of our findings in each of these four biological cases in detail, we aim to pursue them outside the scope of this method article. Nevertheless, the four examples outline the broad utility of CruzDB and its applications in diverse areas of biomedical research.

6 CONCLUSION

We have introduced CruzDB, a parallelizable and intuitive programmatic interface with UCSC genome browser that allows integrative context-dependent analyses of diverse local and remotely hosted datasets, as well as annotation and spatial querying. Some of the functions that make CruzDB a library of broad and general utility are the feature extraction, fast queries and simple syntax. Using the library, one can mirror the UCSC databases to a local SQLite or MySQL database, perform location-based queries and perform integrative analyses combining local and remotely hosted features. We have shown how to create a local copy of selected tables is a single line of code and how having that local copy improves the speed of later analyses.

Funding: NCI Physical Sciences Oncology Center pilot grant (U54CA143798), American Cancer Society grant (ACS IRG 57-001-53) and University of Colorado School of Medicine startup grant (to S.D.).

Conflict of Interest: none declared.

REFERENCES

Cabili,M. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.

Chen,G. *et al.* (2013) Genome-wide analysis of human SNPs at long intergenic noncoding rnas. *Hum. Mutat.*, **34**, 338–344.

Clowney,E. *et al.* (2012) Nuclear aggregation of olfactory receptor genes governs their monogenic expression. *Cell*, **151**, 724–737.

Dittmer,T. and Misteli,T. (2011) The lamin protein family. *Genome Biol.*, **12**, 222.

Dreszer,T. *et al.* (2012) The UCSC genome browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, 918–923.

ENCODE. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ernst,J. and Kellis,M. (2013) Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.*, **23**, 1142–1154.

Grimson,A. *et al.* (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Guelen,L. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948–951.

Hansen,R. *et al.* (2010) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA*, **107**, 139–144.

Hindorff,L. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.

Huang,D. *et al.* (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **27**, 1–13.

Kent,J.W. (2002) Blat–the blast-like alignment tool. *Genome Res.*, **12**, 656–664.

Kent,W.J. *et al.* (2002) The UCSC genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Lee,J.T. (2012) Epigenetic regulation by long noncoding RNAs. *Science*, **338**, 1435–1439.

Mardis,E.R. (2010) The $1,000 genome, the $100,000 analysis? *Genome Med.*, **26**, 84.

Martinez,A. *et al.* (2009) Gap-junction channels Dysfunction in deafness and hearing loss. *Antioxid. Redox Signal.*, **11**, 309–322.

Menca,A. *et al.* (2009) Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat. Genet.*, **41**, 609–613.