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

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

[1]University of Colorado, Anshutz Medical Campus, Department of Medicine 12700 East 19th Avenue, 8611 Aurora, CO 80045
[2]University of Colorado Cancer Center. 13001 E 17th Pl, Aurora, CO 80045

Email: Brent S Pedersen*- bpederse@gmail.com; Ivana V Yang - ivana.yang@ucdenver.edu; Subhajyoti De*- subhajyoti.de@ucdenver.edu;

*Corresponding author

Abstract

The biological significance of genomic features is often context-dependent. We present CruzDB, a fast and intuitive programmatic interface to the UCSC genome browser that facilitates integrative analyses of diverse local and remotely hosted datasets. We showcase the syntax of CruzDB using miRNA-binding sites as examples, and further demonstrate its utility with 3 novel biological discoveries. First, we find that while exons replicate early, introns tend to replicate late, suggesting a complex replication pattern in gene regions. Second, variants associated with cognitive functions map to lincRNA transcripts of relevant function. Third, lamina-associated domains are highly enriched in olfaction-related genes. CruzDB is available at https://github.com/brentp/cruzdb

Rationale

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 [1] 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.

Systematic presentation of large-scale datasets from the ENCODE [1] and other projects in the UCSC genome browser [2] 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 [3]. 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 $1,000 genome, the $100,000 analysis' [4], 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 UC Santa Cruz [3] 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 and novel biological findings that we made using CruzDB.

Implementation

CruzDB utilizes the python programming language and sqlalchemy (SQL-alchemy) library to access publicly available data hosted at the UCSC genome browser database [3] . 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).

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 utilizing built-in parallelization. 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/.

Software Features

CruzDB simplifies common tasks such as those that return upstream or downstream features, exons, introns, UTRs and transcription start sites. Location-based queries can utilize the UCSC bin column [2] 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 up and down-stream 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' while 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 DAS server. One can also obtain a list of BLAT [5] hits for a particular feature.

Using CruzDB, it is possible to mirror a subset of tables from UC Santa Cruz to a local MySQL or SQLite database using a single line of python code. A local copy allows a user to add data that is not in UCSC and then use that new table just as one would 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.

In order 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 vs 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 or not parallelization is used when annotating about 3,300 intervals (timing data is available in Additional File 1). Note that SQLite is very 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.

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.

Examples

Code Example: microRNA targets

Since CruzDB is a library, we show a short code example, using the targetScanS database of predicted miRNA targets [6] available in the UCSC genome browser. 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 Additional File 2. 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 1 miR-96 binding site with a score greater than 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'll 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 [7] 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 [8]; when we look at the ontology enrichment from DAVID (Additional File 3), we see terms associated with synapses and cell-junction which are, in turn, known to be associated with deafness and hearing loss [9]. While our findings in this example are not necessarily novel, it does demonstrate the utility 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.

Replication Timing

DNA replication in the human genome is spatio-temporally segregated such that some genomic regions are replicated early, and some late [10]. 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 [10] 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 (Additional Files 4 and 5), which is consistent with that reported by [10]. In contrast, introns were 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 1 intron, the enrichment goes up to 159% (Additional Files 6 and 7). This novel finding suggests that even though gene-rich regions are replicated early, there are finer-scale replication timing patterns that correlate with intron-exon structures.

LincRNAs

Complex genetic diseases are usually associated with multiple common and rare genetic variants. While 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 [11], and there are more than 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 [12] and overlap with the GWAS Catalog from NHGRI [13] as available in UCSC's gwasCatalog table. The catalog contains a list of 12,194 SNPs that have been associated with one of over 600 traits. After annotating with CruzDB (Additional File 8), we examined SNPs from the GWAS catalog that overlapped a lincRNA, and especially those which were more than 10Kb from the nearest gene. Using this criteria we found 388 SNPs which 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 5 SNPs within a lincRNA, some traits among the highest by this metric are intelligence (5 out of 57 SNPs fall in lincRNAs), and other categories related to cognitive disorders (Additional File 9). Although overlap does not automatically indicate causality, it is consistent with the role of these miRNAs in development. There are several more instances where disease-associated variants overlap with lincRNAs with relevant biological functions.

Using a more relaxed criteria, where a SNP was selected simply if it was closer to a lincRNA than to the nearest gene, we found 2153 SNPs (Additional File 10). Our findings, combined with the recent study showing a lower incidence of SNPs within lincRNAs [14] show the importance of annotating GWAS results with lincRNAs in addition to genes.

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) [15, 16]. 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 to that expected by chance. Overlaying data on lamina associated domains (LADs) from [15], and known genes, we find over 5000 genes overlap completely/partially with the LADs (Additional File 11). 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 [7] we report very 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 (Additional File 12). Our findings are consistent with a recent report [17] 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 [17].

Furthermore, when we filter to genes that are strictly contained within a LAD (not merely overlapping) with a score >0.9, and send that stricter subset of 2,570 genes to DAVID, we find even stronger enrichment of olfaction and related terms (adjusted p <1e-106), g-protein coupled recepter (adjusted p <1e-95) (Additional File 13).

Discussion

We have introduced CruzDB, a parallelizable and intuitive syntax-based 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.

We showcase the programmatic interface of CruzDB using miRNA-binding sites as examples, and further demonstrate its utility using 3 novel biological discoveries. First, we showed the syntax of the library by extracting genes with a target site for the miR-96 microRNA. Second, by integrating exon and DNA replication timing data, we show that even though exons typically replicate early, introns are likely to replicate late during S phase. Our findings suggest a more complex DNA replication landscape than previously appreciated. Third, although current GWAS studies have primarily focused on functional variants affecting protein-coding genes, some variants are likely to affect other functional elements including non-coding RNAs. We report several instances where disease-associated variants overlap with lincRNAs with relevant biological functions. For example those related to intelligence and cognitive disorders. Our findings, combined with the recent study showing a lower incidence of SNPs within lincRNAs [14], highlight the importance of examining GWAS hits in this context. Finally, integrating data on lamina-associated domains and protein-coding regions, we find that olfactory receptor genes are highly enriched in the lamina-associated domains. Our findings are consistent with a recent report that nuclear clustering of olfactory receptor genes governs their monogenic expression [17]. It is suspected that laminB receptor-induced changes in nuclear architecture influence singular transcription pattern of the olfactory receptor genes [17]. While we acknowledge that further work needs to be done 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 paper. Nevertheless, the four examples outline the broad utility of CruzDB, and its applications in diverse areas of biomedical research.

Author's contributions

BSP wrote the software. BSP and SD designed the experiments. BSP, SD and IVY wrote the manuscript.

Acknowledgements

References

1. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**:57–74.
2. Kent W, Sugnet C, Furey T, Roskin K, Pringle T, Zahler A, Haussler D: **The UCSC genome browser at UCSC**. *Genome Res.* 2002, **12**(6):996–1006.
3. Dreszer T, Karolchik D, Zweig A, Hinrichs A, Raney B, Kuhn R, Meyer L, Wong M, Sloan C, Rosenbloom K, Roe G, Rhead B, Pohl A, Malladi V, Li C, Learned K, Kirkup V, Hsu F, Harte R, Guruvadoo L, Goldman M, Giardine B, Fujita P, Diekhans M, Cline M, Clawson H, Barber G, Haussler D, Kent W: **The UCSC genome Browser database: extensions and updates 2011**. *Nucleic Acids Research* 2012, **40**:918–23. [Database Issue].
4. ER M: **The $1,000 genome, the $100,000 analysis?** *Genome Medicine* 2010, **26**:84.
5. WJ K: **BLAT–the BLAST-like alignment tool**. *Genome Research* 2002, **12**(4):656–664.
6. Grimson A, Farh K, Johnston W, Garrett-Engele P, Lim L, Bartel D: **MicroRNA targeting specificity in mammals: determinants beyond seed pairing**. *Molecular Cell* 2007, **27**:91–105.
7. Huang D, Sherman B, Lempicki R: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Research* 2009, **27**:1–13.
8. Mencía A, Modamio-Høybjør S, Redshaw N, Morín M, Mayo-Merino F, Olavarrieta L, Aguirre L, et al: **Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss**. *Nature Genetics* 2009, **41**(5):609–613.
9. Martínez A, Acuña R, Figueroa V, Maripillan J, Nicholson B: **Gap-junction channels dysfunction in deafness and hearing loss**. *Antioxidants & redox signaling* 2009, **11**:309–322.
10. Hansen R, Thomas S, Sandstrom R, Canfield T, Thurman R, Weaver M, Dorschner M, Gartler S, Stamatoyannopoulos J: **Sequencing newly replicated DNA reveals widespread plasticity in human replication timing**. *Proc Natl Acad Sci USA* 2010, **107**:139–144.
11. JT L: **Epigenetic Regulation by Long Noncoding RNAs**. *Science* 2012, **338**(6113):1435–1439.
12. Cabili M, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn J: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**. *Genes and Development* 2011, **25**:1915–1927.
13. Hindorff L, Sethupathy P, Junkins H, Ramos E, Mehta J, Collins F, Manolio T: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits**. *PNAS* 2009, **106**(23):9362–9367.
14. Chen G, Qiu C, Zhang Q, Liu B, Cui B: **Genome-Wide Analysis of Human SNPs at Long Intergenic Noncoding RNAs**. *Human Mutation* 2013, **34**:338–344.
15. Guelen L, Pagie L, Brasset E, Meuleman W, Faza M, Talhout W, van Steensel B: **Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions**. *Nature* 2008, **453**(7197):948–951.
16. Dittmer T, Mistelli T: **The lamin protein family**. *Genome Biology* 2011, **12**:222.
17. Clowney E, LeGros M, Mosley C, Markenskoff-Papadimitriou E, Myllys M, Barnea G, Larabell C, Lomvardas S: **Nuclear aggregation of olfactory receptor genes governs their monogenic expression**. *Cell* 2012, **151**:724–737.

Figures

Figure 1 - Number of intervals annotated per second for local and remote databases.

Parallelization on 4 cores and reading all intervals into memory greatly improves the speed of both remote and local MySQL instances while SQLite is fast in either case.

Tables

Additional Files

Additional file 1 — Table of timing data for local and remote databases

Columns indicate whether the run was local or remote, whether SQLite or MySQL was used, parallelization, and time to perform the queries.

Additional file 2 — Text of code for first example

Code to look for genes with a miR-96 binding site in the 3' UTR

Additional file 3 — Table of DAVID enrichment output

Output of enrichment categories from DAVID enrichment tool on genes with a 3' UTR binding site for miR-96.

Additional file 4 — Table of annotated early-binding regions

Early binding regions from [10] annotated to nearest refGene

Additional file 5 — Table of annotated late-binding regions

Late binding regions from [10] annotated to nearest refGene

Additional file 6 — Table of annotated early-binding regions with at least 1 intron

Early binding regions from [10] annotated to nearest refGene with at least 1 intron

Additional file 7 — Table of annotated late-binding regions with at least 1 intron

Late binding regions from [10] annotated to nearest refGene with at least 1 intron

Additional file 8 — Table of gwasCatalog SNPs annotated to nearest refGene and lincRNA

Table of gwasCatalog SNPs annotated to nearest refGene and lincRNA

Additional file 9 — Table of GWAS traits

GWAS traits sorted by portion of SNPs in that trait that are near a lincRNA and >10Kb from the nearest gene. Only traits with at least 5 lincRNAs are shown.

Additional file 10 — Table of GWAS traits closer to lincRNA

GWAS traits sorted by portion of SNPs in that trait that are closer to a lincRNA than to the nearest gene. Only traits with at least 5 lincRNAs are shown.

Additional file 11 — Table of Annotate LADs

Lamina Associated Domains annotated with nearest refGene feature.

Additional file 12 — Table DAVID enrichment for LAD's with score >0.9

Output from DAVID enrichment tool for genes touching a LAD with a score >0.9

Additional file 13 — Table DAVID enrichment for LAD's with score >0.90 and strict overlap

Output from DAVID enrichment tool for genes completely contained in a LAD with a score >0.90