# Interactive Analytics for Complex Cognitive Activities on Information from Annotations of Prokaryotic Genomes

Raphael D. Isokpehi, Kiara M. Wootson
College of Science, Engineering and Mathematics
Bethune-Cookman University
640 Dr. Mary McLeod Bethune Blvd.
Daytona Beach, Florida 32114, USA
isokpehir@cookman.edu

Dominique R. Smith-McInnis
Environmental Science PhD Program
Jackson State University
1400 JR Lynch Street
Jackson, Mississippi 39217, USA

Shaneka S. Simmons
Jarvis Christian College
P. O. Box 1470
Hawkins, Texas 75765, USA

## ABSTRACT
Several microbial genome databases provide collections of thousands of genome annotation files in formats suitable for the performance of complex cognitive activities such as decision making, sense making and analytical reasoning. The goal of the research reported in this article was to develop interactive analytics resources to support the performance of complex cognitive activities on a collection of publicly available genome information spaces. A supercomputing infrastructure (Blue Waters Supercomputer) provided computational tools to construct information spaces while visual analytics software and online bioinformatics resources provided tools to interact with the constructed information spaces. The Rhizobiales order of bacteria that includes the *Brucella* genus was the use case for preforming the complex cognitive activities. An interesting finding among the genomes of the dolphin pathogen, *Brucella ceti*, was a cluster of genes with evidence for function in conditions of limited nitrogen availability.

## General Terms
Big Data, Human-Computer Interaction, Microbiology, Visualization.

## Keywords
Bacteria; *Brucella*; Cognitive Activities, Genomics; Stress Response; Universal Stress Protein, Visual Analytics

## 1. INTRODUCTION
The automated annotation of genome sequences of bacteria and archaea produces diverse types of data sets including multivariate data on predicted protein-coding genes [1-4]. Examples of variables annotated for protein-coding genes are genome unique identifier, genome name, unique gene identifier (locus tag), coordinates of the start and end position, product description, Enzyme Commission identifier, length of gene sequence, and location of gene on positive or negative strand.

Several microbial genome databases [1, 3, 4] provide collections of thousands of genome annotation files in formats (such as tab delimited) suitable for importing to computational environments that support the performance of complex cognitive activities. In complex cognitive activities (such as analytical reasoning, decision making, knowledge discovery, learning, planning, problem solving, sense making and understanding), humans interact with information to support their information-intensive thinking processes [5-7].


DOI: https://doi.org/10.22369/issn.2153-4136/8/2/5

The goal of the research reported in this article was to develop interactive analytics resources to support the performance of complex cognitive activities on a collection of publicly available genome information spaces. A genome annotation file containing protein-coding genes of a bacterial (eubacteria and archaebacteria) genome could be described as an information space which can be compared or integrated to other information spaces. The complex genomic information space presents diverse opportunities for knowledge generation on microbial genomes that combines the affordances from both the human cognitive system and computing system. The goal of our research was to obtain potentially biologically relevant insights from the microbial genomic information space. Therefore, we have combined (i) the use of a supercomputing environment (Blue Waters Supercomputer) [8] to construct information spaces; (ii) the use of visual analytics software to interact with the constructed information spaces; and (iii) online bioinformatics resources on microbial genomes.

Visual analytics affords humans to analyze huge information spaces in order to support complex cognitive activities such as decision making and data exploration [9]. The interaction with information through visual representations provides a human-centered approach to the performance of cognitive activities [10, 11]. This human-centered approach lowers the barriers to knowledge generation from genome information spaces. In addition, there is potential to increase the number of undergraduate students who are able to engage in genomics research.

An example of genome information space is the PATRIC Bioinformatics Resource, which provides collection of thousands of genome annotation files available for download at ftp://ftp.patricbrc.org/patric2 [4]. The first objective of this research study was to construct an information space on the count of genes assigned to strands [positive (+) or negative DNA strand (-)] in the thousands of genome annotation files. This objective will lead to a reduction in the complexity of the information space for subsequent complex cognitive activities with desktop visual analytics software. The second objective was to perform complex cognitive activities on genomic information from multiple sources. Though, we recognize that some complex cognitive activities often done without clear distinctions.

These objectives are important to our investigation of stress responsive gene clusters that include genes which encode the universal stress proteins (pfam00582) [12, 13]. The genomes sequenced from bacteria in the order Rhizobiales were used to accomplish the research study objectives. Rhizobiales is a diverse order of bacteria that include nitrogen-fixing bacteria associated with leguminous plants and lichens as well as intracellular

pathogens of animals and plants [14, 15]. Examples of genera in the order Rhizobiales (alphaproteobacteria) are *Bartonella*, *Beijerinckia*, *Bradyrhizobium*, *Brucella*, *Cohaesibacter*, *Hyphomicrobium*, *Methylobacterium, Microvirga*, *Methylocystis*, *Phyllobacterium*, *Rhizobium*, *Rhodobium, Rhodopseudomonas* and *Xanthobacter* [16]. Finally, the interactive views can provide opportunities for learning about the genomes of bacteria.

## 2. METHODS

### 2.1 Source of Genome Annotation Files

The genome annotation files (with file extension RefSeq.cds.tab) were downloaded from the PATRIC Bioinformatics Resource at ftp://ftp.patricbrc.org/patric2/genomes_by_species/ to the Blue Waters Supercomputer. Each file is expected to contain a header row and records with annotation for each gene including genome unique identifier, genome name, unique gene identifier (locus tag), coordinates of the start and end position, product description, Enzyme Commission identifier, length of gene sequence, and location of gene on positive or negative strand.

Three additional files (genome_lineage, genome_metadata and genome_summary) were obtained from ftp://ftp.patricbrc.org/patric2/current_release/RELEASE_NOTES/Feb2016/. These files contain fields that can be used accomplish complex cognitive activities. The genome_lineage file includes taxonomic annotation of genomes including kingdom, phylum, order, genus and National Center for Biotechnology Information (NCBI) Taxonomy Identifier. The genome_metadata includes data on habitat, gram stain category and temperature of the microbial isolate source of the genome sequence. The genome_summary file includes data on genome length, gene count and genome sequencing status (e.g. Whole Genome Sequencing, Plasmid and Complete).

### 2.2 Construction of Information Space on Strand Location of Genes

The genome annotation files include annotation on the transcription direction of the gene (location of gene on the positive (+) or negative (-) strand). A set of computer scripts were developed on Blue Waters Supercomputer [17] to extract the transcription direction of each gene in the genome annotation files. The output file was formatted as a tab delimited file with Genome ID, Gene Count for Strand, the Genome Name and the Transcription Direction. This method allowed us to accomplish our objective to construct an information space on the distribution of genes in genome annotation files by transcription direction [location of gene on positive or negative strand].

### 2.3 Development of Interactive Analytics for Complex Cognitive Activities

We developed interactive analytics using guidelines provided for designing interactive visual representations for complex cognitive activities [10, 18]. Therefore to design human-information interaction tools for decision making, the interaction features in the design are expected to include the following action patterns: blending, filtering, linking/unlinking, measuring, sharing and translating [7].

A software for visual analytics, Tableau Desktop Professional (Tableau Software Inc. Washington, USA), was used to design the views for accomplishing the following activities: (i) to identify biases in gene distribution across genomes [sense making]; (ii) to decide on which bacteria genome to investigate based on annotated comments [decision making]; and (iii) to determine the arrangement and functions of a cluster of genes that are transcribed together [analytical reasoning].

## 3. RESULTS

### 3.1 Information Space on Strand Location of Genes

A total of 21,139 genome annotation files were downloaded from the PATRIC Bioinformatics Resource and processed on the Blue Waters Supercomputer. The collection of files provides a data resource for the performance of data analytics. Each file had 16 fields and number of records corresponding to the protein-coding genes annotated for the genome. The total number of gene records obtained from PATRIC was 74,991,894. The derived information space consisted of four fields: Genome ID, Genome Name, Strand and the Gene Count (assign to each strand).

### 3.2 Interactive Analytics for Sense Making on Protein-Coding Genes in Rhizobiales

A total of 547 Rhizobiales genome annotation files were evaluated because of our interest in *Rhodopseudomonas palustris* [19]. Figure 1 shows the number of protein-coding genes (RefSeq annotation) assigned to the strands of the genomes of *Brucella ceti*, a *Brucella* species that cause chronic diseases in marine mammals such as dolphins and whales [20]. The visualizations in Figure 1 and Figure 2 allow for the difference in count of genes assigned to the genome strands to be calculated.

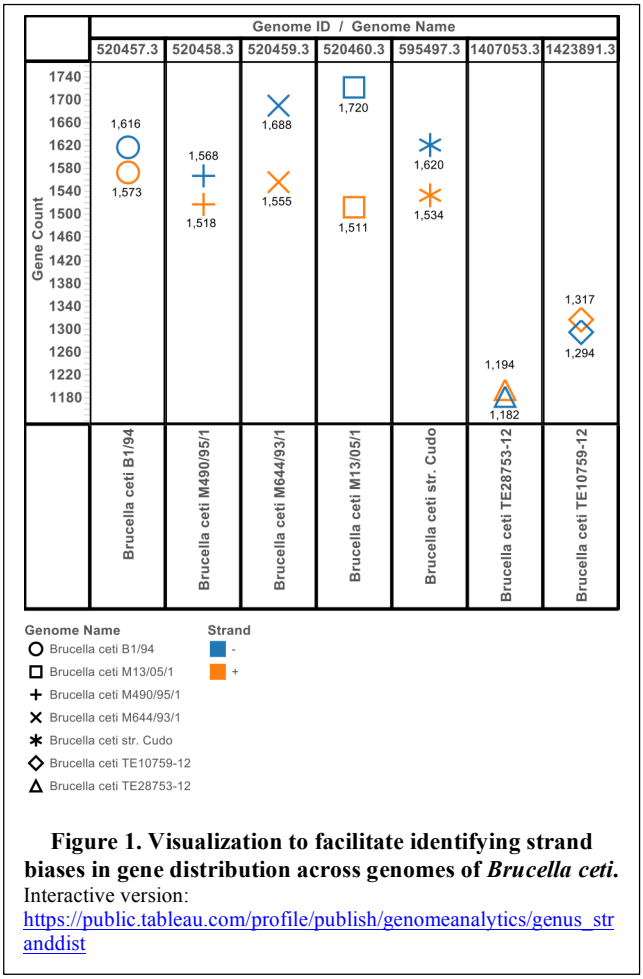



**Figure 1. Visualization to facilitate identifying strand biases in gene distribution across genomes of *Brucella ceti*.** Interactive version: https://public.tableau.com/profile/publish/genomeanalytics/genus_stranddist

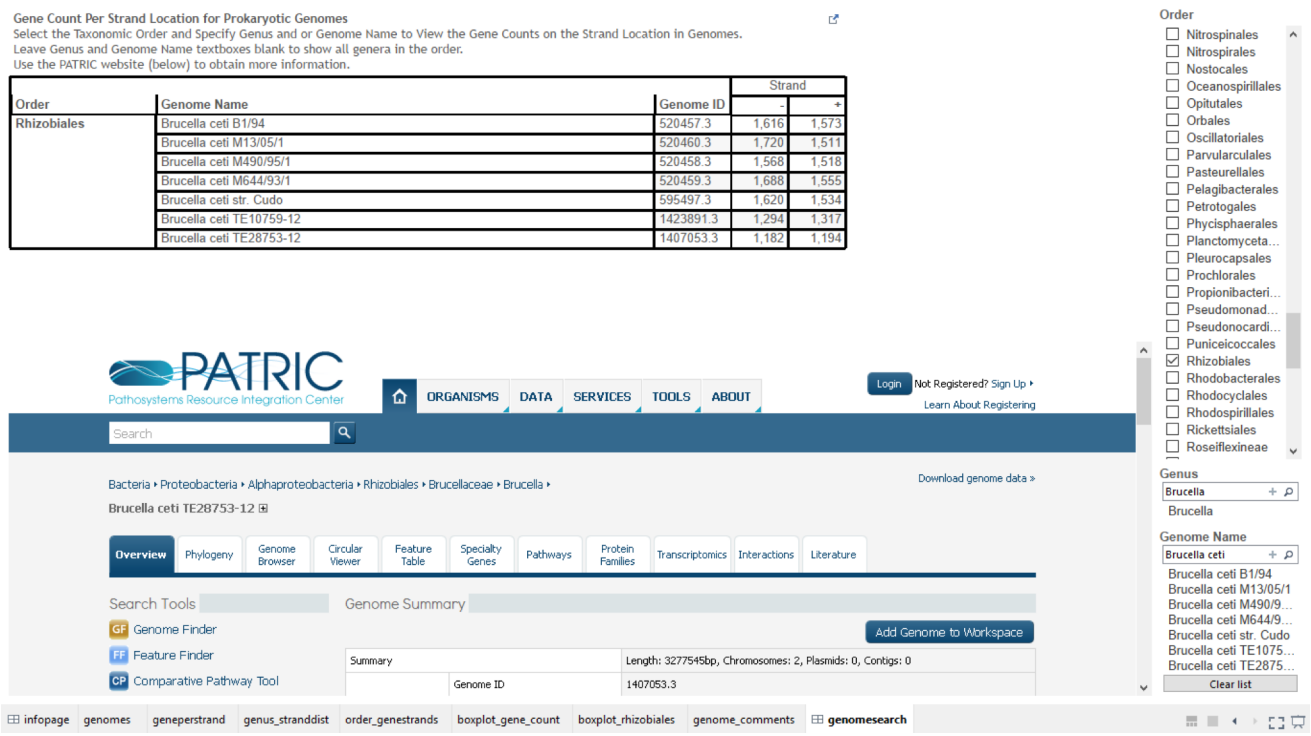



**Figure 2. Dashboard providing access to a bioinformatics resource as well as integrating information on the number of genes assigned to chromosomal strand locations for prokaryotic taxonomic and genome categories.**

Interactive Analytics resource at: https://public.tableau.com/profile/publish/genomeanalytics/genomesearch. User of the resource can perform activities such as sense making and decision making through selection or specifying the taxonomic order, genus or genome name to view the gene counts on the strand location in genomes. Additional information could be obtained through the Pathosystems Resource Integration Center (PATRIC) website. The dashboard can also be used as a resource for learning the distribution of genes to strand location. In the example, the genomes of *Brucella ceti* are the focus of sense making, decision making and learning activities.

## 3.3 Interactive Analytics for Sense Making on Protein-Coding Genes in Rhizobiales

Sense making "is concerned with developing a mental model of an information space about which one has insufficient knowledge" [7]. We used the Box plot visualization technique to compare multiple distributions of the gene counts for genera (*Agrobacterium*, *Bartonella*, *Beijerinckia*, *Brucella*, *Methylobacterium*, *Nitrobacter* and *Rhodopseudomonas*) in the Rhizobiales (Figure 3). Interactive figure is available at https://public.tableau.com/profile/publish/genomeanalytics/boxplot_rhizobiales

The design of the visualization involves blending data fields from (i) the genome_lineage file (contains taxonomic information); (ii) the genome_summary file (contains plasmid count); and (iii) the constructed information space on the strand location of genes. The interactive version allows user to specify the bacteria taxonomic family or families to compare.

In the case of the Rhizobiales genomes, examining the box plot revealed genomes with outlier protein coding sequence within the genus. Outlier values in the box plot were annotated for selected genomes. For example, *Brucella ceti* TE10759-12 has 2,376 protein-coding genes in the RefSeq genome annotation file. The missing genes of TE10759-12 provides a user with information to generate testable hypotheses.

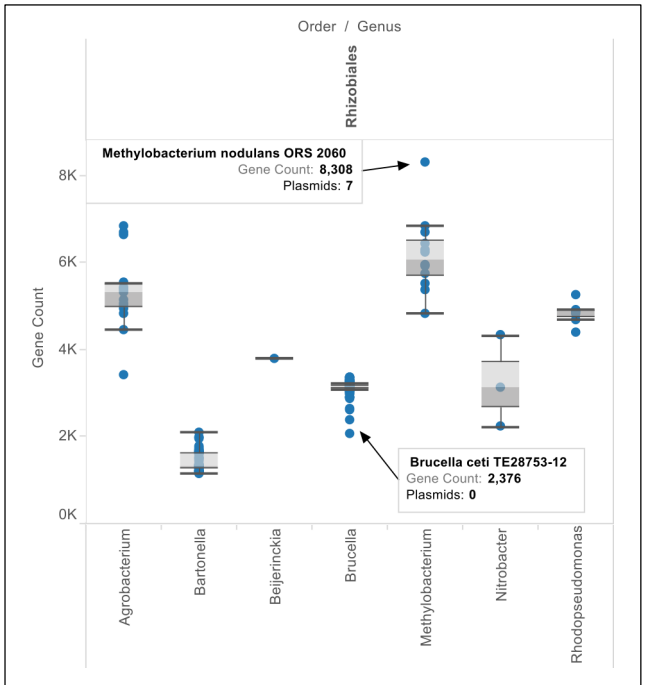



**Figure 3. Visual representation (box plot) to facilitate sense making of protein-coding gene counts for selected genomes of Rhizobiales bacteria.**

Interactive version: https://public.tableau.com/profile/publish/genomeanalytics/boxplot_rhizobiales

## 3.4 Interactive Analytics for Decision Making on Genomes for Investigation

In decision making "the attention that is drawn to emergent features may facilitate the choice of one among a number of alternatives within the information space" [21]. We developed a view from the genome_metadata file to display the comments associated with eight *Brucella ceti* genomes. Four categories of comments were identified (Table 1).

**Table 1. Categories of Comments on *Brucella ceti* genomes**

| ***Brucella ceti* Strains** | **Comment Category** |
|---|---|
| B1/94, M13/05/1, M490/95/1, M644/93/1 | This strain will be used for comparative analysis with other *Brucella* species. |
| B1/94 | Sequencing of *Brucella* species for qPCR assay development. |
| str. Cudo | *Brucella ceti* Cudo was isolated from a bottlenose dolphin (*Tursiops truncatus*). The genome sequence of this organism will provide interesting insights into the evolution of this species. |
| TE10759-12, TE28753-12 | … The aim of the study is the deep characterization of the isolates … |

https://public.tableau.com/profile/publish/genomeanalytics/genome_comments (Interactive version of genome comments).

The comment "*Brucella ceti* Cudo was isolated from a bottlenose dolphin (*Tursiops truncatus*)" facilitated our decision to further conduct gene neighborhood analysis of the universal stress proteins of *Brucella ceti* Cudo. Universal stress proteins contain the protein family (Pfam) domain with Pfam Identifier as PF00582 or pfam00582 [22]. We obtained a list of 1377 genes predicted as encoding universal stress proteins in 348 *Brucella* genomes. The Locus Tags for *Brucella ceti* Cudo universal stress proteins (USP) are BCETI_1000312, BCETI_3000327, BCETI_5000106 and BCETI_7000519. Only BCETI_7000519 was annotated as located on the positive strand (+) location.

We subsequently obtained and used the image of gene neighborhood of each USP gene using the BioCyc Database Collection [23]. The comparison of the gene neighborhood images would help us to confirm the transcription direction and also discover the functions adjacent to the *Brucella* genes for universal stress proteins (Figure 4). We found that BCETI_1000312 USP gene is at the beginning of a four-gene transcription unit (operon) (Figure 4). The other genes (BCETI_1000313, BCETI_1000315 and BCETI_1000316) respectively encode for tryptophanyl-tRNA synthetase (trpS), integral membrane protein (MviN) [renamed Peptidoglycan biosynthesis protein MurJ], and protein-P-II uridylyltransferase (glnD). The gene BCETI_1000311, adjacent to the USP gene BCETI_1000312, encodes a nitrogen fixation related protein. BCETI_1000311 is not predicted to be in same transcription unit with the USP gene (BCETI_1000312).

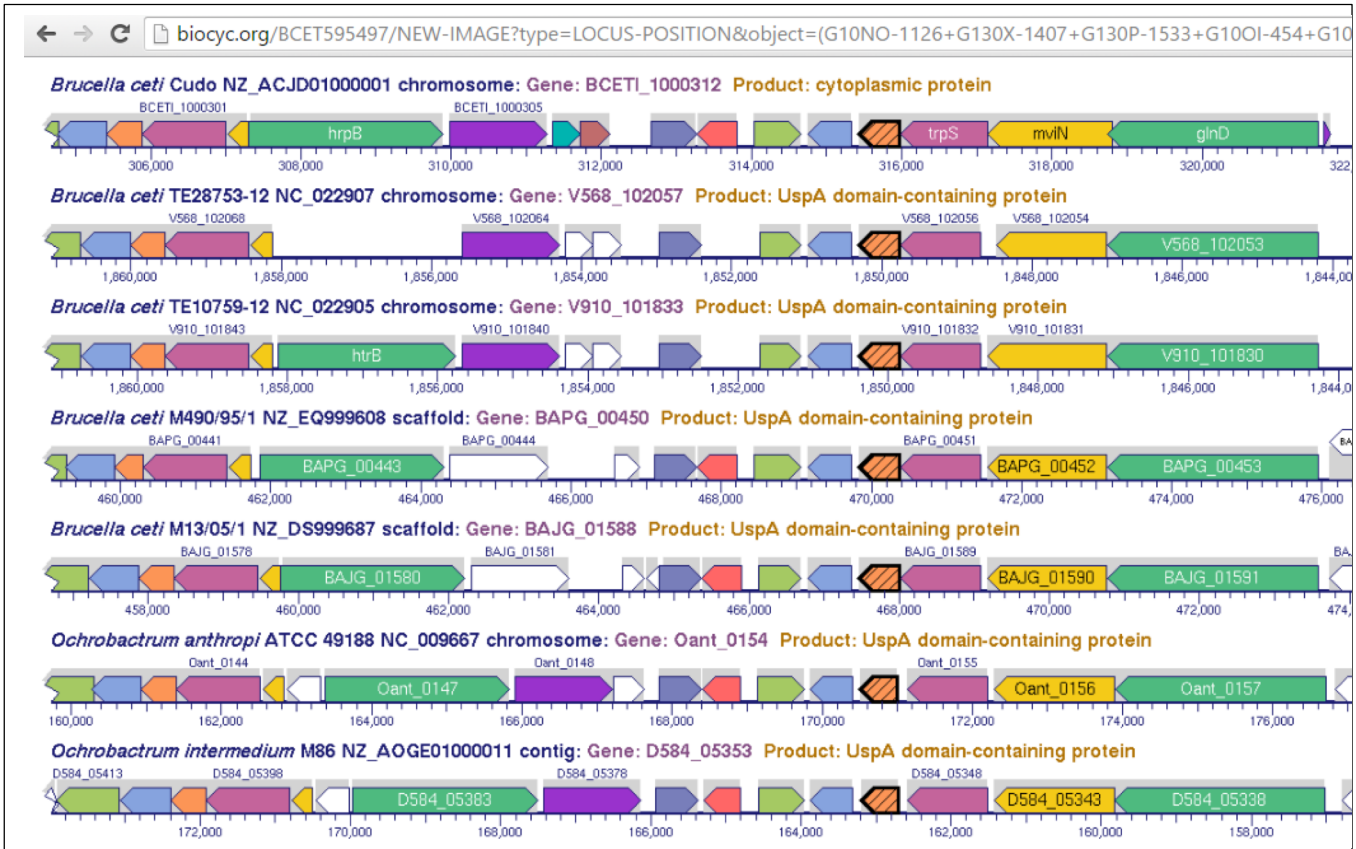



**Figure 4. Multi-Genome alignment of the gene neighborhood of predicted genes for universal stress proteins in genomes of *Brucella ceti* and *Ochrobactrum* species. The genes for universal stress proteins have diagonal lines.**

### 3.5 Interactive Analytics for Analytical Reasoning on *Brucella ceti* Transcription Units containing Gene for Universal Stress Protein

Analytical reasoning "is based on rational, logical analysis and evaluation of information" as well as "a structured, disciplined activity" [7]. We performed analytical reasoning on the multi-genome alignment of the gene neighborhood of 37 *Brucellae* genomes in BioCyc. The interactive alignment is can be constructed at BioCyc.org. We used the *B. ceti* Cudo four-gene transcription unit as template to analyze the presence and composition of transcription units and subsequently evaluate the level of conservation of the genomic region between *Brucella ceti* and *Ochrobactrum* genomes (Figure 4). The finding that BCETI_1000312 and BCETI_1000311 are not an operon was confirmed with a multi-genome alignment of the gene neighborhood. Among the *Brucella ceti* genomes, strain Cudo is unique for having the 4-gene transcriptional unit, which consists of genes for universal stress protein, tryptophanyl-tRNA synthetase, peptidoglycan biosynthesis protein and protein-P-II uridylyltransferase, a regulator of nitrogen status of *Escherichia coli* [24].

## 4. DISCUSSION

### 4.1 Information Space on Strand Location of Genes

We developed a computational workflow that led to a reduction in the complexity of 21,139 genome annotation files from 16 fields to 4 fields. This complexity reduction process implemented involved algorithmic operations including sorting and comparisons that required high performance computing resources. There is growing need for use of supercomputing resources and cloud computing in bioinformatics [25, 26]. The derived information space enabled a variety of complex cognitive tasks to be performed with desktop visual analytics software as well as online bioinformatics software.

Our research used the RefSeq genome annotation files. PATRIC bioinformatics resource includes re-annotated versions of microbial genomes [4]. Therefore, the computational protocols that we have developed on the Blue Waters Supercomputer [17] for deriving new information space on strand location of genes can be adapted for the PATRIC genome annotation files (with extension PATRIC.cds.tab). We expect to obtain additional genomes and gene loci. For example, our information space included 547 Rhizobiales genome annotation files. Based on statistics available at the PATRIC website (patricbrc.org), we expect to have at least 1441 Rhizobiales genomes. A web-based

### 4.2 Interactive Analytics for Sense Making on Protein-Coding Genes in Rhizobiales

As shown in Figure 1, among the seven *Brucella ceti* strains, 3 strains had excess of at least 50 genes mapped to the negative strand. The M13/05/1 strain has the largest difference in number of mapped genes, at 209 genes. This may indicate that certain genes have been recently duplicated, or that groups of genes were transferred from one strand to another, thereby providing a user with information to generate testable hypotheses.

The integration of information space on strand location with other annotation files enabled us to make sense of the distributions of the gene counts for genera in the Rhizobiales (Figure 3). We chose to use the box plot technique since the technique is suitable to visually summarize and compare groups of data [27]. A finding from the box plot visual representation (Figure 3) is that methanol-oxidizing *Methylobacterium nodulans* ORS 2060, the legume (*Crotalaria*) root-nodule-forming and nitrogen-fixing bacteria [28], has at least 7 sequenced plasmids [29]. The possession of an intact 120kb megaplasmid correlated with ability of *Methylobacterium extorquens* DM4 to utilize dichloromethane as sole source of carbon and energy [30]. Comparative analysis of the genes in the plasmids of *Methylobacterium* species could improve understanding of methylotrophy and nitrogen-fixation.

*Rhodopseudomonas palustris* TIE-1 has an upper outlier gene count among the *Rhodopseudomonas*. Further research could investigate the function of the additional genes in the iron oxidizing *R. palustris* strain [31].

### 4.3 Interactive Analytics for Decision Making on Genomes for Investigation

The comments associated with eight *Brucella ceti* genomes (Table 1) helped us decide to further investigate the genome of *Brucella ceti* Cudo, a dolphin associated *Brucella* [32, 33]. In the BioCyc pathway databases, a transcription unit is a set of one or more genes that are transcribed to produce a single messenger RNA [34]. Our research interest is in multi-gene transcription units which include at least one gene for universal stress protein. Four genes for universal stress proteins were observed in the genome of *B. ceti* Cudo. We have not observed reports describing the function of the *B. ceti* USPs. Therefore, this report provides new insights into the organization of transcription units and possible function of *B. ceti* USPs. [35]. The decision making then led to analytical reasoning of the gene neighborhood of *B. ceti* USP transcription units.

### 4.4 Interactive Analytics for Analytical Reasoning on *Brucella ceti* Transcription Units containing Gene for Universal Stress Protein

Among the *Brucella ceti* genomes, strain Cudo is unique for having the 4-gene transcriptional unit, which consists of genes for universal stress protein, tryptophanyl-tRNA synthetase, (pfam 00579), peptidoglycan biosynthesis protein (pfam03023) and protein-P-II uridylyltransferase (pfam08335) (Figure 2). There is a need for research studies to confirm the existence of the 4-gene transcription unit as well as the role of each gene. A common annotated function of the proteins encoded by the transcription unit is metabolism of nitrogen. The universal stress proteins are induced in response to stress conditions including nitrogen starvation [36-38]. Tryptophanyl-tRNA synthetase (TrpRS) ensures the translation of the genetic code for tryptophan, a nitrogen containing amino acid, by catalyzing the activation of tryptophan by adenosine triphosphate (ATP) and transfer to the tryptophanyl-tRNA ($tRNA^{Trp}$) [39].

The peptidoglycan biosynthesis protein in *Escherichia coli* is a lipid II flippase essential for cell wall peptidoglycan synthesis [40]. The protein-P-II uridylyltransferase (GlnD) is involved in glutamine metabolism and primary sensor of nitrogen [41]. In *Mycobacterium tuberculosis*, an intracellular pathogen as *Brucella* species, L- glutamine is a major component of the cell wall [42] and a source of nitrogen in *Brucellae* [43]. An immune response in mammalian cells for the control of intracellular pathogens includes the gamma interferon induced production of indoleamine 2,3-dioxygenase (IDO), an enzyme for the degradation of tryptophan [44]. The transcription direction of the four genes is conserved in the two *Ochrombactrum* genomes (Figure 3). Furthermore the functions for peptidoglycan synthesis and

nitrogen sensing exist as a transcription unit in both *Ochrombactrum* genomes and four of the five *B. ceti* genomes. In summary, there is evidence that the function of the transcription unit in the *Brucella ceti* Cudo genome that contains the gene BCETI_1000312 is for nitrogen stress response.

# 5. CONCLUSIONS

The goal of the research reported in this article was to develop interactive analytics resources to support the performance of complex cognitive activities on a collection of publicly available genome information spaces. Our expectation is that the information spaces and interactive views present opportunities for learning about the microbial genomes. An overview of the resources developed is presented in the figure in the Appendix section.

A supercomputing infrastructure (Blue Waters Supercomputer) provided computational tools to construct information spaces while visual analytics software and online bioinformatics resources provided tools to interact with the constructed information spaces. The Rhizobiales order of bacteria that includes the *Brucella* genus was the use case for preforming the complex cognitive activities. An interesting finding among the *Brucella ceti* genomes was that strain Cudo is unique for a predicted four-gene transcriptional unit that contain genes known to respond to limited nitrogen availability.

# 6. REFLECTIONS

## 6.1 Dominique Smith-McInnis, PhD Candidate in Environmental Science at Jackson State University, Mississippi.

The main goal of my doctoral research is to generate knowledge on the biological processes in *Brucella* species that include universal stress proteins. I am a recipient of career development fellowships from the Institute for Infectious Animal Diseases, a Department of Homeland Security Science & Technology Center of Excellence at Texas A & M University. I have conducted biological research using computational techniques and resources. This manuscript shows examples of knowledge on universal stress proteins of *Brucella* species that were generated using bioinformatics and visual analytics tools. The learning experiences from my doctoral research has equipped me for a career in K-12 education and higher education.

## 6.2 Kiara Wootson, Undergraduate Student in the College of Science, Engineering and Mathematics, Bethune-Cookman University, Daytona Beach, Florida.

I was an intern in Blue Waters Internship Program from May 2015 to April 2016. At the beginning of the internship I attended the two-week Blue Waters 2016 Petascale Institute held at the University of Illinois Urbana-Champaign (UIUC) from May 24th to June 5th 2015. I gained an introduction to high performance computing. During my mentored internship at Bethune-Cookman University I became familiar with command-line instructions for performing computing actions. The internship training has helped me to better understand microbial genomes as well as data visualization techniques. I have a clearer understanding of career pathways that incorporate computational science.

# 7. ACKNOWLEDGMENTS

National Science Foundation Grant Award: HRD-1435186. KMW and RDI acknowledge the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. RDI, DRSM and SSS acknowledge funding support from the U.S. Department of Homeland Security Science and Technology Directorate: 2011-ST-062-000048. DRSM acknowledge Fellowship from National Center for Foreign Animal and Zoonotic Disease Defense and the Environmental Science PhD Program. **Disclaimer:** "The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies".

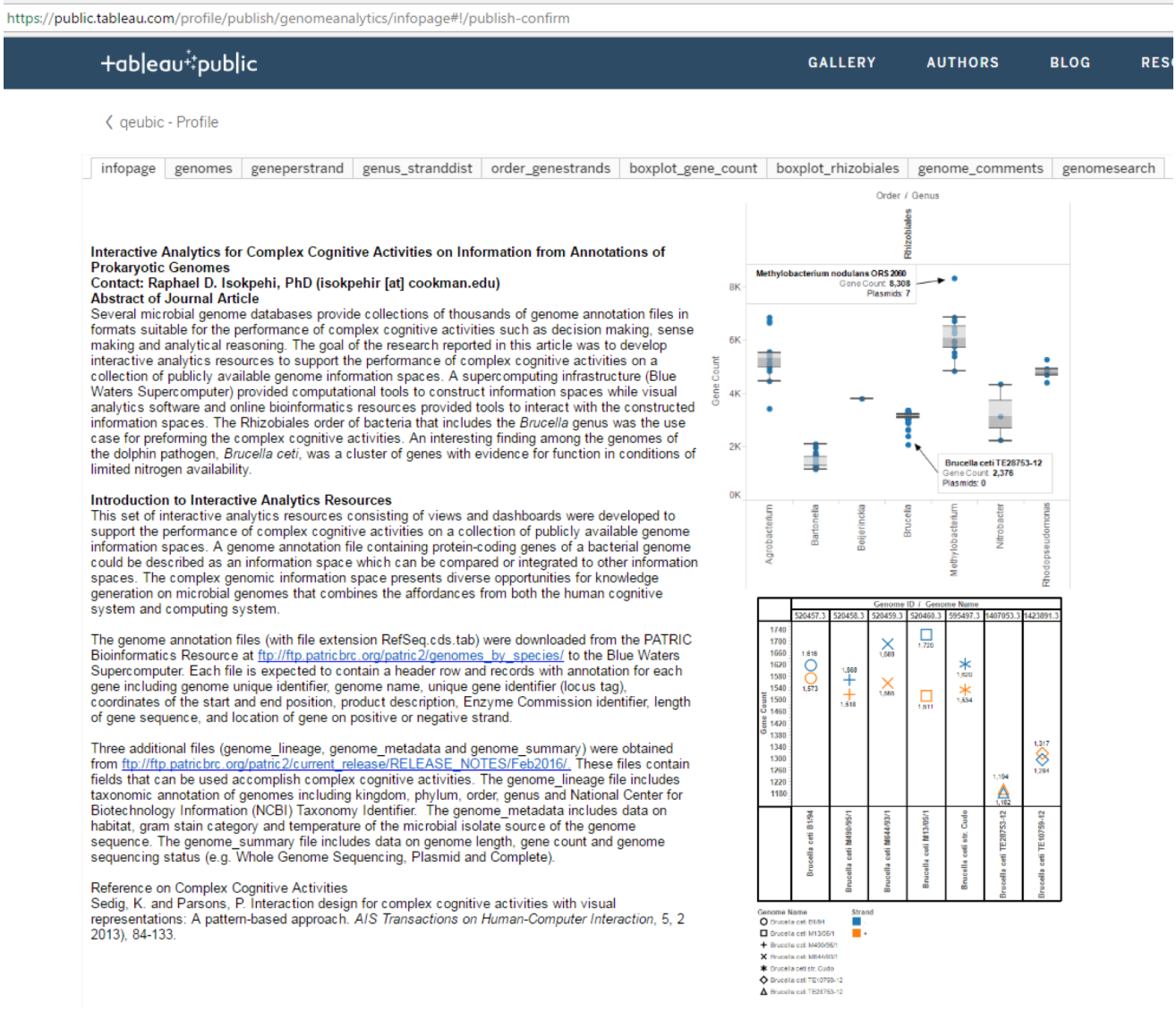
# 8. REFERENCES

1. Chen I-MA, Markowitz VM, Palaniappan K, Szeto E, Chu K, Huang J, Ratner A, Pillay M, Hadjithomas M, Huntemann M: **Supporting community annotation and user collaboration in the Integrated Microbial Genomes (IMG) system**. *BMC Genomics* 2016, **17**:307.
2. Mavromatis K, Ivanova NN, Chen I-MA, Szeto E, Markowitz VM, Kyrpides NC: **The DOE-JGI Standard operating procedure for the annotations of microbial genomes**. *Standards in Genomic Sciences* 2009:63-67.
3. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I: **RefSeq microbial genomes database: new representation and annotation strategy**. *Nucleic Acids Research* 2013:D553-D559.
4. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R: **PATRIC, the bacterial bioinformatics database and analysis resource**. *Nucleic Acids Research* 2013:D581-D591.
5. Albers MJ: **Human–Information interaction with complex information for decision-making**. *Informatics* 2015, **2**(2):4-19.
6. Parsons P, Sedig K: **Distribution of information processing while performing complex cognitive activities with visualization tools**. In: *Handbook of Human Centric Visualization.* Springer; 2014: 693-715.
7. Sedig K, Parsons P: **Interaction design for complex cognitive activities with visual representations: A pattern-based approach**. *AIS Transactions on Human-Computer Interaction* 2013, **5**(2):84-133.
8. Di Martino C, Kalbarczyk Z, Iyer RK, Baccanico F, Fullop J, Kramer W: **Lessons learned from the analysis of system failures at petascale: The case of Blue Waters**. In: *Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on: 2014*: IEEE; 2014: 610-621.
9. Sacha D, Stoffel A, Stoffel F, Kwon BC, Ellis G, Keim DA: **Knowledge generation model for visual analytics**. *Visualization and Computer Graphics, IEEE Transactions on* 2014, **20**(12):1604-1613.
10. Sedig K, Parsons P: **Design of visualizations for human-information interaction: A pattern-based framework**. *Synthesis Lectures on Visualization* 2016, **4**(1):1-185.
11. Sedig K, Parsons P, Dittmer M, Haworth R: **Human-centered interactivity of visualization tools: Micro-and macro-level considerations**. In: *Handbook of Human Centric Visualization.* Springer; 2014: 717-743.

12. Isokpehi RD, Udensi UK, Simmons SS, Hollman AL, Cain AE, Olofinsae SA, Hassan OA, Kashim ZA, Enejoh OA, Fasesan DE: **Evaluative profiling of arsenic sensing and regulatory systems in the human microbiome project genomes**. *Microbiology Insights* 2014, **7**:25-34.
13. Mbah AN, Isokpehi RD: **Application of universal stress proteins in probing the dynamics of potent degraders in complex terephthalate metagenome**. *BioMed Research International* 2013, **2013**:196409.
14. Carvalho FM, Souza RC, Barcellos FG, Hungria M, Vasconcelos ATR: **Genomic and evolutionary comparisons of diazotrophic and pathogenic bacteria of the order *Rhizobiales***. *BMC Microbiology* 2010, **10**:37.
15. Erlacher A, Cernava T, Cardinale M, Soh J, Sensen CW, Grube M, Berg G: **Rhizobiales as functional and endosymbiontic members in the lichen symbiosis of *Lobaria pulmonaria* L**. *Frontiers in Microbiology* 2015, **6**:53.
16. Parte AC: **LPSN—list of prokaryotic names with standing in nomenclature**. *Nucleic Acids Research* 2014, **42**(D1):D613-D616.
17. Bode B, Butler M, Dunning T, Gropp W, Hoe-fler T, Hwu W-m, Kramer W: **The Blue Waters Super-System for Super-Science. Contemporary HPC Architectures, Jeffery Vetter editor**. In.: Sitka Publications, November; 2012.
18. Lurie NH, Mason CH: **Visual representation: Implications for decision making**. *Journal of Marketing* 2007, **71**(1):160-177.
19. Simmons SS, Isokpehi RD, Brown SD, McAllister DL, Hall CC, McDuffy WM, Medley TL, Udensi UK, Rajnarayanan RV, Ayensu WK: **Functional annotation analytics of *Rhodopseudomonas palustris* genomes**. *Bioinformatics and Biology Insights* 2011, **5**:115-129.
20. Moreno E, Guzmán-Verri C, Gonzalez-Barrios R, Hernandez G, Morales JA, Barquero-Calvo E, Chaves-Olarte E: ***Brucella ceti* and brucellosis in cetaceans**. *Frontiers in Cellular and Infection Microbiology* 2012, **2**:3.
21. Parsons P, Sedig K: **Common visualizations: Their cognitive utility**. *Handbook of Human Centric Visualization* 2014:671-691.
22. Isokpehi RD, Simmons SS, Cohly HH, Ekunwe SI, Begonia GB, Ayensu WK: **Identification of drought-responsive universal stress proteins in viridiplantae**. *Bioinformatics and Biology Insights* 2011, **5**:41-58.
23. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Research* 2016, **44**(D1):D471-D480.
24. Tøndervik A, Torgersen HR, Botnmark HK, Strøm AR: **Transposon mutations in the 5′ end of glnD, the gene for a nitrogen regulatory sensor, that suppress the osmosensitive phenotype caused by otsBA lesions in *Escherichia coli***. *Journal of Bacteriology* 2006, **188**(12):4218-4226.
25. Dumancas GG: **Applications of supercomputers in sequence analysis and genome annotation**. *Research and Applications in Global Supercomputing* 2015:149-175.
26. Singh P: **Big genomic data in bioinformatics cloud**. *Applied Microbiology: Open Access* 2016, **2**(1000113):2.
27. Williamson DF, Parker RA, Kendrick JS: **The box plot: a simple visual method to interpret data**. *Annals of Internal Medicine* 1989, **110**(11):916-921.
28. Jourand P, Giraud E, Béna G, Sy A, Willems A, Gillis M, Dreyfus B, de Lajudie P: ***Methylobacterium nodulans* sp. nov., for a group of aerobic, facultatively methylotrophic, legume root-nodule-forming and nitrogen-fixing bacteria**. *International Journal of Systematic and Evolutionary Microbiology* 2004, **54**(6):2269-2273.
29. Marx CJ, Bringel F, Chistoserdova L, Moulin L, Haque MFU, Fleischman DE, Gruffaz C, Jourand P, Knief C, Lee M-C: **Complete genome sequences of six strains of the genus *Methylobacterium***. *Journal of Bacteriology* 2012, **194**(17):4746-4748.
30. Gäll R, Leisinger T: **Plasmid analysis and cloning of the dichloromethane-utilization genes of *Methylobacterium sp.* DM4**. *Microbiology* 1988, **134**(4):943-952.
31. Jiao Y, Kappler A, Croal LR, Newman DK: **Isolation and characterization of a genetically tractable photoautotrophic Fe (II)-oxidizing bacterium, *Rhodopseudomonas palustris* strain TIE-1**. *Applied and Environmental Microbiology* 2005, **71**(8):4487-4496.
32. Wu Q, McFee WE, Goldstein T, Tiller RV, Schwacke L: **Real-time PCR assays for detection of *Brucella* spp. and the identification of genotype ST27 in bottlenose dolphins (*Tursiops truncatus)***. *Journal of Microbiological Methods* 2014, **100**:99-104.
33. Setubal C, Brettin T, Sobral BW, Boyle SM, Tsolis J, Munk C, Tapia R, Han C, Detter J, Bruce D: **Genomes reveals *Brucella* analysis of ten**. *Journal of Bacteriology* 2009, **191**(11):3569.
34. Caspi R, Billington R, Foerster H, Fulcher CA, Keseler I, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q: **BioCyc: Online Resource for Genome and Metabolic Pathway Analysis**. *The FASEB Journal* 2016, **30**(1 Supplement):lb192-lb192.
35. Williams BS, Isokpehi RD, Mbah AN, Hollman AL, Bernard CO, Simmons SS, Ayensu WK, Garner BL: **Functional annotation analytics of *Bacillus* genomes reveals stress responsive acetate utilization and sulfate uptake in the biotechnologically relevant *Bacillus megaterium***. *Bioinformatics and Biology Insights* 2012, **6**:275-286.
36. Gumber S, Taylor DL, Marsh IB, Whittington RJ: **Growth pattern and partial proteome of *Mycobacterium avium* subsp. paratuberculosis during the stress response to hypoxia and nutrient starvation**. *Veterinary Microbiology* 2009, **133**(4):344-357.
37. Kvint K, Nachin L, Diez A, Nyström T: **The bacterial universal stress protein: function and regulation**. *Current Opinion in Microbiology* 2003, **6**(2):140-145.
38. Gustavsson N, Nyström T: **The universal stress protein paralogues of *Escherichia coli* are co-ordinately regulated and co-operate in the defence against DNA damage**. *Molecular Microbiology* 2002, **43**(1):107-117.
39. Doublié S, Bricogne G, Gilmore C, Carter Jr CW: **Tryptophanyl-tRNA synthetase crystal structure reveals an unexpected homology to tyrosyl-tRNA synthetase**. *Structure* 1995, **3**(1):17-31.
40. Ruiz N: **Bioinformatics identification of MurJ (MviN) as the peptidoglycan lipid II flippase in *Escherichia coli***. *Proceedings of the National Academy of Sciences* 2008, **105**(40):15553-15557.
41. Yurgel SN, Rice J, Kahn ML: **Transcriptome analysis of the role of GlnD/GlnBK in nitrogen stress adaptation by**

***Sinorhizobium meliloti*** **Rm1021**. *PloS One* 2013, **8**(3):e58028.
42. Pashley CA, Brown AC, Robertson D, Parish T: **Identification of the *Mycobacterium tuberculosis* GlnE promoter and its response to nitrogen availability**. *Microbiology* 2006, **152**(9):2727-2734.
43. Ronneau S, Moussa S, Barbier T, Conde-Álvarez R, Zuniga-Ripa A, Moriyon I, Letesson J-J: ***Brucella*, nitrogen and virulence**. *Critical Reviews in Microbiology* 2014:1-19.
44. Schaible UE, Kaufmann SH: **A nutritive view on the host–pathogen interplay**. *TRENDS in Microbiology* 2005, **13**(8):373-380.

## 9. APPENDIX




**Interactive Analytics Resources for Complex Cognitive Activities on Information from Annotations of Prokaryotic Genomes**
Website: https://public.tableau.com/profile/publish/genomeanalytics/infopage
This set of interactive analytics resources consisting of views and dashboards were developed to support the performance of complex cognitive activities on a collection of publicly available genome information spaces.