**A. Significance**

***Need for integrative approaches in the post-GWAS era:*** With the advances in high throughput SNP typing and next-generation sequence technologies, the numbers of loci and variants associated with human complex traits and diseases are growing. For example, over the past five years the number of complex trait loci identified by genome-wide association studies (GWAS) has grown exponentially (Figure 1). The NHGRI GWAS catalog currently (as of 7/7/11) lists 933 publications and 4,621 associated SNPs for large scale studies of human diseases and/or traits (Hindorff LA, Junkins HA, Hall PN, Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed 7/7/11). Furthermore, whole-exome and whole genome re-sequencing to identify causes of Mendelian disorders and determine the genome sequence of individuals is now a reality (REFS) and the likelihood of its application for complex traits is imminent. However, the next challenge in the post-GWAS era is to identify the causal genes or sequence features, as many associations are intronic or intergenic, and are likely to have complex and epistatic relationships. Certainly, with the large number of human GWA studies, functional proof of causality of associated SNPs/loci is often provided through data and experimentation in model organisms. Having an integrated cross-species platform for individual variants, traits, phenotypes and diseases can lead to prioritization of candidate genes or pathways as well as choice of the ‘best’ model for functional validation.

**Figure 1.** Number of publications reporting GWAS. PubMed search criteria (as of 7/7/11): (genome-wide association OR GWAS). Review articles excluded.

***Comparative Genomics to link common biology across species:***The primary motivation for comparative mapping and comparative genomics is to leverage knowledge obtained from onr organism and to transfer that knowledge to other organisms of interest. Therefore, simply aligning genomes is not sufficient to link the biology between model organisms; one must also annotate that biology onto the genome of each organism, under the assumption that some genetic mechanisms will play a role in that trait/disease across species. Although met with some initial disbelief, using conserved synteny between genomes [14; 15; 16; 17; 18] to map disease-causing genes or regions from one organism to another has begun to bear fruit. With the availability of additional genome resources for other species, increasing numbers of studies are using cross species comparisons to find causes of diseases, both single gene disorders and complex disease [20; 21]. Concerted efforts such as human GWAS, the mouse collaborative cross (REF) and the sequencing of multiple strains in such model organisms as the rat (Worthey et al) are expanding the resources to map complex disease loci at high resolution, and efforts such as the 1000 genomes project are setting the stage for studying variants in individual human genomes. Integration of the genome sequence with existing mapping data and the biological data attached to those maps, with annotation of a comprehensive catalog of gene products will increase the use of such comparative studies on translational research.

***VCMap links the genomes and their biology:***Most genome browsers (e.g. UCSC (REF), NCBI (REF) and Ensembl (REF)) are very successful at cross-species genome alignments, and allow deeper mining of genomic data in species with more complete annotation. The VISTA [27] and ECR [28] browsers provide multiple-species sequence alignments to facilitate the identification of conserved coding and non-coding regions and putative functional genome elements. However, most comparative genome tools available to date do not systematically annotate and integrate biology across species. To meet this need, Dr. Kwitek and colleagues at the Medical College of Wisconsin developed a novel and powerful integrative genomics environment with markers, maps, and annotations; a source file of the maps, markers, and physiological results; and a cross-organism bioinformatic mapping tool to address these issues [29, REFS]. VCMap (<http://vcmap.bioneos.com/>) is a Java-based, dynamic software tool with a web interface that allows an investigator to dynamically navigate from map to map, genome to genome, and thus uniquely combines and visualizes distinctive information linked to each type of map. In particular, it is a powerful strategy to link physiologically important and genetically influenced traits, phenotypes, and diseases which can, via their genome positions, be displayed and tracked from genetic maps to genomic sequence and thus link to genome browsers for review of candidate gene annotations.

Since its development, the VCMap platform has expanded from rodent-human based comparisons to that of multiple domesticated agricultural species, focused mainly on QTL through a USDA-funded collaborative effort between Drs. Kwitek, Reecy, Shimoyama, and Dwinnel. An infrastructure has been developed to support multiple map types including genetic, RH, and multiple genome assemblies for human, rodents and livestock species, as well as multiple annotation types including genes and QTL linked to respective disease and phenotype ontologies. This project will expand on the current infrastructure to integrate multiple human and model organism GWAS annotations, additional and more refined conserved synteny comparisons, the ability to load and visualize personal datasets such as mRNA-seq and individual variation data, and integrated ontology annotations to serve a broad number of researchers.

***Standardized ontologies for annotating biological data:*** Clinical and model organism phenotype characterization is critical to genetic studies. While the types of measures and methodologies vary, the basic traits being measured are often similar, offering a wealth of information to guide human genetic studies of complex traits. Ontologies provide a convenient and consistent way by which data can be classified, annotated and compared among humans and model organisms. The first large-scale effort at applying bio-ontologies was the Gene Ontology (GO) project, which was created to standardize terminology and capture the characteristics of genes and gene products across multiple databases and for multiple species (Blake et al). The success of GO is demonstrated in part by the dozens of groups which use GO as the primary ontology to annotate genes, and the scores of tools utilizing these annotations to classify, organize and analyze such datasets. Recognizing the utility of ontologies and the need for their expanded use, the NIH provided funding for the creation of the National Center for Biomedical Ontology (NCBO) (http://bioontology.org/) in 2006 (Rubin et al).There are currently 270 ontologies catalogued at NCBO including a number which focus on phenotypes and diseases. Two of the disease ontologies at NCBO, OMIM and Medical Subject Headings (MeSH), have been integrated into a single ontology currently used by the Rat Genome Database and Mouse Genome Informatics for annotating genes. The Mammalian Phenotype Ontology (Smith CL et al) is also used by RGD and MGI to annotate alleles and genes related to abnormal morphological and physiological characteristics.

These and other ontologies have been developed primarily to annotate specific genes, alleles and mutants, while other types of mapped loci such as QTL and SNP associations have been annotated by vocabularies or names developed at individual databases. Unfortunately, as a result of independent development, the vocabularies have limited utility for cross species comparisons and are often ill-suited to loci in which the exact variant is unknown. To standardize QTL and other locus annotations, we are developing the Vertebrate Trait Ontology (VTO). Its focus is to outline the developmental, morphological and physiological characteristics or states that are found in all members of a vertebrate species, as opposed to the pathological states/phenotypes that are the focus of the Mammalian Phenotype (MP) Ontology. Mapped loci underlie the expression of these traits. The assessment of such traits and the resulting measured values are used by researchers and clinicians to *define* abnormal phenotypes and disease states. In general, disease annotations have indicated a relationship between a genomic region or specific variant(s) and individuals or groups of individuals who exhibit the collective phenotypes which define a disease, while the phenotype annotations generally indicate a direct relationship between a single allele or gene and a single abnormal characteristic. By standardizing and integrating these three types of annotations and genomic elements across species, we will be able to provide researchers with Virtual Morbid Maps illustrating all loci and regions of a species which may be involved in various states and processes associated with disease, while maintaining the distinction and meaning of the individual annotation types.

***Summary****:* We envision the new VCMap as a powerful and flexible data exploration tool for integrating genomic maps and biological annotation between human and multiple model organisms. It will allow a user to select the comparison they would like to visualize and then allow dynamic zoom, filter, and extended annotation features to mine deeper into the genomic and biological data associated with a region of interest across multiple species. VCMap will also link to related genome browsers, and provide a means to download associated data in report and image formats. By utilizing a web service architecture and the ability to easily upload one’s own data, VCMap will change from a simple visualization tool to a dynamic visualization and annotation browser and will be a data provider in its own right. We envision VCMap becoming both a tool for the end user and a tool/resource for other bioinformaticians on which to build further tools, visualize and download annotations greatly increasing its impact. In addition to the bioinformatics tools that will result from this proposal, the data that we will collect, curate, and make available to the public also has great value.

**B. Innovation**

* *A highly interactive integration tool to link biological attributes across species*: While most genome browsers and comparative genomics tools include conserved synteny and gene content information, to our knowledge none systematically annotate and integrate mammalian biology across species. We will expand the current software to develop the first comprehensive genome integration platform to link biological annotations of genomic loci across multiple species.
* *An integrated set of ontologies to fully link multiple types of biological annotations.* One of the stated goals of this PAR is to “combine existing software with modern ontologies or libraries of controlled vocabularies”. In order to ensure efficient transfer of biological information between species, three major ontologies will be used for standardizing data, including the established Comparative Disease and Mammalian Phenotype Ontologies as well as a novel Vertebrate Trait Ontology which will allow cross-species QTL and other mapped locus comparisons for the first time.
* *A cross-species morbid map:* Currently loci related to diseases, phenotypes and traits can be found on multiple map types for multiple species. There is the OMIM Morbid Map, the NHGRI GWAS Catalogue as well as gene and QTL maps for human, mouse, rat, cow, pig and other species. Each map uses different vocabularies to indicate relationships to diseases, phenotypes or traits while often providing a view of a single genomic element type. By leveraging three ontologies to standardize annotations of genes, alleles and QTLs from multiple species, direct comparisons across data elements and species is possible. In addition, by mapping such loci using conserved synteny and orthology, virtual maps for human, rat, mouse and other species can be created which will display the related mapped traits, diseases and phenotypes within the genomic context of the species of interest. Because it leverages phenotype and disease loci data from multiple organisms, we will create more complete morbid maps than exist today.
* *A platform that can compare maps and annotations both within and between species:* The tool allows comprehensive and flexible comparative analysis: for species with multiple genome assemblies; for species with preliminary genome sequence but other established genome maps; for individual’s or strain’s personal genome data, and the capability to link these data across species..

**C. Approach**

**C.1 Preliminary Data**

Our intent in this section is to demonstrate our abilities by showing data and other evidence that we are positioned to accomplish the objectives we have outlined for the proposed project. In the following sections we describe our proposed approaches.

***C.1.1 VCMap Visualization tool:***

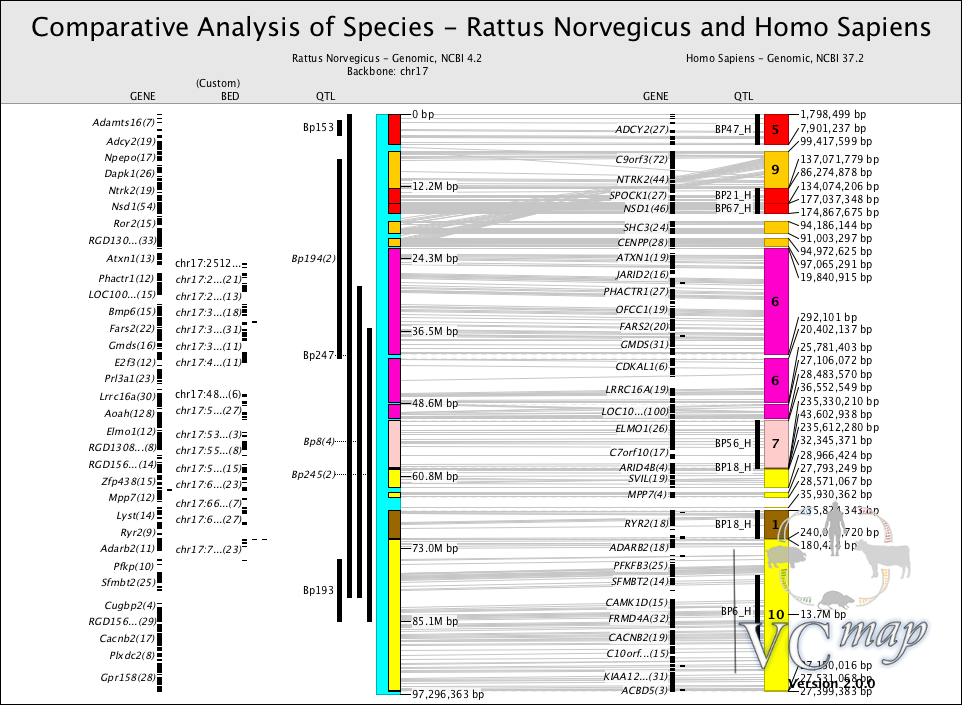
The Virtual Comparative Map (VCMap) tool was originally developed by Drs. Kwitek, Howard Jacob, Simon Twigger and Peter Tonellato at the Medical College of Wisconsin, as part of the Rat Genome Database (REFs). It has been subsequently redesigned through collaborative efforts between Drs. Kwitek, Reecy, Shimoyama and Dwinell, and implemented by Bio::Neos, Inc (<http://bioneos.com/>), as part of a USDA funded project to develop a comparative QTL viewer. The current VCMap tool made the system more user friendly and intuitive, expanding its versatility and utility. We have developed a MySQL database to house all map and annotation data for each species. Map data including publicly available genome assemblies, genetic maps, and radiation hybrid maps for six species are currently in the database, including human, rat, mouse, cow, pig, and chicken. A new Java front end was developed that allows a user to load a backbone map and those of additional species and then overlay selected annotation data as shown in Figures 2 and 4. Current conserved synteny data was obtained from UCSC and utilizes level 1 (top-scoring) chain data. Annotation data includes genes (appearing as default annotation), QTL, STS, and pseudogenes (Table 1). All current annotation data, with the exception of QTL, is obtained from NCBI. Thus far, we have gene and pseudogene annotations from all six species, and QTL data for all species (except mouse) through the curation efforts at ISU and MCW. Mouse QTL curation is obtained through MGD.

**Table 1.** Annotation data in VCMap.



The database structure for the VCMap system is a unique schema that was designed specifically for use with the VCMap front-end client tool. The database primarily stores three major elements, 1) the maps and their associated annotations, not including any sequence data; 2) the relationships between annotations; and 3) the relationships between large blocks of genomic sequence, identified as syntenic blocks. The storage for the maps is fairly straightforward as the database simply records the map type, data source, and number and length of chromosomes. For the annotations for each map, the storage is straightforward as well. Each annotation stores its type, name, position (based on the units of the map for which it is associated), and any additional information that we can gather from the data source. For features that have been mapped across multiple map types, for example STS markers that are part of the genetic linkage map and have also been placed on the genomic sequence map, we create multiple copies of the same information but different positions. In order to handle the relationships between these annotations that represent the same feature in different maps, as well as annotations such as genes, that have been structurally, functionally, or otherwise identified as similar across species, we have a table in the database that creates a many-to-many relationship between rows in our annotation table. The data in this table is generated by identifying features with the same source and reference identifier, such as a matching UniGene ID, as well as using the NCBI Homologene dataset to identify sequence based homologs across supported species. Finally, the relationships between the large syntenic regions of the maps across species are generated by using the net conservation track from UCSC. We identify regions greater than 100 kbps from the level 1 data from the conservation track and store the positions on each map for this data.

The Java front end allows a user to load a backbone map, add those of additional selected species and then overlay annotation data. A user can switch species perspectives on the fly using a “swap backbone” feature, add and filter data as needed, and download location specific features. Annotations are clickable to display additional information and links to source data. VCMap version 2.0 has made significant improvements to the user interface and supported features of the software tool. Usability of the user interface has been improved, especially with respect to the navigation of large-scale regions on the order of multiple megabases. Most importantly, the tool now supports the ability to load custom datasets from text files in a variety of common formats including GFF3, BED, and SAM/BAM. This feature greatly increases the utility of the tool because it provides users with the ability to perform comparative research using their own personal, private research data with VCMap as a framework for navigating and investigating their data. The user also can export publication quality images of the data visualization. For example, Figure 2 is a picture output of rat mRNA-seq data uploaded as a BED file, aligned to rat genes, QTL, the genome map of rat chromosome 17 aligned to the conserved syntenic regions of human and its corresponding genes and QTL. Features are clickable, link to additional source data, and can be downloaded by the user.

****

**Figure 2.** Publication quality figure showing user uploaded mRNA-seq data aligned to rat chromosome 17 and human conserved synteny, along with genes and blood pressure QTL for both species.

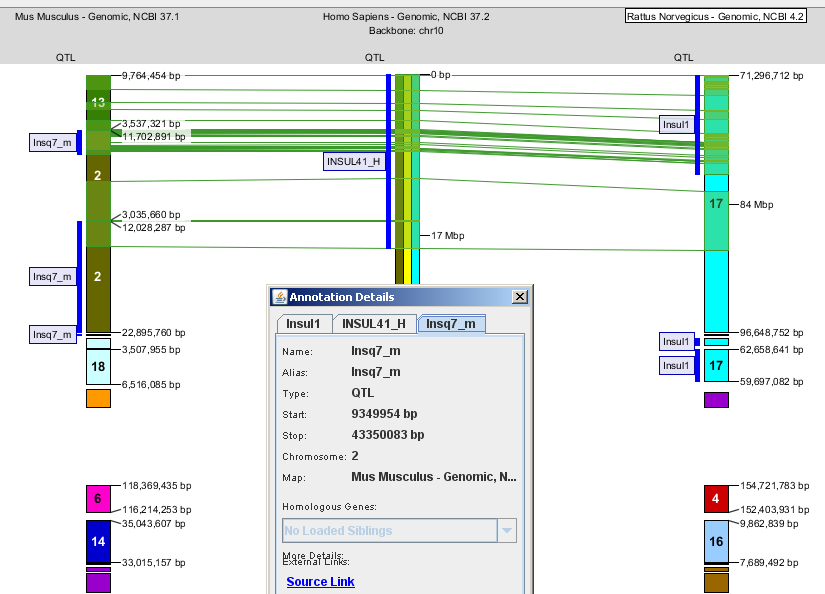
***C.1.2 Curation and ontology for QTL:***

One of the major recent efforts to improve VCMap has been to curate QTL across all species included in the current version. Although QTL data has been available by species through RGD, MGI, and Animal QTLdb, VCMap provides a platform for the integration of data from human, rat, mouse, and livestock. Table 1 indicates the number of QTL for a variety of species included in VCMap. Through concerted efforts by us as well as by RGD, the Animal QTL Database and MGI, VCMap currently contains over 26,000 curated QTL for human, rodents, and several livestock species. Curated QTL typically include information on: 1) trait evaluated 2) chromosome; 3) QTL location, whether denoted by cM position and interval, or by peak/flanking markers; and 4) at least one statistical measure (e.g., p-value, F-statistic, LOD score). Each publication or pre-publication report from which data is to be entered is assigned a unique ID (PubMed ID if available). Curation is divided into three areas: 1) reference information, including article author(s), title, year, journal information, and abstract; 2) experiment information, including sample information experimental design, and analysis method; and 3) QTL details, including trait, QTL position, statistics, QTL effects, and candidate gene information. Traditionally QTL have been annotated using the vocabularies developed by each source making comparisons difficult; therefore, standard annotation to the Vertebrate Trait Ontology (VTO) is a goal of this application.



**Figure 3.** Organizational structure of the Vertebrate Trait Ontology for several organ systems. The + indicates further tree expansion.

The Vertebrate Trait Ontology was created to provide a standardized vocabulary for annotating trait data to facilitate comparison across vertebrate species. Through collaboration between the Rat Genome Database, Animal QTLdb, and Mouse Genome Informatics, the VTO was created to represent developmental, physiological, and morphological states and characteristics for multiple vertebrate species. Unlike disease or phenotype ontologies, the VTO represents the average states and characteristics of an organism rather than those which are abnormal or pathological in nature, making it more suitable for QTL or other mapped loci in which the exact allele or mutation which might result in the pathological state is unknown. In general, the VTO (Figure 3) has followed an organ system approach as have other related ontologies – the Mammalian Phenotype Ontology (Smith and Eppig), Human Phenotype Ontology ( Robinson and Mundlos), the Foundational Model of Anatomy (FMA) (Cook et al), Mouse Adult Gross Anatomy (MA) (Hayamizu et al) and other anatomy ontologies.  In addition, we have included as part of the terms, where possible, those used for qualities as outlined in the Phenotype Quality Ontology (PATO) (Gkoutos et al). This will allow for deconstruction of our “precomposed” ontology into “decomposed elements” and will make it possible to cross reference other ontologies and make connections to them. The VTO uses the Open Biological and Biomedical Ontologies (OBO) format. To date there are over 3,060 terms in the Vertebrate Trait Ontology with definitions provided for each term and synonyms, with external database identifiers included to cross reference related terms in other ontologies. Terms are added to the existing basic structure as QTL are addressed for annotation or mapped trait-related papers are reviewed, and definitions are derived from sources such as medical dictionaries, physiological textbooks and published research papers. The VTO is scheduled for release to NCBO in September 2011 and is currently available at <http://rgd.mcw.edu/pub/ontology/vertebrate_trait.obo>.



**Figure 4.** Screenshot of comparative QTL mapping in VCMap. Shown is the top of human chromosome 10 (middle) aligned to the conserved regions of mouse (left) and rat (right). The shaded regions, with green lines connecting the genomes, at the top are in conserved synteny between the species. QTL for each species are annotated to the left of each respective species. Boxed in QTL involve plasma insulin levels in conserved synteny between all three species.

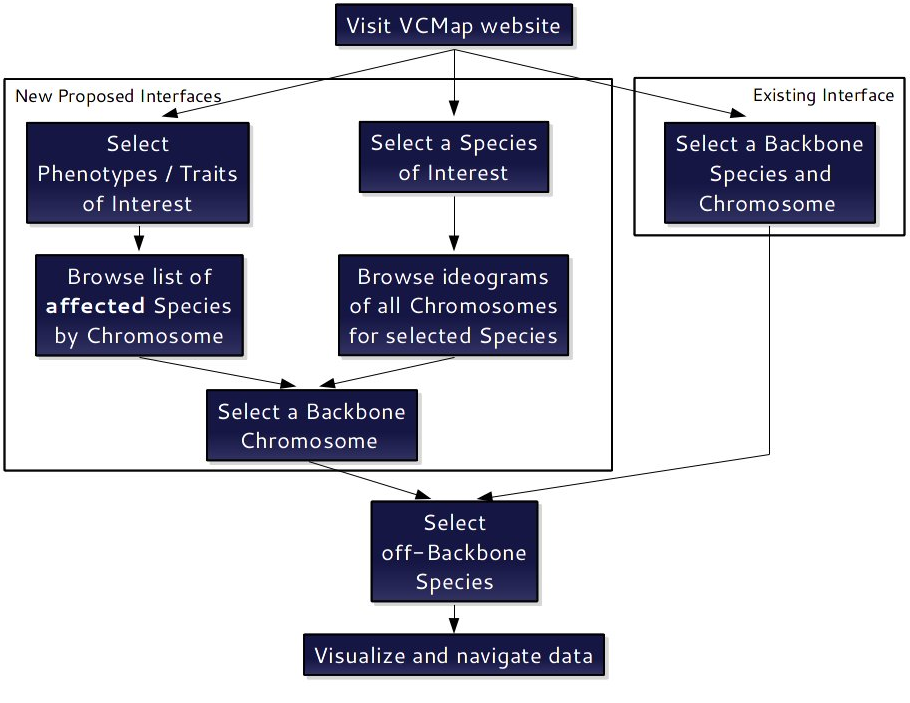
***Summary:***While we have made significant improvements to VCMap, we propose many key improvements that will make this a unique and broad-based platform for the community. First, we will include several additional genome features/annotations, such as SNPs and other genome variants as well as additional maps, including alternative genome assemblies. Second, we will link VCMap to the genome-wide ontology-based viewer, GViewer (REF) to allow a genome wide snapshot of specific traits, phenotypes, and diseases, whose detail can be further investigated at the chromosome and subchromosomal views in VCMap. Third, we will incorporate human and model organism gene associations and allele phenotypes into the annotations. Finally, we need to improve the ontology linking function. For instance, identifying the insulin QTL within shared conserved synteny in Figure 4 required the user to recognize the QTL names as relating to insulin levels. There was no shared ontology term to link them because curators for each species names and annotates them independently. Furthermore, the same intervals also contain QTL for adiposity and body weight. If an investigator is interested in the metabolic syndrome for example, it would be useful to identify these related QTL, as well as any known gene associations for related traits, gene mutation phenotypes, and diseases, with a single query. The new VCMap tool will be designed to address these complex questions by integrating multiple ontology annotations into our curation and developing a user friendly query tool.

**C.2 Specific Aim 1. Generate a comprehensive platform to integrate inter- and intra-species genomic and phenotype/disease information and expand the genomic features incorporated in VCMap.**

***C.2.1 Rationale:*** The VCMap application is a functional, useful tool for facilitating comparative genomics because of its uncommon ability to allow users to begin with a generalized knowledge of a phenotype, trait, or even chromosomal position and interrogate data across species, refining the region of interest and identifying other potential candidates for additional analysis. Most other genome browsers, whether designed for comparative research or single species, expect users to begin with a fairly specific region of interest and allow them to navigate and explore surrounding data from that narrow starting viewpoint. As these browsers are asked to expand their range of visualization, they often perform poorly or present the user with a cluttered, unusable view of the data. Currently, VCMap is limited in the number of supported species, the amount and types of data that it can present, and the interoperability of the tool with other common genomic research tools. Aspects of the user interface have room for improvement as well, especially the initial interface used to launch the tool and select. Furthermore, enhancements to the application itself, and the scripts used to gather data and populate the database, will greatly increase the utility of the tool and make it more generally applicable to the comparative research community.

***C.2.2 Experimental Approach:***

*2.2.1 Generate a multifaceted user interface for loading VCMap:* The objective of this approach is to improve how the user interacts with the application to begin a session. The current user interface for the application requires the user to specify what maps they would like to load, one map at a time, followed by specifying what types of annotation to load, also one type at a time. In addition, the interface is very chromosome-centric, as it requires the user to select a single chromosome for the backbone map as the main starting point for loading all data. This process is tedious and forces users to begin their usage of the system from a specific location. Both of these issues can be addressed through user interface improvements (Figure 5). In order to streamline the loading process and present users with more focused data, the load interface will be redesigned to present the user with a dialog that will allow them to specify all of the species of interest, as well as supporting annotation in one step. This redesigned interface will be more efficient, and more intuitive, allowing users to begin examining the visualization of their data sooner and making the tool more accessible to new users unfamiliar with the system.



**Figure 5.** Current and proposed user work flow..

In addition to this improved load dialog, we plan to add two other entry points to the system to allow for parallel modes of entry into the data visualization aspect of the tool. First, in order to broaden the chromosome-centric focus of the tool, we plan to add visualization for a full genome (as a collection of all of the relevant chromosomes for a species). This interface will be limited to a single species and less resolution, but will still provide the user with a more general entry point into the system than the current interface, for example through markers, bp positions or a QTL name or ID. After specifying a chromosome of interest as a backbone, the system will then allow the user to visualize additional species on top of this backbone. Second, we will add another interface to the system that will allow users to begin using the system from the starting point of a phenotype or trait of interest as opposed to a genomic location. This interface will rely on heavily on ontology terms and associations that we will preload into the VCMap database. The interface will allow users to query on these terms and a single species, similar to the ontology viewer at the Rat Genome Database (RGD) (http://rgd.mcw.edu/tools/ontology/ont\_search.cgi) (Dwinell et al) and present them with a visualization of the locations throughout the genome of a single species where this term is associated. As with the genome-based entry point, the user will be able to visualize additional species after selecting a single chromosome of interest.

*2.2.2 Increase the amount of data supported by VCMap:* We will increase the utility of the VCMap software by adding additional data sources into the preloaded VCMap database, to streamline the current data loading process and supporting database, and to improve the performance of the custom data loading feature to allow the system to handle larger datasets (including next generation sequencing datasets). The current VCMap system contains a MySQL based database backend that we preload with map and annotation data from various public data sources including NCBI, UCSC, and Ensembl as well as QTL data from ISU and MCW. In order to increase the utility of this tool for comparative genomics research with respect to human disease, we plan to add data from the OMIM Morbid map (REF), the GWAS catalog at NHGRI (REF), and dbSNP at NCBI (REF). Furthermore, the Mouse Genome Informatics (www.informatics.jax.org accessed July 29 2011) lists 210,965 phenotype annotations for 44,229 genotypes. These include spontaneous mutations, targeted knockouts, chemically induced mutations and other variations. The phenotype annotations association with gene alleles have been incorporated into RGD and attached to the mouse genes, thus this dataset can easily be incorporated into VCMap. Additional mouse variation-phenotype datasets such as the Mouse Collaborative Cross (see C.4 for details), will be reviewed and pipelines developed for addition to VCMap. Finally in addition to rat genes and QTL, we will incorporate congenic strains with mapped positions and phenotype annotations as well as emerging rat knockout strains with defined phenotypes. We will tie these sources with ontology terms, when possible, that are strongly integrated into the rest of the system, adding filtering and querying capabilities that will be extremely useful to researchers of human diseases. Finally, to supplement the new data, we will improve the current performance of the tool to support datasets in the size range of 10k to 100k per chromosome.

*2.2.3. Improve the import and export formats supported by VCMap:* The objective of this approach is to increase the import and export features for the VCMap tool to support more formats and allow for greater interoperability of this tool with other common genomic research tools. In order to promote increased support from other tools, we will define a standard method for existing tools to automatically launch the VCMap application using specific parameters which will automatically load a subset of data of interest. Since the tool is launched via Java Webstart, we can incorporate this feature by defining a URL and query parameter set standard that we publish to the VCMap website. This will allow existing genome browsers to link directly to the VCMap application, thus enabling users to navigate data in a comparative research context after already using a third party tool to identify a region or feature of interest. In order to improve data export, we plan to incorporate the ability to output a selected set of the annotation features in the General Feature Format v3.0 (.GFF3) format. We currently import data in that format, but only provide data export as comma separated values (.CSV) files. Expanding the data export feature to include .GFF3 support will also open the possibility of incorporating VCMap into a GMOD based workflow. Several of the GMOD applications support this format, (GFF – GMOD. Available at: http://gmod.org/wiki/GFF#GFF\_in\_GMOD Accessed: July 2011) and exporting data in this format will even allow VCMap data to be loaded into any CHADO based database, allowing users to access this data from any GMOD component. In order to improve data import, we will expand the supported data formats to include variant call format (.VCF), the data format created for, and utilized by, the 1000 genomes project. This will enable the visualization of 1000 genomes, and dbSNP data in the VCMap tool (See section C.4.2.2 for additional details). Additionally, supporting the .VCF format should promote the usage of this tool when viewing “personalized” genome data. Combined with the additional ontology terms, improved QTL data, and new GWAS data, this will make for a powerful tool for examining variants *in a comparative genomics environment*. We also plan to add support for the wiggle format (.WIG), another common format for UCSC data tracks. In order to best support this format, we will also expand the modes of rendering data in the system. Currently, all annotation features are rendered as a single contiguous black indicator within vertical data tracks displayed next to the chromosome representations (see for example the QTL tracks in Figure 2). Because the .WIG format defines a value for each position of a feature, it will be important to render this value as opposed to a single binary indicator (presence or absence) as is done currently. This visualization will use a combination of color and size to indicate the specified values provided by.WIG files. When incorporating support for this additional mode of rendering annotation data, we will design a modular framework that will allow the tool to support multiple additional modes of visualization. This will enable future enhancement to the tool by creating a stable framework for adding new annotation rendering modes**.**

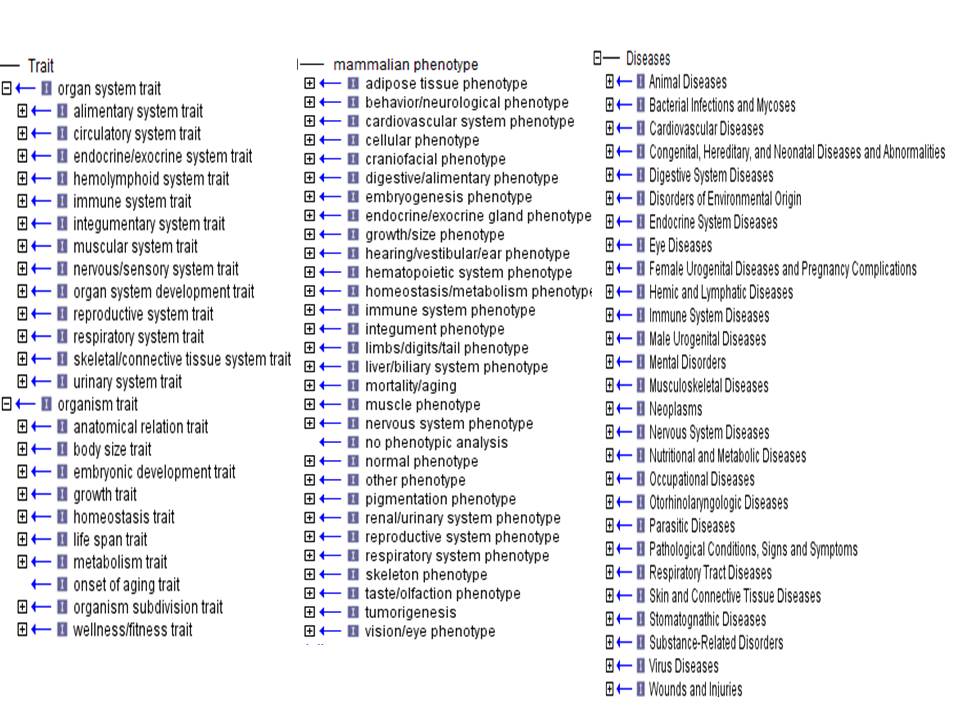
***C.2.4 Anticipated Problems and Solutions*:** The current interface is limited by performance issues when loading over 3,000 features per chromosome into a single data track. While this is sufficient for our current preloaded data (Genes, Pseudogenes, STS markers, and QTL), loading higher density datasets such as dbSNP, or next generation sequencing results will result in performance degradation during loading and zooming operations. The current implementation of VCMap uses an algorithm to group features by position and size when the resolution of the viewport is too small to display high density data, e.g. when the tool is zoomed out to view a full chromosome. This grouping allows for usable navigation of data when viewing large regions, but the performance is at best O(m\*n2) where m is the number of maps loaded, and n is the number of features to group. As n grows, this operation becomes to costly to complete during a zoom operation because it would cause a delay that detracts from the usability of the tool, on the order of 10 to 60 seconds on a typical modern computer. In order to handle larger datasets, we will 1) improve this algorithm with the goal of O(m\*n\*log(n)) performance, and 2) examine reorganization of the implementation to allow this grouping to occur during the load operation, increasing load time, but allowing for a fluid user experience after loading of an entire dataset.

**C.3. Specific Aim 2: Employ and link appropriate ontologies to annotate mapped traits, phenotypes and diseases from multiple organisms to facilitate cross species comparisons and the creation of comprehensive virtual morbid maps.**

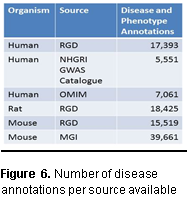
***C.3.1 Rationale:*** Identifying potential trait and disease related regions through comparative genomics approaches is a mainstay of much research today. Standardization of these mapped loci will improve the ability of researchers to compare data across organisms, identify potential sites for manipulation within their own organism of interest and identify potential existing model organisms to validate findings from GWAS and other studies. For example, the availability of data for mouse and rat mutants, transgenics, knockouts and congenic strains on a single morbid map will allow users to query loci of interest from GWAS or other studies and determine whether these fall within the regions defined by the various mouse and rat strains. Links to the originating databases will provide users with information on the source and availability of such strains. Thus, in one easy step, a user will be able to determine whether of gene or SNP of interest lies within the region of rat or mouse for which there is a congenic, transgenic or knockout animal available.

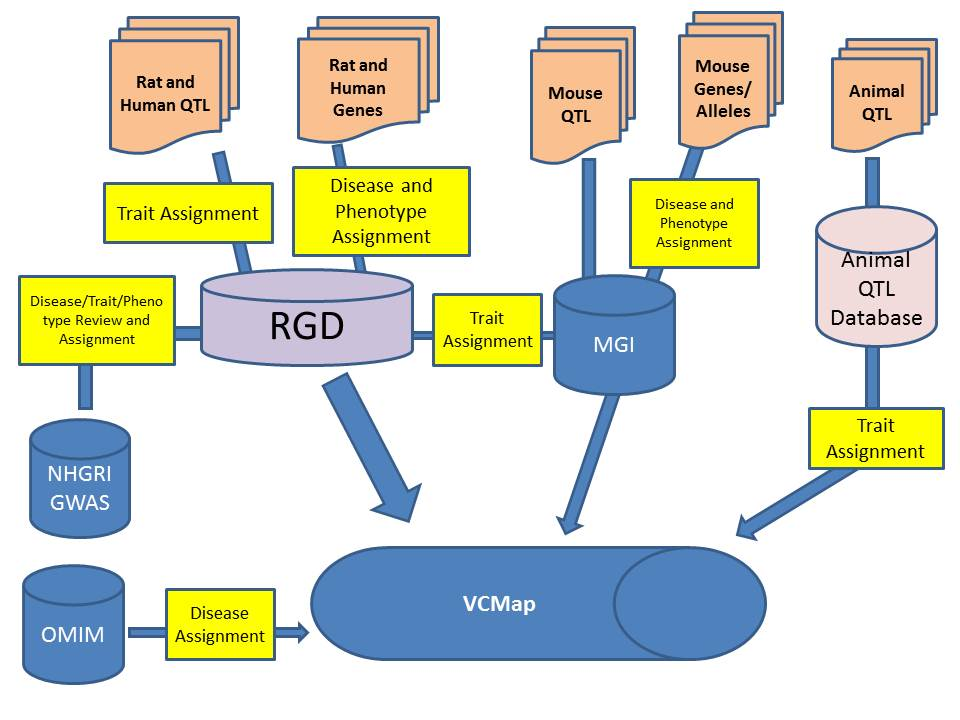
***C.3.2 Experimental Approach:***

*3.2.1 Develop and Implement Ontologies for Cross Species Trait, Phenotype and Disease Comparisons:* Three ontologies will be employed to annotate data with the Vertebrate Trait Ontology (http://rgd.mcw.edu) being used to standardize mapped traits, particularly QTL, the Mammalian Phenotype Ontology for mapped phenotypes, most commonly used for genes and alleles, and a Comparative Disease Ontology ([http://ctd.mdibl.org](http://ctd.mdibl.org/)) for mapped diseases, most commonly used for individual mutations, gene-disease annotations and often in GWAS studies for statistically significant variations (Figure 5). The Vertebrate Trait Ontology has been developed to standardize QTL and other genomic element annotation across multiple organisms with the Animal QTL Database ([www.animalgenome.org/QTLdb/](http://www.animalgenome.org/QTLdb/) ), the Rat Genome Database (RGD) (rgd.mcw.edu) and Mouse Genome Informatics (MGI) ([http://www.informatics.jax.org](http://www.informatics.jax.org/)) participating in its development. Continued development of this initial version to accommodate emerging QTL and other genomic elements will be part of the proposed project. The Mammalian Phenotype Ontology is currently used by RGD and MGI for annotation of genes and alleles for rat, mouse and human. The Comparative Disease Ontology incorporates both the MeSH and OMIM terminologies for disease and is currently used by RGD, MGI and the Comparative Toxicogenomics Database (<http://ctd.mdibl.org/>) (Davis et al). These three ontologies were chosen because of their similarity in structural organization which will facilitate data queries and presentation and their use with the data from organisms represented in VCMap as well as their availability in the widely used Open Biomedical Ontologies (OBO) format (<http://www.obofoundry.org/> ) (Smith et al) *which* will facilitate the development of unified query and display options.

****

**Figure 5** The Vertebrate Trait Ontology, Mammalian Phenotype Ontology and Comparative Disease Ontology will be used to annotate mapped loci of interest to the VC Map project.

*3.2.2 Integrate Trait, Phenotype and Disease Datasets from Multiple Organisms and Multiple Species:* QTL represent a rich source of mapped trait data. For this application, human, rat, and mouse QTL as well as QTL representing traits of interest to human health from organisms such as cow, pig, chicken and sheep will be annotated with the Vertebrate Trait Ontology (VTO) to standardize representation of traits across organisms. Currently, the respective repositories for these QTL utilize differing methods of indicating trait, either through QTL name (MGI) or organism specific trait lists or vocabularies (RGD and AnimalQTLdb). Mapping of the current trait terms to the appropriate VTO terms will facilitate a semi-automated method of annotation for groups of related QTL. The major sources for phenotype annotation datasets for human, mouse and rat are RGD and MGI while those for disease are RGD, MGI, OMIM and the NHGRI‘s Catalog of Genome-Wide Association Studies (www.genome.gov) (Figure 6). Data sets already annotated with the appropriate ontology annotations can be readily integrated into VCMap. These will include all QTL from the AnimalQTLdb and RGD, human QTL curated as part of the initial development of VCMap and RGD curation projects, phenotype annotations for rat genes and disease annotations for rat, human and mouse genes curated at RGD. Mouse genes and alleles also are annotated with the appropriate MP and OMIM terms at MGI for easy integration. For datasets without ontology annotations or for those with annotations from other vocabularies, conversion to the three ontologies outlined here will be accomplished through the ontology curation software in place at RGD (for software description see Laulederkind et al). This software allows uploading of genomic elements and provides curators with the ability to annotate groups of elements with the same ontology associations simultaneously, thus, minimizing the manual resources needed. Figure 7 shows the basic data and ontology assignment flow. Data from the following sources will be directly loaded into VCMap because appropriate ontology assignment is made at the originating source: QTL from the Animal QTL Database, human and rat genes and QTL from RGD, mouse genes and alleles from MGI and mapped loci from OMIM. Mouse QTL and data from the GWAS studies will be reviewed and ontology annotations made at RGD using existing software.



**Figure 7.** Data flow and ontology assignment.

*3.2.3 Create Virtual Morbid Maps for Species Integrated into VCMAP:* Because trait, phenotype and disease data is scattered across multiple resources and organisms, it is difficult for researchers to ascertain the location of these mapped loci on the organism of their choice without conducting multiple queries and alignments, often requiring multiple steps and tools. Using the technology of VCMap, we will build a Virtual Morbid Map for each of the organisms represented by mapping all trait, phenotype and disease loci to their respective positions in the genomic map of the organism of interest. This will allow the user, for example, to compare the Rat Virtual Morbid Map with the rat genomic map for a particular chromosome or region to facilitate the determination of the boundaries for identifying or creating a congenic or knockout model that might correspond to the disease locus identified in a human GWAS or other study. Users would also be able to easily see when disease or phenotype loci from multiple organisms overlap in a region of interest. Because the loci will be annotated with ontologies, users will be able to filter loci on the virtual morbid maps for only cardiovascular related loci, for example, or other criteria. For example we would map all human and mouse QTLs, phenotype and disease annotated genes and alleles to the rat genome via orthology and have the rat specific QTL, congenic strains and disease and phenotype genes as well. These could be color coded to indicate those found in the species of focus and those mapped via orthology. Users will be able to mouse over the elements in the morbid map and see the ontology term and the organism. The morbid maps would only contain the elements with trait, phenotype or disease annotations.

***C.3.3 Expected Results:*** Upon completion of this aim, targeted QTL datasets from multiple organisms will be annotated to the Vertebrate Trait Ontology and disease loci datasets annotated to the Comparative Disease Ontology. Existing datasets from MGD and RGD for phenotype loci will also be integrated into VCMap. Users will be able to easily view trait, phenotype and disease loci from multiple organisms mapped to the organism of their choice and will be able to customize the view of the morbid map through ontology based query filters.

***C.3.4 Anticipated Problems and Solutions:*** Appropriate standardized annotation of QTL and other trait, phenotype and disease associated loci is dependent on the accuracy and availability of annotations and metadata at the originating resource. When absent, access to published literature associated with the loci will provide the necessary information to make the ontology annotations, but may slow the process because of additional manual steps. Because multiple, different, but related, ontologies will be used, it will be necessary to create relationships among traits, related phenotypes and diseases. Because disruption of a single trait may result in multiple phenotypes and many diseases involve multiple phenotypes, these relationships are complex but mappings will be necessary to facilitate querying and filtering for users. Although we will build on the initial efforts to map phenotypes to diseases provided in PhenoMizer and PhenExplorer (www.human-phenotype-ontology.org) in which phenotypes have been mapped to OMIM diseases, we recognize the potential for increased manual efforts which might delay progress. Targeting those areas related to the loci of interest will limit the resources necessary but will also limit the comprehensiveness of the domains of the ontologies mapped. Construction of the virtual morbid maps is reliant on adequate orthology and synteny mappings among organisms. Organisms with inadequate genome assemblies and annotations may limit the mapped loci that can be included.

**C.4 Specific Aim 3:** **Expand the intra-species and inter-species connectivity.**

***C.4.1 Rationale:*** Comparative genomics approaches can provide critical information about gene function that is invaluable in the dissection of genetic mechanism that underlie human disease (8). For example, the ability to detect DNA regulatory regions responsible for gene regulation required multi-species comparisons as demonstrated by the ENCODE project (1, 2). Recent advances in whole genome association studies (Hawkins et al., 2010; Price et al., 2010; Zeggini, 2011) and development of genetic resources, like the Mouse Collaborative Cross (Threadgill et al., 2011), have made it relatively easy to identity specific haplotypes that are associated with traits and diseases of interest (e.g. diabetes, obesity, blood pressure). However, there is a real need to visualize this data to facilitate discovery of causal genetic variants. Furthermore, advances in sequencing technologies (Glenn, 2011) have facilitated the generation of RNA-seq, Chip-seq, and exome and genome re-sequencing, which all provide invaluable information that will help lead to the eventual identification of causal variants. For example, the pilot project of the 1000 genomes project reported the analysis of 179 individuals (The 1000 Genomes Project Consortium, 2010), which identified large numbers of SNP, insertions/deletions, copy number variants, and haplotypes. Unfortunately, there is a lack of visualization tools other than at a gene centric level (<http://browser.1000genomes.org/index.html>). We believe that the utility of VCMap could be expanded and improved to facilitate chromosomal level intra-species comparisons (e.g. comparison of different haplotypes within a species) to fill this obvious visualization gap. In addition to genome sequences, cross-species comparisons are also available in the form of bacterial artificial chromosome, radiation hybrid (RH) and linkage map data (e.g. (9)). Unfortunately, the ability to link this data to genomic maps is not always straightforward. When inter- and intra-species comparison data is not available, e.g. due to new genome builds, multiple genome builds or lack of conserved synteny comparisons, the transfer of genomic and annotation information across species can be difficult or cumbersome. We envision VCMap as a bridge when direct comparative information is not available that will provide the most comprehensive intra- and inter-species links between genome variation and phenotypic outcomes available to date.

***C.4.2 Experimental Approach and Analysis***

*4.2.1: Develop a pipeline to define synteny and homology for unavailable inter- and intra-species comparisons:* We will build a suite of tools that will allow VCMap users to generate links across species, strains/individuals within a species, or with external data to allow comparison of genomic features relative to a reference genome. *Interspecies comparisons:* If a genome has been processed by Ensembl, NCBI, and/or UCSC, it can be loaded into the standard VCMap database. However, in cases where genome assemblies have not been processed by Ensembl, NCBI, and/or UCSC, a pipeline will be modeled off of Ensembl’s Compara conservation pipeline (See letter of support from Paul Flicek) to facilitate genome comparison prior to Ensembl, NCBI, and/or UCSC annotation. Boundary regions for syntenic blocks, as well as annotations from cross-species comparisons will be obtained for use in the VCMap database. Briefly, we will use BLASTZnet to develop tracks to compare genomes (6, 13). Gene homology will be determined using BLASTALL. Annotation information will be build upon existing homology found at NCBI’s Homologene, UCSC and Ensembl by comparing potential homologues to known genes in information-rich species. *Intraspecies comparisons:* Whole genome sequence anchors across strains or individuals will be developed by matching the positions of sequences of SNP/STS markers. Localized comparisons within chromosome blocks will be precomputed using BLASTALL and then stored in the database for later visualization. These comparisons will be designed to focus on targeted sub-chromosomal regions in an effort to define sequence-level differences. Tools will be developed to use haplotype information from programs such as BEAGLE (3) or PLINK (12) to define haplotype blocks across strains. The application will also support publically available HapMap information as well as the haplotypes made available from the parental strains of the collaborative cross (see letter of support from Dr. Pardo-Manuel de Villena) and other mouse strains (e.g. Yang et al., 2011 Nature Genetics 43: 648-655).

*4.2.2: Develop tools to visualize generic inter- and intra-species comparisons using the VCMap GUI:* The objective of this approach is to facilitate visualization of 1) cross-species comparisons of new model species, 2) genomic architecture of multiple strains or individuals within species, and 3) external data in the VCMap GUI. To achieve this goal, we will utilize sequence-aligned data from the inter- and intra-species comparison pipeline (above) as a basis for comparing new model species of interest against existing species in VCMap. To promote the use of additional species in VCMap, we will develop the ability to view generic species in the GUI. This will allow users to privately visualize their own comparative data from different model species than are currently available on VCMap. We will also develop tools to better utilize other map data (e.g. RH, BAC, linkage) to link species to physical maps already within VCMap where possible. To allow intraspecies visualization, we will expand VCMap visualization tools to allow whole chromosome and targeted genome coordinate comparisons across strains. Strains will be clustered by haplotypes to define strains that are more or less similar (See Figure 8 for a proposed view). In addition, we will visualize strains in phylogenies based on haplotype. The goal will be to develop a tool that allows users to easily visualize differences in haplotype structure and the ability to zoom to a sequence level view for haplotypes of interest to view sequence-level differences. A critical objective of this approach will be to facilitate genomic comparisons across strains or individuals within the same species (e.g. the collaborative cross).

**Figure 8.** Individual genomes could be loaded as individual .vcf files and then visualized as shown above in that each individual is represented as a column, with different haplotype represented by different colors. When visualized in concert with SNP/RNA-seq data, it would allow users to quickly identify genome regions of interest for further investigation.

***C.4.3 Expected Results:*** We will greatly enhance the detail with which the VCMap software can compare intra- and inter-species genomes. We expect that it will be possible to use more detailed synteny comparisons at UCSC and also those at ENSEMBL to improve upon the cross-species annotation of genomic regions. We will also enhance the annotation information available for cross-species comparisons. In addition, the interspecies comparative pipeline will allow us to develop generic tools that allow cross-species comparisons to genomes that currently lack the required information. This will allow the tool to be used to improve the detail of the annotations and to better view genomic differences within species. It will also provide suggested corrections to specific genome builds and cross-species conserved synteny comparisons. In addition, it will allow users to update genome comparisons quickly when new genome builds are developed, prior to the time that UCSC, ENSEMBL and NCBI develop there cross-species comparisons. The users will benefit from new genome builds without sacrificing the ability to view comparative genomics information. The pipeline we propose will facilitate the ability to compare new model species to existing genomes in VCMap. The GUI tools developed to facilitate generic inter and intra species comparisons will allow new model genomes or multiple strains to be easily compared to the existing genomes in VCMap. Updates proposed to the VCMap GUI will allow visualization of differences in multiple strains/individuals at the haplotype and sequence level. This will be a considerable asset to projects such as the mouse collaborative cross or human 1000 genomes project. The development of these visualization tools will also be extremely useful for groups with experimental populations used for GWAS studies to view key recombinations and structural differences that define association regions. The final approach presented will allow improved use and application of the tool by other groups who desire to link their tools programmatically to VCMap to mine synteny and comparative annotation information. The overall result of these enhancements to VCMap will result in more detailed annotation and syntenic information within and across species. The improvements are expected to assist researchers mining genomic regions associated with complex diseases.

***C.4.4 Anticipated Problems and Solutions:*** Development of our own synteny and cross-species alignment pipeline may provide ambiguous alignments in some cases. Ideally after a species is sequenced, it will be analyzed by UCSC, NCBI, or Ensembl and will then migrate into the VCMap database. However, in the absence of UCSC, NCBI or Ensembl analysis, this tool would fill a visible gap that will facilitate research in the absence of pre-existing conservation links. Additionally, we will work with NCBI/Ensembl staff during this process to minimize duplication of effort (Please see letter of collaboration). Our team has considerable experience with software development and we expect no problems developing the proposed pipeline and visualization tools. Furthermore, we will establish a working relationship with the Mouse Collaborative Cross Consortium (see letter of collaboration), which will help tremendously with the intra-species comparison aspects of VCMap, as they have a very large intra-species comparison bottleneck.

Rubin, D.L., et al., *National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge.* OMICS, 2006. 10(2): p. 185-98.

Blake, J.A. and M.A. Harris, *The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis.* Curr Protoc Bioinformatics, 2008. **Chapter 7**: p. Unit 7 2.

# Smith, C.L., C.A. Goldsmith, and J.T. Eppig, *The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information.* Genome Biol, 2005. 6(1): p. R7.

# Threadgill DW, Miller DR, Churchill GA, de Villena FP. The collaborative cross: a recombinant inbred mouse population for the systems genetic era. ILAR J. 2011 52(1):24-31.

Glenn, T.C. Field guide to next-generation DNA sequencers. Molecular Ecology Resources May 19. doi: 10.1111/j.1755-0998.2011.03024.x.

MIM Morbid Map. National Center for Biotechnology Information. Available at: <http://www.ncbi.nlm.nih.gov/Omim/getmorbid.cgi> Accessed July 2011.

Hindorff LA, Junkins HA, Hall PN, Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: [www.genome.gov/gwastudies](http://www.genome.gov/gwastudies). Accessed July 2011.

dbSNP Home Page. National Center for Biotechnology Information. Available at: <http://www.ncbi.nlm.nih.gov/projects/SNP/> Accessed July 2011.

**Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, and Kent WJ**. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* 38: D613-619, 2010.

Dwinell MR, Worthey EA, Shimoyama M, Bakir-Gungor B, DePons J, Laulederkind S, Lowry T, Nigam R, Petri V, Smith J, Stoddard A, Twigger SN, Jacob HJ; RGD Team, The Rat Genome Database 2009: variation, ontologies and pathways, Nucleic Acids Res. 2009 Jan;37 (Database issue):D744-9

Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, Rosenstein M, Wiegers T, Mattingly CJ, The Comparative Toxicogenomics Database: update 2011. Nucleic Acids Res 2011 Jan;39 (Database issue):D1067-72

Smith B, Ashburner,M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ: OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S, The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nat Biotechnol. 2007 Nov;25(11):1251-5

Worthey EA, Stoddard AJ, Jacob HJ, Sequencing of the rat genome and databases, Methods Mol Biol, 2010;597:33-53