BMC Bioinformatics

Database

MuTrack: a genome analysis system for large-scale mutagenesis in the mouse

Erich J Baker*[1], Leslie Galloway[2,3], Barbara Jackson[4], Denise Schmoyer[4] and Jay Snoddy[2,5]

Address: [1]Department of Computer Science, Baylor University, Waco, USA, [2]Life Sciences Division, Oak Ridge National Laboratories, Oak Ridge, USA, [3]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, USA, [4]Computer Science and Mathematics Division, Oak Ridge National Laboratories, Oak Ridge, USA and [5]Graduate School in Genome Science and Technology, University of Tennessee and Oak Ridge National Laboratories, Knoxville, USA

Email: Erich J Baker* - Erich_Baker@Baylor.edu; Leslie Galloway - gallowayld@ornl.gov; Barbara Jackson - jacksonbl@ornl.gov; Denise Schmoyer - schmoyerdd@ornl.gov; Jay Snoddy - Snoddyj@ornl.gov

* Corresponding author

Published: 03 February 2004

BMC Bioinformatics 2004, **5**:11

Received: 16 October 2003
Accepted: 03 February 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/11

Abstract

Background: Modern biological research makes possible the comprehensive study and development of heritable mutations in the mouse model at high-throughput. Using techniques spanning genetics, molecular biology, histology, and behavioral science, researchers may examine, with varying degrees of granularity, numerous phenotypic aspects of mutant mouse strains directly pertinent to human disease states. Success of these and other genome-wide endeavors relies on a well-structured bioinformatics core that brings together investigators from widely dispersed institutions and enables them to seamlessly integrate data, observations and discussions.

Description: MuTrack was developed as the bioinformatics core for a large mouse phenotype screening effort. It is a comprehensive collection of on-line computational tools and tracks thousands of mutagenized mice from birth through senescence and death. It identifies the physical location of mice during an intensive phenotype screening process at several locations throughout the state of Tennessee and collects raw and processed experimental data from each domain. MuTrack's statistical package allows researchers to access a real-time analysis of mouse pedigrees for aberrant behavior, and subsequent recirculation and retesting. The end result is the classification of potential and actual heritable mutant mouse strains that become immediately available to outside researchers who have expressed interest in the mutant phenotype.

Conclusion: MuTrack demonstrates the effectiveness of using bioinformatics techniques in data collection, integration and analysis to identify unique result sets that are beyond the capacity of a solitary laboratory. By employing the research expertise of investigators at several institutions for a broad-ranging study, the TMGC has amplified the effectiveness of any one consortium member. The bioinformatics strategy presented here lends future collaborative efforts a template for a comprehensive approach to large-scale analysis.

Background

The rapid diversification of experimental techniques, expertise and public domain data has necessitated a shift away from the traditional institutionally-centric research paradigm. Indeed, an inclination towards comprehensive approaches to biological research on a genome-wide scale dictates that any one single institution may not contain the critical mass of physical and intellectual resources necessary to address certain broad biological questions. We describe herein an approach to this challenge that focuses on the creation of inter-institutional research teams that leverage existing internet technologies to bring together wide-ranging expertise in an efficient and effective analysis system.

While the metaphor of research teams often exists at the institutional or local level they do not exist across several institutions for mostly logistical reasons. Effective distributed collaborations require the implementation of an infrastructure that handles a fundamental array of information processes unique to non-local research communities. Researchers must have mechanisms for exhaustive electronic data storage, curation, and sharing. They must be permitted to make observations about the data and the experimental process, and they must have access to computational tools that assist in the extraction of new knowledge from the common warehouse of shared data. Concurrently, researchers in a distributed collaboration must find the bioinformatics core flexible enough to handle the immense diversity of information produced by modern experimental techniques, and structured enough to enforce machine-readable data types for future analysis. Finally, distributed data systems must meet ease-of-use requirements while simultaneously applying explicit control over who has access to data sets and observations.

The criteria for effective distributed collaborations have been tested in theoretical scenarios [1,2] and as limited implementations of expanded and distributed laboratory information management systems (LIMS) [3,4], but the literature lacks examples of comprehensive bioinformatics systems that support data collection, curation, and analysis. Here, we utilize an opportunity presented by a federally funded attempt to perform a genome-wide survey of heritable mutant phenotypes in the mouse. The test case for our distributed computational system is the Tennessee Mouse Genome Consortium (TMGC) [5]. In contrast to other funded phenotyping efforts, the TMGC is unique among organizations in its attempt to use the geographically distributed resources of consortium members to perform domain-specific phenotypic analysis of mutangenized mouse pedigrees (Figure 1).

The utility of employing the mouse as a model for human disease is well documented [6-9]. Traditional methods of site-directed *in vivo* mutagenesis are tedious and require prior knowledge of gene function and location [10]. Alternative approaches, developed to induce primarily single base pair changes in a genome region of interest [11], are also effective at producing recessive and dominant heritable mutations in the mouse [12,13] but lack the specificity of traditional approaches. As a result, any single mutation event may be silent or effective and may lie within a gene directing a visible phenotypic characteristic, a gene without phenotypic consequence, or in a non-coding region [11]. In order to produce substantive phenotypic anomalies in large-scale germ-cell strategies, such as *N*-ethyl-*N*-nitrosourea (ENU) directed mutagenesis, the *production and phenotypic classification* of vast numbers of mouse pedigrees from birth through senescence and death is required.

The system implemented to satisfy this bioinformatics task is named *MuTrack*, and has evolved into the central mechanism that supports the functions of the broad based TMGC consortium. It resides as a collection of database-backed, on-line analysis tools capable of tracking mouse breeding schemes, the shipment of mutant mice throughout the consortium and the exchange of physical samples, ranging from sperm to histological sections. In total, it collects raw and processed data and observations from the twenty-two discrete phenotype testing domains and provides a real-time statistical analysis of possible phenodeviant mouse lineages based on the collected experimental data. It simultaneously allows member researchers to select mice for secondary and tertiary study to test mutant heritability and provides a means to distribute new mutant strains to researchers outside the collaboration. To date, it has aided in the successful identification of 75 new mutant mouse strains, and has screened more than 22,500 individual mice.

Successful development of heritable mouse mutations will contribute to our understanding of human disease states through the development of new mouse models. Of equal consequence, the implementation of a workable and collaborative data sharing architecture represents a significant advancement in the way researchers bring to bear comprehensive high-throughput analysis in biology's information rich environment.

Construction and Content

Data import and export

MuTrack accepts two types of import formats: web-based forms or direct upload of text in a comma separated value (CSV) format. Husbandry information is generally accepted through web-based forms containing pre-calculated attributes where possible. Domain investigators may submit internet forms or use preformatted spreadsheets that parallel the Microsoft Excel paradigm. Immediately

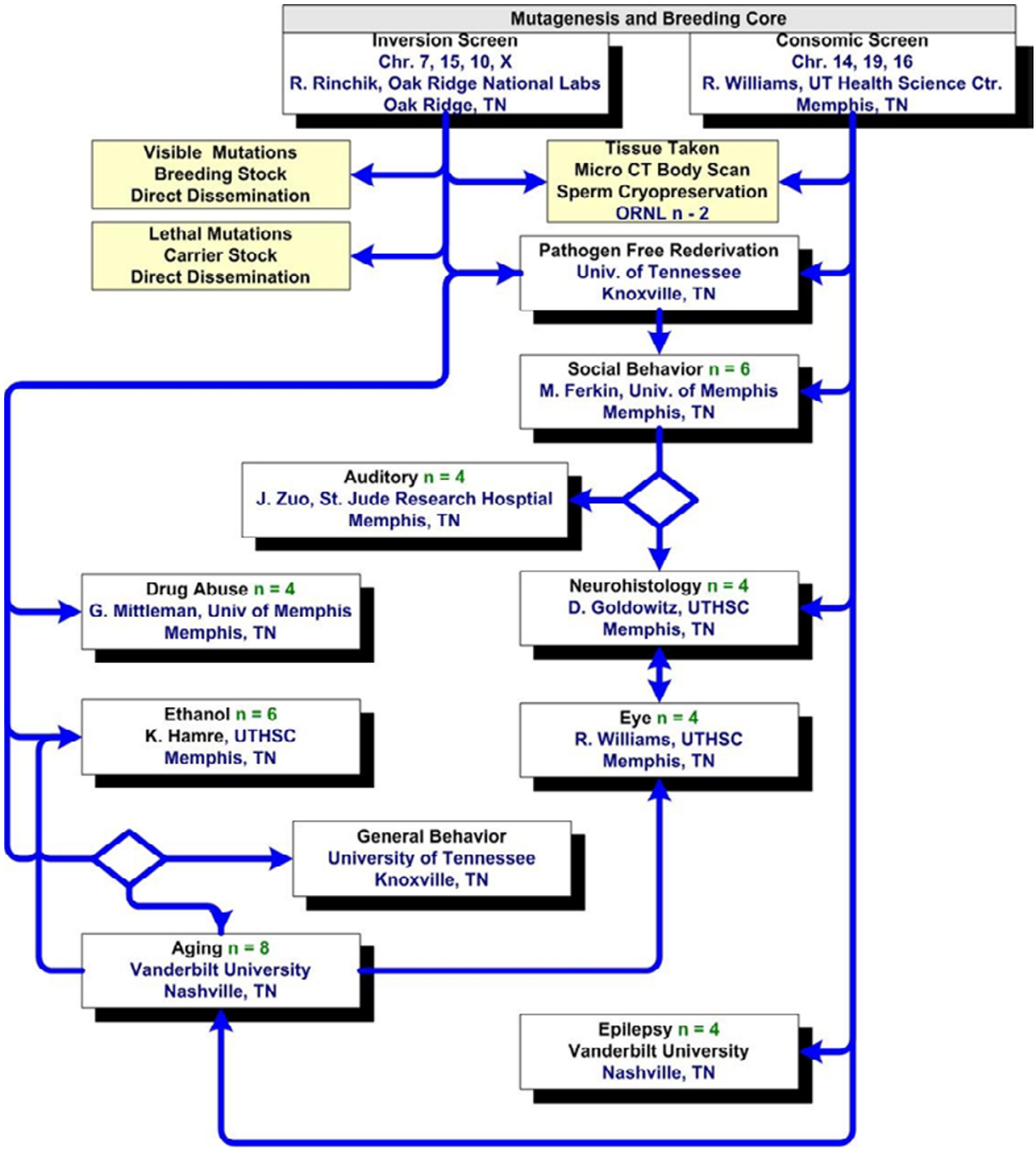

Figure 1

Primary collaboration relationships represented in *MuTrack*. This is a partial depiction of the collaborative effort to study genome-scale mouse mutagenesis and reveals the complexity of distributed collaborations. Mice are mutagenized at two separate institutions within the state of Tennessee and bred according to a scheme that produces visible, non-visible, or lethal mutations. Mice that do not present gross phenotype anomalies are processed throughout the state by experts in mouse genetics, behavior, physiology and aging, among other fields. *MuTrack* is responsible for tracking the complex breeding schemes, the physical location, health and test status of all mice in the testing pipeline, experimental data collection from each testing domain, real-time analysis of results, and free-form observations about pedigrees of interest. Not shown, but also within the scope of *MuTrack*, is the secondary screening template of possible mutants involving other locations (Case Western University, Meharry Medical College) and the system to distribute heritable mouse mutants to the mouse research community.

upon submission, these files are pipelined through an error checking process and uploaded into their respective database relations. The error checking process includes examination for proper formatting, data type constraints, and maintenance of the testing pipeline structure, ensuring the testing of mice in proper chronological order. Domain investigators may likewise search any information associated with their testing domain and export search results in CSV or tab-delimited formats. Image information collected via on-line means from the neural histology and eye cores may be exported in **png** format along with the dynamically generated statistical graphs associated with any mouse pedigree or testing domain.

Statistics

Consortium member statistics

MuTrack seamlessly integrates with the strong analysis tools in the SAS statistical system, allowing incorporation of more complex and highly appropriate data analysis into the simple user interface. Robust estimates of the population mean and standard deviation are calculated from pedigree means using SAS (Version 8.2) [14]to eliminate contamination biases inherent to the detection of unknown mutants from a set of observations. The robust mean is obtained from the Univariate procedure with the Trim option set at 0.25. The Trim option is selected to reduce the influence of phenodeviant pedigrees on the mean and to reduce the movement of means estimates of central tendency as new mice are tested for each domain. By trimming the extremes (i.e. defined or suspected phenodeviants) the central data remaining should be a close unbiased and robust representation of the "normal" mice, thus giving a more stable and accurate population to predict against. The robust standard deviation estimator, Mean Absolute Deviation (MAD) sigma, is obtained using the SAS Univariate procedure robust option. The estimator is insensitive to the inflated variance that results when outlier pedigrees are present.

Each pedigree is averaged and measured against the trimmed population mean for distance. Outliers are flagged (highlighted) at plus or minus 1.645 SD (10% in each tail) and again at 1.96 SD (5% in each tail) from the mean to alert investigators about the possibility of an outlier. Each investigator is expected to take these results and compare it against their own notes about the pedigree. Retests are called based on these results.

Using the methods from above, an investigator may select a testing area of interest, based on experimental domains, and any pedigree for any field exceeding a distance from the mean of 1.645 or greater is included in a two-way table with the appropriate cell highlighted. All scripts are batched on a weekly basis to provide a data management overview, but are also generated dynamically as users engage in database queries.

Histogram Plots

The data for a particular test in a domain are plotted in a histogram with a normal curve overlaid using the Capability procedure in SAS. Drop-down lines indicating 1 and 2 SD from mean are also included. This tool, along with normality assessment statistics generated using SAS Univariate procedure, alerts the investigator to non-normal data distributions and the presence of outliers. Specific data plots, such as those resulting from non-parametric analysis, may be done at a consortium member's request.

Estimating Group Effects

A tool has been added using the T-Test procedure to assess whether blindness in the 33TNK strain has any effect on testing. A list of mice known to be sighted or blind is used to create a dataset with which this comparison is made. Any domain that used any of these mice is eligible for this test.

Cross-domain Analysis

The aging data are evaluated using the Boxplot procedure. The Test Tables for Aging uses the plots to determine growth of a pedigree across time. The investigator uses this to determine weight gain or loss relative to the "family" to check for outliers. Another tool looks at each pedigree within a family together at a particular age to see outliers, as it is believed that slower growing mice live longer.

Program development and data integrity

The bulk of *MuTrack* is written in an object-oriented style in PHP v 4.0 [15], an open-source server-side embedded scripting language explicitly designed to integrate database technology and dynamic HTML presentation. Dynamically generated graphical representations and interfaces utilize the gd.pm module [16] of PERL v 5.6 [17], and Javascript is used to dynamically validate form field information [18]. An Apache server, version 1.3.26, running on a Solaris 8 Sun box, serves the entire system and allows the use of the established apache secure socket layer.

An exception to the open-source paradigm is the choice of database framework. The Oracle 8*i* DBMS comes with extensive redundancies that allow for seamless data recovery of edits or interrupted transactional processing resulting from hardware or software failure or operator error [19]. Data clashes are prevented at the interface level as well as at the database level, ensuring that only one record exists for each data iteration. Log tables transparently save edited data, allowing the recovery of results edited in earlier sessions. *MuTrack* also implements intrinsic

concurrency functions that search the database for duplicate or non-standard records.

Tremendous local expertise and experience in Oracle and SAS technologies was a contributing factor in the decision to avoid open-source alternatives such as PostgreSQL and R, respectively. We believe that future implementations of a comparable system in a complete open-source environment is feasible.

Utility and Discussion

MuTrack [20] was planned from its inception to address requirements dictated by the experimental design. The genome-wide mutagenesis strategy mandated a system that could integrate the logic of mouse breeding strategies and husbandry logistics throughout the testing processes, accommodate diverse phenotype screening strategies, determine statistically significant patterns in the collected data, and provide a logical and transparent means to classify mutant mice. The end product is a centralized database-backed system that separates each process into a separate operational domain (Figures 2 and 3).

User interfaces

Because *MuTrack* is available as a web-based platform, numerous considerations about internet navigation, security and accessibility were addressed. The site maintains a consistent look and feel designed around dynamically generated web-pages. A generalized view of each data representation is located within one click of the main page, and each relation is one click away from any other relation. When a user becomes familiar with one area of *MuTrack* they will, by similarity, be familiar with all areas. Computational tools that deal with statistical analysis and pages designed as areas for free-form textual observation are complex and require specific homepages one level down from the main *MuTrack* page. In these cases, web navigation is menu driven, allowing users to make very specific observations or drill down to a specific statistical test performed on particular mice or mouse pedigrees. Most areas within *MuTrack* are available to the public using a guest password, while specialized sites are limited to TMGC researchers in general and specific domain investigators in particular.

Mouse breeding schema and sample tracking

The TMGC mutagenesis project uses two distinct and well-identified breeding strategies that have been summarized in the recent literature [13,21]. While both strategies differ in their molecular focus, they maintain the need to sustain large stocks of breeding mice for several generations. *MuTrack* begins the process of sample tracking by forcing technicians to input unique mouse information into a ***Mouse ID*** relation for mice of generation zero. Once mice *exist* within the system they are put on a mating schedule based on age and lineage. *MuTrack* tracks the removal of fertilized embryos from test-generation mice and manages their shipment and implantation into immunologically *clean* surrogate mothers located at a different institution. New mice are tracked through their ***Litter*** and are entered into the ***Mouse ID*** relation after ***Weaning***.

During the breeding process sample tissues are often collected and stored for later analysis; the database must likewise account for destroyed mice. Hence, the ***Mouse Disposal*** and ***Tissue Sample*** tools reside within this domain and may be accessed by any privileged user anywhere within *MuTrack*. These represent integral processes in the highly structured chain-of-custody standards enforced at the interface and database levels. Indeed, the primary computational concern of the analysis pipeline is the location, status, and ownership of each mouse or tissue sample generated by the consortium. Adequate appraisal of this information provides project supervisors the ability to maintain a constant flow of animals through the testing domains, and reduces the amount of experimental data lost to logistical oversights.

Phenotype screen strategy

While *MuTrack* enables researchers to identify and catalogue lethal and easily visible mutations throughout the breeding strategy, subtle phenodeviants are initially identified during a regimented screening process (Figure 1), and are currently stored in association with specific mutangenized pedigrees. The distributed screening process, itself, addresses two equally important concerns. First, the interface to collect, store, and curate data must be uncomplicated and durable. This is accomplished by granting primary investigators of each phenotype testing domain the ability to submit raw and processed data to the database via internet technologies that the domain investigators are most comfortable with. Some researches submit data through online HTML forms, others use Microsoft Excel spreadsheets, and others use direct machine generated data sets. Table 1 illustrates the scope of data points collected by each researcher and submitted to *MuTrack*. There are over one million tuples of information stored in *MuTrack* relational tables, representing more than 17 million discrete observations. Domains that utilize a significant amount of image data, the neurohistology and eye cores in particular, may also submit images for warehousing within the database schema.

The second concern addressed by *MuTrack* is that of data integrity and security. Once information has been submitted to the database it can only be removed by the database administrator. Investigators may edit individual data items, but *MuTrack* tracks updated data in mirror log relations, adding another layer to data recoverability. In addition, while other researchers and those entering the site

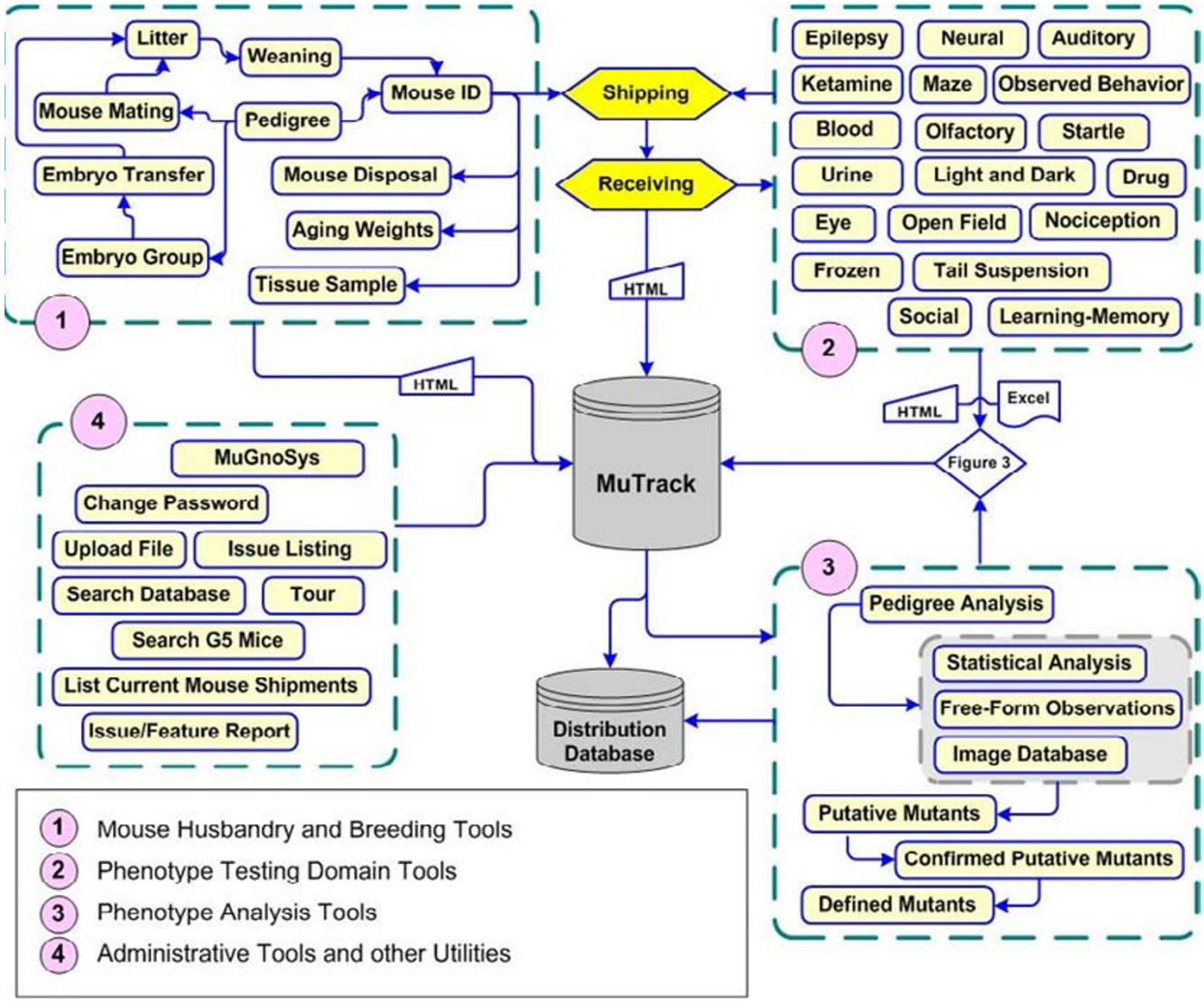

Figure 2

Schematic flow diagram representing central MuTrack architectural. The distinct On-Line Analytical Processes (OLAP) within *MuTrack* are contained within four distinct domains: (1) The *husbandry domain* that contains embedded logic about mutagenesis breeding schemas, tissue storage, test-class mouse pedigrees; (2) the *phenotype testing domain* that constitute the central data collection responsibilities in *MuTrack*. These primary and secondary testing laboratories require unique computational tools that enable them to collect and report raw and pre-processed data to the central database. (3) A domain containing *computational analysis tools* provides a means to discriminate subtle deviant phenotypes based testing domain data and (4) *MuTrack utilities* allow researches to communicate experimental observations among themselves, report bugs and features to the *MuTrack* team, and to perform necessary administrative functions. All of *MuTrack's* domains work within an embedded shipping-receiving system that tracks the movement of mice and intellectual property throughout the consortium.

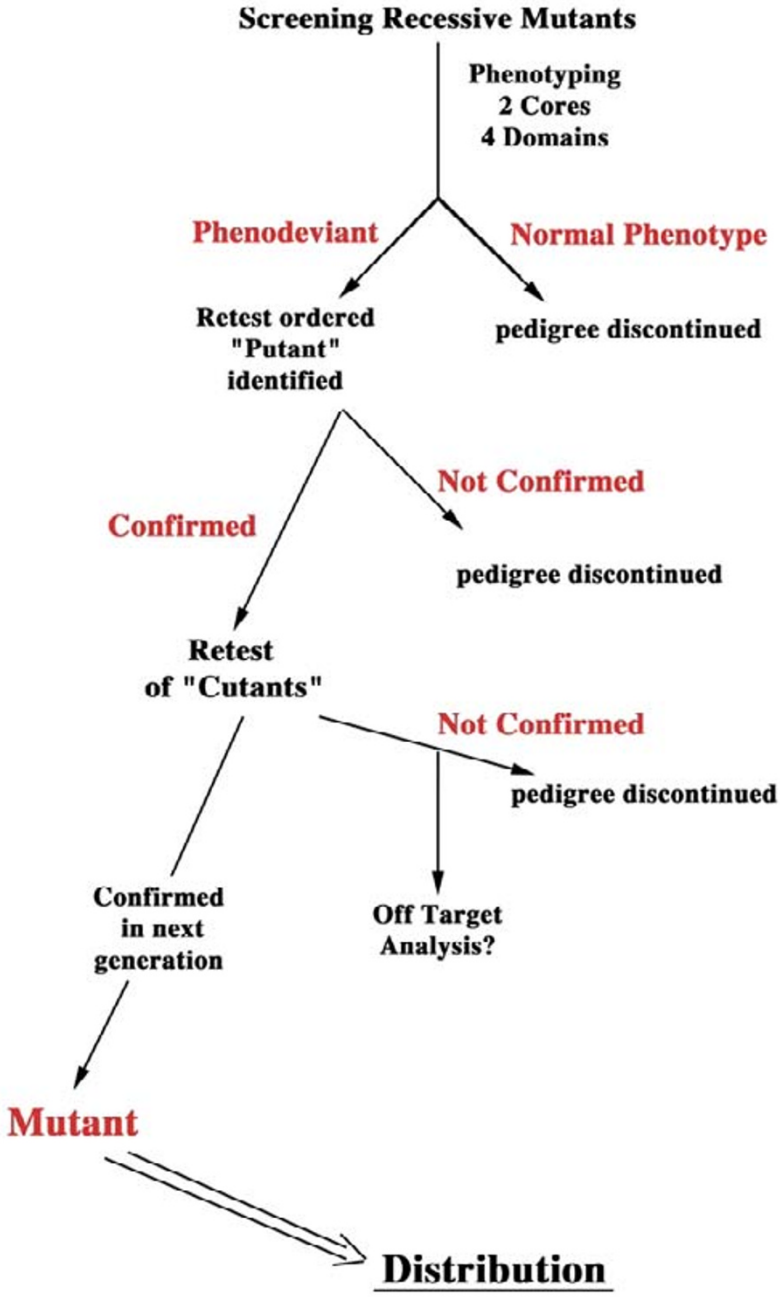

Figure 3
Decision tree for mouse movement through phenotyping domains. The computational tools contained within *MuTrack* are designed to control the flow of mutagenized mice through the different testing domains. Once representatives from a particular pedigree have been screened by all domains, they will be categorized as normal, within the scope of observed behaviour or phenotype of control pedigree members, or phenodeviant. Representatives of the latter categories are classified as *putative mutants*, or "putants", and are retested. Pedigrees continuing to express phenotypic deviations are further classified as *confirmed putative mutants*, or "cutants", and are re-tasked by *MuTrack* for heritability testing. If the phenodeviation persists in subsequent generations, mice from the pedigree are confirmed as *mutants* and are made available for distribution.

Table 1: Size and Scope of MuTrack Database

#	Relation	Number of Attributes	Primary Key(s)	Number of Tuples	Number of Discrete Data Points
1	AGING	134	AGING_ID, MOUSE_ID	1875	251250
2	AUDITORY	41	AUDITORY_ID, MOUSE_ID	497	20377
3	BEHAVIOR_LD	41	BEHAVIOR_LD_ID, MOUSE_ID	625	25625
4	BEHAVIOR_LM	101	BEHAVIOR_LM_ID, MOUSE_ID	15714	1587114
5	BEHAVIOR_MZ	26	BEHAVIOR_MZ_ID, MOUSE_ID	458	11908
6	BEHAVIOR_NO	21	BEHAVIOR_NO_ID, MOUSE_ID	1812	38052
7	BEHAVIOR_OB	40	BEHAVIOR_OB_ID, MOUSE_ID	971	38840
8	BEHAVIOR_OF	46	BEHAVIOR_OF_ID, MOUSE_ID	8484	390264
9	BEHAVIOR_OL	21	BEHAVIOR_OL_ID, MOUSE_ID	1653	34713
10	BEHAVIOR_ST	34	BEHAVIOR_ST_ID, MOUSE_ID	112475	3824150
11	BEHAVIOR_TC	22	BEHAVIOR_TC_ID, MOUSE_ID	99	2178
12	BEHAVIOR_TS	35	BEHAVIOR_TS_ID, MOUSE_ID	20475	716625
13	BLOOD	35	BLOOD_ID, MOUSE_ID	987	34545
14	DRUG	103	DRIG_ID, MOUSE_ID	1968	202704
15	EMBRYO_GROUP	15	EMBRYO_GROUP_ID	1031	13403
16	EMBRYO_TRANSFE	12	EMBRYO_TRAN_ID	1073	12876
17	EPILEPSY	24	EPILEPSY_ID, MOUSE_ID	210	5040
18	ETHANOL	48	ETHANOL_ID, MOUSE_ID	1974	94752
19	EYE	54	EYE_ID, MOUSE_ID	1975	106650
20	FORUM	9	USERID	1706	15354
21	FROZEN	30	FROZEN_ID, MOUSE_ID	261	7830
22	HERITABILITY	22	HERITABILITY_ID	5	110
23	KETAMINE	32	KETAMINE_ID, MOUSE_ID	90	2880
24	MATING	15	MOTHER_ID, FATHER_ID	2341	35115
25	MOUSE	14	MOUSE_ID	22600	35115
26	MOUSE_DISPOSAL	11	MOUSE_ID	12471	137181
27	MUTANT	23	MUTANT_ID	75	1725
28	MUTANT_ORDER	43	MUTANT_ORDER_ID	7	301
29	MUTRACK_LOG	11	LOG_ID	823100	9054100
30	NEURAL	40	NEURAL_ID, MOUSE_ID	14340	573600
31	PEDIGREE	6	FOUND_FEMALE_ID, PEDIGREE_ID	951	5706
32	PUTANT	12	PUTANT_ID	288	3456
33	SHIPPING	13	SHIPPING_ID, MOUSE_ID	1472	19136
34	SHIPTRACK	16	MOUSE_ID, SHIPPING_ID	13012	208192
35	SOCIAL	70	MOUSE_ID, SOCIAL_ID	2248	157360
36	TISSUE	14	MOUSE_ID, TISSUE_ID	5685	79590
37	URINE	27	URINE_ID, MOUSE_ID	1004	27108
38	WEANING	15	WEANING	7050	49350
	TOTALS	**1,276**		**1,083,062**	**17,824,275**

using the public password have access to view and search data, only the primary domain investigator has permission to download, submit, delete or edit information.

Computational tools

The main strength of *MuTrack* lies in its ability to initiate a real-time analysis of phenotype domain data to classify subtle phenodeviants. Analysis tools are designed to compare any particular mouse against members of its same litter, pedigree, generation or against control pedigrees and pedigrees under similar mutational pressure. These processes are entirely dynamic.

In order to ensure that publicly defined mutations are not released before a consensus of their proven heritability has been reached, the computational tools are separated into two different web domains. Public users may enter the "open" statistical pages for *MuTrack*. These pages allow public users to search mice located in any domain for deviants based on standard deviations from the mean. These tests are often used by TMGC researchers as a rudimentary analysis of their submitted data and *do not* represent precise statistical outliers. Figure 4 is a graphical representation of how members of mouse pedigree 047TNJ faired in one particular test in the ***Ethanol*** domain. Other relative individuals from different pedigrees are placed along the horizontal access for compari-

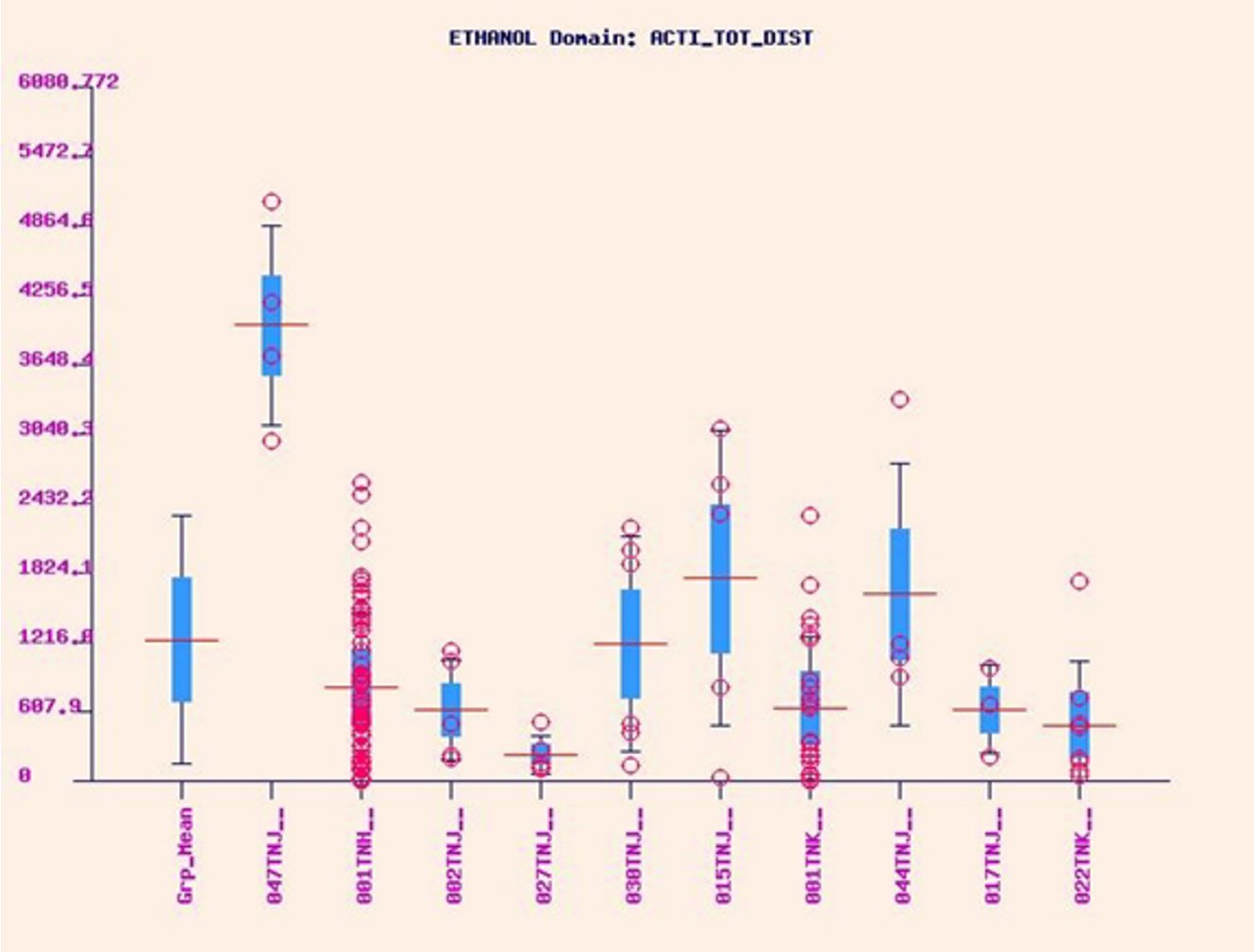

Figure 4

Sample graphical representation of experimental outliers. *MuTrack* contains several different statistical methods for determining phenodeviant pedigrees against a user-defined background population. This rudimentary box and whisker plot represents the distribution of results for each mouse in pedigree 047TNJ in a one test performed by the *Ethanol* testing domain. Comparison pedigrees are placed along the horizontal axis from comparison. Red circles represent individual mice, blue boxes represent one standard deviation (SD) and whiskers represent two SD.

son. Public *MuTrack* access also allows users to view graphical representations of data from areas outside of the statistic pages. Figure 5 demonstrates the weight progress of mice belonging to pedigree 268TNC, available in the main ***Aging Weights*** domain. The red bar may indicate a weight gain that is significantly greater or less than that of comparable mice. This allows researchers to quickly gauge the health of the testing pedigree stock.

TMGC consortium members have access to more complete computational tools as described in the methods section. Tools in this domain also compute dynamic reports with the aim of isolating statistical outliers, but are more robust in sample selection, test selection, and cross-domain test comparisons. In addition, tests located in this controlled space correct for blindness, a side-effect of some breeding strategies, sex, aging and other variables of particular concern to the testing domain. Researchers can create dynamic weekly reports that use trimmed testing sets and can create publication-quality histograms of data sets. An exhaustive list of available administrative and analysis tools is available on the *MuTrack* site.

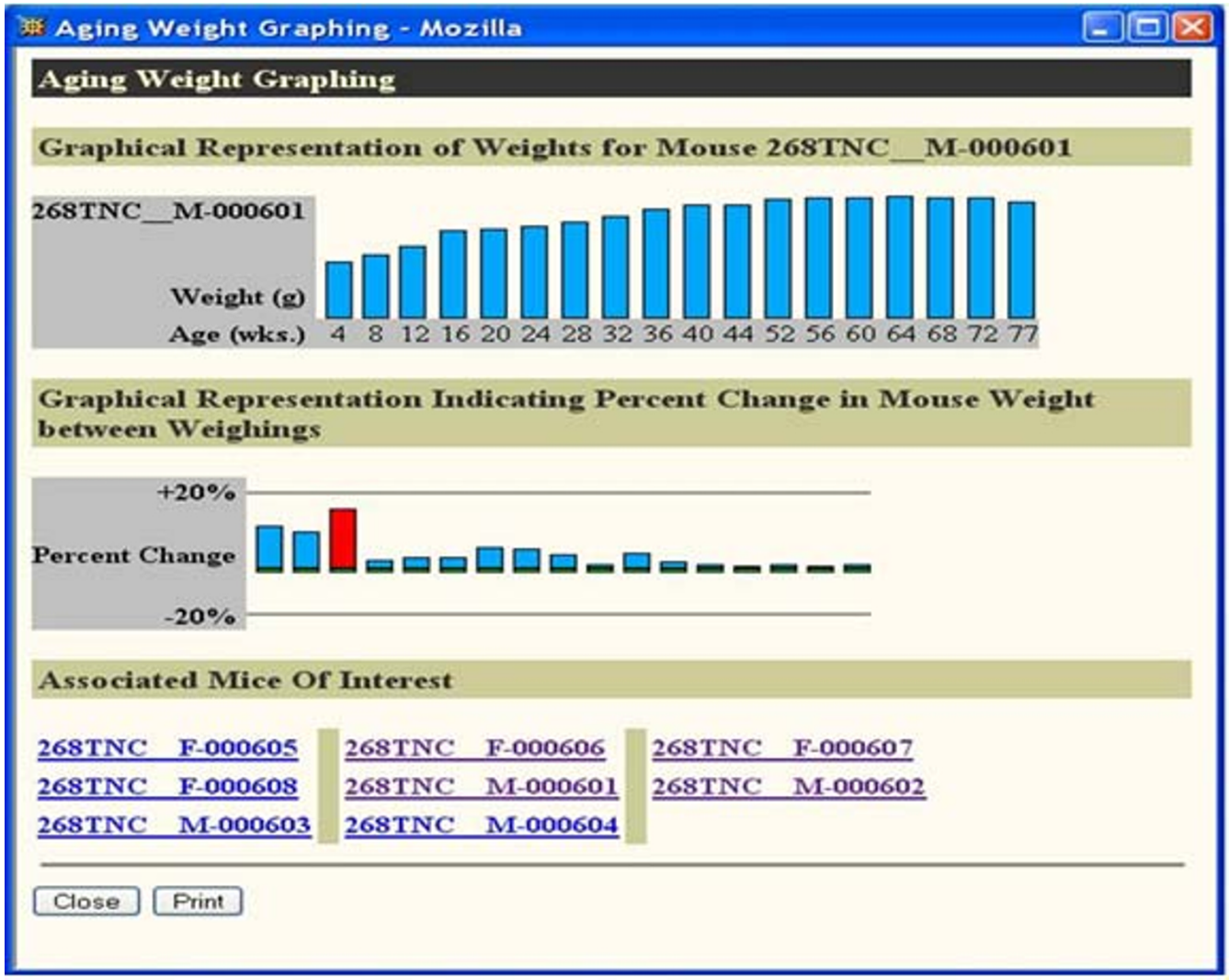

Figure 5

Sample chart of mouse weight by age. *MuTrack* users may access real-time data representations at various locations within the system. In this example, weights for mouse pedigree 268TNC have been examined and compared to mouse populations with similar mutagenic backgrounds. Unhealthy total weights or percentage of weight gains and loses are depicted by red bars. Systems such as these allow *MuTrack* users the ability to easily assess various aspects of mouse health, location, or testing status.

Defining mouse mutants

During primary mouse screening researchers rely on the statistical analysis generated by *MuTrack's* computational tools to make determinations about the deviation of a pedigree's phenotype. Any primary domain investigator may set a "deviation" flag via online switches, indicating that the mouse pedigree is a 'putative mutant', or *putant*. Following a structured decision tree (Figure 3), *MuTrack* initiates an automatic alert and the physical retesting of the putant pedigree. If pedigrees continue to be classified as statistically aberrant, the domain investigator is given the opportunity to promote putants to *cutants*, or 'confirmed mutants'. *MuTrack* then initiates a process to test the phenotype deviation for heritability. Putant and cutant pedigrees return to the same testing domain that first noticed the primary abnormality and, in addition, are tested in secondary domains, some of which are located outside of the consortium. Secondary domains provide alternate methodologies for quantization of phenotype abnormalities that serve to refine phenotype characteristics. *MuTrack* combines and interprets data from primary and secondary testing domains and forwards results to a

TMGC committee that makes the final determination of mutant heritability. Positive mice are determined to be *mutants* and are made available to the mutant mouse distribution effort (along with visual and lethal mutants), located on the Jackson Laboratory website [22]. A listing of current mouse mutants is available at the Tennessee Mouse Genome Consortium homepage [5].

Conclusion

The diversification of experimental techniques in all areas of biological research has caused a trend in laboratory specialization that exceeds the ability of any single primary investigator to provide comprehensive validation of genome-wide investigations. Simultaneously, the excess of quantitative data and empirical observations produced by varying research techniques far outstrips the ability of computational tools to adequately analyze the data for meaningful inferences. These issues combined with finite labor and funding resources have forced large research projects to use bioinformatics techniques to extract a maximum of information at a reasonable cost from geographically dispersed researchers. Researchers at the TMGC are attempting to bring together research teams using a centralized on-line database and analysis toolbox. Because distributed bioinformatics collaborations are relatively unknown quantities in large-scale hypothesis driven research, the TMGC was forced to engineer a system *de novo* to meet its particular needs.

The *MuTrack* system was initially released as the central bioinformatics tool for the TMGC in February, 2001. The database responsible for collecting experimental data and generating dynamic web content, including data analysis and knowledge exchange, has grown by the average rate of 34,000 tuples per month. The system has proven to be flexible, robust, extensible and, most importantly, has to date helped to elucidate 75 new heritable phenotypes.

While the system is fundamentally sound it is not exhaustive. Development continues to incorporate ongoing research as it moves into the molecular characteristics of mouse phenodeviants. Ideally, future mutants will be categorized at both gross and molecular granularity and *MuTrack* will be used to bring together genetic observations and phenotypic effects. Incorporation of primitive phenotyping ontologies will greatly increase our ability to communicate new phenodeviants [23]. Computational systems are under development that will enable *MuTrack* to support recombination analysis, including the examination of quantitative trait loci and make reasonable inferences about molecular networks and gene regulation. Operationally, it is beyond the scope of *MuTrack* to create a panacea for the needs of every mouse-centric research scenario, but it remains our goal to maintain the software flexibility necessary to allow future application development in a variety of concerted research directions.

Lessons learned from *MuTrack* can contribute favorably to future distributed team research directives. First, there is no immediately apparent generic or proprietary solution to every problem encountered during the development of distributed bioinformatics software. Research, by definition, produces either novel data types or requires the novel interpretation of data. Cogent engineering of software must be conducted in conjunction with a clear biological hypothesis to demonstrate progress in either area. Secondly, the compulsory use of *MuTrack's* data collection, analysis and results reporting tools by consortium researchers has greatly aided in the refinement of the system for external users. Bioinformatics systems are capable of producing substantive results only if meaningful data is collected and analyzed, and robust software is only created under real conditions of use. Finally, future large-scale projects that rely heavily on centralized software must allow individual researchers the ability to supplement generalized computational results with free-form observations. To this end, *MuTrack* developers are attempting to incorporate data analysis systems and results-reporting functions with virtual publication areas, where consortium members may collaborate in the construction of publication quality documents.

There are currently several large-scale and genome-wide research projects that rely heavily on bioinformatics for the elucidation of novel observations. *MuTrack* provides a working framework for these projects.

Availability and Requirements

MuTrack is available to members of the TMGC neuromutagenesis phenotyping project. There are currently twenty-two discrete testing and husbandry domains located at seven independent institutions within the state of Tennessee that make daily contributions to, or take advantage of, *MuTrack* data, knowledge, or analysis. Non-members can access a limited number of web interfaces via the TMGC homepage [5] when using the directed public password and username.

Authors' Contributions

Web-based interfaces for the original version of *MuTrack* were developed by EB, with subsequent development by JS, BJ, and LG. Design and implementation of the Oracle 8*i* database that supports *MuTrack* is credited to DS, BJ, JS and EB. Statistical analysis packages, computational tools and knowledge sharing software were designed and implemented by LG and EB. JS and EB were responsible for overall project design and coordination. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Daniel Goldowitz Ph.D. and Gene Rinchik Ph.D. for their critical reading of this manuscript. In addition, we would like to thank Elissa Chesler, Ph.D. for her statistical insights. We would also like to acknowledge the funding and collaborative support provided by the TMGC member institutions.

References

1. Avery P: **Data Grids: a new computational infrastructure for data-intensive science.** *Philos Transact Ser A Math Phys Eng Sci* 2002, **360:**1191-1209.
2. Gantenbein RE: **Designing an Internet-based collaboratory for biomedical research.** *Biomed Sci Instrum* 2002, **38:**399-404.
3. Goh CS, Lan N, Echols N, Douglas SM, Milburn D, Bertone P, Xiao R, Ma LC, Zheng D, Wunderlich Z, Acton T, Montelione GT, Gerstein M: **SPINE 2: a system for collaborative structural proteomics within a federated database framework.** *Nucleic Acids Res* 2003, **31:**2833-2838.
4. Tonini C, Beghi E, Telaro E, Candelise L: **The Cochrane collaboration in neurology: acquisitions, research, and new initiatives.** *Neuroepidemiology* 2001, **20:**153-159.
5. **The Tennessee Mouse Genome Consortium Homepage** [http://www.tnmouse.org/]
6. Buer J, Balling R: **Mice, microbes and models of infection.** *Nat Rev Genet* 2003, **4:**195-205.
7. Rogner UC, Avner P: **Congenic mice: cutting tools for complex immune disorders.** *Nat Rev Immunol* 2003, **3:**243-252.
8. Svenson KL, Bogue MA, Peters LL: **Genetic Models in Applied Physiology: Invited Review: Identifying new mouse models of cardiovascular disease: a review of high-throughput screens of mutagenized and inbred strains.** *J Appl Physiol* 2003, **94:**1650-1659.
9. Watase K, Zoghbi HY: **Modelling brain diseases in mice: the challenges of design and analysis.** *Nat Rev Genet* 2003, **4:**296-307.
10. Shastry BS: **More to learn from gene knockouts.** *Mol Cell Biochem* 1994, **136:**171-182.
11. Justice MJ, Carpenter DA, Favor J, Neuhauser-Klaus A, Hrabe de Angelis M, Soewarto D, Moser A, Cordes S, Miller D, Chapman V, Weber JS, Rinchik EM, Hunsicker PR, Russell WL, Bode VC: **Effects of ENU dosage on mouse strains.** *Mamm Genome* 2000, **11:**484-488.
12. Rinchik EM, Carpenter DA: **N-ethyl-N-nitrosourea-induced prenatally lethal mutations define at least two complementation groups within the embryonic ectoderm development (eed) locus in mouse chromosome 7.** *Mamm Genome* 1993, **4:**349-353.
13. Rinchik EM, Carpenter DA, Johnson DK: **Functional annotation of mammalian genomic DNA sequence by chemical mutagenesis: a fine-structure genetic mutation map of a 1- to 2-cM segment of mouse chromosome 7 corresponding to human chromosome 11p14-p15.** *Proc Natl Acad Sci U S A* 2002, **99:**844-849.
14. **SAS...The Power to Know** [http://www.sas.com/]
15. **PHP Development Website** [http://www.php.net]
16. **GD Graphics Library** [http://www.boutell.com/gd/]
17. **The Perl Directory** [http://www.perl.org]
18. **The Definitive Javascript Resource** [http://www.javascript.com/]
19. **Oracle Corporation** [http://www.oracle.com]
20. **MuTrack Hompage** [http://www.tnmouse.org/mutrack/]
21. Williams RS, Willard HF, Snyderman R: **Personalized health planning.** *Science* 2003, **300:**549.
22. **The Jackson Laboratory** [http://www.jax.org/]
23. **Integrated access to mouse phenotyping projects** [http://www.neuromice.org/]