A fielded wiki for personality genetics

Finn Årup Nielsen
DTU Informatics
Technical University of Denmark
Kongens Lyngby, Denmark
fn@imm.dtu.dk

ABSTRACT

I describe a fielded wiki, where a Web form interface allows the entry, analysis and visualization of results from scientific papers in the personality genetics domain. Papers in this domain typically report the mean and standard deviation of multiple personality trait scores from statistics on human subjects grouped based on genotype. The wiki organizes the basic data in a single table with fixed columns, each row recording statistical values with respect to a specific personality trait reported in a specific paper with a specific genotype group. From this basic data hard-coded meta-analysis can compute individual and combined effect sizes. The meta-analytic results are displayed in on-the-fly computed hyperlinked graphs and tables. Revision control on the basic data tracks changes and data may be exported to comma-separated files or in a MediaWiki template format.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information Services—*Web-based services*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*Architectures*; J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms

Design, Human Factors

Keywords

Wikis, structured data, tables, meta-analysis, bioinformatics, neuroinformatics

1. INTRODUCTION

Neuroinformatics deals with representing neuroscience data and develops tools for analysis and visualization of such data. Most data in the neuroscience field is only reported in scientific papers and database curators have difficulties in keeping up with the increasing amount of data being generated. We have constructed Brede Database[1] based on XML and Matlab for results of neuroimaging data [9]. With information from 186 papers it is smaller than the BrainMap database[2] [7] and SumsDB[3], that each have information from around 2000 papers. However, even these large databases cannot keep up with the papers published in the neuroimaging field [4]. I have advocated for a more collaborative and wiki-oriented approach to overcome the problem of database entry [11], and set up the MediaWiki-based Brede Wiki to record text and numerical information from neuroscience [10]. Neither the Matlab entry interface for the Brede Database nor the raw edit field of the MediaWiki interface present a convenient interface for adding data, and therefore we are exploring alternative ways of data entry. Here, I describe a field-based wiki interface to add and edit data from a very specific area of neuroinformatics/bioinformatics: Personality genetics. This field investigates the relationship—the genetic association—between personality traits and specific genetic variations by letting a group of genotyped subjects complete a personality questionnaire (personality inventory). The researchers typically compute average personality scores across subjects within a genotype group and list the values in tables in their published papers. AlzGene [3], SzGene and PDGene record data from genetic association studies on Alzheimer, schizophrenia and Parkinson diseases, respectively. The original data from the scientific studies are entered and presented in a Web-based environment with meta-analysis results and visualizations. Inspired by these systems the fielded wiki also includes meta-analysis and visualization.

2. THE WIKI

The basic data structure of the wiki is a table where the columns are: row identifier, gene, polymorphism, genotype, inventory, trait, mean value, standard deviation, number of subjects, recruitment information and document identifier, see Figure 1. A revision table has the same columns and further adds columns for revision time, revision user as well as a revision identifier. For the gene field the standard gene symbol is used. Numbering systems for genetic variants associated with a gene—polymorphisms—have emerged, e.g., "rs6311", but the wiki has so far used an alphanumeric system, e.g., "A-1438G" (a format that may vary slightly between researchers). The fourth column records the genotype for example "A/A" for a homozygote subject group with the

[1] http://neuro.imm.dtu.dk/services/brededatabase/
[2] http://brainmap.org/
[3] http://sumsdb.wustl.ed

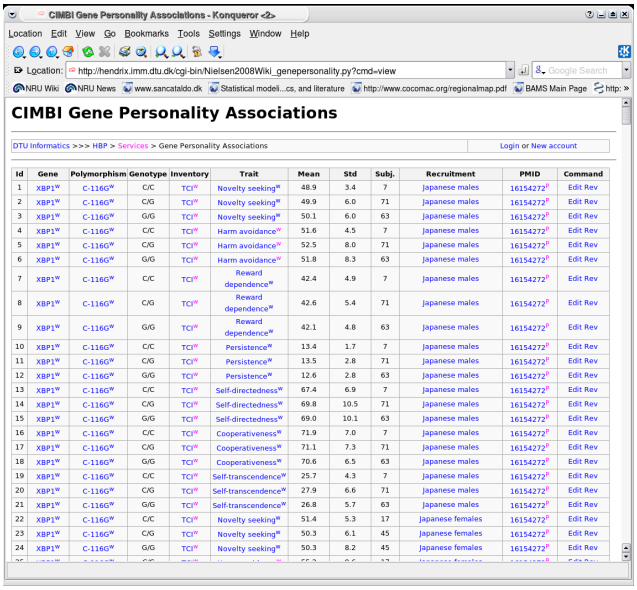

Figure 1: The main window with the basic data of the wiki as organized in a table. Field values are hyperlinked to filtered views or to external wikis: Wikipedia and the Brede Wiki.

A allele for a specific polymorphism. The next two columns record the type of personality test—the so-called personality inventory—and the personality trait. The next three columns record numerical data: The mean value for the score on the personality trait for subject group, the standard deviation of that group as well as the number of subjects in it. The second last field describes the subject group or recruitment information with a string such as "Japanese medical staff". The last column is a document identifier associated with the scientific paper that displays the results, and the wiki uses an external identifier: The PubMed identifier (PMID).

New data is entered in a form displayed as an extra row in the table. Through JavaScript an autocomplete feature copies the corresponding value from the previous row to the input field if the input field is left empty, — except for mean, standard deviation and number of subject fields which will almost always be different between consecutive rows. For the trait field a simple autoreplace expands, e.g., 'N' to 'Neuroticism'. If a paper only report data from a single subject group with a single inventory, then for most rows only the genotype, mean and standard deviation needs to be entered. The rest of the fields will automatically be completed when the user tabs through the fields.

There is little constraint on what the user can enter in the fields, e.g., it is possible to enter a value in the gene field that is not a valid human gene symbol.

Apart from adding new data the wiki interface also allows a user to go back and edit a specific row. A separate form is used for that purpose, see Figure 2. On this form predefined options for the most common choices guide the user to select appropriate values.

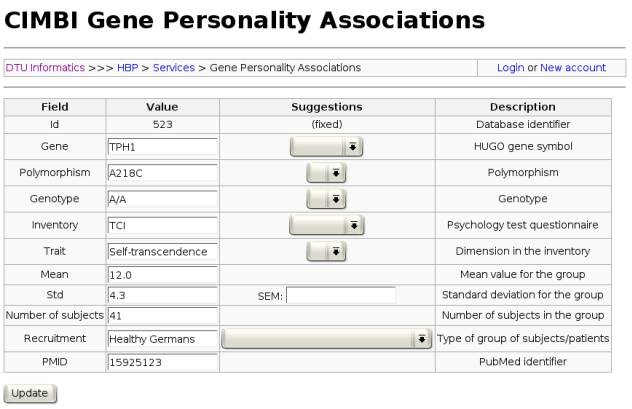

Figure 2: Pressing "edit" on a specific row in the main window will bring up a form-based interface where data associated with a personality score can be changed.

For gene, polymorphism, inventory, trait, recruitment and PMID hyperlinks are automatically constructed so filtered views of the data are presented when the links are followed. Also gene, polymorphism, inventory and trait are hyperlinked to their corresponding pages in Wikipedia and the Brede Wiki, and the PMID is linked to the PubMed database.

The wiki can generate so-called forest plots showing each row in the table as a dot with whiskers: The mean determines the position of the dot, the standard deviation the whiskers, and the number of subjects the size of the dot. A form interface allows to filter which rows are shown.

2.1 Meta-analysis

A statistical framework for meta-analysis has existed for a long time [6], and this framework can be applied to genetic association data. A central concept in meta-analysis is the *effect size* d, which in the present case is defined as a standardized mean difference

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}, \tag{1}$$

where $\bar{x}_1$ is the mean personality score for one group of subjects, $\bar{x}_2$ is the mean for a comparison group and s is the standard deviation within the groups. Slight variations exist for this computation and the wiki uses Hedges' unbiased d for the computation [6].

Presently effect sizes are computed for all homozygotes-homozygotes comparisons for each polymorphism and personality trait examined in each paper, and it is performed when the user follows the "compute effect" link. Values from two rows in the wiki needs to be extracted to form one effect size. The computed effect sizes are added to another persistent database table, and the user can view this table with the Web interface.

Given several studies and their computed effect sizes a meta-analytic estimate of a common effect size is possible. Hedges and Olkin suggest a large sample approximation for the vari-

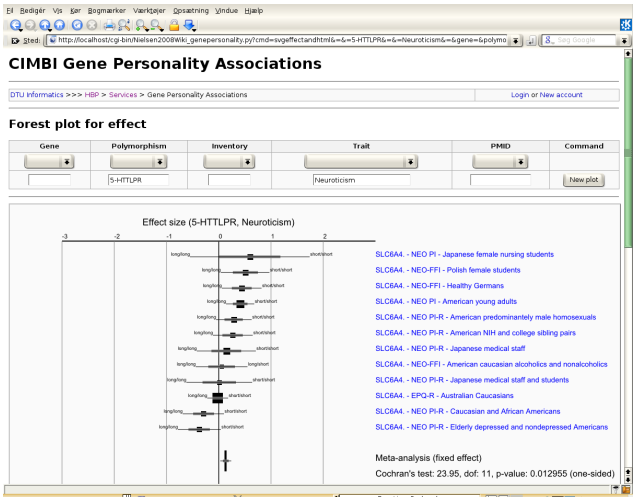

Figure 3: Forest plot of effect sizes in SVG controlled by a form interface with filtering on the 5-HTTLPR polymorphism and the personality trait neuroticism. The combined meta-analytic effect size is plotted at the bottom.

ance of d of the kth study as

$$\hat{\sigma}^2(d_k) = \frac{n_{1,k} + n_{2,k}}{n_{1,k} n_{2,k}} + \frac{d_k^2}{2(n_{1,k} + n_{2,k})}. \quad (2)$$

A combined effect size d_+ among K studies may be found as

$$"d_+ = \hat{\sigma}^2(d_+) \sum_{k=1}^{K} \frac{d_k}{\hat{\sigma}^2(d_k)}, \quad (3)$$

where the variance of the combined effect, $\hat{\sigma}^2(d_+)$, is estimated as

$$\hat{\sigma}^2(d_+) = \left(\sum_{k=1}^{K} \frac{1}{\hat{\sigma}^2(d_k)} \right)^{-1}. \quad (4)$$

With a large sample Gaussian approximation the combined effect size and its variance may be used in a comparison against the normal distribution to obtain P-values and confidence intervals.

The wiki generates a forest plot with the effect sizes (Eq. 1), their standard error (Eq. 2) as well as the combined effect (Eqs. 3 and 4). Through a form interface the user can filter which effect sizes should be included in the visualization and the on-the-fly computation of the combined effect sizes, see Figure 3. The blue annotations in the plot are hyperlinked to their associated entries in the effect size table.

As another type of meta-analytic plot used to check for possible publication bias a so-called funnel plot displays a scatter plot of the effects with, e.g., the effect size on the x-axis and the standard error on the y-axis [5]. The wiki generates these plots for all effects or filtered according to, e.g., polymorphism and trait, see Figure 4. The points in this plot are also hyperlinked to their entries in the effect size table.

Data can be exported in comma-separated files or in a MediaWiki template format, that can be included in the Brede

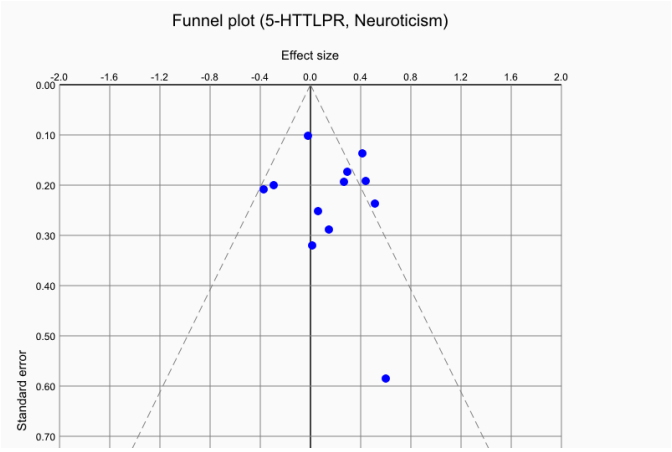

Figure 4: Funnel plot for the polymorphism 5-HTTLPR and the trait neuroticism in an on-the-fly generated SVG file. Each dot represents one of 12 effects and each is hyperlinked to its respect entry in the effect size table.

Wiki (or any other MediaWiki). The wiki exports each row as a MediaWiki template, and in the Brede Wiki these template are formatted to rows in HTML tables.

A simple cookie-based authentication system may allow a user to log in. Otherwise the edit activity is recorded with the IP number in the revision table.

So far data from 23 papers have been entered amounting to 26 different personality traits from 7 different test batteries, 10 different polymorphisms from 7 different genes, 20 different subject groups (patient groups as well as healthy). It adds up to a total of 540 trait/genotype personality score values. This database is not the largest in terms of studies, e.g., a 2003 meta-analysis included 46 studies, though only around 250 personality score values [8].

The wiki is implemented in a single Python script with the SQLite database engine as a backend. The database is separated in two SQLite files: One for the personality genetics data, and one for the wiki user account data. This separation makes it possible to distribute the personality genetics data without exposing user account data. The SciPy Python package is used for statistical testing. The visualizations are generated in Scalar Vector Graphics

3. DISCUSSION

The fielded wiki lacks the standard wiki concept of a page with markup, so it a strict sense it is not a wiki at all. However, it has collaborative Web-based entry with tracking of revisions.

Some of the features of the wiki could be implemented with other structured collaborative systems, e.g., with MediaWiki templates, MediaWiki table extensions (see e.g., DynaTable [1]), Semantic MediaWiki [12] with form interface, OntoWiki [2], Google Fusion Tables or online spreadsheets, such as the spreadsheet in Google Docs. As an example SNPedia[4] is

[4]http://www.snpedia.com

a wiki that uses Semantic MediaWiki to store information about polymorphisms and genetic associations, i.e., data very much like the data in the present wiki. Although such systems may represent the data, most of these systems cannot directly perform the automated hyperlinking, the specialized meta-analysis and visualization implemented in the wiki. Furthermore, the wiki can tailor its interface to support fast entry.

The wiki was designed so the basic data would be represented in an as simple structure as possible. More tables would be needed if, e.g., bibliographic information and subject demographics should be recorded in detail. Although the wiki has been setup for personality inventories and traits, it may also record results from other psychological tests. However, it will not be able to record more complex experimental designs with, e.g., interaction effects.

Using the PubMed identifier as key for a scientific paper is convenient, but unfortunately not always possible: Scientific papers may not be recorded in that database, and papers in press may not have an identifier assigned yet.

The wiki lacks a flexible interface for meta-analysis. In some meta-analyses, e.g. in AlzGene, one chooses to compute the results without the initial study, regarding it as the hypothesis generating study, and the rest as the hypothesis testing studies. One may also split the studies according to subject population, e.g., only analyze Caucasians or healthy subjects. The wiki has not yet such flexibility. For the moment the user may export the data as comma-separated files and make a more tailored analysis in an external program. Alternatively, the functionality of the wiki should be extended, e.g., with a programming-oriented interface or with an interface to specify which values should enter the meta-analysis.

The studies entered in the wiki are so-called candidate gene studies where only few genes and polymorphisms (often just a single) are examined. Newer studies perform genome-wide scans across large numbers of polymorphisms. Results from such studies may not be presented in the paper with enough details to be included in the wiki, and would be to cumbersome to enter anyway. To handle such studies an interface where the researcher could upload the results would be necessary.

The wiki can be regarded as a system of open data in science, where the results included in a meta-analysis are readily available for re-use, and the wiki has potential for large-scale meta-analysis across multiple traits and genetic variants.

4. CONCLUSIONS

A simple fielded wiki for a specific field—personality genetics—has been described. The wiki lacks a flexible representation, but with a simple interface it enables tailored, incremental and collaborative data entry, hyperlinked view filtering as well as meta-analysis and visualizations.

5. ACKNOWLEDGMENTS

Thanks to the Lundbeck Foundation for support through the Center for Integrated Molecular Brain Imaging (CIMBI).

6. REFERENCES

[1] C. Arnold, T. Fleming, D. Largent, and C. Lüer. DynaTable: a wiki extension for structured data. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, New York, NY, USA, 2009. ACM.

[2] S. Auer, T. Riechert, and S. Dietzold. OntoWiki — a tool for social, semantic collaboration. In I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Uschold, and L. Aroyo, editors, *The Semantic Web — ISWC 2006*, volume 4273 of *Lecture Notes in Computer Science*, pages 736–749, Heidelberg/Berlin, 2006. Springer.

[3] L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics*, 39(1):17–23, January 2007.

[4] J. Derrfuss and R. A. Mar. Lost in localization: The need for a universal coordinate database. *NeuroImage*, 48(1):1–7, October 2009.

[5] M. Egger, G. D. Smith, M. Schneider, and C. Minder. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–634, September 1997.

[6] L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, Florida, 1985.

[7] A. R. Laird, J. L. Lancaster, and P. T. Fox. BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics*, 3(1):65–78, 2005.

[8] M. R. Munafò, T. G. Clark, L. R. Moore, E. Payne, R. Walton, and J. Flint. Genetic polymorphisms and personality in healthy adults: A systematic review and meta-analysis. *Molecular Psychiatry*, 8:471–484, 2003.

[9] F. Å. Nielsen. The Brede database: a small database for functional neuroimaging. *NeuroImage*, 19(2), June 2003. Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY. Available on CD-Rom.

[10] F. Å. Nielsen. Brede Wiki: Neuroscience data structured in a wiki. In C. Lange, S. Schaffert, H. Skaf-Molli, and M. Völkel, editors, *Proceedings of the Fourth Workshop on Semantic Wikis — The Semantic Wiki Web*, volume 464 of *CEUR Workshop Proceedings*, pages 129–133, Aachen, Germany, June 2009. RWTH Aachen University.

[11] F. Å. Nielsen. Lost in localization: A solution with neuroinformatics 2.0. *NeuroImage*, 48(1):11–13, October 2009.

[12] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, pages 585–594, New York, NY, USA, 2006. ACM.