Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computational Advances in Cancer Informatics (B)

StickWRLD as an Interactive Visual Pre-Filter for Canceromics-Centric Expression Quantitative Trait Locus Data

Robert Wolfgang Rumpf, Samuel L. Wolock and William C. Ray

The Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, Columbus, OH, USA.

ABSTRACT: As datasets increase in complexity, the time required for analysis (both computational and human domain-expert) increases. One of the significant impediments introduced by such burgeoning data is the difficulty in knowing what features to include or exclude from statistical models. Simple tables of summary statistics rarely provide an adequate picture of the patterns and details of the dataset to enable researchers to make well-informed decisions about the adequacy of the models they are constructing. We have developed a tool, StickWRLD, which allows the user to visually browse through their data, displaying all possible correlations. By allowing the user to dynamically modify the retention parameters (both *P* and the residual, *r*), StickWRLD allows the user to identify significant correlations and disregard potential correlations that do not meet those same criteria – effectively filtering through all possible correlations quickly and identifying possible relationships of interest for further analysis. In this study, we applied StickWRLD to a semi-synthetic dataset constructed from two published human datasets. In addition to detecting high-probability correlations in this dataset, we were able to quickly identify gene–SNP correlations that would have gone undetected using more traditional approaches due to issues of low penetrance.

KEYWORDS: visual analytics, eQTL, gene–SNP correlation

SUPPLEMENT: Computational Advances in Cancer Informatics (B)

CITATION: Rumpf et al. StickWRLD as an Interactive Visual Pre-Filter for Canceromics-Centric Expression Quantitative Trait Locus Data. *Cancer Informatics* 2014:13(S3) 63–69 doi: 10.4137/CIN.S14024.

RECEIVED: May 19, 2014. **RESUBMITTED:** July 30, 2014. **ACCEPTED FOR PUBLICATION:** July 30, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: Authors disclose no funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: wolfgang.rumpf@nationwidechildrens.org

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

The typical approach to cancer drug development has traditionally involved high-throughput screening of in vitro cancer cell lines.[1,2] Studies such as the NCI-60, the Cancer Cell Line Encyclopedia, and the Genomics of Drug Sensitivity in Cancer[2–4] have produced rich data including drug responsiveness and genomic characteristics of numerous cancer cell lines, which allows researchers to begin to correlate genotype to both phenotype and *change* in phenotype induced by pharmaceutical intervention. As may be expected, as the datasets grow larger, the analysis becomes increasingly more complex, requiring more sophisticated approaches. Simple tables of summary statistics rarely provide an adequate picture of the patterns and details within the dataset to enable researchers to make well-informed decisions about the adequacy of the models they are constructing, and it is impractical to apply conventional statistics and graphing approaches to a number of possible two-node ("x vs y") comparisons within such a dataset. A visual analytics approach that supplements the more conventional statistical approaches by simultaneously calculating the relevant statistics and subsequently displaying the significant correlations for all possible two-node comparisons can facilitate such an analysis.

Previous attempts at correlating drug responsiveness and cancer cell line characteristics have been encouraging, suggesting that a machine learning approach could be used to predict the drug response based on the genotype of the cell line and the chemical characteristics of the drug. Machine learning is well established as an assistive tool in drug development (for reviews see Barret and Langdon,[5] Wale[6]), but has only recently been applied to the problem of correlating genotypes, phenotypes, and drug sensitivities. Beerenwinkel et al.[7] created a Support Vector Machine (SVM), geno2pheno (available as an online resource), which can predict HIV-1 drug sensitivity in patients based on the polymerase sequence of the individual's HIV strain. Ruderfer et al.[8] created an SVM that was able to predict the sensitivity of yeast to drugs based on genotype/phenotype with 70% accuracy. Menden et al.[9] using data from the Genomics of Drug Sensitivity in Cancer project,[4] were able to accurately predict the outcomes of drug/cell line interactions for a subset of the data that had been intentionally excluded from the training data using as little as 20% of the GDSC data to train their neural network. By including SMILES descriptors to characterize compounds, they laid the groundwork for predicting cell line/genotype responses to any compound based on chemical characteristics,[10] a move toward both custom therapies, based on individual genotype, and a required streamlining of the screening process – by predicting which chemical characteristics may impact a specific genotype, the number of potential drugs/cell line combinations which must be screened can be significantly reduced.

Regardless of the method used to reconstruct these relationships, correctly identifying genotype/phenotype correlations depends on properly defining the threshold of significance – for example, what is the minimum *P* value that should be used to define significance. In some cases, this threshold is set arbitrarily – to compensate for limitations in processing power, for example, reducing the complexity of the analysis, such as in an eQTL analysis, or in other cases, to make the signal stand out from the noise, or to reduce false positives in multiple-hypothesis testing.

Critical to this process, and regardless of the intent, it is important to recognize that *no* specific threshold can guarantee that all features rejected at that threshold are spurious or unimportant, or that all things that pass the threshold are true positive features of importance. Because both machine-learning and more conventional approaches are made more feasible when only the most informative information is used for the actual analysis, all approaches are accentuated by tools that enable an expert user to make educated choices regarding the maximally informative features of the data. The trick, then, is to define which information is, in fact, the most informative – we approach this problem using a visual-analytics tool that enables browsing and exploration of the complete marginal and joint-distribution space implied by the available data.

We have developed a system, StickWRLD,[11,12] which combines both domain expertise and a variable means of visually displaying relational data. This allows the user to dynamically modify thresholds (*P* value, departure from independence) and visually assess changes in the visual – eg, the appearance of relevant signal, based on specific domain knowledge (eg, knowledge of particular genotype/phenotype characteristics, or other features, collections of features, or patterns of interaction that should appear), or the disappearance of obstructing noise – essentially manually adjusting the signal to noise ratio to an optimum. StickWRLD enables the user to intelligently and dynamically determine the optimal thresholds based on one of the most advanced pattern recognition systems available – the human brain. By using this rapid pre-filtering to identify significant relationships, StickWRLD can help identify data that may safely be eliminated from computationally intense processes such as the training of neural nets or conditional random fields, and simplifies more conventional analysis by allowing the researcher to focus only on the most interesting and/or significant relationships.

Expression Quantitative Trait Locus (eQTL) data is an ideal archetype for developing and validating this type of approach. eQTLs are typically SNPs which modulate the expression of genes. Since eQTL data is, at essence, simply another way of characterizing genotype/phenotype/phenotype correlations, we used StickWRLD to process a previously described eQTL dataset as a proof of concept. We propose that StickWRLD can enable an end-user to rapidly restrict a dataset to that most informative subset which can subsequently be used to optimally train downstream statistical models such as conditional random fields or neural networks.

Methods

Dataset. For demonstration purposes, in this manuscript, we have used a well-characterized semi-synthetic eQTL dataset that has been described previously.[11,13] This small but complex dataset was chosen because it contains several predictable features of importance, while maintaining realistic biological complexity. At the same time, its limited size makes demonstration of the analytical features of our StickWRLD tool practical, where larger datasets that remain tractable in the interactive StickWRLD interface would become unpresentably complex in static printed form.

The dataset combines two published human eQTL datasets, and provided a background of real biological eQTL relationships while also containing a spiked-in known signal as a reference control. The first dataset used in the construction of the semi-synthetic dataset included 193 neurologically and psychiatrically normal postmortem human brain samples with a microarray assay that provides data on gene expression from all known genes, and genomic data comprised genotypes at 500,000 SNP loci.[14] The second dataset consisted of 150 normal and psychiatrically diagnosed postmortem human brain samples with directly analogous gene expression and SNP data.[15] From this combined dataset, we choose a subset containing 500 individuals, with normalized expression values for 15 cadherin superfamily genes as well as the allelic state for 239 SNPs. This subset of the larger dataset was chosen to be of a size easily presentable for publication purposes; in practice, StickWRLD itself is capable of processing many more nodes and is limited practically only by the available screen real estate available to the user. For each variable, each state (eg, no change, upregulated, or downregulated in the case of the loci; or diploid allelic state in the case of SNPs) was represented with an alphabetic letter designation to facilitate entry into StickWRLD. The complete datasets (original as well as binned), as well as the StickWRLD manual and python scripts, are available for download from http://www.stickwrld.org/rwr-ci-2014/

StickWRLD visualization. The binned dataset was loaded into the python-executable version of StickWRLD build 1321 (available from the authors upon request). StickWRLD calculates the P and r values for all possible combinations of values, where P is the canonical P value for significance (set at 0.05 by default), and r is the residual of the observed number of covariates minus the expected number of covariates,[12] and then displays the correlations based on the threshold settings. The initial thresholds for the display of correlations were $P = 0.05$ and $r = 0.1$. This results in StickWRLD displaying only those two-node correlations that exceeded both these values. As the initial display showed too few correlations of interest, the residual threshold was reduced in increments of 0.01 until the display showed additional correlations of interest, specifically those indicating up- or down-regulation of any of the genes in the dataset.

Results

StickWRLD correlation visualizations. Correlations in StickWRLD are displayed as line connecting the edges of the circle; each variable (in this case, a gene or SNP) is located in a stacked column along the circumference of the circle. Each possible value of that variable is represented by a sphere, where the size of the sphere represents the corresponding prevalence of that value – thus, a larger sphere indicates that a particular value occurs more frequently. The color of the spheres is simply tied to the state and has no mathematical relevance. When a line is drawn between two points on the circle, this indicates the presence of a correlation with a corresponding P and r of *at least* the current filter settings. The style of the line indicates whether the correlation is positive (solid line) or negative (dashed line), and the thickness of the line indicates the strength of the correlation, where a thicker line reflects a stronger correlation.

Figure 1 displays the initial view of the dataset when loaded into StickWRLD using the defaults of $P = 0.05$ and $r = 0.1$. Although several correlations are seen using these defaults, these are all in fact correlations between SNPs. This indicates that the co-occurrence of the correlated SNPs occurs at a significant frequency. For example, rs4783754 is correlated to both rs9922615 and rs9924505, suggesting that there are specific alleles of rs4783754 which tend to be co-inherited with specific alleles of rs9922615 and rs9924505. Similarly, CHR16:67319752 and CHR16:67381383 are also correlated to one another. This is completely expected for proximal SNPs where the frequency of recombination between them is low, but can reveal interesting distantly interacting loci when the SNPs are sufficiently distant that the probability of recombination approaches 50%. None of the patterns observed here, however, showed any correlation to changes in expression of the genes in the dataset.

Tuning the residual down by increments reveals additional correlations – again all SNP to SNP – until the residual is reduced to 0.05 (Fig. 2). Here, we see our first significant correlation (with $P = 0.05$) between expression levels of a gene and an SNP – specifically, CDH1 and rs35255374. Dialing the residual down to 0.025 reveals three additional gene to SNP relationships: CHD1 to 16:67369626; PCDH1 to 16:67374748; and CDH22 to rs35255374 (Fig. 3).

At a residual of 0.015, many additional gene–SNP correlations are revealed (Fig. 4). To simplify the visualization, all SNP–SNP edges were removed programmatically so that only correlations of interest (gene–SNP) remain. Of significant interest is the discovery of several cases where the minor SNP allele is correlated to a change in expression (Fig. 4, panel B), and a case where two SNPs affect different genes depending on which allele is present (Fig. 4, panel C).

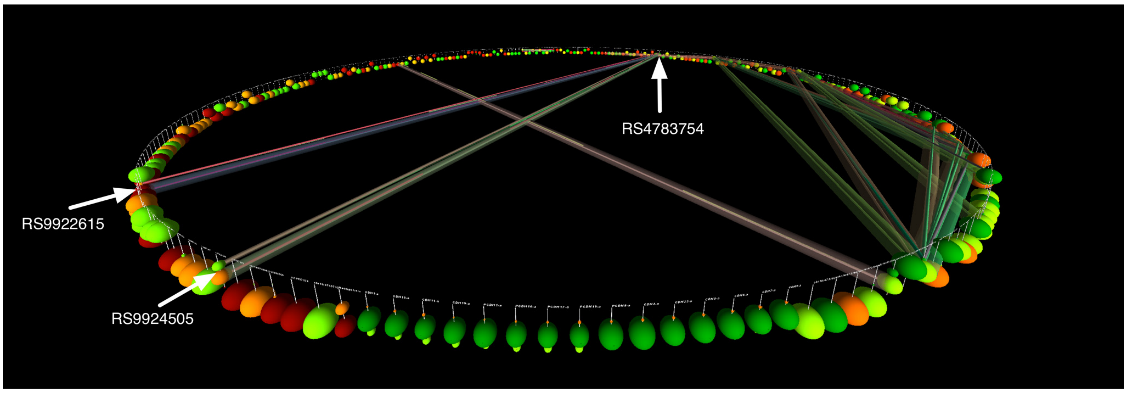

Figure 1. Initial view of dataset in StickWRLD using the default settings for p and residual (r). The fifteen columns in the foreground represent the genes in the dataset (the other columns represent SNPs); the green sphere most prevalent in each indicates that the state of "no change" in expression (up or down regulation) is the most common. At these settings the only correlations which can be seen are SNP to SNP relationships.

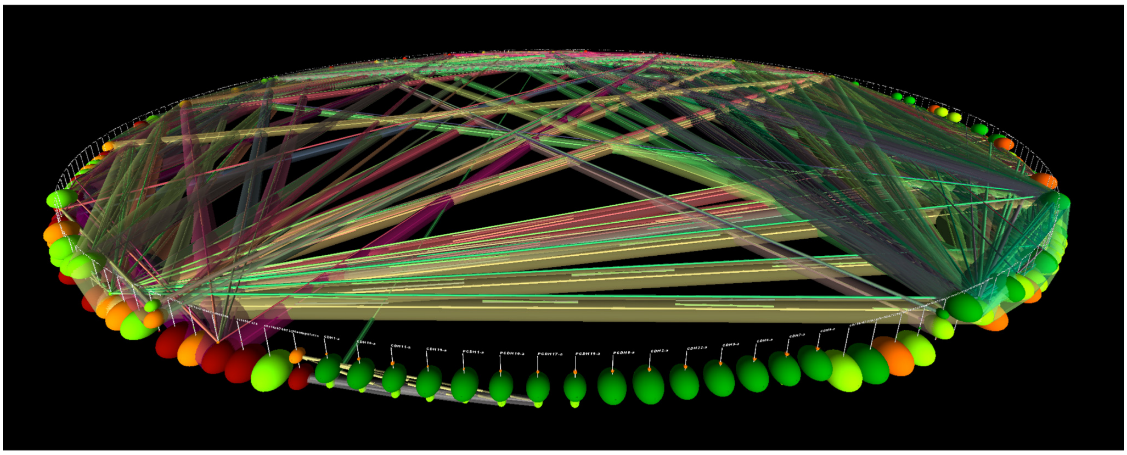

Figure 2. Tuning the residual to a lower value reveals numerous additional correlations, including one between a gene (CHD1) and a SNP (16:67369626). The correlation of interest (gene to SNP) is easily seen by it's sudden appearance when the residual is modified. In a more traditional analysis, this "signal" would easily be overwhelmed by the amount of "noise" – while the SNP to SNP relationship are of interest, they are not the primary concern in this analysis, and the ability to rapidly isolate the signal from the noise visually makes allowed us to quickly determine which relationships to focus on.

Pearson's correlation coefficient. Four of the gene-to-SNP relationships seen above were subjected to a Pearson's correlation test[16] to determine the strength of the correlation as measured through conventional means. As Table 1 shows, the correlation of CDH1 to rs35255374 showed a weak (0.2–0.29) correlation using the Pearson's test; the other three fall into the "negligible" category.

Relationships detected via StickWRLD are highlighted; others are displayed as illustrative of non-correlated results. The CDH1 to rs35255374 relationship shows a weak significance by conventional means of correlation, whereas the other three relationships fall into the negligible range.

The Pearson's correlation SNP-to-SNP relationships were also analyzed; rs4783754 had a negligible (<0.2) negative correlation to both rs9922615 and rs9924505. The correlation between chr16.67319752 and chr16.67381383 was very strong, at 0.599.

Discussion

StickWRLD as an expert-guided hypothesis engine. StickWRLD allows researchers to visually and interactively browse their data, dynamically displaying correlations based on both *P* value and residual (observed–expected) and allowing the user to explore how the relationships change with different settings and perspectives. As such, StickWRLD functions as a visual analytical tool, facilitating rapid hypothesis discovery. We were able to very rapidly explore all the potential relationships in this large dataset and were able to identify several correlations, which could by investigated more fully using conventional statistical approaches.

Another goal in analyzing this dataset was to examine the utility of StickWRLD in pre-filtering an eQTL dataset for downstream analysis – using StickWRLD to identify relationships of interest that could be used to train a neural net, for example, or to perform more mundane and conventional statistics such as the Pearson's correlations performed

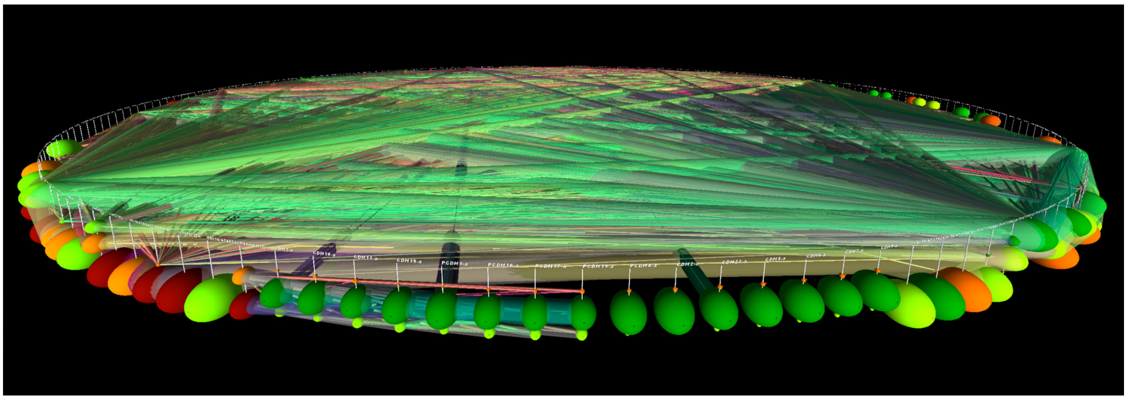

Figure 3. Additional relationships are revealed by further reducing the residual; note that the *P* value remained significant at 0.05 throughout the analysis. Several strong correlations between genes and SNPs can be seen as the bold dark connectors leading from the genes in the foreground to their corresponding SNPs in the background.

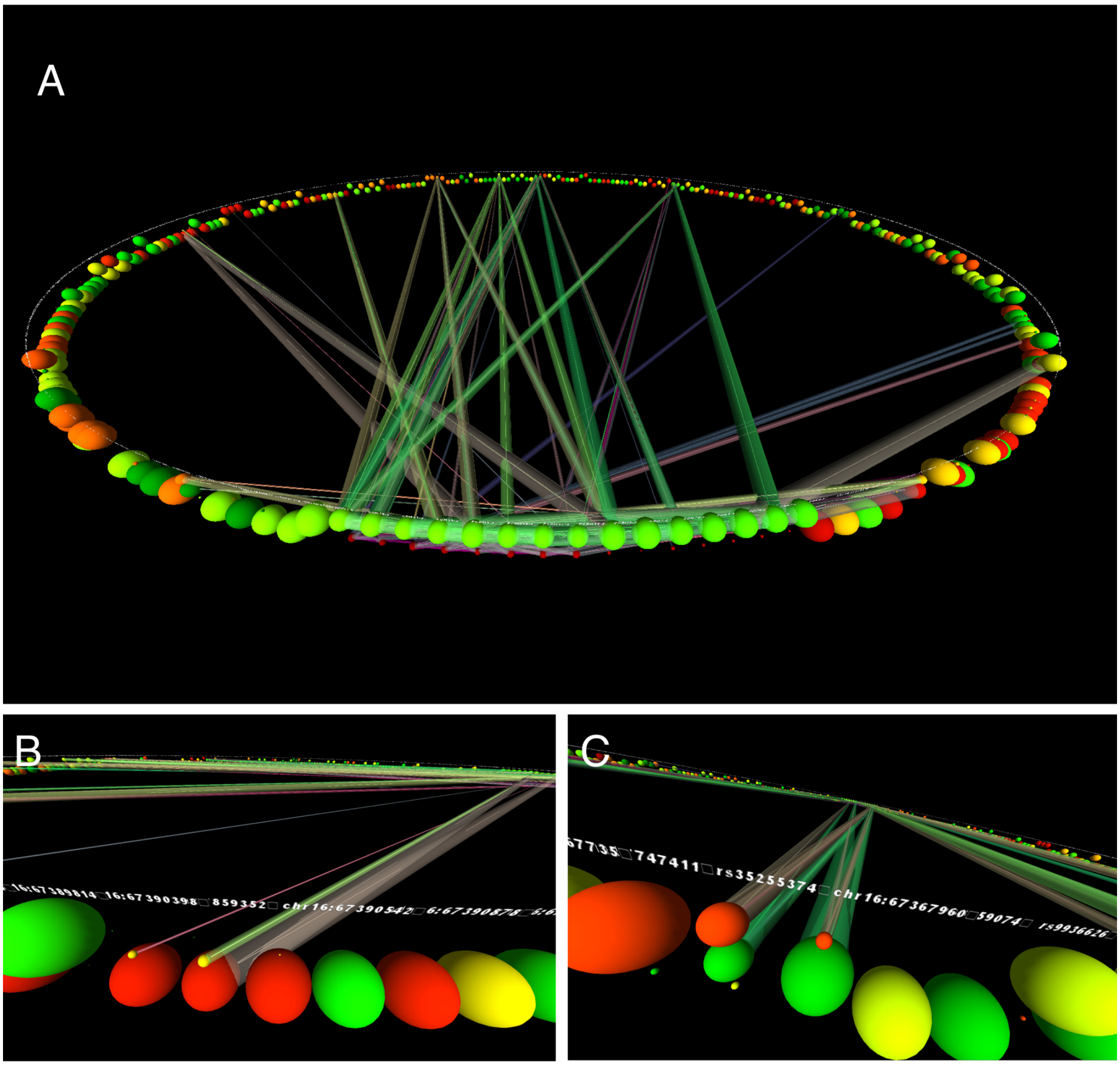
A
B
16:67389814 16:67390398 859352 chr16:67390542 6:67390878 6:6
C
677135 747411 rs35255374 chr16:67367960 59074 rs9936626

Figure 4. Reducing the residual further and eliminating edges which were not of interest revealed additional gene-SNP relationships of interest (**A**). Notably, there are several cases where the minor SNP allele is correlated to a change in expression (**B**), and one (**C**) where specific alleles of two SNPs differentially effect the expression of multiple genes. These effects were not seen at higher values for the residual due to low penetrance.

above. The number of possible two-node relationships in this dataset is extremely large; by identifying the critical relationships, we can greatly simplify the downstream analysis both in terms of computational power and human analysis. In the process of tuning residual to the level required to discover gene to SNP relationships – the correlations of interest as identified by a domain expert (eg, an individual with sufficient knowledge within a field to be able to search for and identify expected and/or meaningful relationships) – we inadvertently discovered other relationships with at least the same level of significance. Having a known relationship to look for allowed us to tune the residual until that relationship was discovered; any other relationships that met the same criteria of significance were automatically "brought along for the ride" as well. Using the resulting group of relationships to train a neural net, for example, would not simply have trained the network to predict gene–SNP relationships – it would also have been able to predict any other dependent relationships. In this case, the use of domain knowledge to define the stop, or threshold, retains significance that otherwise might be overlooked – but allows the user to discard information that is not *at least* as significant as the relationship used as the threshold.

By allowing the user to arbitrarily modify the visualization based on significance, StickWRLD takes advantage of the human ability for pattern recognition. Beginning at a low residual setting, StickWRLD can display all possible relationships

Table 1. Pearson's correlation coefficients for the gene–SNP pairs detected by StickWRLD analysis.

	rs35255374	16.67369626	16.67374748
CDH1	0.276	0.122	−0.001
PCDH1	−0.003	0.012	0.019
CDH22	−0.016	−0.012	0.017

between variables (significant or not). By gradually increasing the residual setting, StickWRLD filters out less-significant relationships, until a pattern emerges – essentially allowing the user to identify the signal from the noise visually. Likewise, an exploration could begin with a low residual threshold, a (numerically) high significance (*P* value) threshold, and iteratively increase the stringency of the statistical threshold while examining the consequences to the retained and rejected sets. When combined with domain expertise, which allows the user to validate that *known* significant relationships are preserved (which may happen at lower than optimal *P* or residual values), the user is guided to discover potential relationships of at least equal significance to already established relationships.

Coinheritance of SNPs. The initial StickWRLD display of the dataset revealed a small number of SNP–SNP correlations that were significant at the default values for *P* and *r*. The strongest correlation (0.599) was found between chr16.67319752 and chr16.67381383. This is unsurprising, given that these two SNPs are approximately 61 kb apart from one another. Similarly, rs4783754, rs9922615, and rs9924505 are all within 41 kb of one another on chromosome 16.

The coinheritance seen between these SNPs reflects the degree of linkage disequilibrium present due to their physical proximity on the chromosome. This in fact validates the use of StickWRLD to detect significant relationships – StickWRLD very clearly displayed the relationship between the SNPs, whereas a more conventional statistical approach (eg, Pearson's correlation coefficient) may have discounted many of them. This gains additional relevance when we consider the nature of the data. Biological data have interdependencies forced upon it by evolutionary and functional requirements, and as such cannot be treated as a traditional dataset where all data points are assumed to be independent. Further, not all data points in a genetic dataset are accessible to a statistical analysis. Many will be removed by natural selection – selected against because the nature of specific combinations will be deleterious and hence removed from view. Thus the canonical "normal" distribution may be skewed when we examine genetic data, which inevitably is subject to linkage to some degree, making detection of correlation require more sensitive techniques.

Gene–SNP correlations. While the *P* value threshold remained set at 0.05 to ensure that only significant relationships were displayed, the detection of any gene–SNP correlations in the dataset required a significantly reduced value for *r*, the residual of (*observed–expected*). This is a result of the low penetrance of the genotypic to phenotype effects – there are typically other factors affecting gene expression that may reduce the impact of the eQTL on the phenotype – and only with 100% penetrance would the effect of the eQTL be 1:1. Assuming 100% penetrance, a relationship with a *P* of 0.05 and a correlation value of 1.0 would be detectable at a residual of 0.25. This is due to the fact that, given a binary state (eg, two possible allelic states for the SNP), at best 50% of one allele could be correlated to one condition, with the remaining 50% correlated to the other condition. As these distribution frequencies drop out of balance, the residual correspondingly drops as a function of them. Factoring penetrance in, if we assume a 10% penetrance, the maximum possible residual drops to 0.025 – much smaller, in fact, than the threshold used to detect the initial set of gene–SNP relationships in the above dataset. Once the residual was tuned down even further – to 0.015 – many additional correlations (ostensibly representing relationships with an even lower penetrance) were seen.

StickWRLD's ability to simultaneously visualize all possible two-node (eg, "x vs y") relationships within a dataset makes it a powerful supplemental tool to use alongside traditional summary statistics. The correlations above were rapidly discovered in StickWRLD, but could have eluded detection using conventional statistical approaches, which can often fail to detect subtle and complex trends within datasets. This is a well-known statistical problem highlighted by Anscombe's Quartet,[17] where four very different datasets display identical summary statistics. Anscombe's Quartet points out the need for visualization approaches (eg, graphing) to reveal patterns and relationships within complex datasets. In many such datasets, summary statistics such as the Pearson's Correlation Coefficient are either inadequate to detect complex but meaningful effects or, because of their precision, require impractically many calculations to define features that may be detected approximately by visualization with relative ease. StickWRLD can be treated as an extension to traditional graphical approaches, allowing the user to simultaneously look at all combinations of "x vs y" comparisons for a dataset. While these comparisons *could* be done using conventional approaches like Pearson's, it would be impractically time-consuming to construct and evaluate all possible combinations of variables and their inter-value ranges. StickWRLD's graphical approach allows the user to perform all these evaluations simultaneously.

This failure of simple statistical thresholds to identify meaningful associations is typical for genetic features with small effects sizes – any fixed threshold risks ablating relationships that may be important. By using StickWRLD to visualize the strength of all of the relationships simultaneously, the user is able to make informed decisions regarding which thresholds to accept, and whether any patterns of statistically weak effects merit further investigation.

Given both the correlation coefficient and the residual required to display the correlation in StickWRLD, an estimate as to the penetrance of that relationship in the dataset's population can be inferred. Again, a traditional statistical analysis may have discounted these relationships; StickWRLD enabled us to rapidly discover relationships of interest and at the same time retain all other statistically important (in terms of both *P* and the residual *r*) relationships.

Conclusions

StickWRLD allowed us to rapidly browse through our dataset looking for correlations of interest. By modifying the retention criteria (eg, *P* and residual *r*), we were able to use domain expertise to quickly identify relationships of interest for further analysis. In doing so, other relationships of at least the same level of significance were retained as well. Allowing users to quickly browse data looking for correlations of interest allows the user to identify relationships for deeper analysis as well as to discover novel unexpected correlations, many of which would have been undetectable using conventional statistical approaches, due to additional factors (eg, penetrance). Lastly, by reducing the complexity of the dataset in this fashion, StickWRLD can act as a pre-filter for neural network or conditional random field training, reducing the need for semi-guided training.

The StickWRLD python executable code as well as the StickWRLD manual and the dataset presented here are available for download from http://www.stickwrld.org/rwr-cri-2014/. Instructions for preparing StickWRLD-formatted datasets, which are essentially any variety of tabular categorical, or bin-able continuous-valued data, can be found in the StickWRLD manual.

Author Contributions

Conceived the concepts: WCR, SLW. Analyzed the data: RWR. Wrote the first draft of the manuscript: RWR. Contributed to the writing of the manuscript: WCR. Agree with manuscript results and conclusions: RWR, SLW, WCR. Jointly developed the structure and arguments for the paper: RWR, WCR. Made critical revisions and approved final version: RWR, WCR. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Sharma SV, Haber DA, Settleman J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nat Rev Cancer*. 2010;10:241–53.
2. Grever MR, Schepartz SA, Chabner BA. The National Cancer Institute: cancer drug discovery and development program. *Semin Oncol*. 1992;19:622–38.
3. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483:603–7.
4. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012;483:570–5.
5. Barret SJ, Langdon WB. Advances in the application of machine learning techniques in drug discovery, design and development. In: Tiwari A, Knowles J, Avineri E, Dahal K, Roy R, eds. *Applications of Soft Computing: Recent Trends*. Springer Berlin Heidelberg; 2006:99–110.
6. Wale N. Machine learning in drug discovery and development. *Drug Dev Res*. 2011;72:112–9.
7. Beerenwinkel N, Däumer M, Oette M, et al. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res*. 2003;31(13):3850–5.
8. Ruderfer DM, Roberts DC, Schreiber SL, Perlstein EO, Kruglyak L. Using expression and genotype to predict drug response in yeast. *PLoS One*. 2009;4(9):e6907.
9. Menden MP, Iorio F, Garnett M, et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One*. 2013;8:4.
10. Kelloff GJ, Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. *Nat Rev Drug Discov*. 2012;11:201–14.
11. Ray WC. StickWRLD: interactive visualization of massive parallel contingency data for personalized analysis to facilitate precision medicine. In Proceeding of 2013 AMIA Workshop on Visual Analytics in Health Care, Washington, DC, Volume; Proceedings of the 2013 Annual Symposium of the American Medical Informatics Association 2013.
12. Ray WC. MAVL and StickWRLD: visually exploring relationships in nucleic acid sequence. *Nucleic Acids Res*. 2004;32(Web Server issue):W5–63.
13. Bartlett CW, Cheong SY, Hou L, et al. An eQTL biological visualization challenge and approaches from the visualization community. *BMC Bioinformatics*. 2012;13(suppl 8):S8.
14. Myers AJ, Gibbs JA, Webster K, et al. A survey of genetic human cortical gene expression. *Nat Genet*. 2007;39(12):1494–9.
15. Liu C, Cheng L, Badner JA, et al. Whole-genome association mapping of gene expression in the human prefrontal cortex. *Mol Psychiatry*. 2010;15(8):779–84.
16. The R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing; 2014.
17. Anscombe FJ. Graphs in statistical analysis. *Am Stat*. 1973;27(1):17–21.