VISUAL INTEGRATION OF RESULTS FROM A LARGE DNA BIOBANK (BIOVU) USING SYNTHESIS-VIEW *

SARAH PENDERGRASS
Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University
507D Light Hall, Nashville, TN 37205, USA
Email: sarah.a.pendergrass@vanderbilt.edu

SCOTT M. DUDEK
Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University
509 Light Hall, Nashville, TN 37205, USA
Email: dudek@chgr.mc.vanderbilt.edu

DAN M. RODEN
Department of Medicine, Department of Pharmacology, Office of Personalized Medicine, Vanderbilt University
536 Robertson Research Building, Nashville, TN 37205, USA
Email: dan.roden@vanderbilt.edu

DANA C. CRAWFORD
Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University
505 Light Hall, Nashville, TN 37205, USA
Email: dana.crawford@chgr.mc.vanderbilt.edu

MARYLYN D. RITCHIE
Center for Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University
509 Light Hall, Nashville, TN 37205, USA
Email: ritchie@chgr.mc.vanderbilt.edu

In this paper, we describe using Synthesis-View, a new method of presenting complex genetic data, to revisit results of a study from the BioVU Vanderbilt DNA databank. BioVU is a biorepository of DNA samples coupled with de-identified electronic medical records (EMR). In the Ritchie et al. study [1] ~10,000 BioVU samples were genotyped for 21 SNPs that were previously associated with 5 diseases: atrial fibrillation, Crohn Disease, multiple sclerosis, rheumatoid arthritis, and type 2 diabetes. In the proof-of-concept study, the 21 tests of association replicated previous findings where sample size provided adequate power. The majority of the BioVU results were originally presented in tabular form. Herein we have revisited the results of this study using Synthesis-View. The Synthesis-View software tool visually synthesizes the results of complex, multi-layered studies that aim to characterize associations between small numbers of single-nucleotide polymorphisms (SNPs) and diseases and/or phenotypes, such as the results of replication and meta-analysis studies. Using Synthesis-View with the data of the Ritchie et al. study and presenting these data in this integrated visual format demonstrates new ways to investigate and interpret these kinds of data. Synthesis-View is freely available for non-commercial research institutions, for full details see https://chgr.mc.vanderbilt.edu/synthesisview.

* This work was supported in part by LM010040 (SAP, SMD, MDR), HG004798 (SAP, DCC, MDR), HL065962 (MDR, DCC)

1. Introduction

The use of results from genome-wide association studies (GWAS) in the emerging field of personal genomics requires the further investigation and characterization of potentially functional single nucleotide polymorphisms (SNPs) originally identified in GWAS. The additional studies required usually characterize less than 100 SNPs, often include multiple and correlated phenotypic measurements, and can include data from multiple-sites, multiple-studies, as well as multiple race/ethnicities. The Vanderbilt University biobank (BioVU)[2] aims to both characterize previously detected SNPs, as well as discover new associations between genetic variation and diseases and phenotypes. BioVU has an "opt-out" system, whereby DNA samples are collected from blood remaining after routine clinical testing at Vanderbilt Medical Center. De-identified electronic-medical record (EMR) data, called the "synthetic-derivative" (SD) is coupled to DNA of the biorepository. Cases and controls for phenotype-genotype association are identified using the synthetic-derivative through the use of electronic phenotyping algorithms developed in by EMR content experts along with biomedical informaticists.

In the Ritchie et al. study [1], the first approximately 10,000 DNA samples collected in BioVU were genotyped for a series of SNPs that each had a previously known and robust association with of one of five common diseases. The goal of this proof-of-concept study was to demonstrate that EMR data can successfully be used to accurately define phenotypes that enable the investigation of genotype-phenotype correlations. In this study the electronic phenotyping algorithms were deployed in the SD to determine cases and controls for atrial fibrillation, Crohn disease, multiple sclerosis, rheumatoid arthritis, and type 2 diabetes in a sample of largely European American descent. A total of 9483 DNA samples were successfully genotyped, and 21 tests of association were performed. Significant associations ($p < 0.05$) were found for 8/14 tests where SNPs had a previously reported odds ratio (OR_{PR}) > 1.25, and 0/7 where SNPs had a lower OR_{PR}. In the initial presentation of the results of this study, the majority of the results were provided in a tabular form. While tabular data provides a record of the exact results of a study, it can be challenging to identify and convey the trends and patterns within a set of results using a tabular data alone.

Visualizing data results such as those of the BioVU study as well as other candidate-gene replication studies that move beyond initial GWAS findings, provides a way to interpret the complex and multi-layered results of these studies in a more integrated way, and allows for rapid comparisons of multiple forms of information not easily achievable through reviewing large tables of numbers. To visualize the results of these forms of studies, we developed the software tool "Synthesis-View" to visually synthesize the results of candidate gene and GWAS replication studies in stacked data-tracks, providing a single image where p-values (or other measures of significance), odds-ratios, allele frequencies, sample sizes, effect size, and direction of effect are all incorporated. While Manhattan plots already exist for the effective visualization of GWAS data, the results of candidate gene studies, studies investigating genetic variation in specific regions in detail, or even isolated GWAS results, are not often presented in visual form. Our tool

provides a unique and direct way to generate accessible visual information from these kinds of data.

2. Methods

The Synthesis-View software tool used herein was developed in Ruby and utilizes the RMagick graphics library. Synthesis-View is available for use through a web interface, and can alternately be used at the command line. Figure 1 shows a screen-capture of the web interface, which allows for the flexible choice of various options for Synthesis-View plots. The required and optional tab-delimited text input file format to produce a Synthesis-View plot are briefly described here, and are also described in greater detail at the Synthesis-View website along with example input files. One file is necessary to produce a standard Synthesis-View plot, a file containing a column for

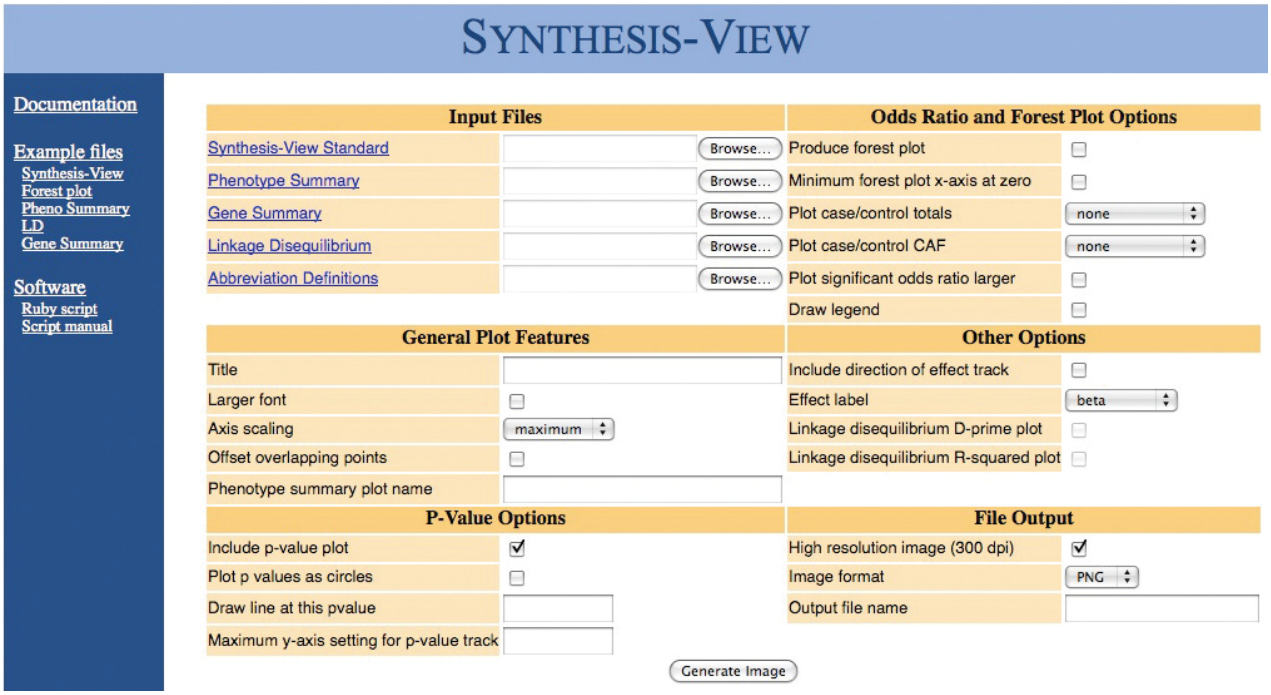

Figure 1 – Synthesis-View web interface screen capture.

SNP identification (such as RS number), a column for which chromosomes the SNPs map to, and a column for SNP genomic location information. The rest of the standard input file can optionally contain information on p-values, odds-ratios, allele frequencies, and sample size, with tracks plotted if data are present. Other files can be provided for Synthesis-View to plot additional tracks of data. If a phenotype summary file is supplied, summary information about continuous phenotypes will be plotted. If a gene summary file is included, information on gene name and location in relation to SNPs plotted will be in a track at the top of the plot. If a linkage disequilibrium file is provided that contains D′ or r^2 correlation data, the data will be plotted in Haploview style format [3]. Finally, if abbreviation definitions are provided, an additional legend

describing plot abbreviations will appear below OR/forest plots when "Draw Legend" is selected. Table 1 describes the various possible settings available in the web interface.

3. Visualization of Results

The focus of the proof-of-concept BioVU study was to both show and characterize the utility of using electronic phenotype algorithms deployed in an EMR linked to a DNA biobank. As described in Ritchie et al. [1], blood samples that showed poor-quality or that yielded insufficient DNA, blood samples from individuals < 18 years of age, a lack of consent-to-treatment form, any indication of opt-out, or discovery of a duplicate sample, resulted in exclusion from the study. In addition, 2% of samples in BioVU are randomly dropped out, further randomizing individuals not included in the biobank and consequent studies. After filtering for exclusions, definite cases of European Ancestry (EA) and probable EA were defined using the administrative information recorded in the EMR. Almost a tenth of the records (9.2%) did not include ancestry information, or recorded the ancestry as "unknown". The data were thus analyzed with cases and controls that indicated EA specifically as the race/ethnicity, and also separately analyzed with cases and controls defined as both EA and individuals characterized as unknown.

To define disease state for case/control status, for one set of association tests, identification of case/control status was solely determined using an electronic phenotyping algorithm (see Ritchie et al. appendix for algorithm details). Content experts were used to develop the algorithm that used disease-specific billing codes and patient encounter information, including records such as medication information, electrocardiogram data, and past medical history from the SD. "Definite" cases were defined by the algorithm as disease present, excluding those with indications of overlapping disease or symptoms, or lack of a clear diagnosis. Controls were defined as those with clear absence of the specific disease used in the case/control association. In the case of multiple sclerosis, algorithm classified cases were also manually reviewed because of the small sample size. In addition to the algorithm defined Definite cases, for rheumatoid arthritis and multiple sclerosis, a set of association tests were separately performed with both Definite cases as well as cases showing indications of overlapping autoimmune diseases and/or symptoms. These cases were described as "Probable".

After defining cases/controls, association tests for the 21 genotyped SNPs were performed. For SNPs associated with atrial fibrillation, Crohn's disease, or Type 2 diabetes, tests of association were performed for both EA with cases Definite cases and EA + Unknown with Definite cases. For SNPs known to be associated with rheumatoid arthritis and multiple sclerosis, tests of association were performed for EA with Definite cases, EA with Definite and Probable cases, EA + Unknown with Definite cases, EA + Unknown with Definite and Probable cases.

3.1 *Synthesis-View Forest Plot*

The results of the association tests of the BioVU study were presented in Table 1 of the Ritchie et al. manuscript [1]. The results for EA alone with Definite cases were presented in a forest plot along with OR_{PR} from previous studies in Figure 1 of the Ritchie et al. manuscript [1,4-10]. In the current paper, Figure 2 is a modified forest plot using Synthesis-View to visualize the results of the BioVU study. From left to right in Figure 2 are tracks with various pieces of data:

1. The first track is a *physical genome track*, displaying the chromosome and relative location of each SNP used in the 21 association tests. Having the SNP data presented in this way visually shows the location of SNPs in reference to other SNPs within the same study. Lines lead from the relative location of each SNP to the SNP identifier.
2. The next track is the *significance track*, showing the p-values of both the original OR_{PR} as well as the results of the Ritchie et al. paper. A single color consistently represents results for the original OR_{PR} (in blue), as well as for the new associations: EA_D (European American, Definite disease classification, in red); EA+U (European American and Unknown, Definite disease, in orange); EA_P (European American, Definite as well as

Table 1. Synthesis-View plotting options

Synthesis-View Option	Description
Title	Title for Synthesis-View plot
Larger font	Produce a plot with larger sized text than the default
Axis scaling	If set to "maximum", axes limits will start and end utilizing the range of the data with tick-marks at regular intervals in-between. If set to "cleaner" the axes will still encompass the range of the data, however the range will begin and end with a multiple of five or ten, and the plot tick-marks will also be a multiple of five or ten.
Offset overlapping points	When points overlap, this setting will include "jitter", whereby overlapping points are offset horizontally to make them more distinguishable.
Phenotype summary plot name	If phenotypic summary data will be incorporated into the Synthesis-View plot, the title for the phenotype summary plot should be specified here.
Include p-value plot	Include plot of p-values
Plot p-values as circles	To plot p-values as circles, instead of triangles that include direction of effect, even if direction of effect information is supplied in the Synthesis-View standard input file.
Draw line at this p-value	Specification of a horizontal red line at a specific p-value of interest.
Maximum y-axis setting for p-value track	Specify the maximum y-axis value for the p-value track in order to limit the range of the y-axis. Any p-value result more significant than this y-axis cutoff value will be plotted at the cutoff value in larger size.
Produce forest plot	To produce a forest plot in Synthesis-View from odds-ratio results
Minimum forest plot x-axis at zero	To set the minimum value of the forest plot x-axis to zero
Plot case/control totals	The total numbers of cases/controls can be plotted either in two separate tracks ("split plot"), or in one track where the total numbers of cases/controls are indicated using open/closed circles ("combined plot").
Plot case/control CAF	The respective coded allele frequency (CAF) for cases/controls can be plotted either as two separate tracks ("split plot"), or in one track where cases/controls are indicated using open/closed circles ("combined plot").
Plot significant odds ratio larger	Plot significant odds-ratio results in larger size
Draw Legend	When an "Abbreviation Defintions" file is provided, and Draw Legend is selected, an additional legend describing plot abbreviations will appear below OR/forest plots
Include direction of effect track	Even if direction of effect information is supplied, this setting allows for inclusion/exclusion of a direction of effect track.
Effect label	Choice of effect size label
Linkage disequilibrium D-prime plot	If linkage disequilibrium information is included as an input file, select this to include a d-prime correlation track.
Linkage disequilibrium R-squared plot	If linkage disequilibrium information is included as an input file, select this to include an R-squared correlation track.
High resolution image (300 dpi)	Select to produce a 300 dpi image, otherwise the image is 72 dpi
Image format	Choices of image format include PNG, JPEG, and TIFF
Output file name	Choice of file name for output Synthesis-View plot

Probable disease, in purple); and EA+U_P (European American and Unknown, Definite as well as Probable disease, in green). Applying a red line at a p-value cutoff of choice is one of the options of Synthesis-View, in this case the vertical red line was applied at a p-value of 0.05, allowing for a more quick detection of values above and below the chosen p-value. For two SNPs, rs6457620 and rs3135388, in studies prior to Ritchie et al. the results were extremely significant at 4E-186[10] and 9E-81[6] respectively. When these two SNPs were originally plotted on the same track as the rest of the p-values, there was compression of other p-values along the bottom of the plot due to the wide spread of the data points. Synthesis-View allows for the choice of a p-value cutoff, whereby any points more significant than that cutoff are plotted at that cutoff value with a larger sized point. Thus, on this plot, after choosing a p-value cutoff of 1E-50, the two points for SNPs, rs6457620 and rs3135388 are plotted at p-value 1E-50 but are larger in size. Also of note, the various BioVU p-value results for each SNP were very similar, thus when initially plotted, the points had considerable overlap. Synthesis-View allows the application of "jitter", where points that overlap are spread out vertically along the "abacus" line leading down from the SNP identification information. Thus the jitter option was applied, providing more visual discrimination between multiple overlapping points.

3. The next four tracks are *odds-ratio/forest-plot tracks*. Each track shows the individual odds ratio (OR) and confidence intervals for each of the separate sets of associations, such as those for EA_D or EA+U. Each OR result is plotted as a square, with a line indicating the upper and lower 95% confidence interval. In this case a specific option in Synthesis-View was used, whereby if the result is significant (the upper or lower boundary of the confidence intervals do not cross 1.0), the square is plotted in larger size. This allows for quick visual identification of significant results in forest plots that may show many results. In the case of the results for the previous studies, the confidence intervals were small enough they were overplotted by the OR square. As the eye moves from left to right, there are visible trends. Results that were not significant in the BioVU study were in the same direction as OR_{PR}. Also, it is easy to determine how similar the results were in the BioVU study, even with inclusion or exclusion of data from Unknown individuals and Probable case data. With Synthesis-View both an overview of the data as well as individual results are available, and a table can be used to look up exact numerical results of interest.
4. The second to last track is the *coded allele frequency track*. Synthesis-View provides the option of either the coded allele frequencies (CAF) of both cases and controls plotted on the same track, with closed circles indicating cases and open circles indicating controls, or the allele frequencies of cases and controls can be plotted in two separate tracks. In either case, colors match those of the groups of the previous tracks, allowing the user to look at the allele frequencies between groups by eye for trends. This can aid in interpreting the potential lack of replication of results.

5. The last track is the *sample size track*. Like the CAF track, case/control sample size can either be plotted with closed circles indicating cases, and open circles indicating controls. The colors match those of the groups of the previous tracks, allowing the user again by eye to look at sample size across groups.

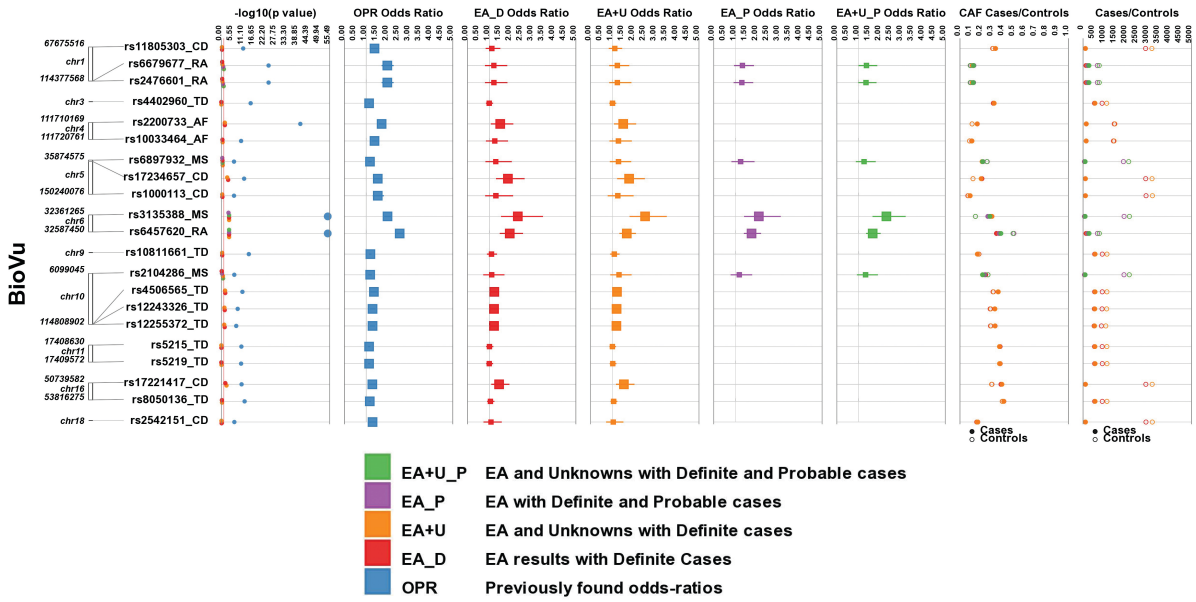

Figure 2 - The results of using the forest-plot option in Synthesis-View and the data of the Ritchie et al. BioVU paper. Moving from left to right, the first track shows SNP location, the next track shows $-\log_{10}$(p-value). Each SNP identifier also has the following abbreviations for associated disease: atrial fibrillation (AF), Crohn disease (CD), multiple sclerosis (MS), rheumatoid arthritis (RA), and type 2 diabetes (TD). The next five tracks are odds-ratio/forest-plots. The abbreviations for these tracks are described in greater detail in the figure legend. The coded allele frequency (CAF) track with allele frequencies for both cases/controls is the second-to-last track. Sample sizes for the cases/controls are plotted in the last track.

3.2 *Synthesis-View Standard Plot*

An alternative way to look at the results of the BioVU study is through stacked tracks where the eye moves from top to bottom (Figure 3). If the "forest-plot" option is not chosen in Synthesis-View, the default data plot is in this format. Again the first track is the *physical genome track*, with chromosome number and the relative location of each SNP with lines leading from the chromosome location track to identification of each of the respective SNPs. The next track is the *significance track*, showing p-value results across groups with an optional horizontal red line at a p-value of 0.05 applied. In this case, again to reduce compression of the p-value results when plotted, a p-value cutoff was chosen (1E-30), with larger points plotted directly at the p-value cutoff. SNPs rs6457620 and rs3135388 and rs2200733 had p-values of 4E-186[10], 9E-81[6], and 3.3E-41[4] respectively. The track below the significance track is an *odds-ratio track*. Unlike the forest-plots of Figure 2, here the ORs are plotted as closed circles. If the OR results are significant, the OR closed circle is plotted in a larger size. So while the confidence intervals are not plotted, it

is still easy to discriminate OR results that are significant. For studies where OR data are omitted, the OR track will not appear. Below the OR track, there is a CAF track. Again, Synthesis-View provides the option of either viewing the allele frequencies of both cases and controls plotted on the same track, with closed circles indicating cases, and open circles indicating controls. The last track is a sample size track plotted in a similar fashion as the CAF track.

There are available Synthesis-View options that were not used in this presentation of the BioVU results. When summary data regarding a continuous phenotype of interest exists, there is an option to add on a summary data plot, which consists of the mean and standard deviation of the continuous phenotype for each group. Future versions of Synthesis-View will incorporate ways to characterize categorical/case-control phenotype summary data. Also, when linkage disequilibrium (LD) data is provided, a D′ or r^2 correlation plot in Haploview style format [3] is plotted.

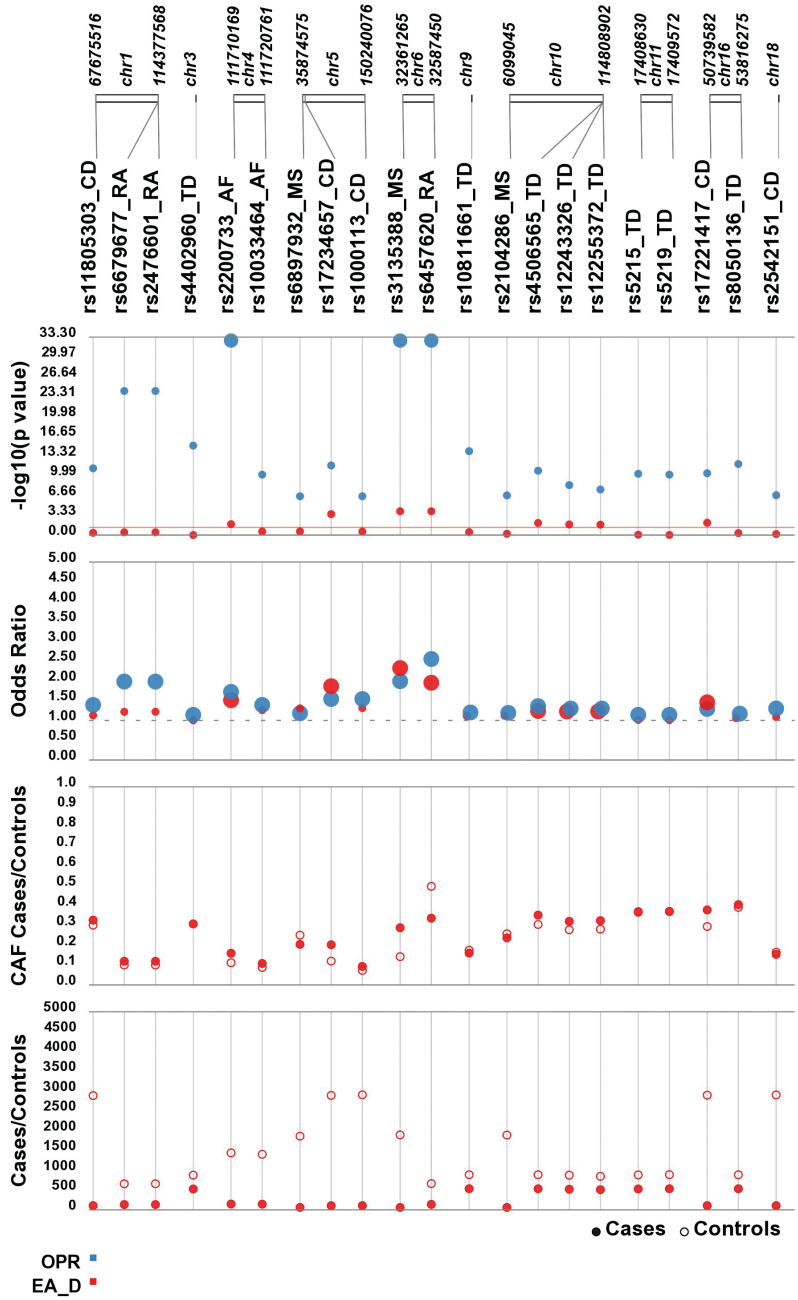

Figure 3 - Default format of the Synthesis-View plot with horizontal data tracks. In red are the results of OR_{PR} (OPR), in blue are results of the BioVU Ritchie et al. study. From top to bottom, data tracks include the physical genome track, odds-ratio track (significant odds-ratio results are plotted larger in size), coded-allele-frequency (CAF) track for cases and controls, and a sample-size track for cases and controls.

4. Conclusions

Synthesis-View was extended from the previous software "LD-Plus". The LD-Plus feature carried through to Synthesis-View is the use of multiple tracks for showing data results, as LD-Plus also

uses a flexible data display format of multiple data "tracks" that can be viewed [11]. However, Synthesis-View allows for visualization of data that is not possible with LD-Plus. In Synthesis-View, through the use of stacked data-tracks, SNP genomic location, presence of the SNP in a specific study or analysis, as well as related data such as genetic effect size and summary phenotype data, are plotted according to user preference. With Synthesis View, trends from many different kinds of information can be visualized in a more integrated way than by using tabular data alone. These multi-faceted views are important to understanding in greater depth the relationships between SNPs, strata, sample size, and phenotypic differences expected with the increasing complexity of emerging datasets.

It is important to note here that we present one set of scenarios where Synthesis-View can be used; however, the software is very flexible and that there are no restrictions to how the data are grouped. The Ritchie et al. paper was able to show proof-of-concept, such that the use of a biobank coupled with EMR data can effectively replicate previously well characterized results. The original results of this paper were largely presented in tabular format, and here we show the utility of Synthesis-View in visualizing these kinds of results. Through using Synthesis-View the larger picture of the data as a whole can be seen, with trends and patterns visually evident, while also allowing a user to determine details about individual results. Tables can then be used as a reference for determining specific numerical results in greater detail after areas of interest are located in the plotted data.

5. Acknowledgments

We would like to acknowledge the following individuals for their suggestions and ideas in designing Synthesis-View: Matthew Thomas Oetjens, Fredrick Schumacher, Janina Jeff, Logan Dumitrescu, and Chris Haiman. This work was supported in part by LM010040 (SAP, MDR), HG004798 (SAP, DCC, MDR), and HL065962 (MDR, DCC).

6. References

1. Ritchie, M.D., *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 86, 560-572 (2010).
2. Roden, D.M., *et al.* Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 84, 362-369 (2008).
3. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263-265 (2005).
4. Gudbjartsson, D.F., *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 448, 353-357 (2007).
5. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-678 (2007).
6. Hafler, D.A., *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med* 357, 851-862 (2007).
7. Groves, C.J., *et al.* Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* 55, 2640-2644 (2006).

8. Zeggini, E.*, et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316, 1336-1341 (2007).
9. Saxena, R.*, et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316, 1331-1336 (2007).
10. Raychaudhuri, S.*, et al.* Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet* 40, 1216-1223 (2008).
11. Bush, W.S., Dudek, S.M. & Ritchie, M.D. Visualizing SNP statistics in the context of linkage disequilibrium using LD-Plus. *Bioinformatics* 26, 578-579 (2010).