# Capabilities and analytical challenges in utilizing mega genomics cohort in the Veterans Healthcare System

Kelly Cho, PhD, MPH,[1,2], David Gagnon, MD, PhD, MPH,[2,3], Hongsheng Wu, PhD,[1,4], J. Michael Gaziano, MD, MPH,[1,2], for The MVP Investigators Team[5,6]

[1]Massachusetts Veterans Epidemiology Research and Information Center; VA Boston Healthcare System, Boston, MA

[2]Division of Aging, Department of Medicine, Brigham and Women's Hospital; Harvard Medical School, Boston, MA

[3]Boston University School of Public Health, Boston, MA

[4]Computer Science and Networking, Wentworth Institute of Technology; all in Boston, MA, USA, Boston, MA

[5]Office of Research & Development, Department of Veterans Affairs, Washington, DC;

[6]VA Connecticut Healthcare System, West Haven, CT;

## Abstract

VA launched the Million Veteran Program, a nationwide genomics resource, which has over 95,000 Veterans enrolled since 2011. This provides a promising opportunity to investigate the connection between VA's longitudinal EMR and genomics data. Our understanding will highly depend on the analytical approaches used to analyze mega genomic resources. Current rapid advancement in tools to collect and extract information from genomics data, such as in GWAS, microarray or proteomics and sequence data, highlights the importance in high dimensional data analysis, including variable selection, multiple testing issues, handling, storage, and computational efficiency. Traditional statistical procedures present eminent challenges in using these data, where the number of parameters $p$ is scalably larger than number of observations $n$. In addition, mega genomics data present a complex relational data structure when interactions and dynamic underlying biological complexities are considered, resulting in ultra-high dimensionality. Further research in statistical accuracy and inference, model interpretability and fitting and computational efficiency and robustness will play a critical role.

**Key Words:** Million Veteran Program, electronic medical record, Veterans Affairs, genomics, high dimensional data

## 1. Introduction

Biorepositories of varying sizes with electronic medical or health record (EMR) systems have enabled the possibility and feasibility of using EMR systems to gain further understanding of gene-disease relationships. Understanding the underlying genetic profile of the Veterans will enable clinical practices to get a step closer toward personalized medicine. In this effort, the U.S. Department of Veterans Affairs (VA) launched the Million Veteran Program (MVP) to create a nationwide genomics resource which has over 95,000 Veterans enrolled in the first year since its inception in early 2011. This provides a promising opportunity for a much awaited investigation of the connection between VA's wealth of longitudinal EMR data and the genomics data. This is timely as it allows for the possibility of "PheWAS" (phenotype-wide association study) of disease-gene associations, the flipside of Genome-Wide Association Scan studies, to determine, for a given genotype, the range of associated clinical phenotypes. The PheWAS approach is a promising and an unbiased way to discover new genetic variants. In addition, the availability and utility of MVP genomic resources opens doors for a nation-wide contribution and collaboration with existing consortiums such as eMERGE[1]. This will provide the largest nationwide sample for validation of common phenotypes as well as become a venue to explore new disease areas and their phenotypic algorithms.

Alongside this promising prospect, our understanding and interpretations will highly depend on the analytical approaches we implement to extract information from mega genomic resources in general. Current rapid advancement in tools developed to collect and extract information from genomic data, such as in genomewide association studies, microarray or proteomics data and sequence data analyses, highlights the importance and challenges in (ultra) high dimensional data analysis, including variable selection, multiple testing issues, handling and storage, computational efficiency, etc. It has been known that many traditional statistical procedures present challenges in their implementation and application using high dimensional genomics, where the number of variables or parameters $p$ is scalably larger than number of observations $n$. In addition, current mega genomics data present a complex relational data structure when interactions and dynamic underlying biological complexities are considered, resulting in ultra-high dimensionality. Continuing research in the area, including increasing efficiency in statistical inference, model interpretability and fitting and making statistical procedure computationally efficient and robust, will play a critical role in translational medicine overall.

In this brief report, we focus on the overview of future capabilities of large genomic cohorts and foreseeable analytical challenges using high dimensional data and potential remedies as the statistical genetics and genomics research fields catch up with rapid advancement in biotechnology.

## 2. Linking large genomic cohorts with longitudinal EMR data

There are numerous genomics cohorts and biorepositories in various stages of development. Much effort has been invested in building genomics resources for further research in this area. Many groups have already established and are in the process of building such resource. Below is a summary list of large cohorts in Europe and North America.

**Table 1:** List of large genomic cohorts in Europe (A) and in North America (B)

| A) Europe | B) North America |
|---|---|
| •Icelandic Biobank and deCODE | •Vanderbilt University BioVU |
| •UK Biobank | •Canadian Consortium [Canada] |
| •Banco Nacional de ADN [Spain] | •dbGaP, NIH [US] |
| •GenomEUtwin | • National Children's Study [US] |
| •Finnish biobank | • Marshfield Clinic [US] |
| •Swedish biobank | •National Health and Nutrition Examinations Surveys [US] |
| •German biobank, KORA | •Kaiser Permanente Northern CA [US] |
| •UK DNA Banking Network & British biobank | •Howard University African Diaspora [US] |
| •Estonian biobank: | •Mayo Clinic |
| •Family-based collections [Nordic] | •ACS |
| •Generation Scotland | |
| •HUNT (cardiovascular)& Biohealth [Norway] | |
| •EPIC, European (cancer) | |
| •Danubian Biobank Consortium | |
| •GATiB Genome Austria Tissue Bank | |
| •Biobank Hungary | |

Many of these cohorts list in Table 1 also have accompanying electronic medical record data in one form or another. Having links between these two sources of data makes these genomic data sets a powerful tool.

### 2.1 Million Veteran Program

In addition to above major efforts throughout Europe and North America (Table 1), a recent initiative from the VA Genomic Medicine Program (GMP) has launched the Million Veteran Project [MVP] in 2011. The MVP is a major research initiative under the GMP that will create a longitudinal cohort of Veterans to study genes and health. VA is the ideal setting for a large 21st century mega-cohort/biobank with its national pool of willing participants, outstanding electronic medical record, diverse expertise and research infrastructure. In general, MVP is a resource to examine the potential of emerging genomic technologies to optimize medical care for Veterans and to enhance development of tests and treatments for diseases.

Specifically, MVP's initial planning includes enrollment up to one million users of the Veterans Health Administration (VHA) into an observational mega-cohort. The resource will include collection of health and lifestyle information as well as blood samples for storage in biorepository for future research. MVP is a major, new research initiative to create one of the largest databases of genomic, military exposure, lifestyle and health information. One of many advantages of the VA user cohort comes with the access to electronic medical record that began in 1997 and the ability to recontact participants under the current approved protocol.

MVP has the potential to better the prevention, diagnosis, and treatment of illnesses in Veterans thus improving their healthcare. MVP voluntary enrollment has begun at 9 vanguard sites and has been followed by nationwide roll-out. Currently there are 40 active MVP sites with a plan to add additional sites in the near future. MVP will allow current Veterans to help transform health care, not only for themselves, but for future generations of Veterans.

### 2.2 Utilization of EMR data in genomics research

Understanding how genes impact disease continues to raise many challenges. While a major challenge is the quality of the sequencing and genotyping data and how to handle it, equally important is the quality of the phenotypic information. Since much of this data will emerge from electronic medical records, a careful approach to the utilization of these data needs to be implemented.

Many of the VA administrative health databases are now being used for biomedical research and provide a great opportunity to conduct useful epidemiological and clinical research. Although there are limitations, including sampling and ascertainment biases, EMR databases are extremely useful in many research settings. With careful evaluation and validation of cases and diagnoses extracted from the EMR databases, many investigators are able to test various hypotheses with existing databases.

Connecting the genomics resource with existing EMR databases play a critical role in creating an environment where genomics research can flourish in the abundance of phenotyping resources. Using a validated and sophisticated algorithm to define a phenotype undoubtedly increases the specificity and sensitivity of genomics findings. Poorly defined or understood phenotypes may not only reduce the power of the genomics analyses but it may lead to erroneous results where there are no replications available to confirm such findings.

Utilizing the EMR data with emerging genomics resources is timely for the possibility of PheWAS analysis of disease-gene associations, the flipside of Genome-Wide Association Scan, determining, for a given genotype, the range of associated clinical phenotypes. The PheWAS approach is a promising and an unbiased way to discover new genetic variants.

## 3 Analytical challenges

In large scale genetic epidemiology studies of complex diseases, there exist many analytical challenges. Some of these include gene-gene and gene-environment interactions, complex phenotypes, rare variants and next generation sequencing, simulations, modeling and computational resources and data management.

In September of 2010, the National Cancer Institute sponsored a workshop[2] with experts in fields of biostatistics, genetics, statistical genetics, genetic epidemiology, epidemiology, and computer science to facilitate discussion on pressing analytical challenges in genetic epidemiology studies of complex diseases. The goals for this workshop were "(1) to facilitate discussions on statistical strategies and methods to

efficiently identify genetic and environmental factors contributing to the risk of complex disease; and (2) to facilitate discussions on how to develop, apply, and evaluate these strategies for the design, analysis, and interpretation of large-scale complex disease association studies in order to guide NIH in setting the future agenda in this area of research"[2].

In their summary report in 2012, their practical and helpful recommendations were published. Among many topics of discussion, the computational resources and data management received a lot of attention and useful recommendations were proposed. Among these were the development of new open-source, user-friendly analytical tools, establishment of new opportunities to support analytical tool development and creation or identification of common, easily accessible, data sets for methods development. Additionally, developing a forum to share lessons learned regarding data management and analysis was widely discussed. These challenges became prominent with the new emerging technologies in genomics research especially with next generation sequencing data. Genomics data are high-dimensional, and some can be formulated as statistical problems with high dimension and low sample size.

## 4. High-dimensional statistical problems: variable selection

One of the main goals of genomic data analysis is to understand how these genomics data are inter-related and how they are related to disease initiation and progress. There are numerous challenges in this area including high-dimensionality, sparsity of data, special structures of the data (local dependency, pathway dependency), integrating prior biological knowledge and complex diseases issues (complex interactions).

As stated earlier, some of the solutions can be formulated as important statistical problems in terms of being able to select a set of important variables from a large number of covariates that are highly correlated. Meanwhile, having the number of covariates being of a much larger order than the sample size worsens the situation. There are many examples of this kind in regression problem, such as in selecting relevant genes, pathways and building predictive models.

Genomics data have high-dimensional and correlated features, for example single nucleotide polymorphisms (SNPs), that can hinder the power of the identification of small to moderate genetic effects in complex diseases. The issue worsens when incorporating other environmental risk factors as effect modifiers or confounders[3]. In general, there are two techniques that aim at addressing the "curse of dimensionality" problem in genomic research: dimensionality reduction and feature selection.

### 4.1 Dimensionality reduction methods

If retaining the original risk factors is not of concern, dimensionality reduction algorithms and techniques can be considered. These techniques involve creating new attributes which are combinations of the old attributes to reduce the dimensionality of data sets[4]. Principal component analysis (PCA)[5] is one of the most widely used dimensionality reduction techniques. By producing new attributes as linear combinations of the original

variables, PCA identifies a relatively small number of independent principal components that capture the maximum amount of variation in the data. Other dimensionality reduction techniques include locally linear embedding (LLE)[6], Laplacian Eigenmaps[7]. However, dimensionality reduction techniques suffer from the major drawback that the components are not easily interpretable.

### 4.2 Feature selection methods

Feature selection methods keep only the most important features and discard the rest without altering the representation of the features. The procedure of feature selection consists of four key steps: subset generation, subset evaluation, checking with stopping criteria, and result validation. One important advantage of feature selection techniques is that model interpretability can be improved by preserving the original attributes and providing better understanding of optimal features. In general, feature selection approaches can be grouped into three categories: 1) filtering methods, 2) wrapper methods and 3) embedded methods[8].

Filtering methods select feature subsets based on their statistical properties. In other words, the implementation of filtering methods does not depend on any classification techniques to remove poor features. Based on certain statistical criteria, all features are ranked and features with lowest rankings are removed. Common filtering approaches include Pearson correlation coefficients and information gain[9]. Although filtering methods are praised for their rapid efficiency[10-12], they ignore the possible interaction among individual features. Thus results may end up with many highly correlated features/SNPs with highly redundant information.

Wrapper approaches[13] "wrap" around a particular learning algorithm that can assess the selected feature subsets in terms of the estimated classification errors and then build the final classifier[14]. There are two groups of wrapper methods:

1) deterministic methods, such as the sequential forward selection and the sequential backward selection[15]
2) randomized methods, such as genetic algorithms[16].

Both groups can help deal with the situation where the whole feature space grows exponentially with the number of original attributes. Compared to randomized methods, deterministic approaches are more likely to get stuck in a local optimum but have a lower risk of over-fitting.

Generally speaking, results provided by wrapper methods are more accurate than those from filter methods. Nevertheless, they do not incorporate knowledge about the specific structure of the classification or regression function[17]. In addition, they are more computationally expensive since they need to use a cross-validation scheme at each iteration.

Similar to wrapper methods, embedded methods[17] also reply on a specific learning algorithm. However, embedded methods embed feature selection as an integral part of the training process, i.e., the learning algorithm can decide whether to keep or remove certain features while building the classifier. Therefore, embedded approaches may be more efficient since they don't have to retrain the classifier from scratch every time they

investigate a subset of features. Some of the advantages include improved computational efficiency and similar performance to wrapper methods. They are asymptotically optimal for high dimensional data. Examples include decision tree learners[18] and, for sparse learning, LASSO and its extensions[19, 20].

## 5. Collaborative utilization of bioinformatics tools and expertise in genomics research

As technology advances in biomedical research, especially with the piles of data generated and available in genomics, one cannot avoid the involvement of researchers in bioinformatics or bioinformaticians in deciphering the meaning of such data. Starting from the initial data transfer from a laboratory that generates such data to a biomedical investigator's computing environment to quality control, analysis and reporting of results, the sheer amount of such data has created an interdisciplinary environment with heavy dependence on the expertise and tools of bioinformatics.

The field of bioinformatics relies heavily on work by experts in statistical methods and pattern recognition. Bioinformaticians come to the discipline from many fields, including mathematics, computer science, and linguistics. By providing algorithms, databases, user interfaces, and statistical tools, bioinformatics makes it possible to do exciting things such as comparing DNA sequences and generating results that are potentially significant. These new tools however, also give one the opportunity to over interpret data and assign meaning where none actually exists.

Bioinformatics is thus the study of the information content and information flow in biological systems and processes and the application of computational and analytical methods to biological problems. But the main goal of bioinformatics isn't developing the most elegant algorithms or the most arcane analyses; the goal is finding out how living things work.

## 6. Discussion

With the increasing technical advancements in genomics research, biomedical researchers are faced with a tsunami of data that provide exciting opportunities for discovery. However, many challenges remain, particularly with data management and interpretation.

It can be also seen that improved analytical methods may lead to new biological insights. Over the last decade, the costs of obtaining genetics data have declined; however, substantial costs for data management and analysis remain. Our understanding of the underlying biology is incomplete and future work should address methods to supplement current knowledge.

Analytical methods are moving from identifying a list of SNPs at a p-value threshold to multivariate methods that model related phenotypes and to model selection (e.g. by grouping SNPs into genes, genes into pathways, or environmental exposures)[21]. Many

had already adapted approaches of high dimensional data from other fields such as computer science, physics and operations research.

To continue to improve and benefit from shared knowledge in tackling these analytical challenges, the scientific community is encouraging the use of common datasets to develop QC standards and methods and to enhance collaboration. In addition, further emphasis and improvement of the way researchers share lessons learned will play an important role in this era of team science and collaboration.

As an emerging mega cohort, the MVP aims to be one of the largest research programs on genes and health in the United States with the goal of improving health care for Veterans. This type of mega genomics cohort will provide great opportunities for researchers and investigators in methodological and translational medicine. However, many analytical and computational challenges remain in dealing with large scale genomics data.

## Acknowledgements

We thank all Veterans for their service and for their participation in the MVP study. We thank all the members of the MVP study and infrastructure teams and the investigators. The Massachusetts Veteran's Research and Information Center is supported by the VA Cooperative Studies Program.

## References


1 . The Electronic Medical Records and Genomics (eMERGE) Network.

2 . Mechanic LE, Chen H-S, Amos CI, Chatterjee N, Cox NJ, Divi RL, et al. Next Generation Analytic Tools for Large Scale Genetic Epidemiology Studies of Complex Diseases. Genetic Epidemiology 2012;36:22-35.

3 . Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet. 2001;2:91-9.

4 . Liu H, Motoda H. Feature Selection for Knowledge Discovery and Data Mining: Kluwer Academic Publishers; 1998.

5 . Jolliffe IT. Principal Component Analysis. 2nd ed: Springer; 2002.

6 . Roweis ST, Saul LK. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science. 2000;290:2323–6.

7 . Belkin M, Niyogi P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering Advances in Neural Information Processing Systems. 2001;14:586–691.

8 . Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the 20th International Conference on Machine Leaning. 2003:856-63.

9 . Bramer M. Principles of Data Mining: Springer; 2007.

10 . Long A, Mangalam H, Chan B, Tolleri L, Hatfield G, Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. J Biol Chem. 2001:19937–44.

11 . Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. Journal of Computational Biology 2001;8:37–52.

12 . Yu J, Chen XW. Bayesian Neural Network Approaches to Ovarian Cancer Identification from High-resolution Mass Spectrometry Data. Bioinformatics. 2005;21 (suppl-1):i487–i94.

13 . Kohavi R, John GH. Wrappers for Feature Subset Selection. Artificial Intelligence. 1997;97:237 – 324.

14 . Inza I, Sierra B, Blanco R, Larranaga P. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. Journal of Intelligent and Fuzzy Systems. 2002;12:25–34.

15 . Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed: Morgan Kaufmann; 2005.

16 . Goldberg DE. Genetic Algorithms in Search: Addison-Wesley Professional; 1989.

17 . Lal TN, Chapelle O, Weston J, Elisseeff A. Embedded methods. Feature Extraction: Foundations and Applications. In Guyon, I., Gunn, S., Nikravesh, M. Zadeh, L. A. ed: Springer; 2006.

18 . Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees: Chapman and Hall; 1984.

19 . Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Statist Soc B. 1996;58:267–88.

20 . Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in Medicine. 1997;16:385–95.

21 . Zhou H, Sehl ME, Sinsheimer JS, Lange K. Association screening of common and rare genetic variants by penalized regression. Bioinformatics. 2010;26:2375–82.