Chapter 7

Big data challenges in genomics

Hongyan Xu*

Division of Biostatistics and Data Science, Department of Population Health Sciences, Medical College of Georgia, Augusta University, Augusta, GA, United States

Corresponding author: e-mail: hxu@augusta.edu

Abstract

With the recent development in biotechnology, especially next-generation sequencing in genomics, there is an explosion of genomic data generated. The data are big in terms of both volume and diversity. The big data contain much more information and also pose unprecedented challenges in data analysis. In this article, we discuss the big data challenges and opportunities in genomics research. We also discuss possible solutions for these challenges, which can serve as the basis for future research.

Keywords: Next-generation sequencing, Data volume, Human genomics, Disease genomics, Bioinformatics, Computational biology, Machine learning, RNA-Seq

1 Introduction

Recent development of biotechnology in genomics has led to revolutionary ways to genome sequencing that are relatively less costly and mostly high-throughput in nature. This revolution in genomics research starts with the genomics sequencing projects. The first sequencing project of model organism is the whole-genome sequencing of bacterium *Haemophilus influenzae Rd* (Fleischmann et al., 1995). The genome of the first eukaryotic organism, Saccharomyces cerevisiae, was sequenced in the next year through the collaboration of 19 countries (Goffeau et al., 1996). At the turn of the century, the flowering plant *Arabidopsis thaliana* was sequenced as the first plant genome project (Arabidopsis Genome Initiative, 2000). Most notably, the Human Genome Project (HGP) started in 1990 and completed in 2003 through the collaboration of 20 institutions and genome centers in 6 countries (International Human Genome Sequencing Consortium, 2004). One of the consequences of these early genome projects is the development of the genomic biotechnologies, which enables the production of high-throughput data at increasingly lower cost. Most notably, the next-generation sequencing (NGS) based approaches

Handbook of Statistics, Vol. 43. https://doi.org/10.1016/bs.host.2019.08.002

has enabled the following genome projects including the Encyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium, 2004), the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010) and the Roadmap Epigenome Project (Roadmap Epigenomics Consortium et al., 2015). More recently, these technologies have been used in larger population-based genomic projects such as the 100,000 Genomes Project in the United Kingdom and the GenomeAsia 100K project. The goal of these later two genome projects is to sequence 100,000 individuals in the United Kingdom and Asia, respectively, in order to understand the health and population structure and relation in these populations.

2 Next-generation sequencing

Among these technologies, next-generation sequencing (NGS) is the most prominent technology that enables current genome projects. NGS is different from the traditional Sanger sequencing in that it can perform sequencing of DNA stand in a highly parallel fashion and can generate millions of short reads, thus substantially increased the sequencing throughput. The technology itself is evolving rapidly since its beginning in mid-2000. The advances have driven the cost of whole-genome sequencing (WGS) below $1000 and are making genome sequencing an increasing available clinical tool for personalized medicine. WGS has been used in the large scale population-based genome projects to study the patterns of genetic variations in worldwide populations and their relationship with various phenotypes including disease risk factors and traits. Besides the improved the throughput and decreased cost, WGS has the advantage of unbiasedness in the survey of genetic variants across genome compared to other high-throughput methods such as genotyping microarrays.

The applications of NGS are not limited to WGS. Exome sequencing and targeted sequencing are also popular approaches for investigating focused genomic regions. By focusing on certain genomic regions, this approach allows higher sequencing depth and larger sample size, which is valuable in characterize rare genetic variants. With the completion of HGP and many genome-wide association studies, rare genetic variants have been shown to be important in explaining "missing heritability" and risk of complex diseases (Gorlov et al., 2008).

Besides the applications of NGS on sequencing of primary DNA sequences, it has been applied to study gene expression and gene regulations with technologies such as ChIP-seq, Methyl-seq, and RNA-seq. ChIP-seq uses immunoprecipitation to pull down the protein-DNA complex and enrich the DNA segments that interact with the protein. The pulled down DNA is then subjected to NGS (Park, 2009). In Methyl-seq, the methylated DNA segments are selective enriched and sequenced with NGS, which has been widely used in epigenetic studies due to its good coverage of the genome. RNA-seq is a popular

approach for evaluating gene expression levels. In a typical RNA-seq experiment, RNA is reverse transcribed to cDNA, which is then subjected to NGS. Compared to microarrays, RNA-seq has good coverage and resolution. It has been used extensively in transcriptome studies to quantify RNA expression and to characterize alternative splicing and RNA isoforms.

Big Data is used to describe the high-throughput data generated by the NGS and related technologies. Big Data creates unique challenges and opportunities characterized by the 5Vs, i.e., Volume, Velocity, Variety, Veracity, and Value. In the following sections, we layout several challenges posed by the Big Data in genomics.

3 Data integration

With current technologies in genomics, data at different layers are available, such as primary DNA sequencing data, DNA methylation data, gene expression data and environmental factors. The goal is of data integration to relate these different types of data to the responses such as disease status, disease progression, and response to treatment.

The relationship of the different types of data is depicted in Fig. 1.

Primary DNA sequences contain all the genetic information—blueprint for life. In order to realize the information encoded in the DNA sequences, they need to be expressed into RNAs and proteins, which has biological functions. All the cells in a human contain the same DNA sequences, yet different cell types have different gene expressions, hence the different morphology and functions of different cell types. In the human population, genetic variations (mutations) lead to variations in gene expression, which then potentially lead to diseases.

DNA methylation was the modification of DNA sequences (epigenetics). Levels of DNA methylation were determined partly by genetic variation and

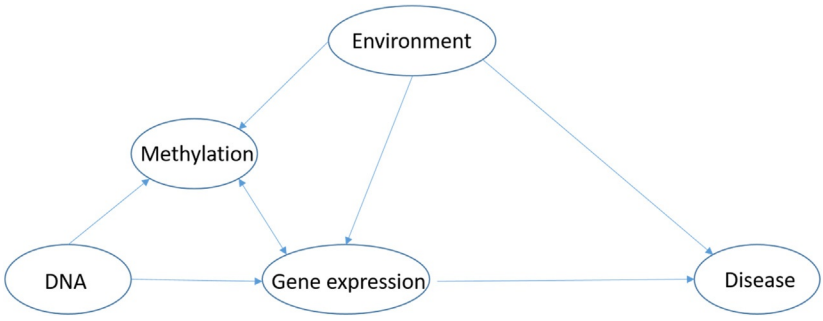

FIG. 1 Relationship of various types of genomic data and diseases. The arrows show the potential direction of the effect.

partly by environmental factors such as smoking and diet. The main effect of DNA methylation was thought to be on gene expression. The hyper-methylation in gene promoter regions has been shown to be inversely related to gene expression. But the relationship between DNA methylation in other genomic regions is not so clear cut. DNA methylation is the result of some proteins (enzymes) in the methylation pathway, whose gene expression levels will affect DNA methylation levels.

As the final piece of the puzzle, gene expression was affected by DNA sequences, DNA methylation, and environmental factors. Variation gene expression will eventually lead to different responses (phenotypes), including diseases.

With all these inter-connected parts, it is highly likely that variations at DNA level was reflected as variations at DNA methylation and gene expression levels and the signal may be amplified at successive levels. Therefore, integrated analysis of the various types of genomic data has the potential of maximizing statistical power by combining information across data types (Ritchie et al., 2015). Current methods for integrated analysis of genomic data could be roughly put into two categories, multi-stage analysis and simultaneous analysis.

In multi-stage analysis, the analysis consists of multiple steps, roughly following the flow of information in Fig. 1. For example, step 1 of the analysis could be the association of sequence variation with the phenotype; genetic variants passed through the first step are then used to filter gene expressions. The expression values of the corresponding genes are tested for association with the phenotype in step 2. The specific statistical methods used for the association will depend on the outcome variable. For continuous outcome variable, linear regression is a commonly used approach. For categorical outcome variable, common methods are based on generalized linear models such as logistic regression. The advantage of the regression based methods is that potential cofounding effects can be adjusted by including covariates in the models. Common covariates in biomedical research include age, sex, race, disease stage, and medications. The multiple stage analysis approach has been successfully applied in recent studies to investigate the genetic basis of drug induced toxicity (Huang et al., 2007, 2008). The disadvantage comes from the arbitrary in the selection criteria usually with P-value cutoffs at each stage. The over-stringency at early stages could lead to missed true signals and therefore overall low statistical power. The optimal strategy for setting the selection criteria has yet to be established.

A specific example of the multiple stage analysis method is the likelihood-based causality model selection (LCMS) (Schadt et al., 2005) to make causal inference of RNA expression on complex diseases. The basic idea is that some common genetic variants are underlying the gene expression and the disease phenotype for the causal relationship. Let L, R, C represents the genetic variant, gene expression and the disease phenotype, respectively.

LCMS uses likelihood-based method to selection from the three models: the causal model (M1), the reactive model (M2) and the independent model (M3). The likelihoods for the three models are

$$\text{M1: } P(L, R, C) = P(L)P(R|L)P(C|R)$$
$$\text{M2: } P(L, R, C) = P(L)P(C|L)P(R|C)$$
$$\text{M3: } P(L, R, C) = P(L)P(C|L)P(R|L,C)$$

Note that in the independent model (M3), $P(R|L,C)$ represents the probability of gene expression given the genetic variants and the disease phenotype, caused by other shared genetic variants and common environment in addition to the genetic variant L. The model selection is based on the Akaike Information Criterion (AIC) value (Sakamoto et al., 1986). Model with the smallest AIC is selected as the best model.

In simultaneous analysis, genomic data of different types are combined in one meta-data set for analysis. The advantage of this approach is potentially multivariate methods could be applied and there is no loss of information since all the data are combined. The disadvantage is that the corresponding model will be more complex with the different data types.

The most straightforward approach for simultaneous analysis is to concatenate various types of genomic data into one big matrix by sample ID. Appropriate statistical methods considering the heterogeneity of the data types can then be applied to the combined data for the analysis. One example of such an approach is a Bayesian integrative model to study the joint effect of genetic variants (SNPs) and gene expressions on a continuous gemcitabine-treatment responses in cancer cell lines (Fridley et al., 2012). The model first specifies the direct effect of SNPs and gene expressions on the response variable with a linear model that includes both SNPs and gene expressions as predictors. Next, the model specifies the effect of SNPs on genes expressions using a linear framework, assuming the gene expressions follows a Normal distribution. Lastly, this approach performs Bayesian variable selection using stochastic search variable selection (SSVS) (George and McCulloch, 1993; Mukhopadhyay et al., 2010) through model averaging and shrinkage of SNP effect toward zero. The prior distribution of the SNP effect is a mixture of two Normal distributions, both centered at 0 but with different variances, to represent the cases of inclusion or exclusion of the SNP in the final model. Another example is the method proposed by Mankoo et al. to perform an integrative analysis of DNA copy number variation, DNA methylation, miRNA and gene expression on time to event (survival time) in ovarian cancer (Mankoo et al., 2011). This method first performs variable selection using least absolute shrinkage and selection operator (LASSO) from the full model with all the different types of independent variables. The selected variables are then used in the Cox regression model to predict the survival time. Because this type of simultaneous analysis combines all the variables, this will increase the number of independent

variables substantially and some types of data reduction methods such as the variable selection method in the two examples would be necessary for further statistical analysis.

One of the difficulties of the concatenation-based method is that different types of genomic data often have very large difference in scales, which can create biases in statistical inference when combined directly. To overcome this problem, several methods have been proposed to transform the data to proper scale before combining them. One example is the graph-based integration approach (Kim et al., 2012) to predict cancer outcomes in brain and ovarian tumors using copy number variation, DNA methylation, miRNA and gene expression data. In this approach an individual graph is generated for each types of genomic data through Graph-based semi-supervised learning (Zhou et al., 2004), in which a node represents a sample and an edge connecting two nodes represents the relationship of the two samples, determined by a Gaussian function of the Euclidean distance between the two samples. The multiple graphs generated from each type of genomic data are then combined through linear combination to generate the final graph for the prediction of cancer outcomes.

In some cases where different types of genomic data are generated from different set of subjects, it is possible to perform the analysis of each data type separately to generate one prediction/classification model for each data type, then perform the integration of the models. An example is the study of driver mutations of melanoma using chromosome copy number and gene expression data (Akavia et al., 2010). In this study, a Bayesian network is constructed using each data type. The resulting Bayesian networks are then combined with a Bayesian scoring function maximizing the overall joint probability of the data and the model structure.

4 High dimensionality

Big data in genomics is characterized by its high dimensionality, which refers both to the sample size and number of variables and their structures. The pure volume of the data brings challenges in data storage and computation. The data volume can be on the order of terabytes for just the raw data of each sample. For the different types of genomic data, it is a good practice to keep the raw data, often in the image file format so that more sophisticated base calling algorithm can be applied later when available for improved accuracy. Data can be stored locally with hard drive arrays and backed up in other more permanent storage media. It is also a good practice to deposit the data into public databases for easy sharing in the scientific communities, such as the Gene Expression Omnibus at the National Center for Biotechnology Information (NCBI) for functional genomics data (Barrett et al., 2013). Cloud storage is another option where the data can be stored and maintained at a central location accessible by all the research communities.

Big Data is characterized by its large number of variables. The traditional algorithms could become instable with the large number of variables in the big data of genomics. The large number of variables also contributes to false positive findings due to multiplicity of statistical testing. Data heterogeneity is also a challenge for big data with the increasing popular international collaborations in order to achieve a large sample size, where data were collected from diverse laboratories and time points. While data heterogeneity is a challenge for big data analysis, it also provides unique opportunities for understanding the unique and common features of each subgroup due to its large sample sizes. For example, one of the popular approach for inferring population structure using genetic data is the Bayesian clustering method STRUCTURE (Pritchard et al., 2000), which can assign proportional ancestry to several populations for admixed individuals. STRUCTURE uses Markov Chain Monte Carlo (MCMC) algorithm. It begins by a random assignment of individuals to a K pre-determined populations. Each population has a distinct genetic allele frequencies from other populations. Genetic allele frequencies are then estimated in each population from the individual genotypes and individuals are re-assigned based on the updated frequency estimates. This iterative process is repeated many times until convergence, typically comprising 100,000 iterations. Upon convergence, we can obtain the final allele frequency estimates in each population and the assign each individual to a particular population according to the posterior allele frequency estimates.

This method has tremendous impact on the research in human genetics, evolutionary genetics and ecology. However, this method is limited with the number of genetic markers and sample size due to computational cost with the MCMC algorithm. This is apparently a limitation for Big Data in genomics. There are several recent works focusing on overcoming this limitations with likelihood-based methods (Alexander et al., 2009; Tang et al., 2005) and assumptions on variations (Raj et al., 2014).

5 Computing infrastructure

The large volumes of Big Data in genomics make the computation infeasible with traditional computing facility. It could take months to finish alignment and annotation of NGS reads for studies with large samples using desktop computers. One solution to this problem is to use the high-performance computing facilities such as computer clusters. The idea is to split the big computing job into small jobs and distribute them to each computing node in the cluster. The result is the highly parallel computing so that the big job can be finished fast (Almasi and Gottlieb, 1989). Most of the computing clusters are of the Beowulf type, where generally identical computers are connected to the header computer in a local-area network. Besides its improved computing power, Beowulf clusters are highly scalable and relatively easy to maintain, which make them especially appealing to Big Data computing needs.

Cloud computing is another potential solution to the challenge in computing facilities. In cloud computing, major computing companies provide services to end users with computing platform, storage, software and CPU times. Current cloud computing services include AWS (Amazon Web Service), Microsoft Azure, Google Cloud Platform, VMware Cloud service, and IBM Cloud. The salient feature of cloud computing is its elasticity and scalability. Users can buy the right service according to the size of the project. The service is available anytime and anywhere with internet collection. It is also maintenance-free and users can assume the platforms are well-maintained with the most recent software packages. Having the data stored at a central database hosted by cloud service providers removes the need of data transfers in separate local databases and thus could save on the time of data transfer, which is usually a bottleneck for Big Data computing.

6 Dimension reduction

Big Data poses challenges in computing and analysis due to its high dimensionality. One solution to this challenge is to use the statistical techniques for dimension reduction. In WGS, the data could be represented by a $n \times d$ matrix, where n is the number of subjects and d is the number of genetic variants. Each entry in the matrix is the respective genotype or genetic score for the subject at the genetic variant. Because of the large number of genetic variants, it is generally infeasible to use the data matrix directly in the standard statistical analysis. The idea of dimension reduction is to reduce the data dimension through linear or non-linear transformation while keeping as much information in the original data matrix as possible.

One common dimension reduction method is principal component analysis (PCA). This is a linear transformation method where it first calculates the eigenvectors of the sample covariance (correlation) matrix. The principal components (first k eigenvectors with the largest eigenvalues) are used to construct a k dimension subspace spanned by the principal components. The original data matrix is then projected to this subspace to obtain a data matrix with $n \times k$ dimension. The reduced data matrix retains a large fraction of the variance in the original data matrix. This approach is been shown to be effective to adjust for population stratification in genome-wide association studies (Price et al., 2006).

With large sample size for genomic studies, direct application of PCA may not be feasible. New methods are needed for efficient dimension reduction with Big Data. One potential method is random projection based on the Johnson-Lindenstrauss lemma, which projects the original data matrix to a subspace that preserves the distance between data points (Johnson and Lindenstrauss, 1984). This method is powerful and computational simple in that its complexity increases linearly with sample size.

7 Data smoothing

NGS data are subject to several problems such as missing values, correlations among neighboring genomic positions, and non-trivial technology-specific noise sources. Many standard statistical methods, as well as some machine learning methods, rely on rather simplistic specifications of correlations and noise—and are not robust if these specifications are not accurate. Data smoothing techniques could be useful in de-noising and obtaining the true signal.

Functional data analysis (FDA), a repertoire of statistical methods that considers data as evaluations of curves (mathematical functions) over a discrete grid, plays a critical role in exploiting the output of NGS assays, and allows sophisticated biological interpretation of shape information. FDA is an appealing option for overcoming the aforementioned problems with NGS data. Actually, correlations among neighboring measurements can be advantageous in FDA—which smooths such measurements into curves, effectively reducing the dimension of the data. Importantly, the dimension of smooth data representations can be controlled selecting the type and number of basis functions employed, while roughness penalties (e.g., on the total curvature of a function) allow continuous control over smoothness. By representing the data as functions, FDA also alleviates the impact of non-trivial noise and "fills in" missing values, improving statistical power. In addition to improving signal-to-noise ratios, and hence power, smoothing can unveil information and biological insights missed by multivariate techniques, as long as the assumption of smoothing is reasonable (Cremona et al., 2019; Frøslie et al., 2013; Ryu et al., 2016).

8 Data security

Genomic data is special in that given enough genetic marker information, it can uniquely identify an individual, much like the fingerprints. Indeed, genetic information has long been used in forensics for individual identification (Jeffreys et al., 1985a,b). Genomic data contains critical information for life. With the availability of Big Data in genomics, it is possible to make predictions of many individual characteristics including major disease risks from the data. Therefore, data security could be a big concern for Big Data in genomics.

Genomic data should be considered as protected health information and be handled according to the regulations of HIPAA (Health Insurance Portability and Accountability Act) in the United States. Common security practices should be implemented such as password protection, data/disk encryption, secure storage, secure transmission, and regular checking of data integrity with checksum analysis. Cloud computing offers convenient access of data and computing services at the same host with continuous support, and hence very popular for Big Data analysis. However, data security is a concern for

cloud computing because the data are hosted externally of the investigator's institution. The cloud computing service providers need to address these concerns by providing corresponding security measures such as controlled access and secure data transfer.

9 Example

One example of utilizing big data analysis in genomics is an integrative genomics project initiated by AstraZeneca. It is a partnership between AstraZeneca, Human Longevity in the United States, the Welcome Trust Sanger Institute in the United Kingdom, and the Institute for Molecular Medicine in Finland. The goal is to use big data analytics on whole-genome sequencing and whole exome sequencing data to identify novel targets for drug discovery. Patients will be matched to the treatments that are mostly likely to be beneficial based on their genomic profiles. In this project, AstraZeneca plan to generate genomic sequences for two million subjects by 2026, including 500,000 subjects from the participants in its various clinical trials. They established a cloud-based warehousing and analysis platform from NANnexus, which can process the raw genomic sequencing data from thousands of subjects per weeks efficiently at a reasonable cost. It also provides a secure platform where genomics data can be combined with clinical data and shared among the collaborators.

10 Conclusion

New biotechnology such as NGS generates Big Data with unprecedented speed and volume in genomics. The Big Data poses challenges in data analysis. We discussed the challenges in data integration, data management and computing facilities, dimension reduction, data smoothing, and data security in the analysis of Big Data in genomics. There are other challenges that are more data specific, such as the analysis single-cell sequencing data, de-novel assemble of sequencing reads, and the analysis of rate genetic variants. Some of the potential solutions are also discussed. These areas are under intense research and will provide important tools and information for the understanding of genomics and life in general.

References

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., McVean, G.A., 2010. A map of human genome variation from population-scale sequencing. Nature 467 (7319), 1061–1073. https://doi.org/10.1038/nature09534.

Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., Pe'er, D., 2010. An integrated approach to uncover drivers of Cancer. Cell 143 (6), 1005–1017. https://doi.org/10.1016/j.cell.2010.11.013.

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19 (9), 1655–1664. https://doi.org/10.1101/gr.094052.109.

Almasi, G.S., Gottlieb, A., 1989. Highly Parallel Computing. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis Thaliana. Nature 408 (6814), 796–815. https://doi.org/10.1038/35048692.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., et al., 2013. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 41, D991–D995. https://doi.org/10.1093/nar/gks1193. Database issue.

Cremona, M., Xu, H., Makova, K., Reimherr, M., Chiaromonte, F., Madrigal, P., 2019. Functional data analysis for computational biology. Bioinformatics (Oxford, England). https://doi.org/10.1093/bioinformatics/btz045.

ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia of DNA elements) project. Science 306 (5696), 636–640. https://doi.org/10.1126/science.1105136.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., 1995. Whole-genome random sequencing and assembly of haemophilus Influenzae Rd. Science 269 (5223), 496–512.

Fridley, B.L., Lund, S., Jenkins, G.D., Wang, L., 2012. A Bayesian integrative genomic model for pathway analysis of complex traits. Genet. Epidemiol. 36 (4), 352–359. https://doi.org/10.1002/gepi.21628.

Frøslie, K.F., Røislien, J., Qvigstad, E., Godang, K., Bollerslev, J., Voldner, N., Henriksen, T., Veierød, M.B., 2013. Shape information from glucose curves: functional data analysis compared with traditional summary measures. BMC Med. Res. Methodol. 13 (January), 6. https://doi.org/10.1186/1471-2288-13-6.

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88 (423), 881–889.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., et al., 1996. Life with 6000 genes. Science 274 (5287), 546. 563–67.

Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., Amos, C.I., 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am. J. Hum. Genet. 82 (1), 100–112. https://doi.org/10.1016/j.ajhg.2007.09.006. pii: S0002-9297(07)00012-2.

Huang, R.S., Duan, S., Bleibel, W.K., Kistner, E.O., Zhang, W., Clark, T.A., Chen, T.X., et al., 2007. A genome-wide approach to identify genetic variants that contribute to Etoposide-induced cytotoxicity. Proc. Natl. Acad. Sci. U. S. A. 104 (23), 9758–9763. https://doi.org/10.1073/pnas.0703736104.

Huang, R.S., Duan, S., Kistner, E.O., Hartford, C.M., Eileen Dolan, M., 2008. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. Mol. Cancer Ther. 7 (9), 3038–3046. https://doi.org/10.1158/1535-7163.MCT-08-0248.

International Human Genome Sequencing Consortium, 2004. Finishing the Euchromatic sequence of the human genome. Nature 431 (7011), 931–945. https://doi.org/10.1038/nature03001.

Jeffreys, A.J., Brookfield, J.F., Semeonoff, R., 1985a. Positive identification of an immigration test-case using human DNA fingerprints. Nature 317 (6040), 818–819.

Jeffreys, A.J., Wilson, V., Thein, S.L., 1985b. Hypervariable 'minisatellite' regions in human DNA. Nature 314 (6006), 67–73.

Johnson, W.B., Lindenstrauss, J., 1984. Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. 26 (189–206), 1.

Kim, D., Shin, H., Song, Y.S., Kim, J.H., 2012. Synergistic effect of different levels of genomic data for Cancer clinical outcome prediction. J. Biomed. Inform. 45 (6), 1191–1198.

Mankoo, P.K., Shen, R., Schultz, N., Levine, D.A., Sander, C., 2011. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS One 6(11), e24709. https://doi.org/10.1371/journal.pone.0024709.

Mukhopadhyay, S., George, V., Xu, H., 2010. Variable selection method for quantitative trait analysis based on parallel genetic algorithm. Ann. Hum. Genet. 74 (1), 88–96. https://doi.org/10.1111/j.1469-1809.2009.00548.x.

Park, P.J., 2009. ChIP-Seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 10 (10), 669–680. https://doi.org/10.1038/nrg2641.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38 (8), 904–909.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155 (2), 945–959.

Raj, A., Stephens, M., Pritchard, J.K., 2014. FastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197 (2), 573–589. https://doi.org/10.1534/genetics.114.164350.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., Kim, D., 2015. Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet. 16 (2), 85–97. https://doi.org/10.1038/nrg3868.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., et al., 2015. Integrative analysis of 111 reference human epigenomes. Nature 518 (7539), 317–330. https://doi.org/10.1038/nature14248.

Ryu, D., Xu, H., George, V., Su, S., Wang, X., Shi, H., Podolsky, R.H., 2016. Differential methylation tests of regulatory regions. Stat. Appl. Genet. Mol. Biol. 15 (3), 237–251. https://doi.org/10.1515/sagmb-2015-0037.

Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1986. Akaike Information Criterion Statistics. D. Reidel, Dordrecht, The Netherlands, 81.

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S.K., et al., 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. 37 (7), 710–717. https://doi.org/10.1038/ng1589.

Tang, H., Peng, J., Wang, P., Risch, N.J., 2005. Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28 (4), 289–301. https://doi.org/10.1002/gepi.20064.

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B., 2004. Learning with local and global consistency. In: Thrun, S., Saul, L.K., Schölkopf, B. (Eds.), Advances in Neural Information Processing Systems 16. MIT Press, pp. 321–328. http://papers.nips.cc/paper/2506-learning-with-local-and-global-consistency.pdf.