YALE JOURNAL OF BIOLOGY AND MEDICINE 88 (2015), pp.397-406.

CASE REPORT YJBM

PART of the WHOLE: A Case Study in Wellness-Oriented Personalized Medicine

Greg Gibson*, Urko M. Marigorta, Elohor R. Ojagbeghru, and Subin Park

School of Biology, Center for Integrative Genomics, Georgia Institute of Technology, Atlanta, Georgia

We describe the Wellness and Health Omics Linked to the Environment (WHOLE†) personalized medicine profile for a 50-year-old Caucasian male living in Atlanta, Georgia. Based on the principle that genomic medicine will be most effective when presented in the context of an individual's clinical and lifestyle data, we propose the use of a "risk radar" that summarizes health risks in eight domains. Rather than providing overwhelming lists of potentially deleterious genetic variants, we argue that profiles should be palatable, actionable, reproducible, and teachable: the PART principle. Genetic risk scores for this individual are strikingly concordant for his height, body mass index (BMI), waist hip ration (WHR), and cholesterol, and blood transcriptome data agrees with and complements his complete blood counts. Despite enjoying currently good health, his risk radar highlights metabolic disease as his major health concern.

INTRODUCTION

While it seems inevitable that genome analysis will one day be incorporated into routine health care, there is as yet no clear strategy for doing so. Leroy Hood's P4 vision that systems medicine should be predictive, preventive, personalized, and participatory [1,2] has gained traction as a statement of intent, but we are not seeing large numbers of generally healthy people having their genomes sequenced and interpreted for clinical purposes. Next generation sequencing is being rapidly adopted in oncology [3,4] and for molecular diagnosis of pediatric congenital abnormalities [5,6]. The next phase will be to implement genome analysis into care of patients with chronic diseases such as diabetes, coronary artery disease, or inflammatory autoimmune conditions [7]. Furthermore, truly preventive and predictive systems medicine will be utilized by healthy adults with an eye to maintaining their wellness into old age [8].

Before this happens, a number of obstacles must be overcome. Perhaps most importantly, utility must be demonstrated, with regard to both cost-effectiveness and capacity to improve outcomes or prevent illness [9]. In the context of lifetime per-individual health care costs that are now in excess of USD $500,000 or of per-visit hospital expenses that can easily exceed $10,000, the cost of sequencing a person's genome is very modest. Yet so long as the expense must come from a person's pocket without reimbursement and given that benefits will often be deferred for years or decades or will entail a period of anxiety and require purposeful modification of behavior, most individuals will not be willing to pay $1,000 for their own genome sequence. It seems more likely that employers may see the benefits of healthier employees in terms of greater productivity and be willing to defray costs, even if only a fraction of people benefit. Metrics need to be developed that establish the utility of personalized medicine in the context of predictive health.

A second set of obstacles relates to the delivery of genomic and comprehensive clinical information to consumers [10,11]. We summarize the challenge with the acronym PART, denoting that information must be provided in a format that is palatable, actionable, repro-

*To whom all correspondence should be addressed: Greg Gibson, School of Biology, Center for Integrative Genomics, EBB1 Building, Georgia Institute of Technology, 950 Atlantic Drive, Atlanta GA 30332; Tele: 404-385-2343; Email: greg.gibson@biology.gatech.edu.

†Abbreviations: WHOLE, Wellness and Health Omics Linked to the Environment; BMI, body mass index; WHR, waist hip ratio; PART, palatable, actionable, reproducible, and teachable; CHDWB, Center for Health Discovery and Well Being; IRB, institutional review board; CAD, coronary artery disease; MI, myocardial infarction; GRS, Genetic Risk Score; SNP, single-nucelotide polymorphism; FDA, U.S. Food and Drug Administration; HDL, high-density liporptoteins; LDL, low-density liporptoteins; BDI, Beck Depression Index; T2D, Type 2 diabetes; BIT, Blood Informative Transcripts; MCHC, mean corpuscular hemoglobin concentration.

Keywords: personalized medicine, genetic risk score, transcriptome profile, wellness

ducible, and teachable. By palatable, we mean that hundreds of thousands of data points must be reduced to a format that is neither overwhelming nor so riddled with negatives that it is demotivating. Actionable means that it should encourage the individual to adopt simple health behavior changes that are likely to make a difference and alternatively that it does not burden them with incidental findings that they are powerless to confront. Reproducible means that the report should be based on transparent algorithms and comparison with large databases that render the conclusions robust to re-evaluation either by others or over time, ensuring a level of trustworthiness. By teachable, we imply recognition that findings are, in most cases, going to be outside the standard knowledge base of consumers, so there must be support for people being able to evaluate the results on their own, learn more, and participate in their own decision making.

Two divergent genomic wellness initiatives are represented by 23andMe and iPOP. The company 23andMe for a time provided customers with an interpretation of their whole genome genotype profile, generated from a saliva sample consisting of mini-genetic risk scores assembled from a handful of markers for each of hundreds of conditions and sent anonymously by courier [12]. This polygenic risk score service is currently suspended pending revision of the status of genomic information as a medical device [13], though provision of highly penetrant rare variants that promote, for example, breast cancer can proceed. At the other end of the spectrum, Stanford Professor Mike Snyder sequenced his own genome and performed transcriptome, metabolome, proteome, immune, and many other diagnostics at regular intervals over a 14-month period [14]. The resultant integrated personal omics profile (iPOP: http://snyderome.stanford.edu/) exposed a pre-diabetic state and thereby argued for the utility of inclusion of functional genomic and clinical information alongside gene sequences. Expense and practicality make comprehensive iPOP less than attractive as a general solution, but it does seem that a compromise is feasible.

In particular, the recognition that personal omic profiles measured from blood are remarkably constant over the interval of a few years, a phenomenon we have called "omic personality" [15], implies that a baseline profile for each person might be generated relatively inexpensively. The Stanford group also has introduced the concept of a genetic Risk-o-Gram, which takes polygenic risk scores and combines them, based on knowledge of how diseases and traits interact, into a single portrayal of the areas of greatest genetic risk for the person [16]. We extended this concept to incorporate an individual's known clinical risk profile, arguing that rather than scaling genetic risk to the population mean, it should be scaled to the risk in a clinically matched sample [17]. That study also discussed how very rare variants identified by whole exome or genome sequencing provide somewhat orthogonal measures of health risks (and possibly also protective factors) to those derived from common variants, which are the focus of this paper. Ultimately as well, risk profiles should incorporate environmental and family history information where available, giving rise to the WHOLE strategy: wellness and health omics linked to the environment [18].

Here, we report a case study of how this WHOLE strategy might be applied following the PART principle. The individual is a Caucasian male, CM763, who visited the Center for Health Discovery and Well Being (CHDWB) at Emory University in Atlanta on three occasions in his late 40s and once more close to his 52nd birthday. He describes himself as being in good health and indeed has an SP36 physical summary score in the top fifth percentile. He also has an Ubble age of 40, which means that his risk of dying of natural causes in the next 5 years (less than 1 percent) is approximately that of a 40-year-old English male (www.ubble.co.uk) [19], based on multi-variate risk-assessment applied to half a million people. (In particular, being a never-smoker who walks fast and feels in good health seem to be major factors in the quick online survey assessment). However, males in his family have not recently lived into their 70s, dying young of various causes, and there are clinical warning signs of impending health concerns in his profile, making CM763 a good candidate for this personalized medicine approach.

MATERIAL AND METHODS

Clinical Measurements

Almost 200 biochemical, anthropomorphic, clinical, and survey measurements, some of which are reported here, were obtained in the course of CM763's participation in the Center for Health Discovery and Well Being study at Emory University, which is described in detail in [20-22]. Blood and urine biochemistry and CBC data points were generated under contract to Quest Diagnostics in Atlanta, GA. Whole body densitometry (DXA scan) and SphygomoCor cardiovascular assessment were performed at the Center, and food and behavioral survey assessments were filled in over the Internet within a week of each visit. Center visits were conducted on January 5, 2011; August 1, 2011; February 6, 2012; February 15, 2013; and July 15, 2015. Percentile scores for the nine traits reported here were generated from the z-score distribution of each trait in 291 males in the study, averaged over an average of 2.7 visits. Framingham risk scores for Diabetes and Cardiovascular Disease were calculated by combining age, gender, BMI, cholesterol, smoking, blood pressure, triglycerides, and fasting glucose as described in [23,24].

Under the terms of the CHDWB institutional review board (IRB) protocols at Emory and Georgia Tech, genetic data cannot be returned to participants, but it is fully available for research. In this case, CM763 volunteered to provide genetic data generated independent of the CHDWB study, and it is combined with clinical data obtained during his participation in the CHDWB study. Data for another 312 participants is reported simply as a histogram

summarizing the distribution of correlations of percentiles of genetic risks and observed traits.

Genetic Data

All of the genotypes reported in this manuscript were derived from CM763's 23andMe profile, which is based on data generated on a custom Illumina genotyping chip with 960,614 SNPs. A total of 576 unique single-nucelotide polymorphisms (SNPs) were chosen for analysis, and these were involved in 604 trait associations. They are listed in Supplementary Table 1. More than half of the 28 genetic effects shared by two or more traits were between cholesterol and triglycerides (8), coronary artery disease (CAD) and myocardial infarction (MI) (4), or Crohn's disease and ulcerative colitis (4), and accordingly, these were the only Genetic Risk Scores (GRS) (see below) that showed even modest correlations.

The list of SNPs was obtained by browsing the Phenotype-Genotype Integrator (PhenGenI) [25] site (http://www.ncbi.nlm.nih.gov/gap/phegeni) in May 2015 for all SNPs associated at $p < 10\text{-}8$ with each of 26 common diseases or phenotypes for which clinical data was available. A further 24 blood pressure-associated SNPs were extracted from Table 1 in [26], two of which are also associated with BMI or CAD. Five hundred ninety-nine of 1,572 SNPs were removed due to linkage disequilibrium ($r2 > 0.2$) with a peak association in the vicinity, producing a list of 973 associations with 27 traits involving 933 unique SNPs. Since PhenGenI only reports the SNP rsIDs, a further manual curation step was required to identify the risk allele. We searched each SNP's profile on SNPedia, which reports the risk SNP, when available, for each relevant study and provides links to the papers as well as population genetic data. For some traits, many of the risk alleles are not reported, and inconsistencies where different studies reported opposite alleles as carrying risk, or ambiguities due to A/T or G/C alleles having similar frequencies, were removed. This left a list of 617 unique SNPs. Most of these were directly reported on the Illumina genotyping chip (a similar version of which, Illumina OmniQuad, was used to genotype the CHDWB cohort), but 179 were imputed with IMPUTE2 [27] against the build 37 phase 1 1000G haplotypes, noting that 37 of the SNPs failed imputation.

Allelic sum genetic risk scores [28] were computed by assigning all risk allele homozygotes for each SNP a value of 2, all heterozygotes a value of 1, and the alternate homozygotes a value of 0, and then summing these scores for each trait in each individual. These scores were computed for CM763 and for each of 317 other Caucasian individuals (males and females) in the CHDWB database. Two hundred eighty out of 192,072 genotypes (604 associations in 318 people) were not available, for a missing value rate of just 0.15 percent. Scores for individuals with incomplete profiles for a given trait were adjusted to the nearest whole number after multiplying their allelic sum by the number of SNPs for the trait and dividing by the number observed in their profile. Thus, someone with a score of 20 based on 19 of 21 SNPs would have an adjusted score of 22. CM763's percentile genetic risk score was then computed by reference to the 318 individual panel and assigned to the midpoint percentile of his bin. Thus, if he had a score of 15 along with 10 percent of the sample, and 21 percent had a score of 14 or lower, his percentile was set at (21+10/2) = 26 percent. The genetic risk scores for all 318 individuals along with CM376's percentile rank are listed in Supplementary Table 2.

Transcriptome Data

All transcriptome data was generated from whole blood samples preserved in Tempus RNA tubes. One microarray sample, corresponding to the baseline January 2011 visit, was generated on the Illumina HT12 human gene expression profiling chip and is included in the data reported at the Gene Expression Omnibus, GSE61672 (sample GG2_0014), processed as described in [29]. Three RNASeq samples were generated from samples collected on April 3, April 10, and April 21, 2015 (CM763 vas travelling overseas on April 14). 100bp paired end short reads were aligned to HuRef GRCh38 release 79 fasta assembly using Kallisto v0.42.2 (http://pachterlab.github.io/kallisto/download.html) to generate transcript-level counts, which were converted to counts per million (cpm) by summing the counts for all transcripts of each of 25,963 genes, dividing by the total counts and multiplying by 10^6. Differential expression relative to the 12 individuals reported in [15], or between the Week 3 and Weeks 1 and 2 samples, was performed in edgeR [30], retaining only 7,720 or 4,357 genes with at least 50 cpm in at least two of the samples, respectively. Preliminary analyses had indicated that Weeks 1 and 2 are much more similar to one another than Week 3, which was close to full recovery, but also during a period of jet lag. For the analysis just of CM763, no further normalization was performed, but for comparison with the 12 CHDWB participants, inter-quartile range transformation was applied to the 39 samples to equilibrate the profiles. This transform adjusts the profiles such that the value of the 25th and 75th percentiles are the same and was performed in JMP Genomics v5.0. Supplementary Table 3 reports the counts and cpm for CM763's three samples.

RESULTS

Consistent with the core principle of the WHOLE approach [18], that genomic information should be supplied in consultation with medical professionals and not just as a report delivered anonymously by email or over the Internet, we start this case study with a brief summary of salient environmental and family history observations. Environment means the combination of behavioral, dietary, lifestyle, and medication influences that are not ostensibly genetic. The main body of the report then places two types of genomic information into this context, namely polygenic risk assessment from common genotypes and gene expression profiling by microarray and RNASeq.

Environmental and Family History Observations

Exercise: CM763 meets the U.S. Food and Drug Administration (FDA)-recommended aerobic exercise goals by walking to work 25 minutes each way, five times a week, and engaging in vigorous exercise (jogging, yoga, formerly swimming) twice a week. He does not use a Fitbit or Runkeeper-style app to log activity.

Diet: CM763 enjoys a diverse, low-fat omnivorous diet based on fresh produce with little fast food. Breakfast and lunch are light, with a single main meal that is taken in the evening without exercise before sleep, which may contribute to his gradual weight gain despite calculations that his total caloric consumption should be less than the number of calories burned. Low-carb diets such as Atkins or South Beach do not agree with him, but reduced carbohydrate intake is always a goal. Portion control and increased exercise has historically been his most effective mechanism for weight loss.

Drugs and Medication: CM763 is a strong proponent of the notion that the body knows best how to heal itself and does not take any medications or supplements. The only exceptions are prescriptions for a few days on a handful of occasions as an adult, and non-steroidal inhalants to treat childhood asthma. He has never smoked. He does, however, consume the standard amount of alcohol for a professional adult, namely one or two glasses of wine or beer a day.

Stressors: CM763 has lived an upper middle-class life, excepting graduate school, without undue stress lasting more than a few months, and has been married for 21 years. He has lived in nine cities on three continents, both in suburbia and the inner city. Aside from a few incidents requiring stitches, he has not experienced any trauma. However, an anomaly is that he has low social function scores, such as the SP12 survey of meaning and peacefulness that places him in the bottom percentile. He does not have children, does not attend church, and is not on Facebook.

Family History: There are no strong indicators of enhanced familial risk for any common or rare diseases in CM763's immediate family. His father died of colon cancer in his 60s, and his mother has survived two separate incidents of breast cancer, but the overall incidence of cancer in the extended family is not unusual. Diabetes and cardiovascular and autoimmune disease are also not indicated in his family. Isolated cases of osteoarthritis and severe depression do not present more than mild concern, as there is again no indication of inheritance in the family.

Genetic Risk Scores

Polygenic risk scores provide an estimate of an individual's risk of disease or sub-clinical abnormality based on the contributions of a collection of common genetic variants. These variants, or single-nucleotide polymorphisms (SNPs), are generated on a genotyping chip that, in this case, included more than 960,000 SNPs. Genome-wide association studies (GWAS) have identified between 10 and 130 SNPs associated with each of the 27 traits and diseases listed in Supplementary Table 1. Approximately three quarters of them were present in the Illumina genotyping files for 317 other Caucasian males in the CHDWB cohort, so bioinformatic imputation methods were used to generate the full set of 592 genotypes for each individual, an average of 22 loci per trait. There are numerous ways to generate genetic risk scores from such data, many of which incorporate effect size estimates into a likelihood estimate of disease [31]. However, we decided that a simple allelic sum [26] is the most transparent and repeatable score at this time, so we report here the percentile rank of CM763 relative to the CHDWB Caucasians for each risk score. We also note that alternative methods for GRS computation may give almost identical (weighted GRS) or moderately (log odds ratios) correlated scores [18], but do not present a comparison here, since for all traits the known genotypes explain less than 20 percent of the genetic variance so in any event should be regarded, at this stage of genomic analysis, as very modest predictors.

Figure 1A shows a remarkable degree of concordance between the genetic risk scores for CM763 and his observed clinical profile. The overall regression of genetic and clinical percentile ranks for nine traits is only marginally significant ($R^2 = 0.47$, $p = 0.04$: dashed line), but if the three metabolic disease-related traits in red are removed, the regression for the other six measures is strongly linear ($R^2 = 0.97$, $p = 0.0003$). The histogram in Figure 1B shows the distribution of correlations for the same nine traits for 313 people (114 men, 199 women) in the CHDWB cohort, with an arrow indicating that CM763's concordance is actually in the top half dozen in the study. His actual WHR, height, blood pressure, BMI, Beck Depression Index, and total cholesterol are very consistent with estimates based on 11, 92, 24, 28, 6, and 43 loci, respectively. The total cholesterol estimate is particularly interesting as he has been in the top few percentiles for this measure his whole life. Both his high- and low-density liporptoteins (HDL and LDL) are unusually high, and while this might be taken as a sign of a familial rare hypercholesterolemia mutation, the GRS implies instead that he has an unusually high polygenic combination of alleles promoting high cholesterol.

At the other end of the plot, he has always been relatively short, and this agrees with a GRS for height at the 14th percentile — if anything, he is taller, at 5'8" (170 cm), than expected. Similarly, his WHR GRS is in just the fourth percentile, and at least until recently, that is reflected in the 20th percentile WHR rank for men. The blood pressure estimate (both systolic and diastolic) is also close to the 28th percentile predicted by the GRS, while his slightly overweight BMI is around the 40th percentile for both measures. The Beck Depression Index (BDI) score [32] is most likely a coincidence, as the score fluctuates widely and for most people is within a standard deviation unit of zero, so a 75th percentile BDI may just reflect willingness to admit occasional anxiety and introspection. The related bipolar depression GRS is based on just six loci that explain very little of the genetic variance,

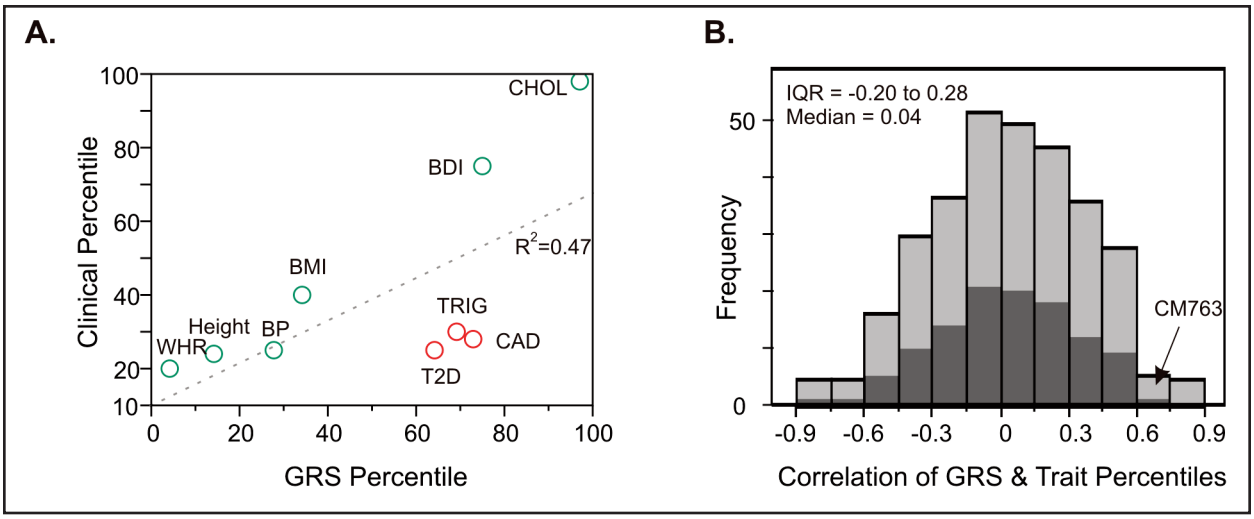

Figure 1. Genetic risk score — Phenotype Correlations. A) Plot of percentile rank of phenotype against percentile rank of allelic sum genetic risk score for the trait (Height, Waist-to-Hip Ratio (WHR), Body Mass Index (BMI), Systolic Blood Pressure (BP), Triglycerides (TRIG), Beck Depression Index (BDI), and Total Cholesterol) or related Framingham Risk Score (for Type 2 Diabetes (T2D) or Cardiovascular Disease (CVD)). Green circles are traits for which the GRS rank closely matches the observed clinical rank for CM763; red points are three outliers. **B)** Histogram of frequencies of Pearson correlations between the percentile ranks for 313 CHDWB participants, including 114 men (dark shading) and 199 women (light shading).

and there is no hint of clinical diagnosis of BPD for CM763. Not shown on the figure is his predicted early age of onset of menarche, since menstruation is not relevant for a male, though he is happy to attribute occasional mood swings to menopause that is predicted to occur in his early 50s.

The discordance in metabolic disease GRS assessments for triglycerides, Type 2 diabetes (T2D), and coronary artery disease, all of which are in the 60th to 75th percentiles, and corresponding clinical measures of triglycerides, and the Framingham T2D and CAD risk measures [23,24], all of which are below the 30th percentile, is interesting given several other clinical data points and the non-genetic data described above. On the one hand, these high GRS along with very high cholesterol is prima facie cause for concern, as is the report from his whole body densitometry DXA scan that places the percent body fat in his tissues between 30 percent and 35 percent, placing him in the top few percentile for men. This result has been consistent over four visits to the CHDWB, but seems implausible given his BMI and is at odds with the alternate percent body fat measure from a Tanita scale, which places him squarely in the normal range. Nevertheless, the possibility of a genetic variant or variants that produce a highly unusual body fat distribution cannot be excluded. On the other hand, CM763 has heart function scores (augmentation index, pulse wave velocity, hyperemia and sub-endocardial viability ratio) that are all on the healthy side of normal, as well as plaque-free arterial CAT-scan images. Given his overall good health, exercise, and diet, CM763 has chosen to regard all of the biomarkers taken together as a potential indicator of future health issues, but not to pursue pharmacological intervention at this time.

Gene Expression Profiles

A second type of genomic measure of particular interest with respect to immune function is peripheral blood gene expression profiling. CM763 had microarray-based transcriptome analysis [33] performed during his initial CHDWB visits, and RNASeq [34] was performed on three samples over a month-long period during a respiratory infection that was slow to resolve in April 2015.

Whole blood RNA is a complex mixture of cell types, numerically dominated by lymphocytes, neutrophils, and monocytes, but with minor contributions from basophils, eosinophils, dendritic cells, macrophages, and other white blood cells. Red blood cells and platelets, being anuclear, have limited RNA repertoires, but a signature of reticulocyte abundance can be detected. We have introduced the use of Blood Informative Transcripts (BIT) as a means to profile immune activity based on the ability of 10 sets of 10 highly co-regulated transcripts to capture the covariance of hundreds of genes related to immune function [35].

The most notable feature of this analysis is the strong concordance between the consistently high lymphocyte counts (90th percentile) and low neutrophil counts (9th percentile) in CM763, and his Axis 1 and Axis 5 scores that respectively capture T-cell signaling and neutrophil/inflammatory signaling and are in the 98th and 4th percentiles respectively [35]. The other Axes that are enriched for gene functions related to B-cell signaling, general cellular processes, and the type 1 interferon response are unremarkable, and there is no corresponding data from the complete blood cell analysis. We have reported recently that each of the BIT scores tend to remain within 10 percent of a person's healthy median value, despite considerable fluctuation in neutrophil counts, so they can be

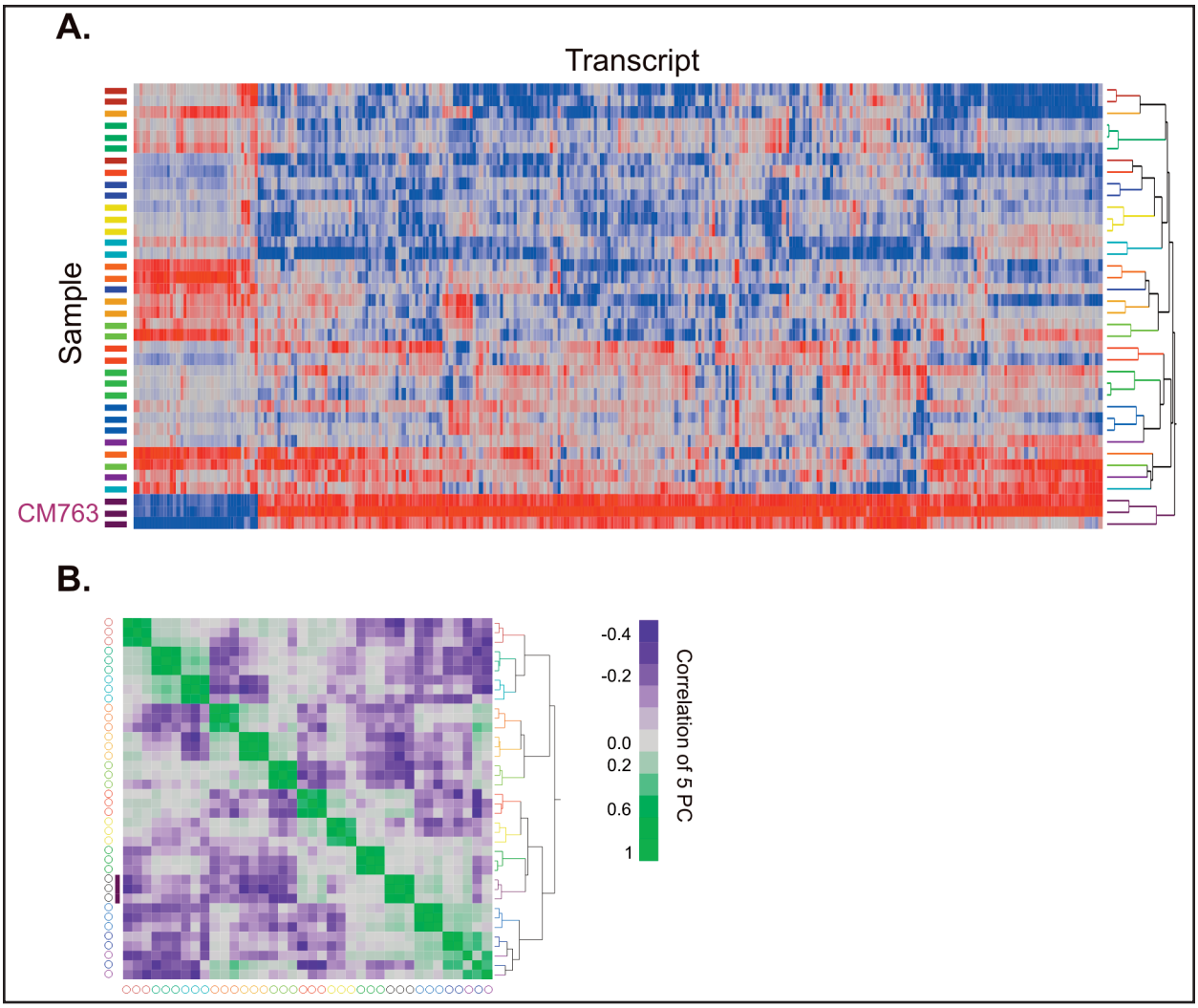

Figure 2. RNASeq-based Transcriptional profiling of CM763 relative to 12 other CHDWB participants. A) Heat map of 300 genes (vertical bars) that are significantly differentially expressed in CM763, in 39 samples (three for each individual, rows) hierarchically clustered in both dimensions by Ward's method, standardizing genes to z-scores (red high expression, blue low expression). Note how the three samples for each individual indicated by color coding of the dendrogram tend to be adjacent, indicating overall conservation of expression, but with four individual samples that are more similar to CM763, who is the bottom set of three samples. **B)** Pairwise correlations of the first five principal components of each sample considering all genes, as a measure of overall profile similarity, demonstrating how CM763's three profiles (highlighted with thin vertical black bars) are embedded within the matrix of 12 other individuals' profiles, each of whom forms a unique cluster. Green high positive correlation, purple negative correlation (range 1.0 to -0.4) indicating very strong to weak profile similarity.

used to establish a baseline profile [15]. Chaussabel et al. [36] have shown that components of these Axes differentiate immune disease-types, so it is possible that the signatures provide biomarkers of immune risks. The data here suggests that CM763 has a strong capacity for adaptive immune response but perhaps less robust innate immune function. He rarely contracts influenza or experiences illness requiring time off, but it remains to be determined whether or how this relates to his unusually high ratio of lymphocytes to neutrophils.

The RNASeq analysis performed during a respiratory infection is also of interest with regard to the gene sets that are enriched both comparing him with 12 other CHDWB participants [15] and comparing Week 3 with Weeks 1 and 2. These lists are documented in Table 1, and the differential expression is shown in the heat map in Figure 2A. First, comparing the third week when he had almost recovered to the first two weeks of low-grade infection whose only symptom was a persistent cough, there was increased expression of multiple genes related to cytokine production, and eight of the 10 Axis 5 blood informative transcripts were also elevated, suggesting a neutrophil response. In contrast, he had reduced expression of three hemoglobins and several genes related to platelet production and mean corpuscular volume. This is in addition to already low general expression of oxygen-binding heme proteins and reticulocyte factors relative to the other 12 individuals, suggesting persistently mild anemia. Furthermore, the microarray analysis had also indicated low Axis

Table 1. Blood Informative Transcript Analysis.

Gene Sets Up-Regulated in CM763 (of 814 genes)
RNA-binding proteins including ribosome constituents (218; FDR 1×10^{-54})
Respiratory electron transport chain (47; FDR 2×10^{-31})
Threonine endopeptidase activity (15; FDR 4×10^{-14})
Oxidoreductase activity (84; FDR 8×10^{-14})
NADH Dehydrogenase activity (17; FDR 5×10^{-10})
Acidosis (30; FDR 2×10^{-5})
Binding sites for ELK1 (107; FDR 2×10^{-14})
Binding sites for STAT1 (26; FDR 2×10^{-3})
Gene Sets Down-Regulated in CM763 (of 36 genes)
Oxygen-binding heme proteins (5; FDR 3×10^{-9})
Reticulocytosis (mouse phenotype) (5; FDR 2×10^{-4})
Thalassemia (3; FDR 3×10^{-7})
Gene Sets Up-Regulated at Week 3 in CM763 (of 45 genes)
Cytokine production (11; FDR 1×10^{-4})
Abnormal macrophage physiology (8; FDR 0.03)
Gene Sets Down-Regulated at Week 3 in CM763 (of 27 genes)
Ribosomal proteins (10; FDR 4×10^{-13})
Oxygen-binding heme (haptoglobulin) proteins (3; FDR 3×10^{-7})
Microcytic anemia (4; FDR 6×10^{-6})
Abnormal platelet number (mouse phenotype) (5; FDR 0.002)
Abnormal mean corpuscular volume (mouse phenotype) (4; FDR 0.008)
Thalassemia (3; FDR 2×10^{-7})

2 score, which is strongly correlated with reticulocyte counts [37] relative to the entire cohort (11th percentile). However, CM763 actually has RBC counts ranging from the 75th to 95th percentile across his CHDWB visits, as well as normal hemoglobin levels, oxygen saturation, and hematocrit levels. By contrast, his mean corpuscular hemoglobin concentration (MCHC) is in the bottom percentile for the study, raising the possibility that it is caused by aberrant reticulocyte-related gene expression detected both by microarray and RNASeq. He is not aware of any adverse health effects related to RBC function, but this profile is cause for attention.

Among the up-regulated peripheral blood genes are hundreds that contain binding sites for the hematopoiesis transcription factors *ELK1* and *STAT1*, as well as hundreds of genes with roles in mitochondrial electron transport, respiration, and ribosomal function. A caveat to this analysis is that the RNASeq of his samples was performed at a different time and on a different sequencer than the other CHDWB samples, and an inter-quartile range normalization was needed to ensure that the profiles have similar overall distributions. Despite this, considering all 13,560 measured transcripts, his three samples are embedded within the overall profile similarity matrix (Figure 2B) forming a unique set, as do each other individual's three samples. This strengthens the inference that the large number of up-regulated genes in Figure 2A and Table 1 has a biological rather than technical basis. It is, though, noteworthy that a cluster of four samples that are most similar to him are from four different individuals at one of their three times, suggesting that perturbation can also give rise to differential expression of the genes that are most divergent in CM763.

DISCUSSION

The results presented above suggest that, at least for CM763, genetic risk and transcriptional profiles can be concordant with and potentially explain clinical attributes of interest. The regression in Figure 1A is actually considerably more significant than theory predicts the relationship between GRS and traits should be [38], given that to date the scores explain generally less than a tenth of the phenotypic variance. Divergence in mean phenotypes is expected at the extremes of the distribution, but not to the degree observed, and this is confirmed by the observation that by chance, he is in the top half dozen individuals of 313 considered in parallel. Nevertheless, the only really discordant measured point, aside from the Framingham risk scores, was triglycerides, but rather than considering it an outlier, the 70th GRS percentile is actually within range of the observed clinical 30th percentile given the small amount of variance explained by genotypes. Generation of similar profiles for thousands of people will be required to establish how unusual CM763's regression is and to define patterns of concordance from real data. It is also important to emphasize that whole genome sequencing would be expected to uncover dozens of very rare variants that severely impair the function of specific genes and may sometimes have larger effects than the cumulative common variant scores, which is a major limitation of the current study.

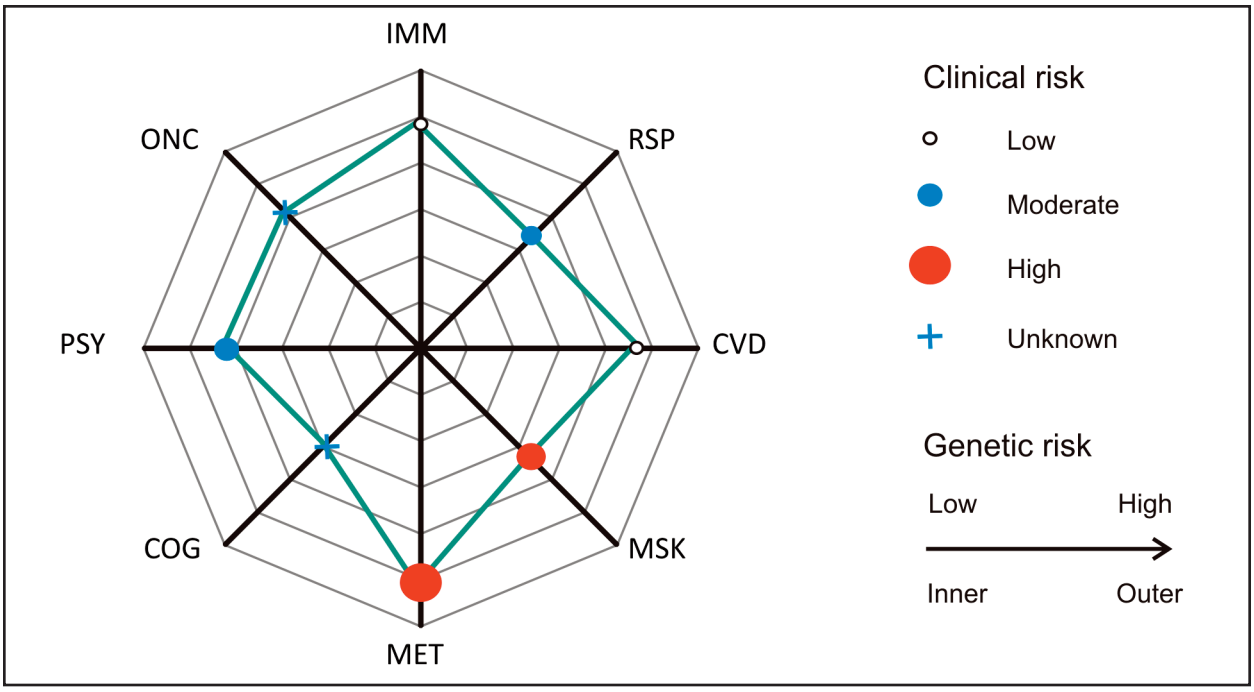

Figure 3. Risk Radar for CM763, showing his percentile rank for genetic risk on the rays and summary of clinical risk as the size of the filled circle in each of the seven health domains. Genetic risk ranges from zero (inner web) to 100 (outer web), as average of up to half a dozen traits in each domain. The objective is not to provide a precise statement of risk for individual conditions, but rather to contrast which domains are concordant for high or low genetic and clinical risk. See Discussion for full explanation. The idea is to provide a simple representation of genetic risk compared with existing clinical risk, in eight domains of disease: IMM, immunological; RSP, respiratory; CVD, cardiovascular; MSK, musculoskeletal; MET, metabolic; COG, cognitive; PSY, psychological; ONC, oncological. Other possible domains that could be added include reproductive health, or organ and tissue aging. See [17] for presentation of how more detailed analysis of genotypes within each domain may generate actionable behavioral or other interventions.

Assuming for now that genetic risk scores continue to improve, and that within 5 years they will be able to explain half of the genetic variance for a wide range of traits, the next issue becomes how to present the data in a palatable, actionable, reproducible, and teachable manner. Figure 3 shows our proposal for how a Risk Radar plot may work [17]. The radiating arms of the radar each represent a domain of health and disease, clockwise from the top: immunity, respiratory, cardiovascular, musculoskeletal, metabolic, cognitive, psychological, and malignancy/oncology. Percentiles of genetic risk are indicated by each successive web, from high risk (100th percentile) on the outside to low risk (0th percentile) at the inner ring. The green line joins a general estimate of risk in each domain, while filled circles represent biochemical or survey measurements that summarize observed clinical risk or health status. Large red circles show high clinical risk, blue intermediate, and small white ones low risk. The two plus signs over the cognitive and oncology domains indicate that in the absence of disease, clinical risk is unknown.

Environmental and family history data is not directly represented on the plot, but can be interpolated by the individual as he or she decides, possibly in consultation with health care professionals, how to respond to the profile. For CM763, the two most striking conclusions are that he has multiple signs of metabolic disease risk that is concordant with high GRS for cholesterol and above average risk for high triglycerides and T2D. Given that his blood glucose is at the high end of normal and that he is reasonably physically active and controls caloric consumption, these metabolic domain scores serve as a warning to keep paying attention. Currently, there is insufficient genotype information to formulate a GRS in the musculoskeletal domain, so it was set to 50 percent, but early signs of arthritis along with bone mineral density in the bottom quartile are mild concerns, also lessened by the fact that CM763 has a low bodily pain index.

Two other domains suggest some discordance between genetic risk and clinical observation. The GRS for CAD and MI are moderately high, but that for blood pressure is not, and aside from very high cholesterol, all other measures of cardiovascular function are healthy. Analysis of exome and eventually whole genome sequence data may also uncover rare variants that alter protein function and may contribute in deleterious or protective ways to CM763's health profile. In the immunological domain, his GRS for a variety of autoimmune and inflammatory conditions are above average, but he seems to have a strong immune system that is supported by high adaptive immune activity from his transcriptome profile, despite the

slow resolution of a persistent cough on this one occasion. Perhaps the biggest surprise from this analysis was the abnormality of his red blood cell related transcriptional profiles, both relative to others and at the end of the mild illness, particularly given that he has never complained of anemia or a blood-related disorder.

This profile will appear to be unduly simplified to many, but the point is to reduce a potentially overwhelming body of data to broad areas that the patient may choose to focus on. As argued in our earlier paper [17], more detailed analyses within each domain, including genetic measurements, can help to refine the risk assessment. Readers may also object to the use of a simple allelic sum to generate GRS, but again the point is not to provide a quantitative point estimate of risk (which, in most cases, will have a very large error of plus or minus at least 20 percentage points), but to place each person's summary profile in context. Allelic sums are actually highly correlated with odds ratio estimates, since effect sizes are generally quite similar and small, and have the advantage that they are not affected by error in the estimate of the effect sizes across populations, and heterozygote effect estimates are not influenced by allele frequencies (as they can be in odds ratio methods). It is, however, important to recognize that the distribution of GRS can vary among populations, so evaluations should be made relative to the appropriately matched ethnicity. Clearly, GRS will change over time as more and more SNPs are discovered, presumably converging on more accurate estimates as more variance is explained. Thus, they should not currently be seen as satisfying the robust repeatability criterion of our PART assessment.

Ultimately, the utility of a systems biology risk profile lies in the willingness of the patient to act on the information received, his or her preparedness for dealing psychologically with the information, and commitment to adopt suitable health behaviors [39]. Genetic information is by nature often negative, risks of not responding to information are usually deferred to a time in the future, and the information is only weakly predictive or difficult to comprehend. Nevertheless, if it helps a patient to take more interest in his or her own health and provides a foundation for a dialog with medical professionals that moves beyond mere phenomenology, then there is potential for benefits that may maintain wellness. Private mutations currently are attracting much of the attention in genomic medicine, but we would argue that integrative genomic approaches that also incorporate common variants, perhaps including methylation and metabolomic profiles, have an important place in the future of medical care.

Acknowledgments: We gratefully acknowledge the support of the Center for Health Discovery and Well Being and the Atlanta Clinical and Translational Science Institute for the clinical profiling reported here. Genotypes were obtained from 23andMe. This research was partially supported by NIH P01 GM099568 Project 3 to GG.

REFERENCES

1. Auffray C, Charron D, Hood L. Predictive, preventive, personalized and participatory medicine: back to the future. Genome Med. 2010;2(8):57.
2. Hood L, Auffray C. Participatory medicine: a driving force for revolutionizing healthcare. Genome Med. 2013;5(12):110.
3. Kalia M. Biomarkers for personalized oncology: recent advances and future challenges. Metabolism. 2015;64(3 Suppl 1):S16-21.
4. Cronin M, Ross JS. Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology. Biomarkers in Medicine. 2014;9(7):293-305.
5. Biesecker LG, Burke W, Kohane I, Plon SE, Zimmern R. Next generation sequencing in the clinic: Are we ready? Nat Rev Genet. 2012;13(11):818-24.
6. Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. Nat Rev Genet. 2013;14(6):415-26.
7. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015;385(9975):1305-14.
8. Khoury MJ, Evans JP. A public health perspective on a national precision medicine cohort: balancing long-term knowledge generation with early health benefit. JAMA. 2015;313(21):2117-8.
9. Wade JE, Ledbetter DH, Williams MS. Implementation of genomic medicine in a health care delivery system: a value proposition? Am J Med Genet C Semin Med Genet. 2014;166C(1):112-6.
10. Cesario A, Auffray C, Russo P, Hood L. P4 medicine needs P4 education. Curr Pharm Des. 2014;20(38):6071-2.
11. Marusina K. Genomic singularity is near. Genet Eng Biotech News. 2014;34(21):38-40.
12. Do CB, Hinds DA, Franke U, Eriksson N. Comparison of family history and SNPs for predicting risk of complex disease. PLoS Genet. 2012;8(10):e1002973.
13. Annas GJ, Elias S. 23andMe and the FDA. N Engl J Med. 2014;370(11):985-8.
14. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012;148(6):1293-1307.
15. Tabassum R, Sivadas S, Agarwal V, Tian H, Arafat D, Gibson G. Omic personality: Implications of stable transcript and methylation profiles for personalized medicine. Genome Med. 2015;7(1):88.
16. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. Lancet. 2010;375(9725):1525-35.
17. Patel CJ, Sivadas A, Tabassum R, Preeprem T, Zhao J, Arafat D, et al. Whole genome sequencing in support of wellness and health maintenance. Genome Med. 2013;5(6):58.
18. Gibson G. Wellness and health omics linked to the environment: the WHOLE approach to personalized medicine. Adv Exp Med Biol. 2014;799:1-14.
19. Ganna A, Ingelsson E. 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study. Lancet. 2015. In Press.
20. Brigham KL. Predictive health: the imminent revolution in health care. J Am Geriatr Soc. 2010;58(Suppl 2):S298-302.
21. Rask KJ, Brigham KL, Johns MME. Integrating comparative effectiveness research programs into predictive health: A unique role for academic health centers. Acad Med. 2011;86(6):718-23.
22. Tabassum R, Cunningham L, Stephens EH, Sturdivant K, Martin GS, Brigham KL, et al. A longitudinal study of health improvement in the Atlanta CHDWB wellness cohort. J Pers Med. 2014;4(4):489-507.
23. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med. 2007;167(10):1068-74.

24. D'Agostino RB Sr., Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation. 2008;117(6):743-53.
25. Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, et al. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur J Hum Genet. 2014;22(1):144-7.
26. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478(7367):103-9.
27. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44(8):955-9.
28. Aulchenko YS, Struchalin MV, Belonogova NM, Axenovich TI, Weedon MN, Hofman A, et al. Predicting human height by Victorian and genomic methods. Eur J Hum Genet. 2009;17(8):1070-5.
29. Wingo AP, Gibson G. Blood gene expression profiles suggest altered immune function associated with symptoms of generalized anxiety disorder. Brain Behav Immun. 2015;43(1):184-91.
30. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.
31. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14(7):507-15.
32. Back AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clin Psych Rev. 1988;8(1):77-100.
33. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet. 1999;21(1):10-4.
34. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.
35. Preininger M, Arafat D, Kim J, Nath AP, Idaghdour Y, Brigham KL, et al. Blood-informative transcripts define nine common axes of peripheral blood gene expression. PLoS Genet. 2013;9(3):e1003362.
36. Chaussabel D, Quinn C, Shen J, Patel P, Glaser C, Baldwin N, et al. A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. Immunity. 2008;29(1):150-64.
37. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, et al. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci USA. 2003;100(4):1896-1901.
38. Visscher PM, Gibson G. What if we had whole-genome sequence data for millions of individuals? Genome Med. 2013;5(9):80.
39. Morton K, Beauchamp M, Prothero A, Joyce L, Saunders L, Spencer-Bowdage S, et al. The effectiveness of motivational interviewing for health behaviour change in primary care settings: a systematic review. Health Psychol Rev. 2015;9(2):205-23.