

Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

Rui Chen,^{1,11} George I. Mias,^{1,11} Jennifer Li-Pook-Than,^{1,11} Lihua Jiang,^{1,11} Hugo Y.K. Lam,^{1,12} Rong Chen,^{2,12} Eliana Miriami,¹ Konrad J. Karczewski,¹ Manoj Hariharan,¹ Frederick E. Dewey,³ Yong Cheng,¹ Michael J. Clark,¹ Hogune Im,¹ Lukas Habegger,^{6,7} Suganthi Balasubramanian,^{6,7} Maeva O'Huallachain,¹ Joel T. Dudley,² Sara Hillenmeyer,¹ Rajini Haraksingh,¹ Donald Sharon,¹ Ghia Euskirchen,¹ Phil Lacroute,¹ Keith Bettinger,¹ Alan P. Boyle,¹ Maya Kasowski,¹ Fabian Grubert,¹ Scott Seki,² Marco Garcia,² Michelle Whirl-Carrillo,¹ Mercedes Gallardo,^{8,10} Maria A. Blasco,⁹ Peter L. Greenberg,⁴ Phyllis Snyder,¹ Teri E. Klein,¹ Russ B. Altman,^{1,5} Atul J. Butte,² Euan A. Ashley,³ Mark Gerstein,^{5,7,8} Kari C. Nadeau,² Hua Tang,¹ and Michael Snyder^{1,*}

¹Department of Genetics, Stanford University School of Medicine

²Division of Systems Medicine and Division of Immunology and Allergy, Department of Pediatrics

³Center for Inherited Cardiovascular Disease, Division of Cardiovascular Medicine

⁴Division of Hematology, Department of Medicine

⁵Department of Bioengineering

Stanford University, Stanford, CA 94305, USA

⁶Program in Computational Biology and Bioinformatics

⁷Department of Molecular Biophysics and Biochemistry

⁸Department of Computer Science

Yale University, New Haven, CT 06520, USA

⁹Telomeres and Telomerase Group, Molecular Oncology Program, Spanish National Cancer Centre (CNIO), Madrid E-28029, Spain

¹⁰Life Length, Madrid E-28003, Spain

¹¹These authors contributed equally to this work

¹²Present address: Personalis, Palo Alto, CA 94301, USA

*Correspondence: mpsnyder@stanford.edu

DOI 10.1016/j.cell.2012.02.009

SUMMARY

Personalized medicine is expected to benefit from combining genomic information with regular monitoring of physiological states by multiple high-throughput methods. Here, we present an integrative personal omics profile (iPOP), an analysis that combines genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles from a single individual over a 14 month period. Our iPOP analysis revealed various medical risks, including type 2 diabetes. It also uncovered extensive, dynamic changes in diverse molecular components and biological pathways across healthy and diseased conditions. Extremely high-coverage genomic

INTRODUCTION

Personalized medicine aims to assess medical risks, monitor, diagnose and treat patients according to their specific genetic composition and molecular phenotype. The advent of genome sequencing and the analysis of physiological states has proven to be powerful (Cancer Genome Atlas Research Network, 2011). However, its implementation for the analysis of otherwise healthy individuals for estimation of disease risk and medical interpretation is less clear. Much of the genome is difficult to interpret and many complex diseases, such as diabetes, neurological disorders and cancer, likely involve a large number of different genes and biological pathways (Ashley et al., 2010; Grayson et al., 2011; Li et al., 2011), as well as environmental contributors that can be difficult to assess. As such, the combination of genomic information along with a detailed molecular analysis of samples will be important for predicting, diagnosing

(Mischel et al., 2003), and RNA sequencing (RNA-Seq) has identified breast cancer transcript isoforms (Li et al., 2011; van der Werf et al., 2007; Wu et al., 2010; Lapuk et al., 2010). Although transcriptome and RNA splicing profiling are powerful and convenient, they provide a partial portrait of an organism's physiological state. Transcriptomic data, when combined with genomic, proteomic, and metabolomic data are expected to provide a much deeper understanding of normal and diseased states (Snyder et al., 2010). To date, comprehensive integrative omics profiles have been limited and have not been applied to the analysis of generally healthy individuals.

To obtain a better understanding of: (1) how to generate an integrative personal omics profile (iPOP) and examine as many biological components as possible, (2) how these components change during healthy and diseased states, and (3) how this information can be combined with genomic information to estimate disease risk and gain new insights into diseased states, we performed extensive omics profiling of blood components from a generally healthy individual over a 14 month period (24 months total when including time points with other molecular analyses). We determined the whole-genome sequence (WGS) of the subject, and together with transcriptomic, proteomic, metabolomic, and autoantibody profiles, used this information to generate an iPOP. We analyzed the iPOP of the individual over the course of healthy states and two viral infections (Figure 1A). Our results indicate that disease risk can be estimated by a whole-genome sequence and by regularly monitoring health states with iPOP disease onset may also be observed. The wealth of information provided by detailed longitudinal iPOP revealed unexpected molecular complexity, which exhibited dynamic changes during healthy and diseased states, and provided insight into multiple biological processes. Detailed omics profiling coupled with genome sequencing can provide molecular and physiological information of medical significance. This approach can be generalized for personalized health monitoring and medicine.

RESULTS

Overview of Personal Omics Profiling

Our overall iPOP strategy was to: (1) determine the genome sequence at high accuracy and evaluate disease risks, (2) monitor omics components over time and integrate the relevant omics information to assess the variation of physiological states, and (3) examine in detail the expression of personal variants at the level of RNA and protein to study molecular complexity and dynamic changes in diseased states.

We performed iPOP on blood components (peripheral blood mononuclear cells [PBMCs], plasma and sera that are highly accessible) from a 54-year-old male volunteer over the course of 14 months (IRB-8629). The samples used for iPOP were taken over an interval of 401 days (days 0–400). In addition, a complete medical exam plus laboratory and additional tests were performed before the study officially launched (day 123) and blood glucose was sampled multiple times after the comprehensive omics profiling (days 401–602) (Figure 1A). Extensive sampling was performed during two viral infections that occurred during this period: a human rhinovirus (HRV) infection beginning on

day 0 and a respiratory syncytial virus (RSV) infection starting on day 289. A total of 20 time points were extensively analyzed and a summary of the time course is indicated in Figure 1A. The different types of analyses performed are summarized in Figures 1B and 1C. These analyses, performed on PBMCs and/or serum components, included WGS, complete transcriptome analysis (providing information about the abundance of alternative spliced isoforms, heteroallelic expression, and RNA edits, as well as expression of miRNAs at selected time points), proteomic and metabolomic analyses, and autoantibody profiles. An integrative analysis of these data highlights dynamic omics changes and provides rich information about healthy and diseased phenotypes.

Whole-Genome Sequencing

We first generated a high quality genome sequence of this individual using a variety of different technologies. Genomic DNA was subjected to deep WGS using technologies from Complete Genomics (CG, 35 nt paired end) and Illumina (100 nt paired end) at 150- and 120-fold total coverage, respectively, exome sequencing using three different technologies to 80- to 100-fold average coverage (see Extended Experimental Procedures available online) and analysis using genotyping arrays and RNA sequencing.

The vast majority of genomic sequences (91%) mapped to the hg19 (GRCh37) reference genome. However, because of the depth of our sequencing, we were able to identify sequences not present in the reference sequence. Assembly of the unmapped Illumina sequencing reads (60,434,531, 9% of the total) resulted in 1,425 (of 29,751) contigs (spanning 26 Mb) overlapping with RefSeq gene sequences that were not annotated in the hg19 reference genome. The remaining sequences appeared unique, including 2,919 exons expressed in the RNA-Seq data (e.g., Figure S1A). These results confirm that a large number of undocumented genetic regions exist in individual human genome sequences and can be identified by very deep sequencing and de novo assembly (Li et al., 2010).

Our analysis detected many single nucleotide variants (SNVs), small insertions and deletions (indels) and structural variants (SVs; large insertions, deletions, and inversions relative to hg19), (summarized in Table 1 and Experimental Procedures). 134,341 (4.1%) high-confidence SNVs are not present in dbSNP, indicating that they are very rare or private to the subject. Only 302 high-confidence indels reside within RefSeq protein coding exons and exhibit enrichments in multiples of three nucleotides ($p < 0.0001$). In addition to indels, 2,566 high-confidence SVs were identified (Experimental Procedures and Table S1) and 8,646 mobile element insertions were identified (Stewart et al., 2011).

Analysis of the subject's mother's genome by comprehensive genome sequencing (as above) and imputation allowed a maternal/paternal chromosomal phasing of 92.5% of the subject's SNVs and indels (see Extended Experimental Procedures for details). Of 1,162 compound heterozygous mutations in genes, 139 contain predicted compound heterozygous deleterious and/or nonsense mutations. Phasing enabled the assembly of a personal genome sequence of very high confidence (c.f., Rozowsky et al., 2011).

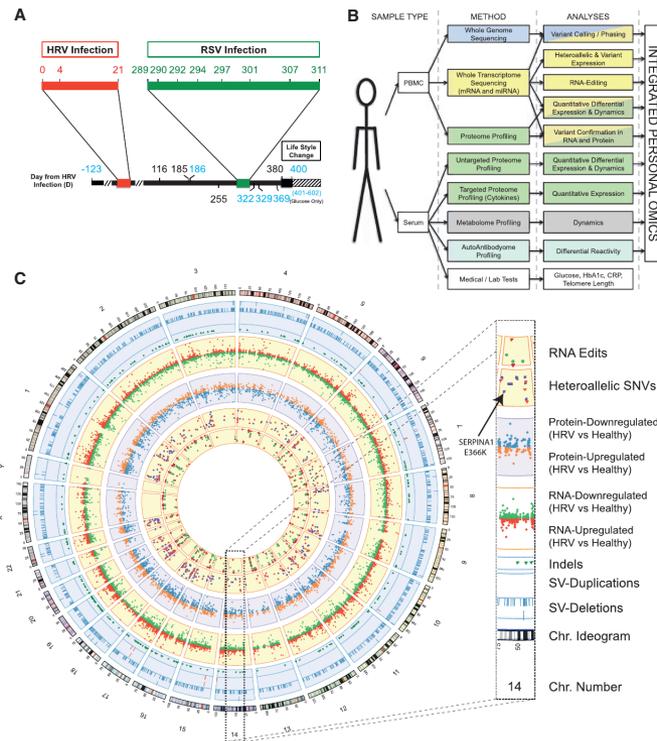


Figure 1. Summary of Study

(A) Time course summary. The subject was monitored for a total of 726 days, during which there were two infections (red bar, HRV; green bar, RSV). The black bar indicates the period when the subject: (1) increased exercise, (2) ingested 81 mg of acetylsalicylic acid and ibuprofen tablets each day (the latter only during the first 6 weeks of this period), and (3) substantially reduced sugar intake. Blue numbers indicate fasted time points.

(B) iPOP experimental design indicating the tissues and analyses involved in this study.

(C) Circos (Krzyszowski et al., 2009) plot summarizing iPOP. From outer to inner rings: chromosome ideogram; genomic data (pale blue ring), structural variants > 50 bp (deletions [blue tiles], duplications [red tiles], indels [green triangles]; transcriptomic data (yellow ring), expression ratio of HRV infection to healthy states; proteomic data (light purple ring), ratio of protein levels during HRV infection to healthy states; transcriptomic data (yellow ring), differential heteroallelic expression ratio of alternative allele to reference allele for missense and synonymous variants (purple dots) and candidate RNA missense and synonymous edits (red triangles, purple dots, orange triangles and green dots, respectively).

See also Figure S1.

WGS-Based Disease Risk Evaluation

We identified variants likely to be associated with increased susceptibility to disease (Dewey et al., 2011). The list of high confidence SNVs and indels was analyzed for rare alleles (<5% of the major allele frequency in Europeans) and for changes in genes with known Mendelian disease phenotypes (data summarized in Table 2), revealing that 51 and 4 of the rare coding SNV and indels, respectively, in genes present in OMIM are predicted

to lead to loss-of-function (Table S2A). This list of genes was further examined for medical relevance (Table S2A; example alleles are summarized in Figure 2A), and 11 were validated by Sanger sequencing. High interest genes include: (1) a mutation (E366K) in the *SERPINA1* gene previously known in the subject, (2) a damaging mutation in *TERT*, associated with acquired aplastic anemia (Yamaguchi et al., 2005), and (3) variants associated with hypertriglyceridemia and diabetes, such as *GCKR*

Table 1. Summary and Breakdown of DNA Variants

Type	Total Variants	Total High Confidence	Heterozygous High Confidence	Homozygous High Confidence
Total SNVs	3,739,701	3,301,521	1,971,629	1,329,892
Total gene-associated SNVs	1,312,780	1,183,847	717,485	466,362
Total coding/UTR	49,017	44,542	27,383	17,159
Missense	10,592	9,683	5,944	3,739
Nonsense	83	73	49	24
Synonymous	11,459	10,864	6,747	4,117
5' UTR	4,085	2,978	1,802	1,176
3' UTR	22,798	20,944	12,841	8,103
Intron	1,263,763	1,139,305	690,102	449,203
Ts/Tv	—	2.14	—	—
dbSNP	3,493,748	3,167,180	—	—
Candidate private SNV	245,953	134,341	—	—
Indels (−107~ +36 bp)	1,022,901	216,776	—	—
Coding	3,263	302	—	—
Structural variants (>50 bp)	44,781	2,566	—	—
In 1000G project ^a	4,434	1,967	—	—

High confidence values are from variants identified across multiple platforms (Illumina and CG) and/or Exome and RNA-Seq data. Annotations were based from variant call formatted (vcf) files for heterozygous calls: 0/1, reference (ref)/alternative (alt); 1/2, alt/alt and homozygous calls: 1/1, alt/alt; 1/(alt/alt-incomplete call). Polyphen-2 was used to identify the location of the SNVs.

^a1000G (1000 Genomes Project Consortium, 2010).

(homozygous) (Vaxillaire et al., 2008), and *KCNJ11* (homozygous) (Hani et al., 1998) and *TCF7* (heterozygous) (Erich et al., 2009).

Genetic disease risks were also assessed by the RiskOGRAM algorithm, which integrates information from multiple alleles associated with disease risk (Ashley et al., 2010) (Figure 2B). This analysis revealed a modest elevated risk for coronary artery disease and significantly elevated risk levels of basal cell carcinoma (Figure 2B), hypertriglyceridemia, and type 2 diabetes (T2D) (Figures 2B and 2C).

In addition to coding region variants we also analyzed genomic variants that may affect regulatory elements (transcription factors [TF]), which had not been attempted previously (Data S1). A total of 14,922 (of 234,980) SNVs lie in the motifs of 36 TFs known to be associated with the binding data (see Experimental Procedures), indicating that these are likely having a direct effect on TF binding. Comparison of SNPs that alter binding patterns of NFκB and Pol II sites (Kasowski et al., 2010), also revealed a number of other interesting regulatory variants, some of which are associated with human disease (e.g., *EDIL*) (Sun et al., 2010) (Figure S1B).

Medical Phenotypes Monitoring

Based on the above analysis of medically relevant variants and the RiskOGRAM, we monitored markers associated with high-risk disease phenotypes and performed additional medically relevant assays.

Monitoring of glucose levels and HbA1c revealed the onset of T2D as diagnosed by the subject's physician (day 369, Figures 2A and 2C). The subject lacked many known factors associated with diabetes (nonsmoker; BMI = 23.9 and 21.7 on day 0 and day 511, respectively) and glucose levels were normal for the first

part of the study. However, glucose levels elevated shortly after the RSV infection (after day 301) extending for several months (Figure 2D). High levels of glucose were further confirmed using glycated HbA1c measurements at two time points (days 329, 369) during this period (6.4% and 6.7%, respectively). After a dramatic change in diet, exercise and ingestion of low doses of acetylsalicylic acid a gradual decrease in glucose (to ~93 mg/dl at day 602) and HbA1c levels to 4.7% was observed. Insulin resistance was not evident at day 322. The patient was negative for anti-GAD and anti-islet antibodies, and insulin levels correlated well with the fasted and nonfasted states (Figure S2C), consistent with T2D. These results indicate that a genome sequence can be used to estimate disease risk in a healthy individual, and by monitoring traits associated with that disease, disease markers can be detected and the phenotype treated.

The subject contained a *TERT* mutation previously associated with aplastic anemia (Yamaguchi et al., 2005). However, measurements of telomere length suggested little or no decrease in telomere length and modest increase in numbers of cells with short telomeres relative to age-matched controls (Figures S2A and S2B). Importantly, the patient and his 83-year-old mother share the same mutation but neither exhibit symptoms of aplastic anemia, indicating that this mutation does not always result in disease and is likely context specific in its effects.

Consistent with the elevated hypertriglyceridemia risk, triglycerides were found to be high (321 mg/dl) at the beginning of the study. These levels were reduced (81–116 mg/dl) after regularly taking simvastatin (20 mg/day).

We also examined the variants for their potential effects on drug response (see Extended Experimental Procedures). Among the alleles of interest, (Figure 2A and Table S2B) two genotypes affecting the *LPIN1* and *SLC22A1* genes were associated with

Table 2. Summary of Disease-Related Rare Variants

Category	Count
Total high confidence rare SNVs	289,989
Coding	2,546
Missense	1,320
Synonymous	1,214
Nonsense	11
Nonstop	1
Damaging or possibly damaging	233
Putative loss-of-function SNVs ^a	51
Total high confidence rare indels	51,248
Coding indels	61
Frameshift indels	27
miRNA indels	3
miRNA target sequence indels	5
Putative loss-of-function indels ^a	4

^aIn curated Mendelian disease genes.

favorable (glucose lowering) responses to two diabetic drugs, rosiglitazone and metformin, respectively.

We followed the levels of 51 cytokines along with the C-reactive protein (CRP) using ELISA assays, which revealed strong induction of proinflammatory cytokines and CRP during each infection (Figures 2E and 2F). We also observed a spike of many cytokines at day 12 after the RSV infection (day 301 overall). These data define the physiological states and serve as a valuable reference for the omic profiles integrated into a longitudinal map of healthy and diseased states described in the next sections.

We also profiled autoantibodies during the HRV infection. Plasma and serum samples from the first four time points (days -123, 0, 4 and 21), along with plasma samples from 34 healthy controls were used to probe a protein microarray containing 9,483 unique human proteins spotted in duplicate. A total of 884 antigens with increased reactivity (Data S2) in the candidate plasma relative to healthy controls were found ($p < 0.01$, Benjamini-Hochberg $p < 0.01$). Among the potentially interesting results was high reactivity with DOK6, an insulin receptor binding protein (NCBI gene database). These results demonstrate that autoantibodies can be monitored and that information relevant to disease conditions can be found.

Dynamic Omics Analysis: Integrative Omics Profiling of Molecular Responses

We profiled the levels of transcripts, proteins, and metabolites across the HRV and RSV infections and healthy states using a variety of approaches. RNA-Seq of 20 time points generated over 2.67 billion uniquely mapped 101b paired-end reads (123 million reads average per time point) and allowed for an analysis of the molecular complexity of the transcriptome in normal cells (PBMCs) at an unprecedented level. The relative levels of 6,280 proteins were also measured at 14 time points through differential labeling of samples using isobaric tandem mass tags (TMT), followed by liquid chromatography and mass spectrometry (LC-MS/MS) (Cox and Mann, 2010; Theodoridis

et al., 2011). A total of 3,731 PBMC proteins could be consistently monitored across most of the 14 time points (see Figure S3A and Data S3). In addition, 6,862 and 4,228 metabolite peaks were identified for the HRV and RSV infection, and a total of 1,020 metabolites were tracked for both infections (see Figure S4 and Data S4, [3]). Finally, as described below, we also analyzed miRNAs during the HRV infection.

This wealth of omics information allowed us to examine detailed dynamic trends related directly to the physiological states of the individual and revealed enormous changes in biological processes that occurred during healthy and diseased states. For each profile (transcriptome, proteome, metabolome), we systematically searched for two types of nonrandom patterns: (1) correlated patterns over time and (2) single unusual events (i.e., spikes that may occur at any given time point defined as statistically significantly high or low signal instances compared to what would be expected by chance). To perform this analysis, we developed a general scheme for integrated analysis of data (see Figure S5 and Extended Experimental Procedures for further details). We used a Fourier spectral analysis approach that both normalizes the various omics data on equal basis for identifying the common trends and features, and, also accounts for data set variability, uneven sampling, and data gaps, in order to detect real-time changes in any kind of omics activity at the differential time points (see Supplemental Information). Autocorrelations were calculated to assess nonrandomness of the time-series ($p < 0.05$ one-tailed based on simulated bootstrap nonparametric distribution by sampling with replacement of the original data, $n > 100,000$), with significant signals classified as autocorrelated (I). The remaining data was searched for spike events, which were classified as spike maxima (II) or spike minima (III) ($p < 0.05$ one-tailed based on differences from simulated, $n > 100,000$ random distribution of the time-series). After classification, the data were agglomerated into hierarchical clusters (using correlation distance and average linkage) of common patterns and biological relevance was assessed through GO (Ashburner et al., 2000) analysis (Cytoscape [Smoot et al., 2011], BINGO [Maere et al., 2005] $p < 0.05$, Benjamini-Hochberg [Benjamini and Hochberg, 1995] adjusted $p < 0.05$) and pathway analysis (Reactome [Croft et al., 2011] functional interaction [F]), networks including KEGG [Kanehisa and Goto, 2000; Smoot et al., 2011], $p < 0.05$, FDR < 0.05). The unified framework approach was implemented on all the different data sets both individually and in combination, and our results revealed a number of differential changes that occurred both during infectious states and the varying glucose states.

We first analyzed the different individual transcriptome, proteome (serum and PBMC) and metabolome data sets; the proteome and metabolome results are presented in the Supplemental Information (Figures S3, S4, S6 and Data S3–S6). A total of 19,714 distinct transcript isoforms (Wang et al., 2008) corresponding to 12,659 genes (Figure S1C) were tracked for the entire time course, and their dynamic expression response was classified into either autocorrelated (I) and spike sets, further subdivided as displaying maxima (II) or minima (III) (Figure 3). The clustering and enrichment analysis displayed a number of interesting pathways in each class. In the autocorrelated group (Figure 3B, [I]; see also Figure S6A and Data S6, [1 and 2]), we

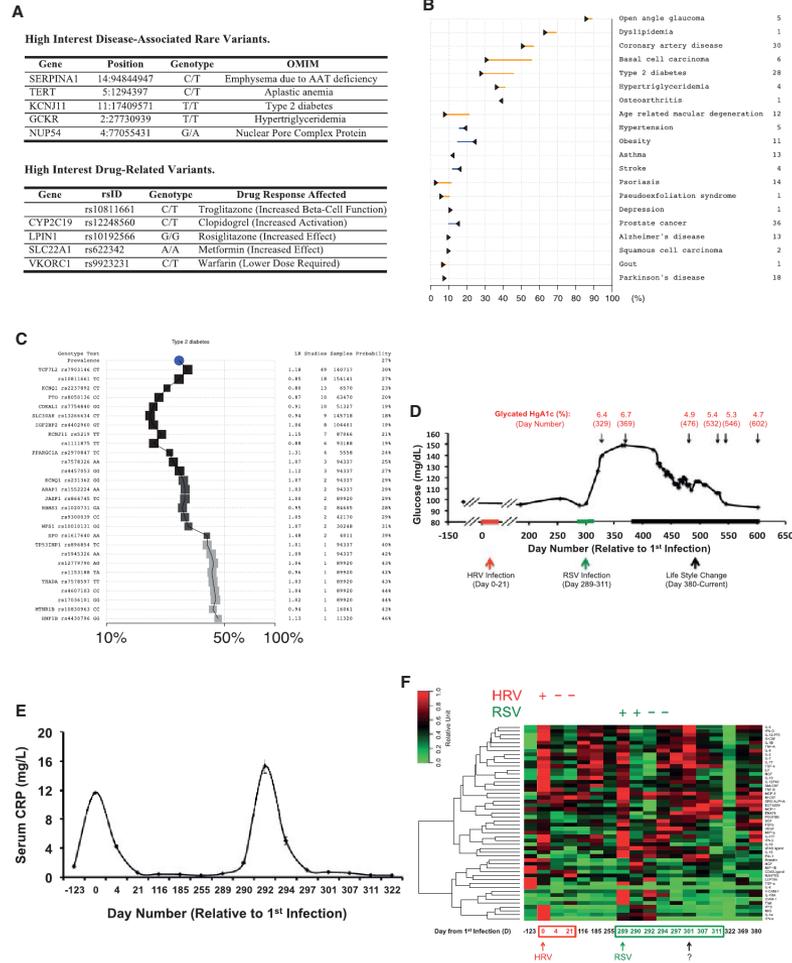


Figure 2. Medical Findings

(A) High interest disease- and drug-related variants in the subject's genome.
 (B) RiskGraph of the top 20 diseases with the highest posttest probabilities. For each disease, the arrow represents the pretest probability according to the subject's age, gender, and ethnicity. The line represents the posttest probability after incorporating the subject's genome sequence. Listed to the right are the numbers of independent disease-associated SNVs used to calculate the subject's posttest probability.
 (C) RiskOgram of type 2 diabetes. The RiskOgram illustrates how the subject's posttest probability of T2D was calculated using 28 independent SNVs. The middle graph displays the posttest probability. The left side shows the associated genes, SNVs, and the subject's genotypes. The right side shows the likelihood ratio (LR), number of studies, cohort sizes, and the posttest probability.
 (D) Blood glucose trend. Measurements were taken from samples analyzed at either nonfasted or fasted states; the nonfasted states (all but days 186, 322, 329, and 369 and after day 400) were at a fixed time after a constant meal. Data was presented as moving average with a window of 15 days. Red

found two main trends: an upward trend (2,023 genes), following the onset of the RSV infection, and a similar coincidental downward trend (2,207 genes). The upward autocorrelated trend revealed a number of pathways as enriched ($p < 0.002$, $FDR < 0.05$), including protein metabolism and influenza life cycle. Additionally, the downward autocorrelation cluster showed a multitude of enriched pathways ($p < 0.008$, $FDR < 0.05$), such as TCR signaling in naive CD4+ T cells, lysosome, B cell signaling, androgen regulation, and of particular interest, insulin signaling/response pathways. These different pathways, which are activated as a response to an immune infection, often share common genes and additionally we observe many genes hitherto unknown to be involved in these pathways but displaying the same trend. Furthermore, we observed that the downward trend, that began with the onset of the RSV infection and appeared to accelerate after day 307, coincided with the beginning of the observed elevated glucose levels in the subject.

In the dynamic spike class we again saw patterns that were concordant with phenotypes (Figure 3B, [II] and [III]; see also Figure S6A and Data S6, [3–14]). A set of expression spikes displaying maxima (547 genes), that are common to the onset of both the RSV and HRV infections are associated with phagosome, immune processes and phagocytosis, ($p < 1 \times 10^{-4}$, $FDR < 6 \times 10^{-3}$). Furthermore, a cluster that exhibits an elevated spike at the onset of the RSV infection involves the major histocompatibility genes ($p < 7 \times 10^{-4}$, Benjamini-Hochberg adjusted $p < 0.03$). A large number of genes with a coexpression pattern common to both infections in the time course have yet to be implicated in known pathways and provide possible connections related to immune response. Finally, our spike class displaying minima showed a distinct cluster (1,535 genes) singular to day 307 (day 14 of the RSV infection), associated with TCR signaling again, TGF receptors, and T cell and insulin signaling pathways ($p < 0.02$, $FDR < 0.03$). Overall, the transcriptome analysis captures the dynamic response of the body responding to infection as also evidenced by our cytokine measurements, and also can monitor health changes over long periods of time, with various trends.

To further leverage the transcriptome and genome data, we performed an integrated analysis of transcriptome, proteomic and metabolomics data for each time point, observing how this corresponded to the varying physiological states monitored as described in the above sections. Because of the availability of many time points through the course of infection, we examined in detail the onset of the RSV infection, as well as extended our complete dynamics omics profile during the times that our subject began exhibiting high glucose levels. Figure 4 shows an integrated interpretation of omics data (see also Figure S6B and Data S7), where all trends are combined for each omics data set and the common patterns emerge providing complementary information. In addition to the common patterns

observed in our transcriptome analysis, new patterns emerged, some unique to protein data, some to metabolite, and some common to all. In particular we found the following interesting results: for autocorrelated clusters we found the same trends as observed in the transcriptome, additionally augmented with concordant protein expressions. Pathways such as the phagosome, lysosome, protein processing in endoplasmic reticulum, and insulin pathways emerged as significantly enriched ($p < 0.002$, $FDR < 0.0075$), and showed a downward trend post-infection, and further accelerated after ~3 weeks following the initial onset of the RSV infection (this cluster comprised of 1,452 transcriptomic and 69 proteomic components, corresponding to 1,444 genes). The elevated spike class showed a maxima cluster on day 18 post RSV infection (one time point after the cytokine maximum), with enrichment in pathways such as the spliceosome, glucose regulation of insulin secretion, and various pathways related to a stress response ($p < 1 \times 10^{-4}$, $FDR < 0.02$)—this cluster included 1,956 transcriptomic, 571 proteomic and 23 metabolomic components, corresponding to 2,344 genes. Even though current proteomic information is more limited than the full transcriptome because it follows fewer components, as evidenced in Figure 4 (II), several pathways, including the glucose regulation of insulin secretion pathway, clearly emerge from the proteomic information and would not have been observed by only monitoring the transcriptome. Additionally, in this cluster we find significant GO enrichment in splicing and metabolic processes ($p < 6 \times 10^{-47}$, Benjamini-Hochberg adjusted $p < 10^{-45}$). Furthermore, inspection of metabolites reveals 23 that show the same exact trend (i.e., spikes at day 18 post RSV infection); at least one, lauric acid has been implicated in fatty acid metabolism and insulin regulatory pathways (Kusunoki et al., 2007). Finally, we observe minima spikes as well, with yet another interesting group on day 18, which showed downregulation in several pathways ($p < 0.003$, $FDR < 0.05$), such as the formation of platelet plug. This cluster displayed a high degree of synergy between the various omics data, comprised of 3,237 transcriptomic and 761 proteomic components corresponding 3,400 genes and 83 metabolomic components.

In summary, our integrated approach revealed a clear systemic response to the RSV infection following its onset and postinfection response, including a pronounced response evident at day 18 post RSV infection. A variety of infection/stress response related pathways were affected along with those associated to the high glucose levels in the later time points, including insulin response pathways.

Dynamic Omics Analysis: Extensive Heteroallelic Variation and RNA Editing

The considerable amount of transcriptome and proteome data allowed us to analyze and follow changes in allele-specific

and green arrows and bars indicate the times of the HRV and RSV infections, respectively. Black arrows and bars indicate the period with life style changes.

(E) C-reactive protein trend line. Error bars represent standard deviation of three assays.

(F) Serum cytokine profiles. Red box and day number, HRV infection; green box and day number, RSV infection; question mark, elevated cytokine levels indicating an unknown event at day 301. Red is increased cytokine levels.

See also Figure S2.

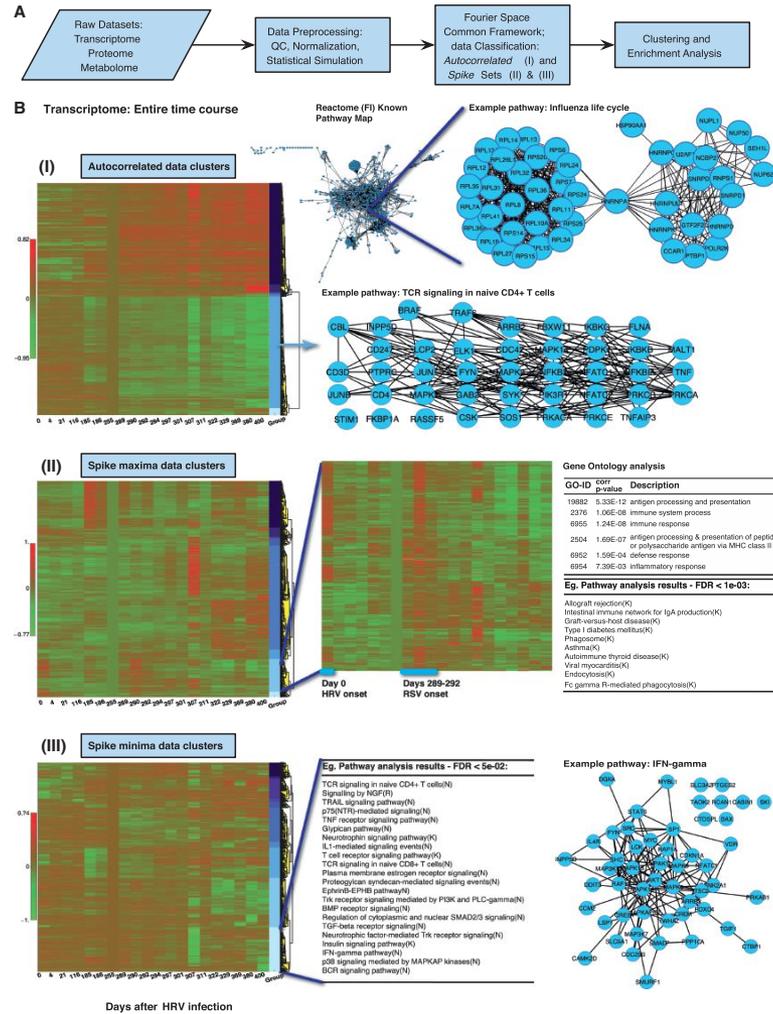


Figure 3. Transcriptome Time Course Analysis
 (A) Summary of approach for identification of differentially expressed components. The various omics sets were processed through a common framework involving spectral analysis, clustering, and pathway enrichment analysis.
 (B) Pattern classification. The different emergent patterns from the analysis of the transcriptome for the entire time course are displayed for the autocorrelation (I), spike maxima (II), and spike minima (III) classes. For different clusters, examples of gene connections in selected pathways based on Reactome (Croft et al., 2011) FI (Cytoscape plugin [Smoot et al., 2011]) are shown as networks. Example GO (Ashburner et al., 2000) enrichment analysis results from Cytoscape (Smoot et al., 2011) BiNGO (Maere et al., 2005) plugin and pathway enrichment results (Reactome FI [Croft et al., 2011]) are included.
 See also Figures S5 and S6.

expression (ASE), splicing, and editing at the RNA and protein levels during healthy and diseased states.

Of the 49,017 genomic variants associated with coding or UTR regions (Table 1), 12,785 (26%) were expressed in PBMCs (≥ 40 read coverage; Table S3). A total of 8,509 of the variants are heterozygous (1,113 missense) and the remainder (4,686; 684 missense) are homozygous. Eight of the 83 nonsense mutations were expressed indicating that not all nonsense mutations result in transcript loss.

The numerous heterozygous variants allowed an analysis of the dynamics of differential ASE, (shrunk ratios, Experimental Procedures; Figures 5A and S7B) in PBMCs during healthy and diseased states. We found 497 and 1,047 genes that exhibited differential ASE during HRV and RSV infection, respectively (posterior probability ≥ 0.75 , beta-binomial model; ≥ 40 reads, ≥ 7 time points); many of these are immune response genes, e.g., *PADI4* and *PLOD1* (Figure 5B). Among the differential ASE sites 100 and 218 were specific to HRV and RSV infected states, respectively (Figures 5C and 5D). Differential ASE genes in the HRV compared to healthy phase were enriched for those encoding SNARE vesicular transport proteins (DAVID analysis; Benjamini $p < 0.05$). Summing over all computed ASE alternative to total ratios revealed that nonreference heteroallelic variants were expressed at 98% of reference variants. The expression of over 50 heterozygous variants, including some of the rare/private SNVs (which form 0.72% of the genomic total), and differentially expressed variants (*SVIL* and *TRIM5*), was confirmed by Sanger cDNA sequencing and/or digital PCR (Hindson et al., 2011) of cDNA (Figures 5B and S7). Overall, these results demonstrate that differential ASE is pervasive in humans and is particularly distinct during healthy and infected states, with many of these changes residing in immune response genes.

The depth of our RNA-Seq data enabled us to re-evaluate the extent of RNA editing (Figure 6 and Data S8 and S11A), typically an adenosine to inosine (A-to-I) conversion (Li et al., 2009b) or infrequently cytidine to uridine (C-to-U), in normal human cells. We found 2,376 high-confidence coding-associated RNA edits, including 795 A-to-I (A-to-G) and 277 C-to-U deamination-like edits (Figure 6A). A total of 587 edits in 175 genes were predicted to cause amino acid substitutions (Polyphen-2 [Adzhubei et al., 2010]); the remainder were nonsense (11), synonymous (435), or located in 5'/3' UTRs (103/1,240). Ten edited bases causing amino acid substitutions were validated by Sanger cDNA sequencing and/or digital droplet PCR, as well as by identification of their peptide counterparts by mass spectrometry (Figure 6B). Interestingly, we identified A-to-G edits (Figure 6B), e.g., *IGFBP7*, *BLCAP*, and *AZIN1* in PBMCs that were known to occur in other tissues (Gommans et al., 2008; Levanon et al., 2005), indicating that the same RNA can be edited in other cell types. *BLCAP* exhibited two edited changes (Figure 6C) with edited/total ratios of 0.12–0.2 and 0.18–0.31, respectively, comparable to the 0.21 ratio previously observed in the brain (Galeano et al., 2010).

Furthermore, we found and validated two missense-causing edits, U-to-C in *SCFD2* and G-to-A in *FBXO25* (Figure 6D), indicating an amination-like RNA-editing mechanism, previously not observed in human cells. Our results reveal that a large number of edits occur and exhibit dynamic and differential changes in

populations of PBMCs (Figure 6B). The total number of edited RNAs, while extensive, is significantly lower than that reported in human lymphoblastoid lines and very different in its distribution (Li et al., 2011). We believe that in addition to tissue-specific variation, the observed differences are also likely due to overcalling of false-positive SNVs, a problem we corrected with deep exome sequencing, removal of repeat regions and pseudogenes, and strings of close-proximity variants (Data S11A).

Finally, to determine whether the nonreference allele and edited RNAs serve as templates for protein synthesis, we generated proteome databases for 4,586 missense SNVs and all 30,385 edits and used them to search our mass spectra from the untargeted protein profiling experiments as well as in a targeted approach to directly search for 500 edited proteins (see Extended Experimental Procedures). Peptides for 48 SNVs and 51 edits were identified (FDR < 0.01 and requiring one unique peptide per protein; Data S9 and S11B). A total of 17/17 selected SNVs (100%) were validated by Sanger sequencing. Seven peptides derived from the SNV and six peptides derived from edited transcripts were unique to a single protein in the IPI database (Kersey et al., 2004) and classified as high confidence. These results indicate that a large fraction of personal variants are expressed as transcripts and a number of these are also translated as proteins.

miRNA Variant Analysis

In addition to the omics profiling above, we identified 619–681 known miRNAs from PBMCs per time point (>10 reads, days 4, 21, 116, 185, and 186), 106 of which showed dynamic changes (e.g., Figures S2D and S2E). Examination of miRNA editing revealed 50 edited miRNAs (C-to-U or A-to-I) with stringent criteria (edited reads $> 5\%$ of total reads or > 399 modified reads) indicating that at least $\sim 4\%$ of expressed miRNAs are potentially edited. Eighteen miRNAs contain edits located within the functionally critical “seed sequences,” potentially affecting their mRNA targets. Interestingly, expression of SNV-containing miRNAs was generally higher compared to SNV-free miRNA (Figures 6E and 6F). In addition to edits, analysis of the SNVs located in miRNAs revealed that most (25 of 31) SNV-containing miRNAs were not expressed. These miRNAs were among those discovered in cancer cell lines (Jima et al., 2010) and may not normally be highly expressed in PBMCs from healthy individuals.

DISCUSSION

To our knowledge, our study is the first to perform extensive personal iPOP of an individual through healthy and diseased states. It revealed extensive complex and dynamic changes in the omics profiles, especially in the transcriptomes, between healthy states and viral infections, and between nondiabetic and diabetic states. iPOP provides a multidimensional view of medical states, including healthy states, response to viral infection, recovery, and T2D onset. Our study indicates that disease risk can be assessed from a genome sequence and illustrates how traits associated with disease can be monitored to identify varying physiological stages. We show that large numbers of molecular components are present in blood samples and can

Although we analyzed a single individual, insights were gained by integrating the multiple omics profiles associated with distinct physiological states. Through examination of molecular patterns, clear signatures of dynamic biological processes were evident, including immune responses during infection, insulin signaling response alterations after the RSV infection. Indeed, careful monitoring of omics changes across multiple time points for the same individual revealed detailed responses, which might not have been evident had the analyses been performed on groups due to interindividual variability. Hence, we expect that our longitudinal personalized profiling approach provides valuable information on an individual basis.

We focused on a generally healthy subject who exhibited no apparent disease symptoms. This is a critical aspect of personalized medicine, which is to perform iPOP and evaluate the importance and changes of all the profiles in ordinary individuals. These results have important implications and suggest new paradigm shifts: first, genome sequencing can be used to direct the monitoring of specific diseases (in this study, aplastic anemia and diabetes) and second, by following large numbers of molecules a more comprehensive view of disease states can be analyzed to follow physiological states.

Our study revealed that many distinct molecular events and pathways are activated both through viral infection and the onset of diabetes. Indeed, the monitoring of large numbers of different components revealed a steady decrease of insulin-related responses that are associated with diabetes-insulin response pathways occurring from the early healthy state to a high glucose state. Although many of the activated and repressed pathways could be detected through transcript profiling, some were detected only with the proteomics data and some with the combined set of data. In addition a large number of connections with diabetes and insulin signaling using metabolites, miRNAs, and autoantibodies were observed. One particularly interesting response detected with the proteomics data was the onset of the elevated glucose response that was tightly associated with the RSV infection and a particular subclinical response at day 12/18 postinfection. It is tempting to speculate that the RSV infection and/or the associated event at day 12/18 triggered the onset of high glucose/T2D. Although viral infections have been associated with T1D (van der Werf et al., 2007), we are unaware of viral infection associated with T2D. Inflammation and activated innate immunity have been associated with T2D (Pickup, 2004), and we speculate that perhaps RSV triggered aberrant glucose metabolism through activation of a viral inflammation response in conjunction with a predisposition toward T2D. Although this cannot be proven with the analyses from a single individual, this study nonetheless serves as proof-of-principle that iPOP can be performed and provide valuable information. Because diabetes is a complex disease there may be many ways to acquire high glucose phenotype; longitudinal iPOP analysis of a large number of individuals may be extremely valuable to dissecting the disease and its various subtypes, as well providing information into the molecular mechanism of its onset.

Finally, we believe that the wealth of data generated from this study will serve as a valuable resource to the community in the developing field of personalized medicine. A large database with the complete time-dynamic profiles for more individuals

that acquire infections and other types of diseases will be extremely valuable in the early diagnostics, monitoring and treatment of diseased states.

EXPERIMENTAL PROCEDURES

The subject and mother in this study were recruited under the IRB protocol IRB-8629 at Stanford University. Full methods and associated references can be found in the Extended Experimental Procedures section.

WGS was performed at Complete Genomics and Illumina. High-confidence SNVs were mostly correct as evidenced by: (1) Illumina Omni1-Quad genotyping arrays (99.3% sensitivity), (2) a Ti/Tv ratio of 2.14 as expected (1000 Genomes Project Consortium, 2010), (3) Illumina capture and DNA sequencing (92.7% accuracy), and (4) Sanger sequencing of 36 randomly selected SNVs (36/36 validated, Table S1). In contrast, the low confidence SNVs had a Ti/Tv of only 1.46 and an accuracy of 63.8% (19 of 33 confirmed by Sanger sequencing, Table S1A). Similarly, the majority of the 216,776 high-confidence indels are likely to be correct as (1) Sanger sequencing validated 14 of 15 (93%) tested indels and (2) exome-sequencing validated most indels (4,706, 82%); meanwhile the 806,125 low confidence indels had a low validation rate (5,225, 0.65%). SVs were called using: (1) paired-end mapping (Chen et al., 2009) (2) read depth (Abyzov et al., 2011), (3) split reads (Ye et al., 2009), and (4) junction mapping (Lam et al., 2010) to the breakpoint junction database from the 1000 G (Mills et al., 2011). A total of 2,566 were found by two different methods or platforms (CG or Illumina) and were called high confidence; >90% of these were in the database of genome variants.

Strand-specific RNA-Seq libraries were prepared as described previously (Parkhomchuk et al., 2009) and sequenced on 1–3 lanes of Illumina's HiSeq 2000 instrument. The TopHat package (Trapnell et al., 2009) was used to align the reads to the hg19 reference genome, followed by Cufflinks for transcript assembly and RNA expression analysis (Trapnell et al., 2010). The Samtools package (Li et al., 2009a) was used to identify variants including single nucleotide variants (SNV) and indels. Small RNAs were prepared from PBMCs for the first five time points; sequencing was performed according to Illumina's Small RNA v1.5 Sample Preparation Guide.

The Luminex 51-plex Human Cytokines assay was performed at the Stanford Human Immune Monitoring Center. For mass spectrometry, proteins were prepared from PBMC cell lysates, labeled at lysines using the TMT isobaric tags by Pierce, and digested with trypsin and analyzed using reverse phase LC coupled to a Thermo Scientific (LTQ)-Orbitrap Velos instrument. In order to profile serum, 14 major glycoproteins were first removed using the Agilent Human 14 Multiple Affinity Removal System (MARS) column in order to analyze the less abundant constituents. Metabolites were extracted by four times serum volume of equal mixture of methanol, acetonitrile, and acetone and separated using our Agilent 1260 liquid chromatography. Hydrophobic molecules were profiled using reversed phase UPLC followed by APCI-MS and hydrophilic molecule were analyzed using HILIC UPLC followed by ESI-MS in either the positive or negative mode.

For the integrated analysis, per omics set, for each time-series curve the Lomb-Scargle transformation (Hocke and Kämpfer, 2009; Lomb, 1976; Scargle, 1982, 1989) for unevenly sampled gapped time-series data was implemented (Ahdemski et al., 2007; Glynn et al., 2006; Van Dongen et al., 1999; Yang et al., 2011; Zhao et al., 2008). This allowed us to obtain a periodogram, which was used to calculate autocorrelations and then reconstruct the time-series with even sampling, allowing standard time-series analysis and performing data clustering, while taking the time intervals into account (see Extended Experimental Procedures).

Autoantibodyome profiling was performed using the Invitrogen ProtoArray Protein Microarray v5.0 according to the manufacturer's instructions.

ACCESSION NUMBERS

The SRA accession number for the WGS sequence reported in this paper is SRP008054.4. The GEO accession number for the RNA-Seq and miRNA-Seq data sequence reported in this paper is GSE33029. See Extended Experimental Procedures for data dissemination details.