Extended Experimental Procedures

Clinical protocol

Healthy volunteers over the age of 18 were enrolled on the National Institutes of Health (NIH) protocol 09-H-0239 (Clinicaltrials.gov NCT01191853), approved and monitored by NIH institutional review boards in accordance with the Declaration of Helsinki. Healthy volunteers were screened for protocol enrollment with a medical history, physical examination, and clinical laboratory studies (CBC with differential, blood chemistry, coagulation and thrombosis screens (PT, PTT, D-dimer), cholesterol panel, urinalysis, and pregnancy test). Pregnant individuals and those who had received vaccines or taken immune modifying medications within six months of study entry were excluded. Individuals meeting inclusion criteria signed written consent and were vaccinated with the 2009 Fluvirin seasonal influenza (Novartis), and H1N1 pandemic (Sanofi-Aventis) vaccines, both without adjuvant. Blood samples were obtained at day -7 (50ml) prior to vaccination, day 0 (150ml) immediately before vaccination, day1 (40ml), day7 (150ml), and day70 (150ml) post vaccination. All blood samples were drawn between 8am and 11am from fasting individuals and were processed within 30 minutes of drawing.

Sample processing

Blood specimens were collected in heparinized syringes and peripheral blood mononuclear cells (PBMC) were isolated using Leucosep tubes with Ficoll-Paque Plus density gradient media and centrifugation according to standard operational procedures (SOP) (http://www.nhlbi.nih.gov/resources/chi/documents/SOP-Isolation.pdf). Fresh isolated PBMC for RNA isolation were directly lysed in 700ul of Qiazol and stored at -80^{o}C till isolation. Human PBMCs used for deep phenotyping by 15 color Flow Cytometry analysis and B cell ELIspot were cryopreserved in freezing medium consisting of the intra-cellular cryo-protecting agent dimethylsulfoxide (DMSO) 10% plus 90% heat inactivated Fetal Bovine Serum. Cells suspended in the freezing medium were cryopreserved using Planer 750Plus Controlled Rate Freezer at control-rate steps to a temperature of -120°C to minimize cell damage, and were then transferred into a liquid nitrogen (LN2) tank in the vapor phase of the tank at -156°C (+/- 20°C) until use. Serums were collected at each time points using 8ml SST tubes (BD) according to SOPs (http://www.nhlbi.nih.gov/resources/chi/documents/SOP-Serum-SSTTubes.pdf). The use of frozen samples permitted all time-points for a given individual to be examined within a single analytic run for each assay, thus minimizing batch effects.

RNA isolation and microarray hybridization

Total RNA was isolated from 1e7 freshly isolated PBMC using miRNeasy kit (Qiagen) according to the manufacturer's instructions and eluted in 40ul of elution buffer with the addition of 1ul of RNase inhibitor (Invitrogen). RNA quality was assessed by Agilent Bioanalyzer and quantified by Nanodrop 2000. 300ng of total RNA were amplified using Ambion WT expression kit (Invitrogen) according to the manufacturer's instructions. Fragmented single-stranded sense cDNA were terminally labeled and hybridized to Human 1.0 ST GeneChip arrays (Affymetrix) and stained on a Genechip Fluidics Station 450

(Affymetrix), according to the respective manufacturers' instructions. Arrays were scanned on a GeneChip Scanner 3000 7G (Affymetrix).

Flow cytometry for immunophenotyping

Frozen PBMCs were washed and re-suspended in PBS. Viability staining was performed for 15 min in presence of LIVE/DEAD Aqua fixable viability dye (Life Sciences, Carlsbad, CA) followed by a wash in FACS staining buffer (PBS supplemented with 1% Normal Mouse serum, 1% Goat serum and 0.02% sodium azide) (Gemini Bioproducts, West Sacramento, CA)). Unstimulated cells were stained according to published protocols (Biancotto et al., 2011) (Biancotto et al., 2012) for 5 tubes of the CLIP panel (T_{reg}, Th_{17}, Th_1/Th_2, $B_{naive/memory}$ and a modified DC tube with addition of CD14-Qdot 655 to identify monocytes). To minimize variability, staining and acquisition of all time points from a given patient were performed on the same day, using a common antibody mixture. Data were acquired on a Becton Dickinson LSRFortessa (BD, San Jose, CA) equipped with five lasers (355 nm, 407 nm, 488 nm, 532 nm, and 633 nm wavelengths) with 22 PMT detectors, optimized as described by Perfetto et al. (Perfetto et al., 2006). Data were acquired using DIVA 6.1.2 software (BD). A minimum of 50,000 CD4 T cells, 20,000 B cells or 20,000 monocytes was recorded from corresponding tubes in order to accurately assess minor cell populations. Compensation was performed with unstained cells and BD compensation beads particle sets (using positive beads only). Compensation was used during acquisition of the specimen to ensure recording of enough events for the populations of interest. In addition, a final compensation matrix was calculated using FlowJo 9.4 (Treestar Inc., San Carlos, CA) during post-acquisition analysis. Debris and doublets were excluded using light scatter measurements and major cell populations were identified based on their forward and side scatter properties as described in (Biancotto et al., 2011). Subsequently, cells were gated using the viability stain and CD45 to ensure that only viable lymphocytes were included for analysis. Each cell population was represented as percentages of the parent population (i.e T cells as % of CD45+ live cells, Treg as % of total CD4+ T cells etc) as indicated in the tables and figures.

Influenza microneutralization titers

Virus-neutralizing titers of pre- and post-vaccination sera were determined in a microneutralization assay based on the methods of the pandemic influenza reference laboratories of the Centers for Disease Control and Prevention (CDC) (Hancock et al., 2009) using low pathogenicity vaccine viruses and MDCK cells. The X-179A virus is a 5:3 reassortant vaccine containing the HA, NA, and PB1 genes from A/California/07/2009 (H1N1pdm09) and the 5 other genes from A/PR/8/34 were donated by the high growth virus NYMC X-157. Immune sera were also tested for neutralization titers of the seasonal vaccine strains H1N1 A/Brisbane/59/07, H3N2 A/Uruguay/716/07, and B/Brisbane/60/2001. Internal controls in all assays were sheep sera generated against the corresponding strains at the Center for Biologics Evaluation and Research, FDA, Bethesda, MD. All individual sera were serially diluted (2-fold dilutions starting at 1:10) and were assayed against 100 TCID50 of each strain in duplicates in 96-well plates (1:1 mixtures). The titers represent the highest dilution that completely suppressed virus replication.

ELISpots for plasmablasts and memory B cells

Total and influenza-specific IgG/A frequencies of antibody-secreting cells (ASC) were measured by ELISpot assays as previously described (Ho et al., 2011), with the following details and modifications. Cryopreserved PBMCs were thawed, washed twice and the number of viable cells enumerated with an automated FACS-based instrument (Millipore). A portion of the PBMCs were examined by flow cytometry using CD19, CD20, CD27, CD21, CD10 and CD95 or CD19, CD20, IgM, IgD, IgG and IgA for immunophenotyping of B-cell markers and normalization of B cell numbers. Coating antigens for ELISpots included: anti-human light chain Abs to enumerate total ASCs; seasonal or H1N1 influenza vaccines to enumerate influenza-specific ASCs; and keyhole limpet hemocyanin (KLH) for a background control. For effector ASC, frequencies measured at Days 0 and 7, PBMCs were incubated directly on the ELISpot plates for 5 h, followed by detection of spots. For memory ASC frequencies measured at Days 0 and 70, PBMCs were first cultured for 4 days with B-cell polyclonal stimuli *Staphylococcus aureus* Cowan I (SAC) particles (EMD Biosciences) and CpG oligonucleotide (ODN2006, Operon) (Crotty et al., 2004). Cultured cells were then washed once, enumerated, immunophenotyped with CD19, CD20, CD27, IgG, IgA and IgM, and incubated on ELISpot plates, as described above.

Data analysis and modeling

1. Overview

Characterizing and utilizing inter- and intra-subject variations is at the core of our approach for integrating, analyzing and interpreting the multi-modal data sets in this study. We utilize inter-subject variation for building predictive models and finding correlates of immune response end-points; we assess intra-subject variation—especially of predictive parameters—to inform whether a given variable is temporally stable and therefore likely reflects stable immune states. A temporally stable parameter is one that retains its relative level within an individual over time. These parameters are biologically interesting because they are more likely to reflect baseline immune states as opposed to acute events such as an infection. Given a parameter measured in multiple subjects and baseline time points, the total observed variation can be attributed primarily to three components: 1) baseline differences across individuals (inter-subject baseline variation), 2) temporal changes around the baseline within subjects (intra-subject variation), and 3) experimental/measurement noise. In general, parameters that exhibit large inter-subject variation can better enable correlate discovery. Perturbing the immune system with vaccination introduces another type of subject-to-subject variation: a subset of the parameters (e.g., transcript levels, cell population frequencies, antibody titers and antigen-specific B-cell responses) can increase or decrease from the respective baselines with *variable* amounts across individuals; this response variation together with baseline variation can be utilized to determine post-vaccination correlates.

One key issue in the context of developing predictive models and finding correlates is that the parameters (both at pre- and post-perturbation) can be intricately linked (and therefore correlated) with each other. For instance, as was found in previous studies (Bucasas et al., 2011; Sasaki et al., 2008), we have detected strong correlation between baseline and response titers in our data. We further saw that B-cell responses on day 7 also strongly depend on initial titers and B-cell statuses (Figure 4). Thus, without accounting for the effect of initial titers on post-vaccination responses, it would be unclear whether any potential correlate from days 1 and 7, for example, is a true "root" correlate or whether the correlation was merely driven by cross-correlation with day 0 titers. Another example involves age, which is known to influence vaccination responses, and was associated with titer responses (Figure 5B) and a host of genes and cell population frequencies in our cohort (data not shown), even though we have relatively few older individuals. Similarly, response variations following vaccination can be attributed to differences in baseline, thus baseline contributions need to be modeled appropriately when determining post-vaccination correlates. Our analysis framework for building predictive models and finding robust correlates of responses was designed to mitigate confounding by such tangling of variables from multiple time-points.

Specifically, we systematically identified contributors to end-point variation in a stepwise manner, first by examining "intrinsic" factors including age and gender, then by analyzing parameters over the progression of time (Figure 5A). In this analysis progression, the variables we analyzed first could affect subsequent variables, but not the other way around. Thus, once we have accounted for the effects of "earlier" variables on end-point variation, we can remove their effects from subsequent variables to ensure that contributions from earlier variables would not be "counted" twice. For example, we first modeled simultaneously the effects of multiple intrinsic variables, including age, gender, and ancestry. We next removed the effects of these intrinsic variables from baseline parameters (gene expression and cell population frequencies) to ensure that any baseline correlates we subsequently identified would not be due to the tangling of correlation among intrinsic variables, baseline data, and titer responses. In this example, the direction of information flow goes from intrinsic variables to baseline data because the other direction is impossible.

An overview of the analysis steps we undertook are outlined below together with brief descriptions of motivations for individual analyses and how various analyses were connected to each other. Although some of these were discussed briefly in the main text, the descriptions below include more specific technical information. Our aim is to provide a more cohesive description with emphasis on the methodological details (and how certain decisions were made) to facilitate potential adoption of our approaches to related settings. Detailed descriptions of individual methods and analysis procedures can be found in the "methodological details" section.

2. Overview of analysis steps

2.1 Quantify inter vs. intra-subject variations

We first developed an approach to assess the relative contributions of inter- versus intra-subject variation to the total observed variation in the immune parameters we have measured. One of the

goals of this analysis is to quantitate the range of subject-to-subject differences in a large number of previously less-characterized immune parameters in healthy individuals (e.g., certain cell subsets). Our approach pooled data from the two baseline time-points (days -7 and 0) as well as day 70—day 70was used as a baseline time-point in this analysis because aside from serological and vaccine-specific B-cell responses, substantial changes in other parameters were not observed. We fitted an Analysis of Variance Model (ANOVA) to assess what fraction of the observed variance (expressed as total R^2 (or sum-of-squares)) can be explained by differences among subjects (see **quantifying inter- versus intra-subject variations**). Here we made the assumption that the residual from the fitted model provides an estimate of the within-subject (or temporal) stability—i.e., the higher the variance across the three baseline time-points for a given subject, the lower the temporal stability. We used the same approach to analyze transcript abundance (Figures 2B and 2C). We also quantified observed variability at the level of gene sets (pathways) by using a metric we devised called "pathway activity score**"** (see **computing the individualistic pathway activity score**), which reflects the degree to which the expression levels of constituent genes from a given gene set/pathway for a particular subject are coordinately deviated from the respective means of the entire cohort.

2.2 Characterize the implications of inter-subject baseline variations

We assessed the implications of baseline variations from two angles. The first was to evaluate whether qualitatively-consistent (or conserved) responses among subjects in the cohort can be detected despite baseline variations (see **detecting coherent changes**) – we were especially interested in assessing cell subpopulations, many of which have not yet been evaluated in the context of influenza vaccination. The second angle was to evaluate whether inter-subject variation in baseline immune states can explain titer response variations by building predictive models and finding robust correlates of antibody responses using baseline immune parameters.

Given the extensive variation in serological and B-cell parameters at baseline, we first developed approaches to assess the effect of pre-existing immunity on day 70 titer responses. In addition, we examined early gene expression and cell population responses at days 1 and 7 to ask whether they correlate with initial serology. We tested several analytical approaches for answering this question, including computing both parametric and non-parametric correlations between day 0 titers and the response variables. We found that the most revealing approach was by binning the individuals into groups using a combination of baseline serological and B-cell parameters, followed by evaluating whether the degree of change in a given parameter (e.g., day 1 vs. day 0, day 7 vs, day 0) was significantly different across the subject groups (**see assessing and comparing coherent changes within subject groups segregated by initial serological and B-cell parameters**).

2.3 Build predictive models of anti-body responses using baseline parameters

Next we followed the predictive modeling framework discussed above (Figure 5A) to identify robust correlates and build predictive models of antibody responses. Given the strong influence of day 0 titers, one approach we considered was to include it explicitly in our models. However, due to the non-linear relationship between day 0 and day 70 titers (Figures 4A and S4), the use of linear modeling alone is

insufficient to capture and therefore mitigate the effect of this type of correlation. Instead, we developed two complementary end-point metrics (MFC and adjMFC) where the non-linear correlation between initial and response titers was removed in adjMFC (see **titer definitions**).

With these two end-point metrics, we first fitted a linear model to assess the contribution from intrinsic variables including age, gender, and ancestry using the entire cohort; we later included day 0 titers and memory B cell statuses (from ELISpot assays) to simultaneously evaluate contributions from these variables also. The model we fitted (in R notation) was: *adjMFC/MFC ~ age + gender + ancestry + day 0 titer* and *adjMFC/MFC~ age + gender + ancestry + day 0 titer + day 0 memory (seasonal) + day 0 memory (pH1N1)* (see **titer definitions**). Here, "day 0 memory" refers to the memory IgG B cell ELISpot values (as a percentage of total IgG) specific for the indicated vaccine antigen. Both memory variables tested were not significantly associated with MFC or adjMFC, hence the result of the second model was not shown in Figure 5B.

Since the previous analyses considered the contributions of intrinsic variables (and the effect of day 0 titers was accounted for by adjMFC), we next removed age, gender, and ancestry effects from the baseline gene expression and cell population frequency data to ensure that these effects do not contribute to downstream predictive modeling and correlate identification. This was done by first fitting the model: *x ~ age + gender + ancestry*, where *x* is the transcript level of a gene or the frequency of a cell population, then retaining the sum of the residual and intercept from this fit for downstream analyses (see **removing intrinsic effects from data**).

Using the baseline gene expression and cell population data with intrinsic effects removed, we derived four parameter sets comprising gene expression data and/or cell population frequencies for cross-validation based predictive modeling: 1) the frequency of 126 cell populations; 2) the expression of all annotated genes (~22,000) from the microarray; 3) the activity score of 183 annotated pathways derived earlier for the analysis of inter-individual variations, and 4) both the cell population frequencies and the pathway activity scores. Cell population and gene expression data were integrated by using the pathway activity score because the numbers of pathways and cell populations were comparable, whereas if data from individual genes were used, the large number of gene expression signals could overwhelm those from cell populations (see **predictive modeling**).

We have also considered the use of the whole cohort vs. the use of only the extreme (high/low) responders for predictive and correlate analyses. The main question was which approach would offer better statistical power for revealing biologic parameters associated with response differences across subjects. In general, discrete classification models (i.e., predicting high vs. low responders) are more robust for detecting non-linear correlations, and they are easier to interpret, evaluate, and adapt to different cohorts/settings in comparison to regression-based models that try to predict continuous outcomes. We also used the plasmablast gene expression signature on day 7 as a positive control and assessed which approach (i.e., continuous correlation measures derived from the whole cohort vs. differences between the high and low responder groups) provided better enrichment signals. We found that the discrete/extreme responder approach generally provided better signals. Furthermore, we have considered using a single cutoff (i.e., the subjects with titer values above or below the 80^{th} percentile mark were defined as high or low responders, respectively). However, this would result in highly

uneven high/low responder groups, which in turn could lead to issues in assessing predictive performance (e.g., a model that simply predicts "low" all the time would have an average accuracy of ~80%); and again, the inclusion of the "mid" responders in the low group would "dilute" the true biological differences between the high and low responders.

2.4 Characterize temporal stability of baseline parameters

Once we confirmed that predictive models could be built using baseline data, we identified the primary contributing parameters and assessed whether these parameters were temporally stable. With data obtained from a cohort of subjects over multiple time-points, we consider a parameter statistically stable if the amount of variation among measurements within subjects (i.e., those obtained from different time-points) is significantly less than the differences across subjects (i.e., subject-to-subject variation). We accomplished this assessment by fitting an ANOVA model to our data (see **quantifying inter- versus intra-subject variations**).

2.5 Validation of predictive model using data obtained from independent baseline time-points

Our goal was to evaluate whether predictive models built using data from one baseline time-point can be used to predict outcome using data from a distinct time-point. A model was built using data from day 0 (as training set by using all subjects); this model was then applied to independently-obtained data from different baseline time-points (days -7 and 70) (see **predictive modeling**).

2.6 Post-hoc prediction of antibody responses

By using the same cross-validation approaches, we took advantage of response variations to test whether day 1 and day 7 predictive models can be built. Since baseline states could still be retained in data obtained from post-vaccination time-points, we used baseline subtracted data to mitigate contributions from the baseline.

3. Methodological details

Low-level microarray data processing

The raw intensities of gene probesets in the Affymetrix CEL files were processed using Affymetrix Power Tools (http://www.affymetrix.com/support/developer/powertools/index.affx). The "apt-probeset_summarize" script was used for background correction and normalization (RMA with sketch quantile normalization) (command: apt-probeset-summarize --console-off -a rma-sketch -p HuGene-1_0-st-v1.r4.pgf -m HuGene-1_0-st-v1.r4.mps -c HuGene-1_0-st-v1.r4.clf -b HuGene-1_0-st-v1.r4.bgp --qc-probesets HuGene-1_0-st-v1.r4.qcc -o temp *.CEL). The processed probe intensities were then extracted from the CEL files using R to enable probe-level analysis of the data. In total, 315 microarrays were processed (63 subjects over 5 time-points).

To conduct quality assessment (QA), the Bioconductor package "arrayQualityMetrics" (Kauffmann et al., 2009) was used to determine the quality of the microarray data. Low quality arrays (i.e., outliers from other arrays) were detected by the "Distances between arrays", "MA plot", and "Boxplot" diagnostic metrics provided by the package. Arrays were marked as "low quality" if they failed

to pass the quality assessment in two out of the three metrics using the default setting. Arrays that failed QA were excluded from further analysis. In total, 23 out of 315 arrays were excluded. Excluded arrays were treated as missing data in subsequent analyses.

To detect batch effects, linear regression was used to assess correlation between the expression values of every probe set with batch variables (by using the linear model in R: *probe set intensity~batch variable 1+batch variable 2+…* for every gene). We tested hybridization date, technician, and sample collection time as potential batch variables. Hybridization date was found to be significantly correlated with most probe sets, and its effect was removed from every probe set using linear regression to retain the residual values (i.e., for every probe set *i*, fit $probe\ set\ intensity_i \sim hybridization\ date$ and only retain the residual from this fit). This batch removal procedure was applied to the raw and to the baseline subtracted data (i.e., with day 0 values subtracted), after pooling data from all time-points. The resulting gene expression values were then log_2 transformed for subsequent analyses.

Low-level flow cytometry data processing

The frequencies (expressed as percentage of parent population) of the manually gated cell populations were first transformed by $\log_{10}$ (all zero values were set to 0.01 before transformation). The tight correlation between control samples (Figure S1B) suggested that batch effects were minimal. Samples and cell populations were assessed for data quality independently for each time point. A sample is excluded from further analysis if the median of viable cell frequencies across all five tubes (Figure S1A) for a particular time point is less than 70%. A cell population is filtered out if it has less than 20 cells in 80% of the samples. Removed values were treated as missing data in subsequent analyses.

Pathway enrichment analysis (PEA)

A standard approach for enrichment analysis is to assess, using the hypergeometric test, whether a given set of transcripts (e.g., those that are differentially expressed on day 1) is enriched for genes from a particular pathway. To increase sensitivity, approaches such as gene set enrichment analysis (Subramanian et al., 2005) that do not use a fixed cutoff and instead rank all genes based on a statistic (e.g., *-log*(p-value) from differential expression) and then evaluate whether the distribution of genes from a particular pathway/gene set is enriched at the top of the ranked list. Here we used a variation of this strategy by using the geneSetTest function in Limma (Goeman and Buhlmann, 2007). The annotated pathways we used were derived from those annotated by Ingenuity Systems (www.ingenuity.com). The p-values were corrected for multiple-testing using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

Computing the pathway activity score for individual subjects

The goal was to compute, for every subject and a given pathway (defined by a gene set), a score that reflects the degree to which the expression levels of genes in the pathway for the subject are coordinately deviated from the respective means of the entire cohort. To achieve this goal, the gene expression values of every gene (across subjects) were first standardized by the z-score transformation (i.e., centered by the cohort mean and scaled by the standard deviation). Thus, the transformed expression value for a given gene and individual reflects the scaled difference between the transcript

level of that gene in the individual and that of the cohort mean. This standardization also ensures that these values are comparable across genes. Next, for each individual, the genes were ranked by their transformed expression values in descending order (i.e., from genes with the highest deviations from the respective cohort means to those with the least). PEA is then applied to this ranked list to obtain the enrichment p-value; the "pathway activity score" (PAS) for the individual and the given pathway is obtained as $-\log_{10}(p)$. Thus, as desired, this score quantifies the degree to which genes from a given pathway tend to deviate (positively and/or negatively) from their respective cohort mean for a given subject. This procedure was then repeated for all individuals and pathways to obtain the pathway by subject "pathway activity score" matrix (PASM). Note that for genes represented by multiple probe sets, we chose the probe set that is most correlated to the first principle component of these probe sets across all arrays.

Quantifying inter- versus intra-subject variation

For each parameter (e.g., transcript or frequency of a cell subset), we pooled data from all individuals from days -7, 0 and 70. As discussed in the text, day 70 was used as an additional baseline time-point to increase statistical power because we did not observe substantial changes by day 70 (except for titers and antigen specific B-cells). Here we have a data vector $X=\{x_{ij}\}$ (individual *i*, time-point *j*) where each x_{ij} is associated with a subject and each subject has a total three measurements obtained from each of the time-points. We then fitted the ANOVA model (in R notation: X ~ Subject) to evaluate 1) total variance (total R^2 – sum of squares), 2) whether subject-to-subject differences (inter-subject variation) is statistically significant, and 3) the fraction of variance explained by subject relative to the residual of the fit. Here we made the assumption that the variance explained by the residual of this fit provides an estimate of the within-subject variation (or temporal stability as depicted in Figure 6D). While this assumption ignores the potential contribution of measurement noise, variance from measurement noise can add to both types of variation, therefore the relative estimated fraction of inter- vs. intra-variation should be more robust to contributions from experimental/technical sources. Also, our assessment of technical variance for cell population frequencies indicated extremely high reproducibility (Figure S1B), therefore experimental noise likely made only minor contributions to the total observed variance of cell frequencies. In addition, we have also tested different (correlation based) metrics to assess temporal stability and intra-subject variation and found essentially the same set of cell population parameters being temporally stable (data not shown; method: we evaluated whether the values of a parameter across individuals were correlated over a pair of time-points).

For the gene expression heatmap shown in Figure 2B, the log_2 transformed values were used to select the 500 most variable genes (with genes on the Y chromosome excluded), followed by hierarchical clustering (Euclidean distance and complete linkage) to visualize transcript pattern heterogeneity across subjects. For visualizing heterogeneity within pathways, we restricted to only genes from a given pathway (Figure 2C). To quantify variability at the pathway level, we used the PASM (see above) to compute the standard deviation of the pathway activity score of each pathway across subjects in the cohort. The pathways were then ranked by this metric to determine the most variable pathways (Figure S2C). For cell population frequency distributions shown in Figure S2B, subsets were

ranked by the interdecile range (IDR - which is the difference between 9^{th} and 1^{st} deciles (90% and 10% quartiles)).

Detecting coherent changes

For gene expression data, paired test from the Bioconductor package "limma" (Smyth, 2004) was used to determine genes changed coherently across individuals (i.e., differentially expressed (DE) genes post-vaccination). Probesets with very low variation (with an inter-quantile range of less than 0.15) and without annotated gene symbols (based on Affymetrix's latest annotations downloaded from their website) were filtered out. In addition, probesets associated with multiple genes were removed. When a gene is associated with multiple probesets, we computed the principle components of these probesets across all arrays and only use the probeset that correlated the most to the first principle component. After these filtering steps, 16885 genes were left. To avoid potential confounding by batch variables that were not explicitly measured, we used Surrogate Variable Analysis of Gene Expression (SVA) (Leek and Storey, 2007) to analyze and remove hidden batch effects. Briefly, after accounting for the primary effects (in this case, the "Day" variable (for days -7, 0, 1, 7, 70) is the primary effect), the significant principle components from the residual matrix were regressed out of the gene expression data (for details see Leek et al.). Data from day 0 were used as the baseline and data at other time-points were analyzed against baseline for DE genes. P values were adjusted for multiple-testing using Benjamini and Hochberg's method to estimate the false discovery rate (FDR) (Benjamini and Hochberg, 1995). Genes passing both the 5% FDR and absolute-$\log_2$-fold-change>0.2 cutoffs were considered DE genes.

To detect post-vaccination changes in cell population frequencies, we subtracted the log-transformed measurement at the pre-vaccination baseline (day 0) from a post-vaccination time-point (day x, for x = 1 or 7) and tested if the median of the log-fold changes ($\log_{10}\left(\frac{day\ x}{day\ 0}\right)$) of individuals were significantly different from zero using a Wilcoxon signed rank test (Rice, 2007). We used this non-parametric, paired test to detect reliable changes in a variable between two days because the flow cytometry data for all time-points of any individual was collected in the same experimental batch. We used the *wilcox.test()* function in the R statistical software to compute the p-values, and for each post-vaccination time-point, we only reported those with p-value < 0.05 (after multiple-testing correction by Benjamini-Hochberg FDR procedure) and a fold change magnitude (in either direction, before log_{10}) of least 1.1.

Assessing and comparing coherent changes within subject groups segregated by initial serology and B-cell parameters

The day 0 serological and B-cell variables were first standardized using the z-score transformation applied to data from days 0 and 70 (because these two time-points revealed the baseline dynamic range (low and high) of these variables). The standardization was required to ensure that the different variables can be combined in clustering analysis. Next we used both k-means and hierarchical clustering (Euclidean distance with complete linkage) to assess clustering structure and found that two clusters can best explain the data (similar to Figure 4B with two group of subjects segregated by high vs. low initial titers and B-cell statuses). To evaluate the robustness of this observation, we sampled 1000 random

subsets of subjects (each containing 85% of the subjects) followed by k-means clustering (k=2) to determine whether individuals were consistently grouped into the same cluster. Using this data, we determined the frequency (out of the 1000 iterations) a particular subject was assigned to the "high" vs. the "low" groups. Only subjects (all except four) who were consistently assigned to a particular group in more than 80% of the iterations were included in subsequent analysis (Figures 4B and C).

To evaluate whether the post-vaccination responses were significantly different between the two groups, we first computed the baseline-subtracted responses (e.g., day7-day0 and day1-day0) for all parameters (cell frequencies, gene expression, B-cell ELISpots) and then assessed whether these responses are significantly different between the two groups using the Wilcoxon test (see above on **Detecting coherent changes**) followed by multiple testing correction. For transcripts, we also applied pathway enrichment as above. While we saw a number of marginally significant cell populations and pathways with distinct responses between the two groups, only the most prominent ones surviving stringent multiple-testing correction were reported in Figure 4C (those involving plasmablasts and vaccine-specific B cells.)

Titer definitions

The MFC and adjMFC metrics were computed using subjects with both gene expression and cell population data on day 0. Given the strong non-linear correlation between day 0 and day 70 titers for each individual virus (Figures 4A and S4A), we tested various approaches to mitigate the effect of this correlation on our prediction analyses. We first tested a linear regression approach similar to the one used by Bucasas et al. (Bucasas et al., 2011). While the linear component was removed as expected, the non-linear ones were not (e.g., the significantly higher variance in the response was retained for individuals with low day 0 titers). Thus, we aimed to devise an approach that could further mitigate this non-linear relationship. Since there were four viruses, we quantitated both the pre-existing and response serological states by taking the maximum across the four viruses, given that none of our transcriptomic and cell population frequency measurements reflect specificity to particular antigens.

To make individual viral titers and their post-vaccination responses comparable so that the maximum is meaningful, we first standardized day 0 titer by: $TV_i^v(d) = \frac{MN_i^v(d) - median(\boldsymbol{MN^v(d)})}{mad(\boldsymbol{MN^v(d)})} with\ d = 0$, where $MN_i^v(d)$ is the microneutralization titer for virus *v*, subject *i*, day *d*; $mad(\boldsymbol{x})$ denotes the median absolute deviation of a vector $\boldsymbol{x}$ as defined by $mad(\boldsymbol{x}) = median(|x_i - median(\boldsymbol{x})|)$. Here we used median subtracted values and scaling by the *mad()* function instead of mean subtraction and scaling by the standard deviation to avoid undesirable influences from outliers. We then defined the day 0 titer state for individual *i* as $TV_i(0) = max\left\{TV_i^{Uruguay}, TV_i^{A/Brisbane}, TV_i^{B/Brisbane}\right\}$ (only the seasonal viruses are included because the median variation (*mad*) of pH1N1 is 0 and hence the TV_i^{pH1N1} estimates are undefined for many subjects; plus it is usually the lowest before standardization).

In a similar manner, we used the same median/mad-based standardization to standardize the day 70 titer response based on fold-change: $FC_i^v = \frac{MN_i^v(70)}{MN_i^v(0)}\ and\ TR_i^v = \frac{FC_i^v - median(\boldsymbol{FC^v})}{mad(\boldsymbol{FC^v})}$. Next the

response for individual *i* was defined as $TR_i = max\left\{TR_i^{Uruguay}, TR_i^{A/Brisbane}, TV_i^{B/Brisbane}, TR_i^{pH1N1}\right\}$. We then applied the inverse normal transform (INT) to the TR_i values for subsequent analyses (Maritz, 1995). We used the INT to avoid the extreme skews in the distribution as INT uses the rank to transform the values such that the resulting distribution is Gaussian. Since we were most interested in the categorical response ("high" vs. "low" responders), this approach mitigates the undesirable effects of outliers, skewness, and non-normality without losing critical information. We have tested the alternative procedure of not applying the INT and found the results largely consistent. In our diagnostics (see below), we also always examine both the transformed and untransformed values. From herein, let $TR_i = INT(TR_i)$.

Next we plotted $TV_i(0)$ and TR_i values against each other across subjects, and as expected, we saw a strong inverse correlation between them. We noticed that the key to remove this non-linear correlation is to standardize the variable mean and variances as a function of the initial titer (TV_i) values (similar stratification approaches have been discussed in the vaccine trials literature, e.g., (Nauta, 2011).) To achieve this goal, we binned the subjects based on their $TV_i(0)$ values and let $\boldsymbol{TR^j}$ be the vector of TR values from subjects in bin *j*. For each individual, we next computed the decorrelated response as: $TR_i^{decor} = \frac{TR_i - median(\boldsymbol{TR^j})}{mad(\boldsymbol{TR^j})}$ where subject *i* belongs to bin *j*. We have tested multiple binning cutoffs and found the result to be robust to exact cutoffs. The results reported were based on grouping the subjects into two bins: (1) $-10 \leq TV_i(0) < 1$; (2) $1.1 \leq TV_i(0) < 50$ (we also found that the results are robust to small changes in the binning boundaries or increasing the number of bins to three, as long as there is a reasonable number of data points within each bin and there is no detectable correlation between initial and response titers within each bin). Finally, the TR_i^{decor} were discretized to "low", "mid", and "high" responders as the subjects with values lower than or equal to the 20th percentile, between the 20th and 80th percentile, and above or equal to the 80th percentile values. We tested changing the cutoffs to (25th, 75th), (30th, 70th), and (35th, 65th) for building predictive models. In most cases, the core predictive parameters remained similar. As expected, as the boundaries between "low" and "high" became closer, noise was introduced and thus caused the identity of predictive parameters and predictive performance to fluctuate. MFC and adjMFC were defined as the discretized TR_i and TR_i^{decor} with the (20th, 80th) cutoff, respectively.

After applying the decorrelation procedure, we also performed a number of diagnostic tests to ensure that our approach had indeed achieved the desired effects. We have checked that 1) the correlation between $TV_i(0)$ and the discretized TR_i^{decor} was indeed removed based on the Spearman rank correlation (for both the continuous and discretized (high/low responders) forms of TR_i^{decor}); 2) at the individual virus level, the correlation between the day 0 titer (of a single virus) and TR_i^{decor} was removed (i.e., before aggregation of the individual viruses via *max*); 3) the identity of the virus that contributed to the day 0 (or day 70 fold-change) maximum does not correlate with TR_i^{decor}. In all of the above, two variables were considered not correlated if the Spearman correlation p value is greater than *0.1*.

Removing intrinsic effects from data

For both cell population frequencies and transcript abundance, the log-transformed data was pooled (i.e., from all time points together) and a linear model was fitted including age, gender, and ethnicity (race) as factors. The intrinsic effect-removed values were residuals from this model with the intercept component added. Using the same procedure, intrinsic effects were removed from the baseline-subtracted data.

Predictive modeling

Monte-Carlo cross-validation was used to train and test prediction models. During each iteration, 75% of the subjects (rounded to the nearest integer) were randomly selected as a learning set, and the rest (25%) were used as a test set. By using this approach we ensured that all training/test sets have the same number of subjects. During each iteration, the standard unpaired t-test was used to compute the t-statistic and corresponding p-value for each feature in the training set *only* by comparing the feature's values in the "high" and "low" groups; the features were then ranked based on the p-values and the top *k* feature were included for model building/training. We have tested *k*=2, 5, 10, 30, and 60. A total of 5000 iterations were performed for each *k*.

We utilized R/Bioconductor's CMA package (Slawski et al., 2008), which allowed us to test and compare multiple types of predictive models. During each iteration we recorded the misclassification rate (ratio of number of false predictions to total number of predictions) and AUC (area under the empirical Receiver Operator's Characteristic (ROC) curve) to estimate prediction performance. AUC is our preferred metric to report because it is not biased by the ratio between the number of "high" and "low" individuals ; the misclassification rate, on the other hand, could be biased because, for example, in cases where most of the subjects are "high" responders, a "dumb" classifier simply has to always predict "high" to achieve an apparently low misclassification rate – but the general performance of this classifier is bad because of its poor performance in cases where the number of "high" subjects is low.

We tested several classifiers implemented in CMA and found that the diagonal linear discriminant analysis (DLDA) model consistently achieved the best (or near the best) performances. DLDA (Dudoit et al., 2002) is a type of naïve Bayes classifier that assumes independent Gaussian distribution of features in each class. In our case, DLDA classifies a sample characterized by the feature vector $x = (x_1, \dots, x_P)$ as either a "high" (H) or a "low" (L) responder. As discussed above, the features can be cell population abundances or pathway activity scores. Given a sample, the model classifies it as *L* iff $\sum_{i=1}^{P} \frac{(x_i-\mu_{H,i})^2}{\sigma_i^2} \geq \sum_{i=1}^{P} \frac{(x_i-\mu_{L,i})^2}{\sigma_i^2}$, where $\mu_{k,i}$ is the mean value of feature *i* for group k, σ_i^2 is pooled variance computed as $\sigma_i^2 = \frac{(x_i-\mu_{H,i})^2+(x_i-\mu_{L,i})^2}{N-2}$.

To obtain the contribution, or "weight", of a feature in the model, during each iteration/training set, the weight of feature *i* was calculated as the difference between the mean in each class (H vs. L) divided by the pooled standard deviation: $w_i = \frac{(\mu_{H,i}-\mu_{L,i})}{\sqrt{\sigma_{H,i}^2+\sigma_{L,i}^2}}$. Note that the weights were calculated using the training set only during each iteration. For each feature *i*, we summarized its weight by computing its mean weight across all trials that included feature *i* in its model.

To assess the contribution of a feature to prediction, we computed two metrics: (1) the number of times the feature was selected to be included for modeling (i.e., in the top *k* for each *k*); (2) the normalized mean weight across all iterations - during each iteration, we normalized the weight of each feature to the maximum weight across all features.

As discussed in the main text, using baseline parameters, only the models involving cell population frequencies were predictive. To determine which cell populations contributed most to prediction, we used a normalized mean weight cutoff of 0.4. This cutoff was chosen because when we plotted the mean weight of all cell populations ranked from the lowest to the highest, we noticed an inflection point (or "elbow") at 0.4, indicating that 0.4 is a transition point potentially separating the informative features for prediction from those that are not.

We also tested building predictive models using gene expression data only where each feature is a gene (Figure S4C). We tested a number of modeling approaches provided by CMA and the plsgenomics packages, including support vector machine (SVM), random forest, partial least squares followed by linear discriminant analysis (PLS-LDA), neural network (Slawski et al., 2008). In general, PLS-LDA model achieved the best predictive results when individual genes were used. To select genes as features for PLS-LDA modeling, during each random trial/iteration we selected *k* genes using the *variable.selection()* function from the plsgenomics package (essentially based on a gene's correlation with outcome ("high" and "low" groups defined by adjMFC/MFC), or according to the package's documentation: "feature selection was done by ordering the transcripts according to the absolute value of the weight defined by the first PLS component. This ordering is equivalent to the ordering obtained with the F-statistic and t-test with equal variances (Boulesteix, 2004)."). We tested multiple models with different number of genes as predictors (*k*=2, 5, 10, 20, 40, 80, 160, 320, and 640). We conducted 5000 random trials for both MFC and adjMFC.

Finding robust pathway correlates of day 70 antibody titer responses

We first used Wilcoxon ranksum test to test whether genes showed significant differences between individuals with "high" and "low" titer responses at day 70 (for MFC and adjMFC), and then used the resulting –log(P) values to rank genes and apply the PEA described above. This yielded a set of enriched pathways for each end-point. To assess the robustness of each such pathway enrichment signature, we randomly selected 75% of the individuals over 100 iterations and counted the number of times the pathway appeared as enriched (at unadjusted $P<0.05$) in these iterations. Note that for genes represented by multiple probe sets, we chose the probe set that is most correlated to the first principle component of these probe sets across all arrays.

Correlation between cell population frequencies and transcript levels

We have correlated the selected flow populations' abundance (from days 0 and dayX-day0) with gene expression data using the standard Pearson correlation metric. Only the probesets/genes passing the filters described in the **detecting coherent changes** section above were used. Correlation p-values were computed by Student's t distribution, and two-tailed p-values were produced by doubling one-tailed p-values (as implemented in MATLAB's *corr* function). The calculated p-values were adjusted for multiple

testing with random sample permutation (200 permutations). We have reported population-genes pairs with adjusted p-value less than 0.01. To test for pathway enrichments, the genes correlated with each cell population were ranked and PEA as described above was used to compute the enrichment p-values. Pathways shown on Figures 7C and D were those enriched in at least one cell population with an adjusted p-value lower than 0.01 (same criterion for cell populations – those shown were ones having at least one pathway with an adjusted p-value lower than 0.01). Note that for genes represented by multiple probe sets, we chose the probe set that is most correlated to the first principle component of these probe sets across all arrays.

Supplemental References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) *57*, 289-300.

Biancotto, A., Dagur, P.K., Fuchs, J.C., Langweiler, M., and McCoy, J.P., Jr. (2012). OMIP-004: in-depth characterization of human T regulatory cells. Cytometry A *81*, 15-16.

Biancotto, A., Fuchs, J.C., Williams, A., Dagur, P.K., and McCoy, J.P., Jr. (2011). High dimensional flow cytometry for comprehensive leukocyte immunophenotyping (CLIP) in translational research. Journal of immunological methods *363*, 245-261.

Boulesteix, A.L. (2004). PLS dimension reduction for classification with microarray data. Stat Appl Genet Mol Biol *3*, Article33.

Bucasas, K.L., Franco, L.M., Shaw, C.A., Bray, M.S., Wells, J.M., Nino, D., Arden, N., Quarles, J.M., Couch, R.B., and Belmont, J.W. (2011). Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans. The Journal of infectious diseases *203*, 921-929.

Crotty, S., Aubert, R.D., Glidewell, J., and Ahmed, R. (2004). Tracking human antigen-specific memory B cells: a sensitive and generalized ELISPOT system. J Immunol Methods *286*, 111-122.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION *97*, 457.

Goeman, J.J., and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics *23*, 980-987.

Ho, J., Moir, S., Wang, W., Posada, J.G., Gu, W., Rehman, M.T., Dewar, R., Kovacs, C., Sneller, M.C., Chun, T.W*., et al.* (2011). Enhancing effects of adjuvanted 2009 pandemic H1N1 influenza A vaccine on memory B-cell responses in HIV-infected individuals. AIDS *25*, 295-302.

Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics *25*, 415-416.

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS genetics *3*, 1724-1735.

Maritz, J.S. (1995). Distribution - Free Statistical Methods (Chapman and Hall).

Nauta, J. (2011). Statistics in clinical vaccine trials (Springer Berlin Heidelberg).

Perfetto, S.P., Ambrozak, D., Nguyen, R., Chattopadhyay, P., and Roederer, M. (2006). Quality assurance for polychromatic flow cytometry. Nat Protoc *1*, 1522-1530.

Rice, J.A. (2007). Mathematical statistics and data analysis, 3rd ed. edn (Belmont, Calif. ; United Kingdom: Thomson/Brooks/Cole).

Sasaki, S., He, X.S., Holmes, T.H., Dekker, C.L., Kemble, G.W., Arvin, A.M., and Greenberg, H.B. (2008). Influence of prior influenza vaccination on antibody and B-cell responses. PloS one *3*, e2975.

Slawski, M., Daumer, M., and Boulesteix, A.L. (2008). CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. BMC Bioinformatics *9*, 439.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol *3*, Article3.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545-15550.