Small Sample-Big Data: Integrative Indexed Systems Biology Reveals Dramatic Molecular Ontogeny over the First Week of Human Life

Abstract

Systems biology provides a powerful approach to unravel complex biological processes yet it has not been applied systematically to samples from newborns, a group highly vulnerable to a wide range of diseases. Published methods rely on blood volumes that are not feasible to obtain from newborns. We optimized methods to extract transcriptomic, proteomic, metabolomic, cytokine/chemokine, and single cell immune phenotyping data from <1ml of blood, a volume readily obtained from newborns. Furthermore, indexing to baseline and applying innovative integrative computational methods that address the challenge of few data points with many features enabled identification of robust findings within a readily achievable sample size. This approach uncovered dramatic changes along a stable developmental trajectory over the first week of life. The ability to extract information from 'big data' and draw key insights from such small sample volumes will enable and accelerate characterization of the molecular ontogeny driving this crucial developmental period.

INTRODUCTION

The first week of life is characterized by heightened susceptibility to infections and a major determinant of overall health for the entire human lifespan [1-3]. Knowledge of the molecular drivers involved in these processes in newborns (defined as those ≤28 days of life) is fragmentary [4]. In this context, systems biology methods employing high-dimensional molecular and cellular measurements along with unbiased analytic approaches enable the identification of molecular networks and signatures relevant to host defense. Although systems biology approaches have already increased understanding of baseline and altered molecular states in adults [5], such approaches have yet to be applied systematically to characterize molecular ontogeny in early life [6]. This likely relates to the limited amount of biosample that can be obtained [6 7, 8] and the many rapid physiological changes around birth [1, 2], resulting in variance of biological measurements necessitating a large sample size, which increases complexity and cost [9].

To overcome these hurdles, we developed a robust experimental and analytical approach feasible with <1ml of newborn blood. Our data represent the most comprehensive systems biology study yet performed. We found that despite substantial between-subject variation, normalizing ('indexing') all samples from a given newborn enabled identification of robust within-subject changes over the first week of life across the entire cohort. Furthermore, data integration using multiple independent strategies confirmed signatures across methodologically- and biologically-distinct datasets. Such convergence of signals assessing different layers of information validated key findings internally. The results highlight that contrary to the relatively biological steady-state observed in healthy adults [7, 10], the first week of human life is highly dynamic; however, despite these dramatic changes, ontogeny appears to follow a common path. This method now enables future studies to focus on the impact of perturbations such as immunization and infection on this developmental trajectory.

RESULTS

Blood Processing SOP

A primary objective of this project was to develop a robust standard operating procedure (SOP) to allow the extraction and analysis of data using systems biology ('big data') approaches from small blood sample volumes that can be obtained for research study purposes from newborns and infants (Figure 1; see also 'Protocol'). Our experimental SOP utilized important sample-sparing modifications whereby we obtained samples for immune phenotyping, transcriptomic, proteomic, and metabolomic analysis from <1ml of blood (see also Supplementary Text) [11, 12].

Immune Phenotyping across 1st week of life

Determining the cellular composition of blood samples has been recognized as crucial in systems biology studies, not only because the relative and absolute cell numbers predict e.g. vaccine responses with high accuracy [13], but also because knowledge about their relative proportion helps deconvolute OMICS data [9, 14]. Analysis of our pre-defined targeted cell populations revealed substantial between-subject variability (Supplementary Figure 1, Supplementary Text, and Supplementary MIFlowCyt document). However, consistent within-subject changes over the first week of life

amongst the entire cohort of 30 newborns emerged when samples were indexed to the average value obtained from all time points for a given subject. Principle component analysis (PCA) identified the following discriminating features over the first week of life: Basophils, plasmacytoid dendritic cells (DC), natural killer cells, and neutrophils decreased, while myeloid DCs increased from Day of Life 0 (DOL0). We also detected dramatic but consistent changes in soluble immune markers including plasma cytokines and chemokines over the first week of life (Supplementary Figure 2). Based on the component coefficients from the relevant PCA, we found that plasma concentrations of ADAM11, CXCL10, IL-17A, IFNγ, and Flt3L increased, while IL-10, CCL5, granulocyte colony stimulating factor 2 (GCSF), IL-6, and TGFα decreased with age.

The ability to reveal a robust developmental trajectory by indexing each subject was surprising, given the biological heterogeneity across subjects at any given time point. We thus applied a similar analytic approach to the other OMICS analyses as well.

Transcriptomic analysis across 1st week of life

Analysis of gene expression by mRNA quantification is common to most systems biology studies [7, 15]. We found that we needed ≥ 500 μl of adult blood to consistenly obtain sufficient high-quality RNA, but as little as 100 μl of newborn blood sufficed (Supplementary Figure 3A and 3B). This likely reflects that newborn blood has a relatively high content of WBC as well as nucleated RBC, which contain abundant globin mRNA (Supplementary Figure 3C)[16]. The higher yield and quality of total RNA extracted from newborn vs. adult whole blood was confirmed across different RNA extraction platforms (RNALater and PAXGene; both of which yielded similar results (data not shown)). We chose to conduct the rest of our study using RNALater (Supplementary Text).

Similar to the immune phenotyping data, we detected substantial between-subject variability in our RNA-Seq data (Supplementary Figure 3D-G). This was resolved by indexing each subject to baseline, which allowed dramatic yet consistent developmental signals across the entire cohort to emerge that relate to age (i.e. ontogeny). In comparing DOL1 vs. DOL0, there were few (12) differentially expressed (DE) genes, however, dramatic developmental changes started to appear when comparing later days of life to DOL0 (Figure 2). Specifically, for DOL3 vs. DOL0 we detected 1125 DE genes, while on DOL7 vs. DOL0, there were 1864 DE genes. All DE genes, pathway enrichment and statistics are listed in Supplementary Table 1 and Supplementary Text. In particular, genes with decreased expression as a function of age are involved with cellular responses to stress, detoxification of reactive oxygen, as well as heme biosynthesis and iron uptake. Conversely genes involved in interferon signaling, negative regulation of RIG-I and complement activation were up-regulated over the first week of life.

Proteomic analysis across 1st week of life

We analyzed changes in the plasma proteome across the first week of life utilizing 5 μl of plasma in a 96-well plate format, allowing large numbers of samples to be analyzed in parallel. A total of 684 different proteins were identified across samples (false discovery rate (FDR) <1%). Of these, 199 proteins met our criteria (which included the detection of

at least two unique peptides per protein) for further detailed quantification. Substantial between-subject variability was again noted in the plasma proteomic analysis (Supplementary Figure 4A), but indexing each subject to baseline enhanced the detection of signatures that again indicated a common developmental trajectory over the first week of life, with differences in plasma protein composition compared to DOL0 increasing with increasing age (Supplementary Figure 4B). Differentially abundant plasma proteins and their respective pathways are listed in Supplementary Table 2 and Supplementary Text. At DOL3 vs. DOL0, three pathways were up-regulated that center around the complement cascade. At DOL7 vs. DOL0, five additional pathways were up-regulated including scavenging heme from plasma and signaling to RAS.

Metabolomic analysis across 1st week of life

We used a mass spectrometry-based global discovery metabolomics approach to analyze ~50 μl of plasma, which enabled detection of ~700 metabolites (Supplementary Figure 5A). While initial analysis revealed substantial between-subject variation, transforming the analysis from a between-subject to a within-subject comparison over the first week of life revealed a steady but dramatic developmental trajectory (Supplementary Figure 5B and C). Few differences in plasma metabolites were discernible comparing DOL1 vs. DOL0, but increasing differences were noted when contrasting DOL3 and DOL7 vs. DOL0; interestingly the differences detected focused on metabolic pathways related to sphingolipid biosynthesis, carbohydrate and androgen metabolism likely reflecting rapid cell proliferation in newborns (Supplementary Table 3 and Supplementary Text).

Data integration

Each methodologically- and biologically-distinct data type we examined revealed substantial changes over the first week of life. We next sought to determine if the observed changes were related to one another across data types, reflecting consistent age-dependent changes in functional pathways. In order to minimize limitations inherent in any single analytical approach, and to detect the most robust signatures, we addressed data integration using three independent strategies. To this end, we employed a function-based strategy based on biologically known Molecular Interactions Networks using NetworkAnalyst [17]; the data-driven (unbiased) multivariate matrix factorization approach DIABLO (Data Integration Analysis for Biomarker discovery using Latent cOmponents) [18-20]; and the previously described multi-scale, multifactorial response network (MMRN) that estimates correlations across data types [7, 21].

Molecular Interaction Networks using NetworkAnalyst

NetworkAnalyst enables the creation of networks based on a framework of known PPI captured in publically curated databases (specifically InnateDB/IMeX)[22]. Minimum connected networks were constructed from seed nodes (i.e. from genes or proteins that changed with age in our data set), as well as first-order interactors that served to connect the seed nodes with each other. To include metabolomics data, metabolic enzymes (synthetic and degradative) that would determine the levels of differentially detected metabolites were used as seed nodes in the network construction. Overall metabolomics, proteomics and transcriptomics data fit well into a single functional network (Figure 3), indicating that these techniques reported on different facets of the

same biological processes. This PPI-based integration strategy recapitulated many key findings that had been identified for each of the individual data types, confirming our expectation that many but not all findings would be validated by different OMICS methods (Supplementary Table 4). For example, integrating transcriptomic with proteomic data confirmed the rapid increase in type 1 interferon (IFN)-related functions over the first week of life. Our integration however also revealed new biological insights not revealed by any single-data domain analysis. For example, only following the integration of transcriptomic and metabolomic data did we identify down-regulation of very long-chain fatty acyl-CoA synthesis activities across the first week of life.

DIABLO

DIABLO is a multivariate approach to address two of the major concerns faced when integrating multilevel datasets: the dimensionality of the data, particularly with few samples, each with many observations, and the heterogeneous nature of our data measured on different scales and technological platforms [18-20]. DIABLO constructs components (linear combinations of the original features) that are maximally correlated across any number of input data types and a specified outcome variable (in this case, DOL), while simultaneously performing feature selection via L1 penalization [23] (Supplementary Text). We created matrices from our five data types as input to DIABLO to identify major ontogeny-related features (Figure 4 and Supplementary Table 5). The resulting model discriminated well between DOLs, with component one of the model separating birth (DOL0) from all other time points, while component two separated DOL1, 3, and 7 from each other (Figure 4C).

We next investigated the relationship between features selected by DIABLO across data types and visualized the selected features in an integrative network (Supplementary Text). We compared this integrative network (Figure 4A vs. 4B) to one derived from features identified using a non-integrative sparse discriminant analysis approach (Supplementary Text). The integrative network was more densely connected (global clustering coefficient = 0.91 vs. 0.68) and composed of few, more tightly connected modules (network modularity = 0.26 vs. 0.09), indicating that DIABLO selected features that were discriminant and well correlated across data types, while the non-integrative approach favored features that were discriminant but not well correlated across data types. The two components of the DIABLO model were composed of distinct sets of features (Figure 4D, blue bars), representing distinct biology (Figure 4E, blue bars; Supplementary Text).

MMRN

Multi-scale, multifactorial response networks (MMRNs) are a recently published framework for data integration [7]. Using MMRN, we found that associations between data types were strongest at DOL1 and decreased across the first week of life, as partial least squares regression scores were significantly higher at DOL1 compared to all other time points (Student's t-test, p-value << 0.01). Stable clusters were more strongly associated with DOL when compared to transient ones (Figure 5C and Supplementary Text), with 15/21 clusters significantly associated with DOL being part of stable networks. This confirmed our already noted robust trajectory of development.

Most significant clusters were transcriptomic (16/21), but we also identified metabolomic (1/21) and flow cytometry-derived (4/21) clusters associated with DOL (Supplementary Table 6). The stable clusters most significantly associated with DOL were composed of various B-cell sub-populations, blood transcriptomic modules (BTMs) related to dendritic cells and monocytes, CCR1, CCR7, cell signaling, heme biosynthesis, TLR and inflammatory signaling, as well as metabolic pathways such as purine metabolism (Figure 5D) [7].

META-INTEGRATION

To assess convergence of signals across methodologically-distinct integration methods, we simplified the output of each integration approach to a list of features associated with DOLs. We then carried out gene set enrichment to determine which biological processes were identified by each method (Supplementary Tables 7 and 8). The selected features for Molecular Interactions Networks (NetworkAnalyst) were the component nodes of minimum connected networks when differentially abundant features (transcripts, proteins and metabolites) for DOL3 vs. DOL0 and DOL7 vs. DOL0 were used as seed nodes (2317 features). For DIABLO, the selected features were the component nodes of minimum connected molecular interaction networks when the features of the two model components were used as seed nodes (184 features). For MMRN, selected features were those that composed the small stable network shown in Figure 5D (564 features). We observed limited overlap at the individual feature level (Figure 6A). However, assessing gene set enrichment against the Reactome annotation system (674 pathways), 472, 251 and 122 pathways were significantly over-represented in the NetworkAnalyst, DIABLO and MMRN feature lists, respectively (all with Benjamini-Hochberg corrected FDR ≤ 0.05, Supplementary Table 7). Importantly, 162 and 26 pathways were identified by NetworkAnalyst and either DIABLO or the MMRN, respectively, and 72 pathways were identified by all 3 approaches, demonstrating convergence of signals across different analytical platforms (Figure 6B). This degree of overlap was unlikely to occur by chance alone as determined by bootstrapping (p-value < 0.001; Supplementary Text). The specific pathways identified by meta-integration indicated as driving molecular ontogeny over the first week of life were interleukin 1 signaling, Toll-like receptor signaling, the NOTCH signaling axis, the DHX RNA helicase pathway connecting to type 1 interferon gene expression, as well as regulators of the complement cascade (Supplementary Table 8). Given the striking convergence observed, we next assessed relevant functional interactions between the in silico identified pathways. To this end, we used NetworkAnalyst and found that 80% of the selected common features fit into minimum connected networks of experimentally-validated interactions (Figure 6C). This suggests that our meta-integration approach identified functional biological interactions relevant to neonatal ontogeny.

DISCUSSION

Here we present a holistic suite of complementary methods that address key hurdles of applying systems biology to newborns in the following ways: *i)* overcome limitations in sample volume via an efficient, field-compatible, sample-sparing SOP to process peripheral whole blood; *ii)* determine the developmental trajectories from each subject's own baseline (indexing) and *iii)* reduce the dimensionality of the dataset by integrating methodologically- and biologically-distinct data types (multi-OMICS integration), which

analytically validated signatures of important pathways. This approach revealed for the first time a dramatic molecular ontogeny over the first week of human life.

Multi-OMICS integration poses an exceptional problem with the increase in dimensionality (*p* features) relative to the typically achievable sample size (*n* samples). We here addressed this p>>>n problem by reducing the dimensionality of our data through selection of only a subset of variables (MMRN, DIABLO)[24], and to impose functionally-relevant, known molecular interaction information (NetworkAnalyst) [21]. Furthermore, as the integrative methods we applied focused on extracting aggregated information from each data set *independently* of each other, confidence in the results, if they independently converged on the same set of key molecular features, was high. This approach confirmed that integrating across multiple technological platforms representing different biological data types can dramatically increase robust biological insight. For example, this approach identified already known changes in early life, including those related to the composition of hemoglobin mRNA, increases in complement protein C9 [25], and reductions in plasma steroids [26], supporting the validity of our findings. However, our approach also identified pathways never before identified as relevant to ontogeny. For example, many pathways consistently identified by all three of our analytical strategies had previously been shown to be relevant for host defense, but they have never before been identified as central to early human development. These pathways include interleukin 1 signaling, Toll-like receptor signaling, NOTCH signaling, the DHX RNA helicase pathway, as well as regulators of the complement cascade [1, 27-30]. Data integration also led to the identification of novel and surprising but biologically plausible findings regarding ontogeny. For example, prostaglandin-endoperoxide synthase 2 (PTGS2 or PGHS2 or COX-2) appeared as centrally important in all of our integrative networks, but not in any of the single OMICS data types analyzed individually. PTGS2 is expressed abundantly in hematopoeitic progenitors [31] and is clinically relevant during premature labor and in necrotizing enterocolitis [32, 33]. Taken together, our findings being to outline a developmental trajectory that may serve as a reference akin to the stable steady state in adults [7, 10, 13]. Deviations from this developmental trajectory could potentially identify subjects at risk prior to the onset of disease or guide the re-establishment of homeostasis once derailed.

In summary, our method presents an integrative analysis across the broadest range of multivariate data sets published to date. Our integration over time (indexing data longitudinally for each newborn) and biological space (multi-OMICS integration) allowed exploration of the dynamic molecular and cellular developmental characteristics of early life. In parallel, we identified signals of potential physiological importance using a modest sample size readily obtainable in most studies. Coupled with our field-tested SOP for sample-sparing pre-analytical processing, we thus provide the methods to overcome key challenges in applying systems biology to neonates. Our data revealed a compelling 'biological narrative', where out of the apparently noisy age-dependent change across the first week of life, a consistent developmental trajectory emerged. These approaches and observations will serve as a crucial backdrop for future studies that characterize the impact of a broad array of factors, including genetics, epigenetics, maternal influences, microbiota, diet, and disease, as well as biomedical interventions such as vaccination.

METHODS (ONLINE)

Peripheral blood processing

30 healthy, term newborns were enrolled at the Medical Research Council (MRC) Unit The Gambia in accordance with a local Ethics Committee-approved protocol (SCC 1436). Following informed consent, mothers were screened for HIV-I and -II and Hepatitis B with positivity for either virus representing an exclusion criterion. Inclusion criteria were a healthy appearing infant as determined by physical examination, born by vaginal delivery at gestational age of ≥36 weeks, 5 minute Apgar scores ≥8, and a birth weight of ≥ 2.5 kilograms. Peripheral blood samples were obtained from all infants on the day of birth (DOL0) and then again either at DOL1, DOL3 or DOL7, in order to reduce venipuncture to a maximum of twice in the first week of life. Peripheral venous blood was drawn from infants via sterile venipuncture directly into heparinized collection tubes (Becton Dickinson (BD) Biosciences; San Jose, CA, USA). Aliquots (200 μl) were immediately placed in RNA-later (Ambion ThermoFisher, Waltham, MA, USA) with the remaining blood kept in the collection tubes at room temperature until further processing within 4 hours. All samples were processed at the MRC Unit The Gambia as described below and subsequently shipped to collaborating laboratories on dry ice, under temperature controlled and monitored conditions (World Courier, New Hyde Park, NY, USA).

Indexing

In this study, we profiled the blood of each subject twice over their first week of life, at DOL0 (baseline) and additionally at either DOL 1, 3, or 7, and sought to identify variables that differed between the baseline and latter time points across all subjects. That is, we were interested in the within-subject variation associated with DOL observed consistently across the entire cohort. For univariate analyses, we considered (indexed) paired differences, either implicitly (e.g. paired t-test), or, in the case of the flow cytometry and luminex cytokine data, explicitly by transforming the data beforehand using a multilevel approach to separate the between- and within-subject variation as described [34]. The same approach was used for all multivariate analyses.

Immune Phenotyping

Flow Cytometry (FCM). Whole blood was centrifuged on site at 500 *xg* for 10 minutes at room temp and plasma harvested and stored at -80°C for later analysis of plasma cytokines, proteins and metabolites. The amount of plasma removed from the whole blood after centrifugation was subsequently replaced with RPMI. For assessment of cellular composition by FCM, EDTA (0.2 mM final concentration) was added to the whole blood/RPMI mixture to ensure adherent cells were not lost. In parallel, cells were stained with fixable viability dye (FVD) at 4°C for 15 min prior to red blood cell lysis followed by storage at -80°C in Smart Tube reagents (Smart Tube Inc.; San Carlos, CA, USA). At the immunophenotyping laboratory samples were thawed, washed in staining buffer (PBSAN; 0.5% BSA, 0.1% sodium azide) and stained on ice in PBSAN with a cocktail of anchor markers to determine frequency of cell populations contained in peripheral blood (for list of cell types and anchor markers, see Supplemental Figure 1; for list of clone/ fluorochrome combination see MIFlowCyt section in Supplemental Information). Flow cytometric analysis employed a custom-built LSRII (for machine

settings and compensation settings see the MIFlowCyt section in Supplemental Information)[35]. Our gating strategy is shown in the Supplemental Figure 1. FCM data was analyzed in an automated fashion using R/BioConductor packages (Supplemental Figure 1). Specifically, *flowCore* supported the analysis in single files according to the Flow Cytometry Standard (FCS), providing the infrastructure to support sub-setting of data, data transformations and gating [36, 37]. Cell population identification was then conducted using *flowDensity*, a supervised gating tool, that was customized to provide threshold calculations designed for each cell subset based on expert knowledge of hierarchical gating order and one-dimensional density estimation [38]. Lastly, *flowType/RchyOptimyx* identified cell populations that correlated with outcome, in this case DOL at the time of blood draw [39]. *flowType* uses partitioning of cells, either manually or by clustering, into positive or negative for each marker to enumerate all cell types in a sample. *RchyOptimyx* measures the importance of these cell types by correlating their abundance to external outcomes, such as DOL, and distils the identified phenotypes to their simplest possible form.

Luminex. Plasma (25 μl) was used to measure cytokine concentrations using a custom-designed multi-analyte Cytokine Human Magnetic Panel bead array, (Invitrogen/Life Technologies, Carlsbad; CA) consisting of CCL2, CCL3, CCL5, CXCL8, CXCL10, GM-CSF, IFN-a2, IL-10, IL-12p40, IL-12p70, IL-1β, IL-6, and TNFα. Results were obtained with a Flexmap 3D system with Luminex xPONENT software version 4.2 (both from Luminex Corp.; Austin, TX, USA). Cytokine concentrations were determined using Milliplex Analyst software (version 3.5.5.0, Millipore).

Immunophenotypic Analysis. For flow cytometric as well as Luminex raw values were normalized with a 1 + Log2-transformation. WithinVariation matrices were computed for each data matrix using the WithinVariation function in R package *mixOmics* version 6.1.2. The Wilcoxon rank-sum test (wilcox.test in base R) was used to determine differentially regulated features within each data type, using the WithinVariation values for each feature. P-values were adjusted for each data type separately using the Benjamini-Hochberg method (p.adjust function, base R). Features were considered statistically different from DOL0 if their adjusted p-values were below 0.1. All analyses were performed in R version 3.3.2 (2016-10-31).

RNA-Seq

Total RNA was extracted from each sample using the RiboPure RNA purification kit (Ambion ThermoFisher; Waltham, MA, USA) following the manufacturer's protocol. Quantification and quality assessment of total RNA was performed using an Agilent 2100 Bioanalyzer (Santa Clara, CA, USA). Poly-adenylated RNA was captured using the NEBNext Poly(A) mRNA Magnetic Isolation Module (catalogue no.: E7409L, NEB; Ipswich, MA, USA). Strand-specific cDNA libraries were generated from poly-adenylated RNA using the KAPA Stranded RNA-Seq Library Preparation Kit (cat. no.: 07277253001, Roche; Basel, Switzerland). All cDNA libraries were prepared at the same time and sequenced on the HiSeq 2500 (Illumina; San Diego, CA, USA), using one Rapid v2 and two lanes of High Output single-read run of 100 bp-long sequence reads (+ adapter/index sequences). Sequence quality was assessed using FastQC v0.11.5 and MultiQC v0.8.dev0 [40]. The FASTQ sequence reads were aligned to the hg19 human

genome (Ensembl GRCh38.86) using STAR v2.5 and mapped to Ensembl GRCh38 transcripts [41]. Read-counts were generated using htseq-count (HTSeq 0.6.1p1)[42]. All data processing and subsequent differential gene expression analyses were performed using R version 3.3.0 and DESeq2 version 1.14.1 [43]. Genes with very low counts (with less than 10 counts in eight or more samples, or the smallest number of biological replicates within each treatment group) and globin transcripts were pre-filtered and removed *in silico*. Differentially expressed genes were identified using paired analysis with the Wald statistics test and filtering for any genes that showed 2-fold change and adjusted p-value < 0.05 (cut-off at 5% FDR) as the threshold. Functional discovery of pathway enrichment and network analyses was performed using Sigora 2.0.1 and *NetworkAnalyst*, respectively [17, 44].

Plasma Proteomics

Plasma samples were prepared for proteome analysis using the in-house (Boston Children's Hospital; Steen laboratory)-developed MStern blotting sample processing and trypsinization protocol [45], which was adapted for plasma samples [46]. To this end, 5 µL plasma was first diluted in 100 µL sample buffer (8 M urea in TRIS-HCl, pH 8.5). Protein disulfide bonds were then reduced with dithiothreitol (10 mM final concentration) for 30 min, and alkylated with iodoacetamide (50 mM final concentration) for 30 min in sample buffer. Three µL (approximately 10 µg) of this protein solution was then transferred to a 96 well plate with a polyvinylidene fluoride (PVDF) membrane at the bottom. Protein digestion was performed with sequencing-grade modified trypsin (V5111, Promega; Madison, WI, USA) at a nominal protease to protein ratio of 1:25 w/w. After incubation for two hours at 37°C, the peptides were eluted from the PVDF membrane, and concentrated to dryness in a vacuum centrifuge. To monitor retention time stability and system performance, iRT peptides (Biognosys, Schlieren, Switzerland) were spiked into all samples. Samples were analyzed using a nanoLC system (Eksigent; Dublin, CA) equipped with a LCchip system (cHiPLC nanoflex, Eksigent) coupled online to a Q Exactive mass spectrometer (Thermo Scientific; Bremen, Germany). From each sample, 0.2 µg peptide material was separated using a linear gradient from 93% solvent A (0.1% formic acid in water), 7% solvent B (0.1% formic acid in acetonitrile) which was increased to 32% solvent B over 60 min. The mass spectrometer was operated in data-dependent mode, selecting up to 12 of the most intense precursors for fragmentation from each precursor scan. Label-free protein quantitation analysis employed MaxQuant 1.5.3.30 [47]. Raw-data were downloaded and used to build a matching library and searched against the UniProt Human Reference Proteome as described [48]. Standard search settings were employed with the following modifications: Max missed cleavage 3; variable modifications Deamidation (NQ) and Oxidation (M) [49]. A revert decoy search strategy was employed to filter all proteins and peptides to < 1% FDR[50]. The list of proteins was further processed in Perseus 1.5.5.3, log2-transformed, and proteins with less than two peptides (razor) were filtered out as described [51]. The samples were grouped according to DOL, and proteins which were not quantifiable in at least five of the samples in any day were filtered out. Remaining missing values were imputed using numbers drawn from a normal distribution with the standard parameters in Perseus (width 0.3, downshift 1.8) to simulate signals from low abundant proteins [52] The R-script *ComBat* was used to correct for batch effects for samples run on different LC-MS columns [53]. Proteins with a statistically significant change of abundance between

different DOL were identified by paired two-samples t-test. To correct for multiple hypothesis testing, permutation-based false positive control was applied using standard parameters in Perseus (FDR=0.05, s0=0.1)[54]. Significant proteins were further analyzed in Cytoscape, SIGORA [44] with the Reactome [55] gene annotation system and DAVID [56].

Plasma Metabolomics

Plasma samples were analyzed using the non-targeted metabolomics platform of Metabolon Inc. (Durham, NC, USA). Samples were extracted and prepared for analysis using Metabolon's solvent extraction method [57]. Each plasma sample was stored at -80^{o}C and accessioned into the Metabolon Laboratory Information Management System (LIMS). Recovery standards were added at the first step in the extraction process to allow quality control for the entire process. Protein removal employed methanol precipitation (Glen Mills GenoGrinder 2000) followed by centrifugation. The supernatants were divided into five fractions: two for analysis by two separate reverse phase (RP)/UPLC-MS/MS methods with positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS with negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI, and one sample was reserved for backup using Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. Compounds were identified by comparison to Metabolon library entries of purified standards or recurrent unknown entities. Furthermore, biochemical identifications were based on three criteria: retention index (RI) within a narrow RI window of the proposed identification, accurate mass match to the library ± 10 ppm, and MS/MS forward and reverse scores between the experimental data and authentic standards found in the LIMS. MS/MS scores were based on a comparison of the ions present in the experimental spectrum to the ions present in the library spectrum. Exact molecular mass data from redundant m/z peaks corresponding to the formation of different parent and product ions were first used to help confirm the metabolite molecular mass. Metabolon's MassFragmentTM application manager (Waters MassLynx v4.1, Waters corp.; Milford, USA) was used to facilitate the MS/MS fragment ion analysis process using peak-matching algorithms and quantified using an area-under-the-curve in-house algorithm. Finally, the identities of the specific metabolites were confirmed by comparing their mass spectra and chromatographic retention times with commercially available reference standards.

Metabolomic data were then pre-processed to original scale values by Bradford protein assay of cellular protein levels and to volume extracted values obtained on a per sample basis. Data was normalized to the interquartile range, in which the metabolites with an IQR of zero are uninformative and can be deleted from further analyses. Missing values below the threshold level were imputed with the minimum value obtained for the respective biochemical from within-all-groups. Each biochemical value was then transformed using generalized logarithmic transformation for distributional assumptions of chemometrics. Statistical analyses were performed using R statistical packages and MetaboAnalyst 3.0 [58]. Paired t-tests with $p \leq 0.05$ were employed to identify altered metabolites between paired samples. Regulation of individual metabolites was compared across groups and time using repeated measures ANOVA and generalized linear mixed model (GLMM) for longitudinal models. Multiple comparisons were

accounted for by estimating the false discovery rate (FDR) using q-values [59]. Significantly altered pathways were determined by pathway set enrichment analysis within Metabolon Pathway Analysis (MPA) software which was determined by the following equation: # of significant metabolites in pathway (*k*)/total # of detected metabolites in pathway (*m*)/total # of significant metabolites (*n*)/total # of detected metabolites (*N*) or (*k*/*m*)/(*n*/*N*). Pathways with a higher number of experimentally regulated compounds relative to the overall study in a follow-up day (e.g., DOL7 vs. DOL0) received a pathway impact score > 1 suggesting that these pathways may be of interest to the metabolic perturbations observed. Pathway analysis and visualization employed the KEGG Pathway [60], The Human Metabolome Database Version 3.6 (HMDB) [61] and Metabolync Pathway Analysis using Cytoscape plugin [62]. In addition, metabolite IDs were converted to HMDB [61], KEGG [60], and PubChem IDs via the Metaboanalyst ID mapping tool [58]. Metabolites were mapped to their directly interacting enzymes via BioCyc (executed in R) [63]. Metabolites were also mapped to their directly interacting enzymes via KEGG (executed in R via functional web scraping). The union of genes from the previous two steps was then used to proceed with metabolome data integration. Two pathway mapping databases were used, as neither appeared to contain a complete list of all metabolites and reactions. The list of proteins derived from the metabolite data processing (above) and lists of differentially expressed genes or proteins were then combined. Pathway enrichment analysis employed SIGORA [44] using the Reactome gene annotation system [55].

Data integration

Data obtained *via* the immune phenotyping (cellular composition and plasma cytokines), transcriptomic, proteomic and metabolomic methods was integrated to identify correlations of signatures across these methodologically- and biologically-distinct datasets, since convergence of signatures across such diverse biological domains provides an independent assessment and approaches functional validation [7]. In addition, we aimed to determine whether we could derive any novel biological information via the integration of multiple OMICS data types that was not revealed in a single dataset alone. To cross-validate results we employed 3 data integration platforms, each applying a different analytical strategy; these independent but complementary data-driven *vs.* knowledge/network-driven strategies, pursued in parallel, decreased the chance of false discoveries. Specifically, data integration strategies included: *i)* a novel method for integrating multiple OMICS data types into known protein-protein interaction networks (Molecular Interaction Network) using NetworkAnalyst that provided context based on annotated of molecular interactions [17]; *ii)* a new approach to identify the underlying key drivers of ontogeny using specialized multivariate methods capable of identifying relevant features from high-dimensional datasets, namely sparse generalized canonical correlation discriminant analysis *via* the '*Data Integration Analysis for Biomarker discovery using a Latent component method for Omics*' (DIABLO) framework, which is part of the mixOmics R package [18-20]; and *iii)* the recently published multi-scale, multifactorial response network (MMRN), querying the informatically derived correlations within a network for statistically significant features [7]. Finally, we also 'integrated the integration' by combining the protein-protein-interaction driven approach (NetworkAnalyst) with the feature-mining power of DIABLO ('meta-integration').

Molecular Interaction Network. NetworkAnalyst integrates data using protein-protein interactions as a biological framework. Metabolomic data required pre-processing in order to integrate into a network with transcriptomic and proteomic data, as metabolites must be associated with proteins that are involved in their creation and/or degradation. To identify such proteins, the following steps were taken: Metabolite IDs were converted to HMDB, KEGG, and PubChem IDs via the Metaboanalyst ID mapping tool [58]. Metabolites were mapped to their directly interacting enzymes via BioCyc (executed in R)[63]. Metabolites were also mapped to their directly interacting enzymes via KEGG (executed in R via functional web scraping). The union of genes from the previous two steps was then used to proceed with metabolome data integration. Two pathway-mapping databases were used, as neither alone appeared to contain a complete list of all metabolites and reactions. The list of proteins derived from the metabolite data processing (above) and lists of differentially expressed genes or proteins were then combined using NetworkAnalyst to produce a zero-order or minimum-connected network, depending on the size of the dataset [17]. Networks consist of seed nodes (proteins/genes that were used as input to build the network), novel nodes (proteins/genes that are 1st order interactors of the seed nodes and are used as "scaffholding" to build the network) and edges (links that join the nodes together and are indicative of a molecular interaction between nodes). Our focus was on the identification of novel nodes as emergent information that can be derived from a biological network. Node lists were then downloaded to identify median degree of connectivity for nodes of each data type and to identify novel nodes stemming from data integration. Transcriptomes were integrated via their respective encoded gene products.

A minimum connected network for transcriptomic, proteomic and metabolomic data was constructed using NetworkAnalyst based on differentially present genes/proteins on each DOL vs DOL0. Novel nodes were identified as those, which were not present in the set of differentially present genes/proteins originally entered (i.e. not part of the seed nodes). Note that for DOL3 vs DOL0 and DOL7 vs DOL0, the transcriptomic datasets were simplified by first producing a zero order network, and only using these nodes for minimum connected network construction. Next, pairwise lists of each data type were constructed and minimum connected networks were constructed. Novel nodes in each network were identified as nodes that were not in either set of differentially present genes/proteins or either minimum connected network constructed from a single data type in the pairwise set. A minimum connected network for all data types was then constructed and novel nodes were identified as nodes that were not in any differentially present list of genes/proteins/metabolites, any novel nodes in a minimum network constructed from any single data type, or any novel nodes in a minimum network constructed from any pairwise data. Pathway enrichment analysis was run using SIGORA software [44] with the Reactome ontology system on the node lists from each minimum connected network built during the novel node identification process outlined above. For each data type alone, (transcriptomics, proteomics and metabolomics), the number of unique pathways identified were counted as novel pathways. For each pairwise combination of data types, previously identified pathways that had been identified in either data type alone were subtracted. For the three-way combination, all pairwise pathways and unique data type pathways were subtracted.

DIABLO. Prior to integration with DIABLO, data were transformed. Specifically, cell proportions from the flow cytometry were normalized to total cell counts; the resulting relative cell proportions were then transformed using centered log-ratios [64]. Normalized transcriptomic, proteomic, cytokine, and metabolomic data were log-transformed. We further decomposed the within-subject from the between-subject variance in the data sets to account for repeated measures [65]. This is analogous to normalizing all samples to their DOL0 time point. Finally, a broad, unsupervised, variance-based filter was applied to the transcriptomic data, retaining the 50% most relevant features. To integrate across data types, we applied sparse generalized canonical correlation discriminant analysis via the DIABLO framework, part of the mixOmics package [18, 19, 66]. DIABLO constructs components across any number of input matrices, maximizing their covariance with each other and a given response variable (in this case, DOL), while simultaneously performing feature selection by applying an L1 penalty (LASSO) on the model loadings [23]. Importantly, DIABLO identifies key drivers associated with the response variable of interest across all input data matrices jointly. Cross-validation (20 x 5-fold) was used to determine the optimal model hyperparameters (number of components, number of features per component), as well as to provide an estimate of the ability of the model to generalize to new data. Selected model features, i.e. key correlates of change across DOL, were subjected to pathway over-representation analysis against the Reactome pathways database (obtained via MSigDB [67]) and BTM [21] annotated gene set libraries, using the hypergeometric test. Obtained p-values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure [68], either separately per component, or after enrichment to include their first degree neighbours sub-graph in the protein:protein interaction data [17].

Multi-scale, Multifactorial response Network. Examples of systematic integration across many large OMICS datasets remain rare. Recently, Li and others [7] used multi-scale, multifactorial response networks to study the immune response to varicella zoster vaccine across four OMICS datasets. We wanted to compare insights generated by our chosen approaches to theirs and, to that end, created a multi-scale interaction network as they described, with the following important modifications. First, instead summarizing the higher dimensional data blocks (transcriptomics, metabolomics) to either gene modules or metabolic pathways (Blood Transcriptomic Modules (BTM])[69]; Metabolon Inc. annotation, respectively) using a simple average, we used eigen-gene summarization[70], a weighted average based on the 1st principal component of the data, in order to maximize the variation explained. Second, clusters were identified at each DOL using Euclidean distance and Ward's method[71], as recommended by Li (personal communication). The optimal number of clusters was determined using the elbow criterion[72]. Stable clusters were identified by comparing cluster membership using the Szymkiewicz-Simpson coefficient [73]. Finally, cluster association with DOL was assessed using the Correlation Adjusted MEan RAnk (Camera) gene set test [74].

Meta-Integration. Minimum connected networks for both components 1 and 2 of the DIABLO model were constructed using NetworkAnalyst, using pairwise lists of each data type for both components to construct a minimum connected network.

Identifying convergence across data integration strategies. The various data integration approaches resulted in outputs that were not directly comparable. We simplified their outputs to lists of features of interests to enable direct comparisons. NetworkAnalyst: list of features that made up the minimal connected network. DIABLO: features selected by the model with hyperparameterization that minimized cross-validation error rate. MMRN: features that made up the inter-connected node-networks that were stably identified across DOLs and significantly associated with DOL. We then compared the identified feature sets and carried out gene set enrichment analysis against the Reactome collection.

FIGURE LEGENDS.

Figure 1. Blood Processing Overview. Newborn peripheral venous blood was drawn directly into heparinized collection tubes. Aliquots (200 μl) were removed for transcriptomic analysis. Plasma was then harvested after a spin, and cryopreserved for cytokine, proteomic and metabolomic analysis. The remaining cellular fraction was diluted to replace the plasma removed, and 100 μl aliquots from this mixture were processed for single cell immunophenotyping by flow cytometry. With a starting volume of 1 ml, this SOP still left the cellular fraction contained in 400 μL of blood that could be used for other analysis.

Figure 2. Transcriptomic analyses identified dramatic increases in the number of differentially expressed genes during the first week of life. In **A**, up- and down-regulated differentially expressed genes are plotted by DOL (vs DOL0) and numbers of genes are listed above each point except for down-regulated genes at DOL1 vs DOL0, where the number is zero. **B** shows zero order interaction networks for genes that differentially expressed at DOL3 vs DOL0 and DOL7 vs DOL0. Within networks, up-regulated nodes are displayed in red and down-regulated nodes in green. **C** shows selected pathways enriched among up- and down-regulated genes along with Bonferroni-corrected *p*-values for DOL3 and DOL7.

Figure 3. Integration of Multiple Data Types via NetworkAnalyst provided novel insight. In **A**, the number of *novel nodes* derived from molecular networks integration from each combination of data types on each DOL vs DOL0 are shown. **B** shows minimum connected networks containing all three individual data types, where nodes derived from the transcriptome are shown in blue, nodes from the metabolome in red, and nodes from the proteome in green. Novel nodes are shown in orange. **C** shows enrichment of *novel pathways* on DOL1, 3 and 7 vs DOL0 from each data type and their respective combinatorial analysis. **D** lists selected novel pathways along with their Bonferroni corrected *p*-values and the data type combinations from which they were derived.

Figure 4: DIABLO uncovered biologically relevant features by integrating information across data types. Schematic representation of two contrasting integration approaches using multi-variate techniques: **A** shows that DIABLO selects features jointly across data types, resulting in the identification of features with strong associations across data types. Conversely, as shown in **B**, ensembles of multi-variate models, constructed independently of each other, result in a selection of features that are poorly associated across data types. This is visualized in correlation heatmaps of the selected features (middle) and corresponding networks (right), with dense sub-graphs, or network modules, encircled. In particular, the network modules identified in **A (right)** include a number of features selected from all data types. This is not the case in **B (right)**. The minimal set of features selected by DIABLO across data types as shown in **C** could discriminate between DOL and distinct sets of these features separated birth from all other DOLs (DIABLO component 1) and DOL 1, 3, and 7 (DIABLO component 2). Features identified by DIABLO (blue bars) were largely distinct from those identified

by more traditional single-OMICS multi-variate approaches (red bars; overlaps in grey); shown in **D** using an UpSet plot. Moreover, features identified by DIABLO were more strongly enriched for known biological (functional) pathways; shown in **E** using an UpSet plot (blue vs. red bars). Horizontal bars are mapped to the number of elements in each set of features being compared. Vertical bars correspond to the number of elements in the intersections when carrying out various set comparisons.

Figure 5. Multi-scale, multifactorial response networks identified a small set of biological processes consistently associated with each other across all DOLs. Networks were constructed separately at each DOL and stable clusters identified. The partial least squares regression scores, a measure of the strength of association between data types, were highest at DOL 1 and decreased across the first week of life are shown in **A**; the transcriptomic and metabolomic data were most strongly associated as indicated in **B**, followed by the proteomics and cytokine data, metabolomic and proteomic, and flow cytometry and transcriptomic. Association of stable and transient clusters with DOL was assessed using the Correlation-Adjusted Mean RAnk (CAMERA) method. **C** shows the comparison of the distribution of the observed Benjamini-Hochberg false discovery rates to expected quantiles is shown in a Quantile-quantile (QQ) plot. Stable clusters were more strongly associated with DOL compared to transient ones. Finally, a small set of stable clusters were consistently associated at all 3 DOLs, forming a stable sub-network; the features that make up this stable sub-network are visualized in a heatmap in **D**.

Figure 6. Meta-integration confirmed findings across analytical methods. **A** is an UpSet plot that depicts the concordance in features selected by the three different integration methods. Features identified by at least two methods are shown in blue, while features identified by all three methods are shown in red. The concordance between significantly enriched pathways (Benjamini-Hochberg FDR ≤ 0.05) identified by the three methods are shown in the upset plot in **B**, with gene sets identified by at least 2 of the 3 (blue) or all 3 methods (red) highlighted. Minimum connected networks containing luminex (purple nodes), proteome (green nodes) and transcriptome (blue nodes) features from DIABLO components 1 and 2 are shown in **C**.

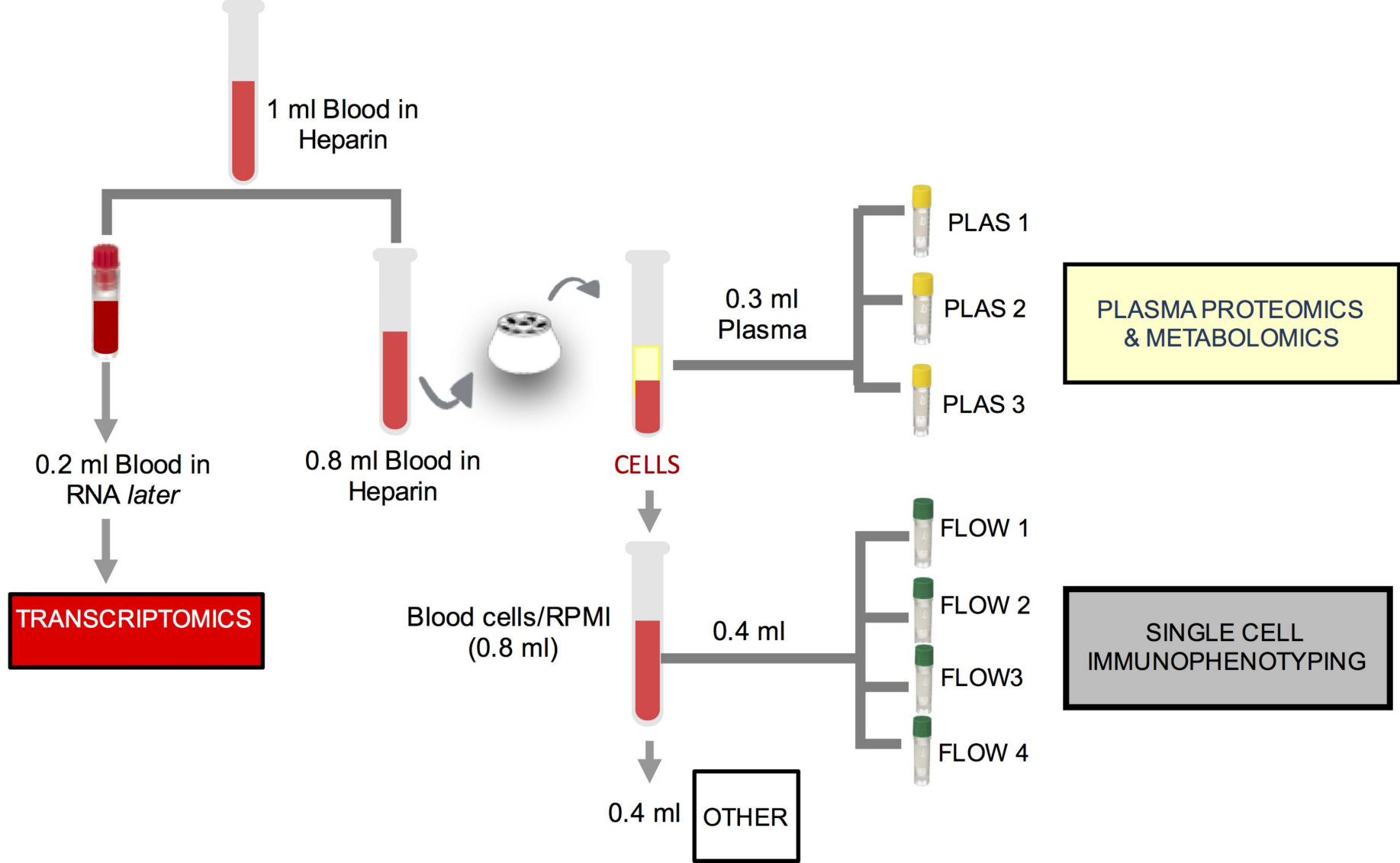
REFERENCES

1. Kollmann, T.R., Kampmann, B., Mazmanian, S.K., Marchant, A. & Levy, O. Protecting the Newborn and Young Infant from Infectious Diseases: Lessons from Immune Ontogeny. *Immunity* **46**, 350-363 (2017).
2. Zhang, X., Zhivaki, D. & Lo-Man, R. Unique aspects of the perinatal immune system. *Nature reviews* (2017).
3. Balbus, J.M. et al. Early-life prevention of non-communicable diseases. *Lancet* **381**, 3-4 (2013).
4. Amenyogbe, N., Kollmann, T.R. & Ben-Othman, R. Early-Life Host-Microbiome Interphase: The Key Frontier for Immune Development. *Frontiers in pediatrics* **5**, 111 (2017).
5. Chaussabel, D. & Pulendran, B. A vision and a prescription for big data-enabled medicine. *Nat Immunol* **16**, 435-439 (2015).
6. Amenyogbe, N., Levy, O. & Kollmann, T.R. Systems vaccinology: a promise for the young and the poor. *Philos Trans R Soc Lond B Biol Sci* **370** (2015).
7. Li, S. et al. Metabolic Phenotypes of Response to Vaccination in Humans. *Cell* **169**, 862-877.e817 (2017).
8. Howie, S.R. Blood sample volumes in child health research: review of safe limits. *Bulletin of the World Health Organization* **89**, 46-53 (2011).
9. Tsang, J.S. Utilizing population variation, vaccination, and systems biology to study human immunology. *Trends Immunol.* **36**, 479-493. doi: 410.1016/j.it.2015.1006.1005. Epub 2015 Jul 1014. (2015).
10. Carr, E.J. et al. The cellular composition of the human immune system is shaped by age and cohabitation. *Nat Immunol* **17**, 461-468 (2016).
11. Smolen, K.K. et al. Single-Cell Analysis of Innate Cytokine Responses to Pattern Recognition Receptor Stimulation in Children across Four Continents. *J Immunol* (2014).
12. Smolen, K.K. et al. Pattern recognition receptor-mediated cytokine response in infants across 4 continents. *The Journal of allergy and clinical immunology* **133**, 818-826 e814 (2014).
13. Tsang, J.S. et al. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* **157**, 499-513 (2014).
14. Shannon, C.P. et al. Two-stage, in silico deconvolution of the lymphocyte compartment of the peripheral whole blood transcriptome in the context of acute kidney allograft rejection. *PloS one* **9**, e95224 (2014).
15. Smith, C.L. et al. Identification of a human neonatal immune-metabolic network associated with bacterial infection. *Nature communications* **5**, 4649 (2014).
16. Henry, E. & Christensen, R.D. Reference Intervals in Neonatal Hematology. *Clinics in perinatology* **42**, 483-497 (2015).
17. Xia, J., Gill, E.E. & Hancock, R.E. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature protocols* **10**, 823-844 (2015).
18. Le Cao, K.A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics* **12**, 253 (2011).
19. Singh, A. et al. DIABLO - an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv* (2016).

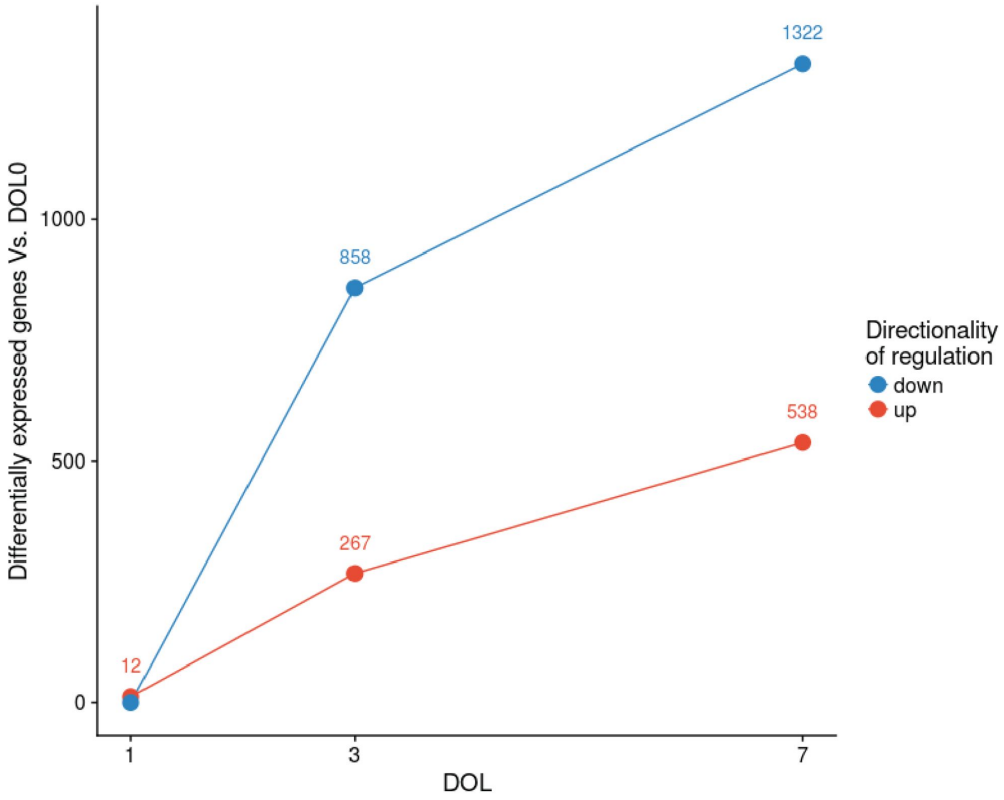
20. Rohart, F., Gautier, B., Singh, A. & Le Cao, K.-A. mixOmics: an R package for 'omics feature selection and multiple data integration. *bioRxiv* (2017).
21. Chaussabel, D. & Baldwin, N. Democratizing systems immunology with modular transcriptional repertoire analyses. *Nature reviews* **14**, 271-280 (2014).
22. Breuer, K. et al. InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic acids research* **41**, D1228-1233 (2013).
23. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. (1996).
24. Fan, J. & Lv, J. A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica* **20**, 101-148 (2010).
25. Pettengill, M.A., van Haren, S.D. & Levy, O. Soluble mediators regulating immunity in early life. *Frontiers in immunology* **5**, 457 (2014).
26. Ismail, A.A., Walker, P.L., Macfaul, R. & Gindal, B. Diagnostic value of serum testosterone measurement in infancy: two case reports. *Annals of clinical biochemistry* **26 (Pt 3)**, 259-261 (1989).
27. Radtke, F., Wilson, A., Mancini, S.J. & MacDonald, H.R. Notch regulation of lymphocyte development and function. *Nat Immunol* **5**, 247-253 (2004).
28. Abdelhaleem, M. RNA helicases: regulators of differentiation. *Clinical biochemistry* **38**, 499-503 (2005).
29. Loo, Y.M. & Gale, M., Jr. Immune signaling by RIG-I-like receptors. *Immunity* **34**, 680-692 (2011).
30. Schmidt, C.Q., Lambris, J.D. & Ricklin, D. Protection of host cells by complement regulators. *Immunological reviews* **274**, 152-171 (2016).
31. Romero-Moya, D. et al. Cord blood-derived CD34+ hematopoietic cells with low mitochondrial mass are enriched in hematopoietic repopulating stem cell function. *Haematologica* **98**, 1022-1029 (2013).
32. Lugo, B., Ford, H.R. & Grishin, A. Molecular signaling in necrotizing enterocolitis: regulation of intestinal COX-2 expression. *Journal of pediatric surgery* **42**, 1165-1171 (2007).
33. Reinebrant, H.E. et al. Cyclo-oxygenase (COX) inhibitors for treating preterm labour. *Cochrane Database Syst Rev*, Cd001992 (2015).
34. Westerhuis, J.A., van Velzen, E.J., Hoefsloot, H.C. & Smilde, A.K. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics : Official journal of the Metabolomic Society* **6**, 119-128 (2010).
35. Lee, J.A. et al. MIFlowCyt: the minimum information about a Flow Cytometry Experiment. *Cytometry* **73**, 926-930 (2008).
36. Hahne, F. et al. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics.* **10:106.**, 10.1186/1471-2105-1110-1106. (2009).
37. Spidlen, J. et al. Data File Standard for Flow Cytometry, version FCS 3.1. *Cytometry A* **77**, 97-100 (2010).
38. Malek, M. et al. flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics.* **31**, 606-607. doi: 610.1093/bioinformatics/btu1677. Epub 2014 Oct 1016. (2015).
39. O'Neill, K., Jalali, A., Aghaeepour, N., Hoos, H. & Brinkman, R.R. Enhanced flowType/RchyOptimyx: a BioConductor pipeline for discovery in high-dimensional cytometry data. *Bioinformatics* **30**, 1329-1330 (2014).

40. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048 (2016).
41. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
42. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
43. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
44. Foroushani, A.B., Brinkman, F.S. & Lynn, D.J. Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures. *PeerJ* **1**, e229 (2013).
45. Berger, S.T. et al. MStern Blotting-High Throughput Polyvinylidene Fluoride (PVDF) Membrane-Based Proteomic Sample Preparation for 96-Well Plates. *Molecular & cellular proteomics : MCP* **14**, 2814-2823 (2015).
46. Bennike, T.B. & Steen, H. High-Throughput Parallel Proteomic Sample Preparation Using 96-Well Polyvinylidene Fluoride (PVDF) Membranes and C18 Purification Plates. *Methods Mol Biol* **1619**, 395-402 (2017).
47. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics : MCP* **13**, 2513-2526 (2014).
48. Keshishian, H. et al. Multiplexed, Quantitative Workflow for Sensitive Biomarker Discovery in Plasma Yields Novel Candidates for Early Myocardial Injury. *Molecular & cellular proteomics : MCP* **14**, 2375-2393 (2015).
49. Bennike, T. et al. A normative study of the synovial fluid proteome from healthy porcine knee joints. *Journal of proteome research* **13**, 4377-4387 (2014).
50. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368-375 (2003).
51. Bennike, T.B. et al. Proteome Analysis of Rheumatoid Arthritis Gut Mucosa. *Journal of proteome research* **16**, 346-354 (2017).
52. Deeb, S.J., D'Souza, R.C., Cox, J., Schmidt-Supprian, M. & Mann, M. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. *Molecular & cellular proteomics : MCP* **11**, 77-89 (2012).
53. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)* **8**, 118-127 (2007).
54. Camargo, A., Azuaje, F., Wang, H. & Zheng, H. Permutation - based statistical tests for multiple hypotheses. *Source code for biology and medicine* **3**, 15 (2008).
55. Croft, D. et al. The Reactome pathway knowledgebase. *Nucleic acids research* **42**, D472-477 (2014).
56. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).
57. Evans, A.M., DeHaven, C.D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification

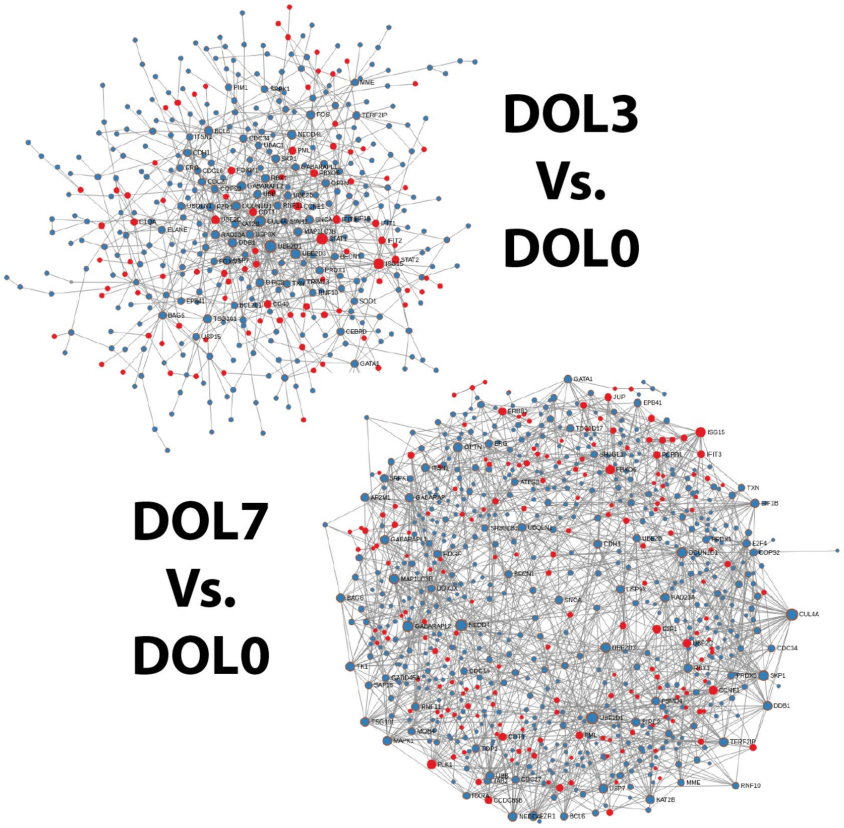
of the small-molecule complement of biological systems. *Analytical chemistry* **81**, 6656-6667 (2009).
58. Xia, J., Sinelnikov, I.V., Han, B. & Wishart, D.S. MetaboAnalyst 3.0--making metabolomics more meaningful. *Nucleic acids research* **43**, W251-257 (2015).
59. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**, 9440-9445 (2003).
60. Ogata, H., Goto, S., Fujibuchi, W. & Kanehisa, M. Computation with the KEGG pathway database. *Bio Systems* **47**, 119-128 (1998).
61. Wishart, D.S. et al. HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic acids research* **41**, D801-807 (2013).
62. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432 (2011).
63. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research* **42**, D459-471 (2014).
64. Aitchison, J. The Statistical Analysis of Compositional Data. (1982).
65. Liquet, B., Le Cao, K.A., Hocini, H. & Thiebaut, R. A novel approach for biomarker selection and the integration of repeated measures experiments from two assays. *BMC bioinformatics* **13**, 325 (2012).
66. Singh, A. et al. Identifying Molecular Mechanisms of the Late-Phase Asthmatic Response by Integrating Cellular, Gene, and Metabolite Levels in Blood. *Annals of the American Thoracic Society* **13 Suppl 1**, S98 (2016).
67. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
68. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. (1995).
69. Li, S., Rouphael, N., Duraisingham, S., Romero-Steiner, S. & Presnell, S. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. **15**, 195-204 (2014).
70. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology* **1**, 54 (2007).
71. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58**, 236-244 (1963).
72. Thorndike, R.L. Who belongs in the family? *Psychometrika* **18**, 267-276 (1953).
73. Shannon, C.P. et al. SABRE: a method for assessing the stability of gene modules in complex tissues and subject populations. *BMC bioinformatics* **17**, 460 (2016).
74. Wu, D. & Smyth, G.K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research* **40**, e133 (2012).


1 ml Blood in Heparin
0.2 ml Blood in RNA later
TRANSCRIPTOMICS
0.8 ml Blood in Heparin
CELLS
0.3 ml Plasma
PLAS 1
PLAS 2
PLAS 3
PLASMA PROTEOMICS & METABOLOMICS
Blood cells/RPMI (0.8 ml)
0.4 ml
FLOW 1
FLOW 2
FLOW3
FLOW 4
SINGLE CELL IMMUNOPHENOTYPING
0.4 ml
OTHER

A.



B.



C.

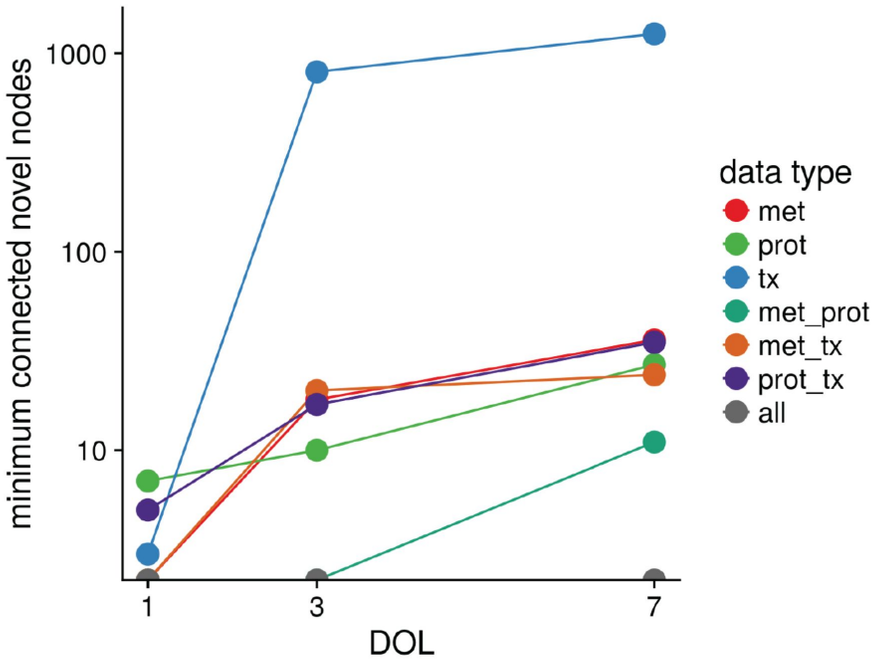
DOL3 Vs. DOL0 Enriched Pathways

Up-regulated		Down-regulated	
Pathway Name	**Bonferroni**	**Pathway Name**	**Bonferroni**
Interferon alpha/beta signaling	1.46E-86	Heme biosynthesis	6.53E-35
Negative regulators of RIG-I/MDA5 signaling	2.35E-31	Cellular responses to stress	8.31E-27
Interferon gamma signaling	8.18E-23	Macroautophagy	2.36E-22
ISG15 antiviral mechanism	1.19E-03	Iron uptake and transport	5.70E-11
		Downregulation of SMAD2/3:SMAD4 transcriptional activity	5.76E-07

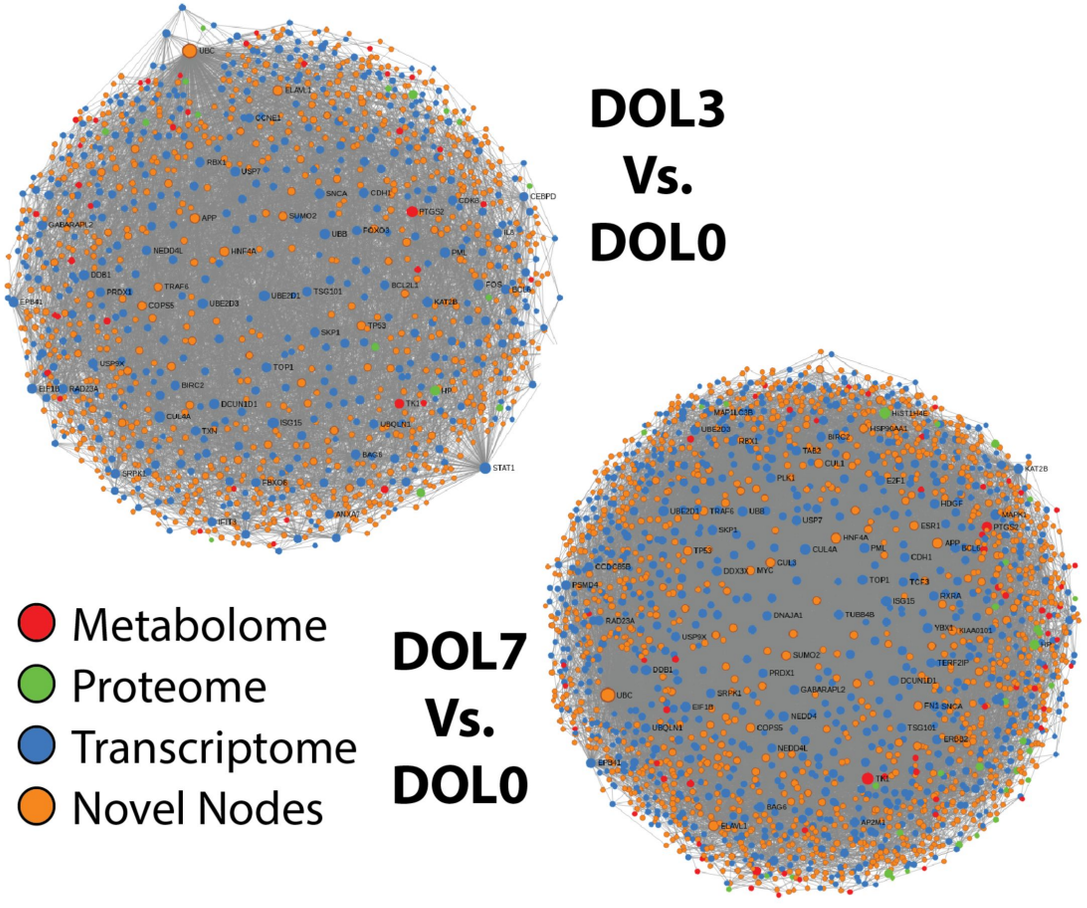
DOL7 Vs. DOL0 Enriched Pathways

Up-regulated		Down-regulated	
Pathway Name	**Bonferroni**	**Pathway Name**	**Bonferroni**
Interferon alpha/beta signaling	1.91E-72	Detoxification of Reactive Oxygen Species	4.01E-31
Classical antibody-mediated complement activation	3.76E-19	Golgi Associated Vesicle Biogenesis	4.83E-23
Signaling by FGFR1	7.13E-04	MyD88-independent TLR3/TLR4 cascade	1.86E-09
Chemokine receptors bind chemokines	1.74E-03	Antigen processing: Ubiquitination & Proteasome degradation	6.51E-09
SHC1 events in EGFR signaling	1.47E-02	Interleukin-6 signaling	2.19E-05

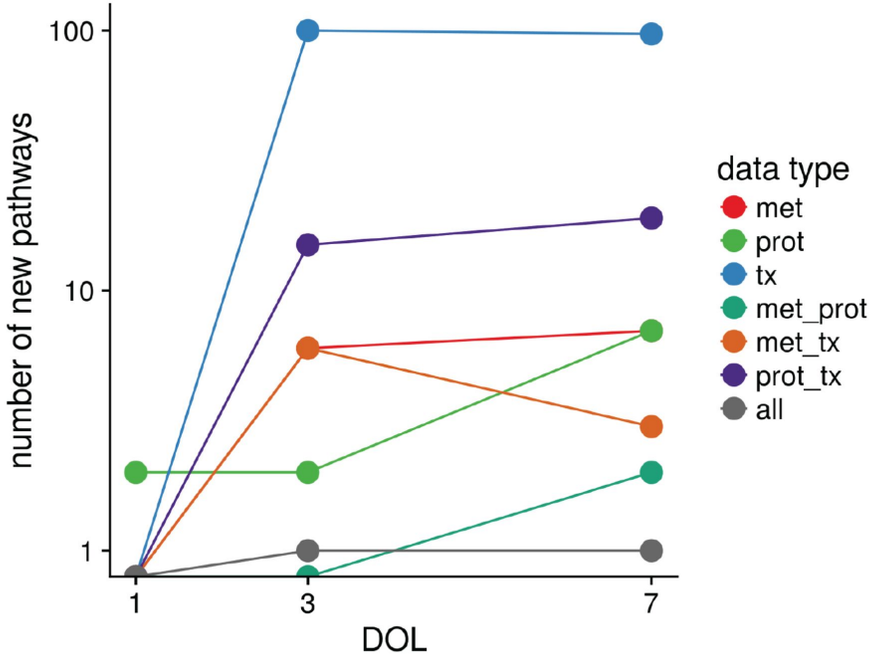
A.



B.



C.

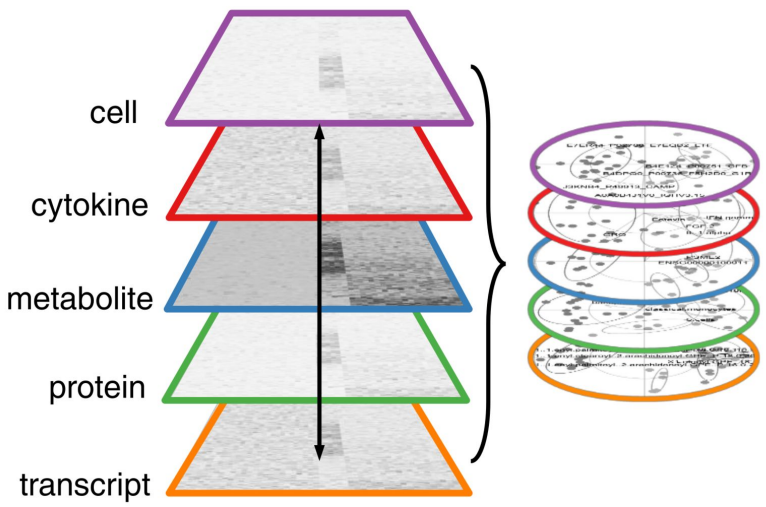


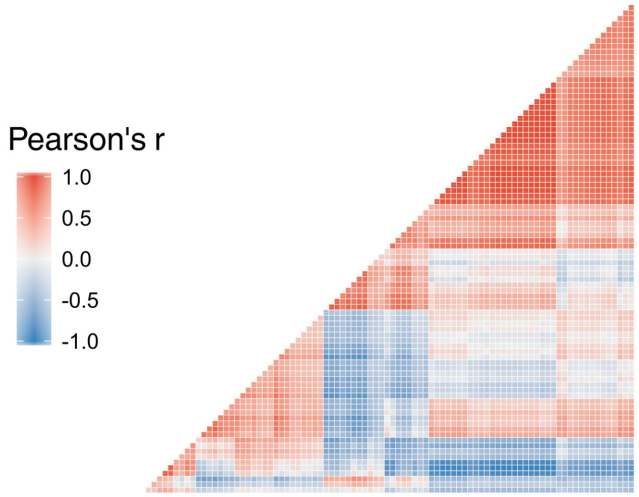
D.

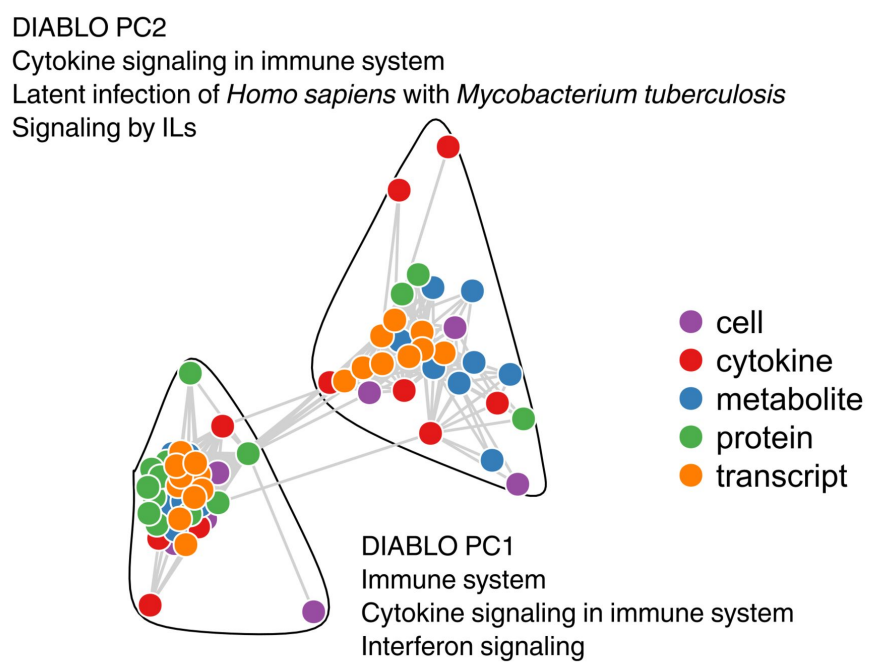
Novel Pathways from DOL3 Vs. DOL0 Data Integration		
Pathway Name	**Bonferroni**	**Data Types**
Transcriptional regulation of white adipocyte differentiation	2.03E-13	Prot + Tx
Metabolism of porphyrins	1.14E-07	Met + Tx
RHO GTPases activate PKNs	1.82E-04	Prot + Tx
Chromatin organization	1.96E-03	Prot + Tx
Sialic acid metabolism	1.62E-02	Met + Tx
Cytosolic sensors of pathogen-associated DNA	3.71E-02	All

Novel Pathways from DOL7 Vs. DOL0 Data Integration		
Pathway Name	**Bonferroni**	**Data Types**
Endosomal/Vacuolar pathway	3.65E-11	Met + Prot
Clathrin derived vesicle budding	1.04E-06	Prot + Tx
Signaling by EGFR	1.47E-06	Prot + Tx
GLI3 is processed to GLI3R by the proteasome	3.50E-02	Prot + Tx
Death Receptor Signalling	4.16E-02	Prot + Tx
Platelet activation, signaling and aggregation	4.60E-02	Prot + Tx

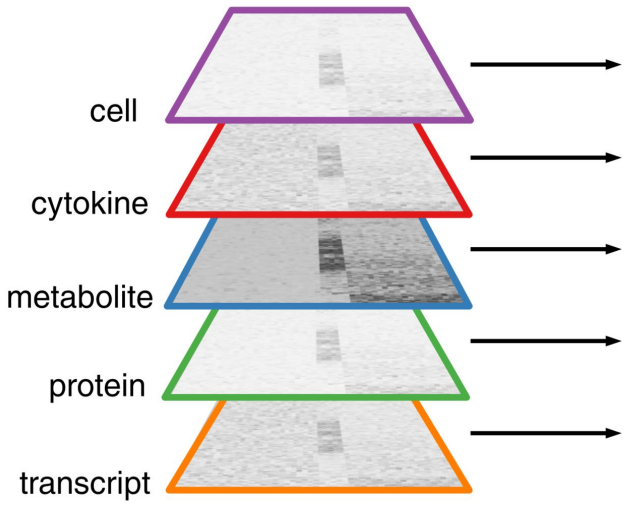
A.

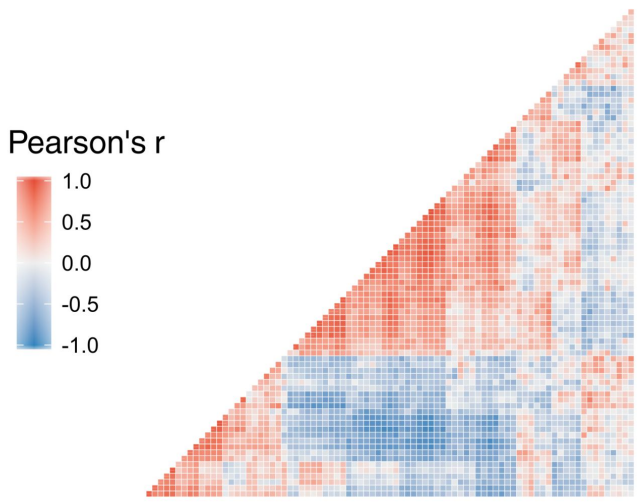

cell
cytokine
metabolite
protein
transcript

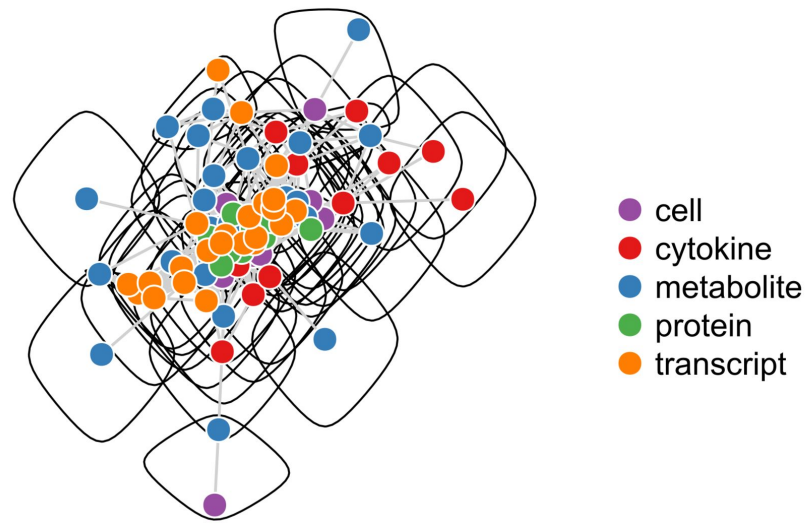

Pearson's r
1.0
0.5
0.0
-0.5
-1.0


DIABLO PC2
Cytokine signaling in immune system
Latent infection of Homo sapiens with Mycobacterium tuberculosis
Signaling by ILs
DIABLO PC1
Immune system
Cytokine signaling in immune system
Interferon signaling
cell
cytokine
metabolite
protein
transcript

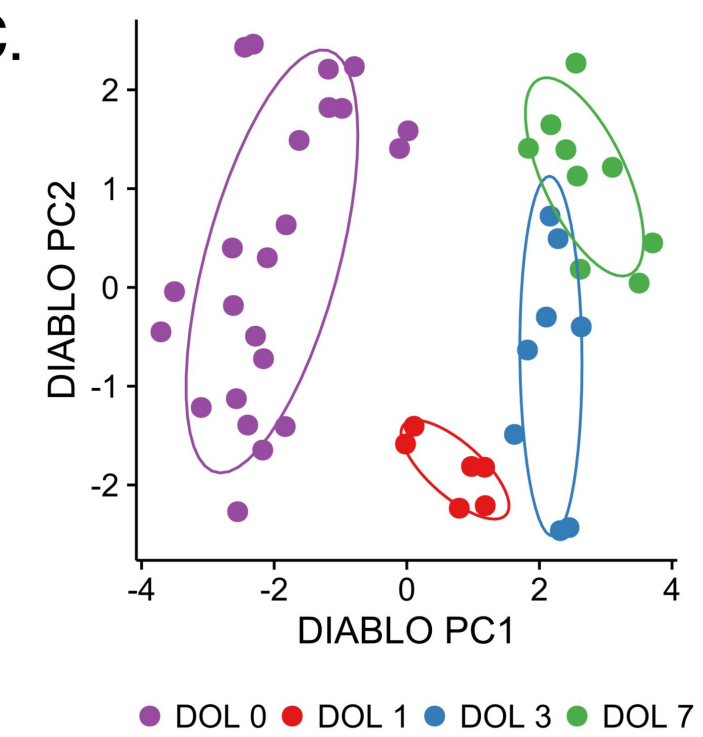
B.


cell
cytokine
metabolite
protein
transcript

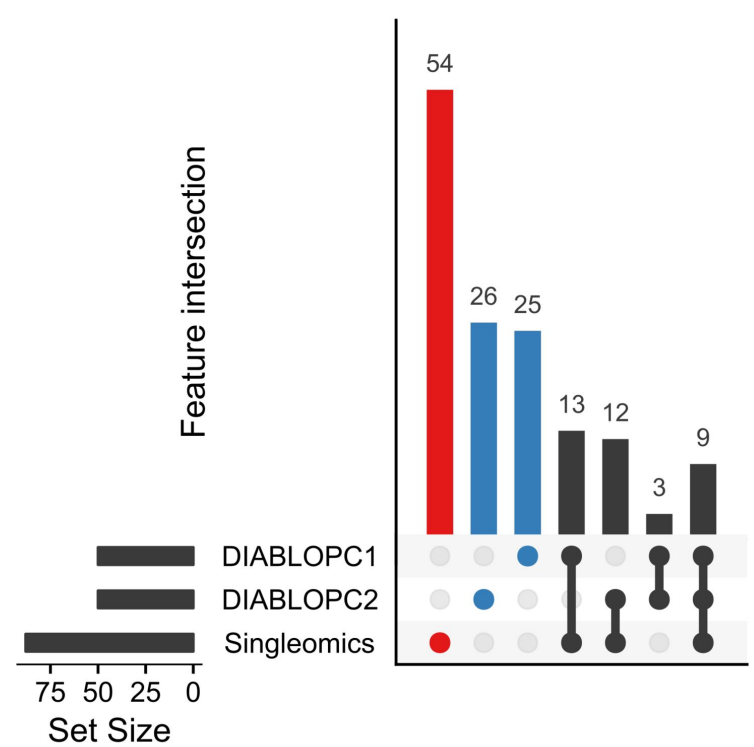

Pearson's r
1.0
0.5
0.0
-0.5
-1.0


cell
cytokine
metabolite
protein
transcript

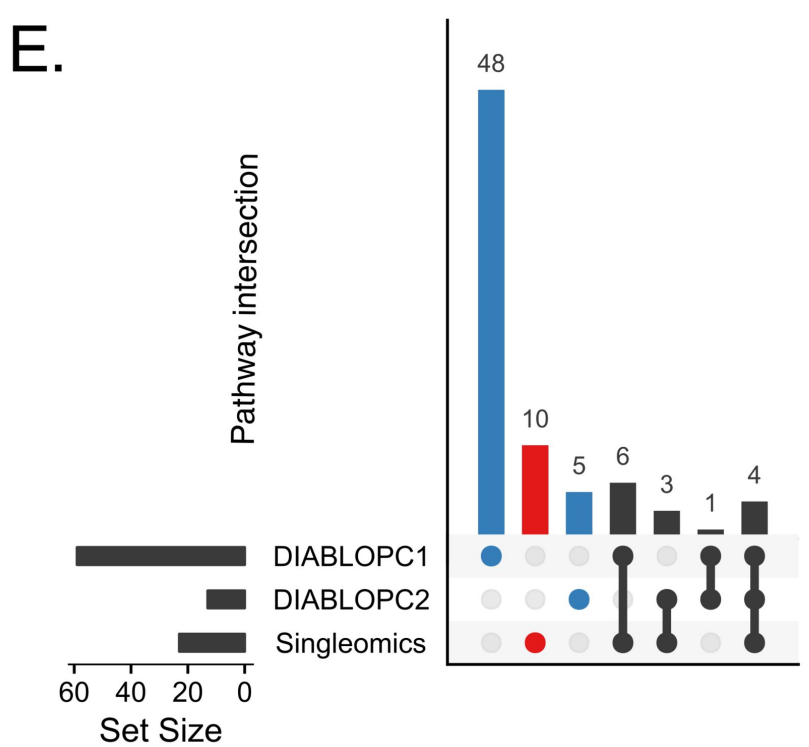
C.


DIABLO PC2
DIABLO PC1
DOL 0
DOL 1
DOL 3
DOL 7

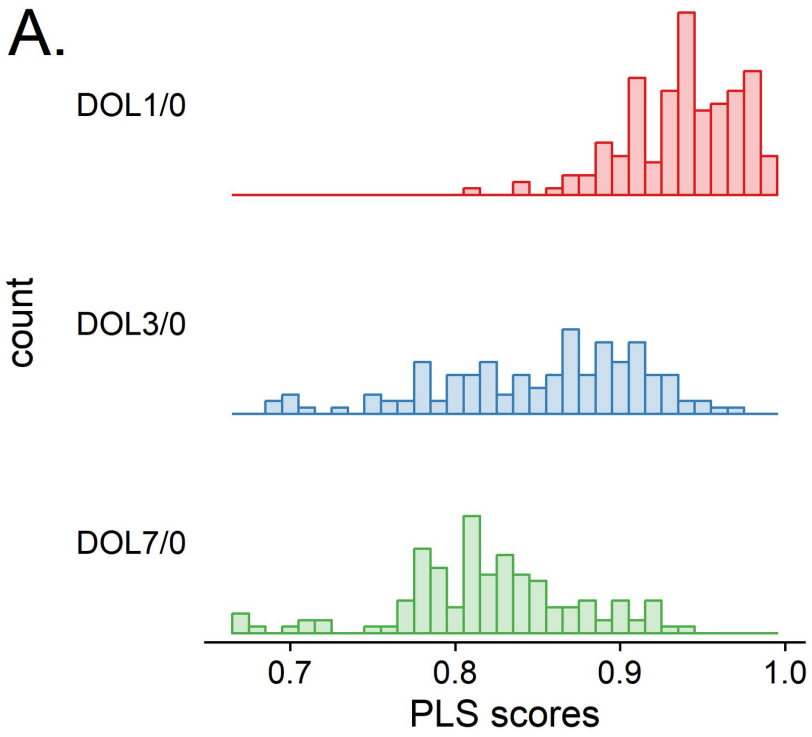
D.


Feature intersection
54
26
25
13
12
3
9
DIABLOPC1
DIABLOPC2
Singleomics
75 50 25 0
Set Size

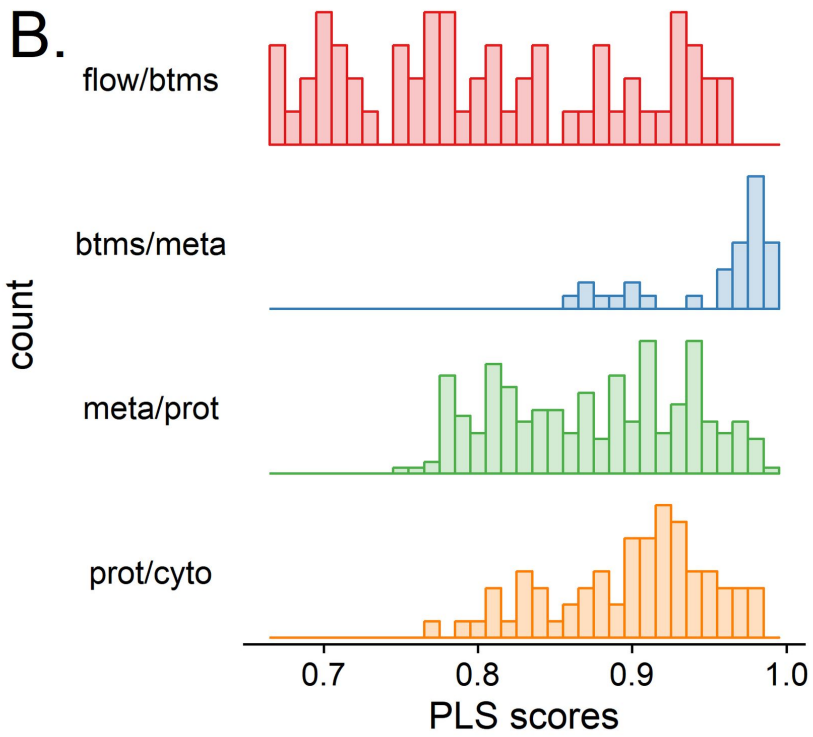
E.


Pathway intersection
48
10
5
6
3
1
4
DIABLOPC1
DIABLOPC2
Singleomics
60 40 20 0
Set Size

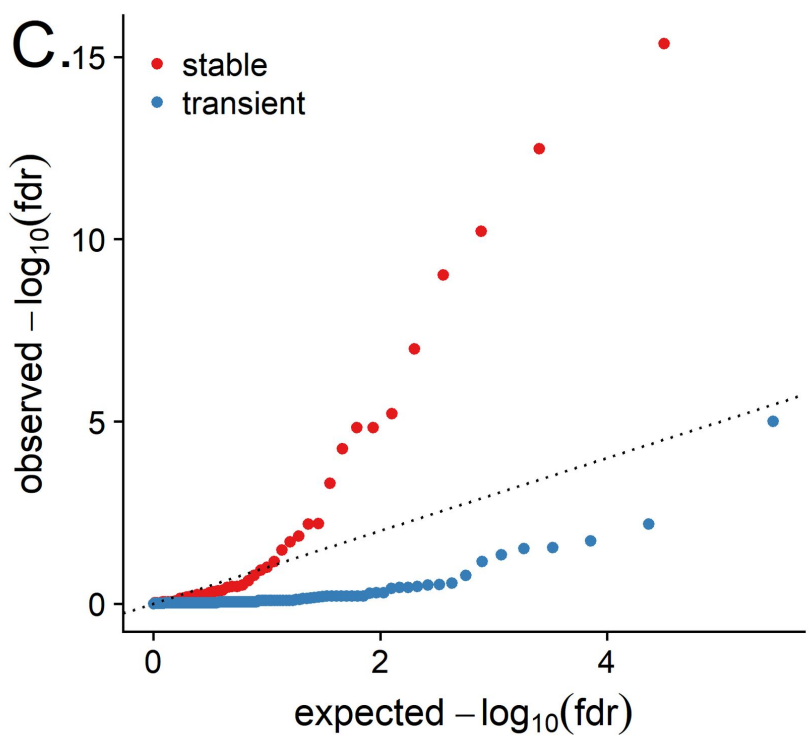
A.


DOL1/0
DOL3/0
DOL7/0
count
0.7
0.8
0.9
1.0
PLS scores

B.


flow/btms
btms/meta
meta/prot
prot/cyto
count
0.7
0.8
0.9
1.0
PLS scores

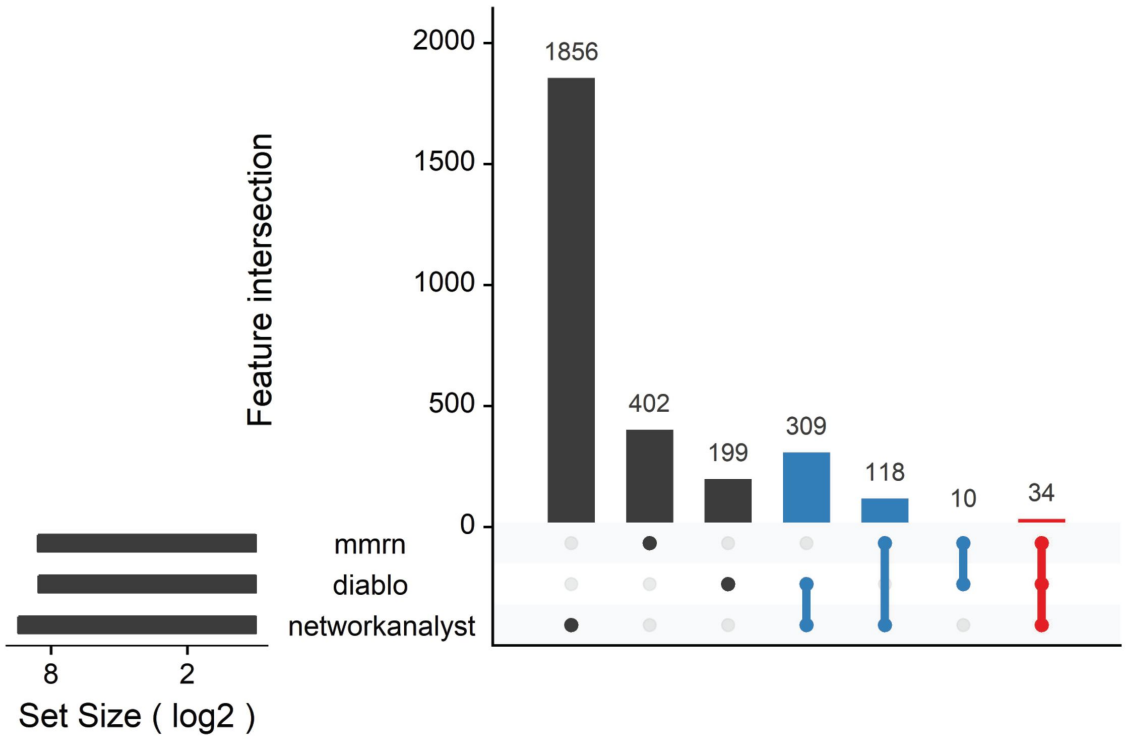
C.


stable
transient
observed −log10(fdr)
expected −log10(fdr)
0
5
10
15
0
2
4

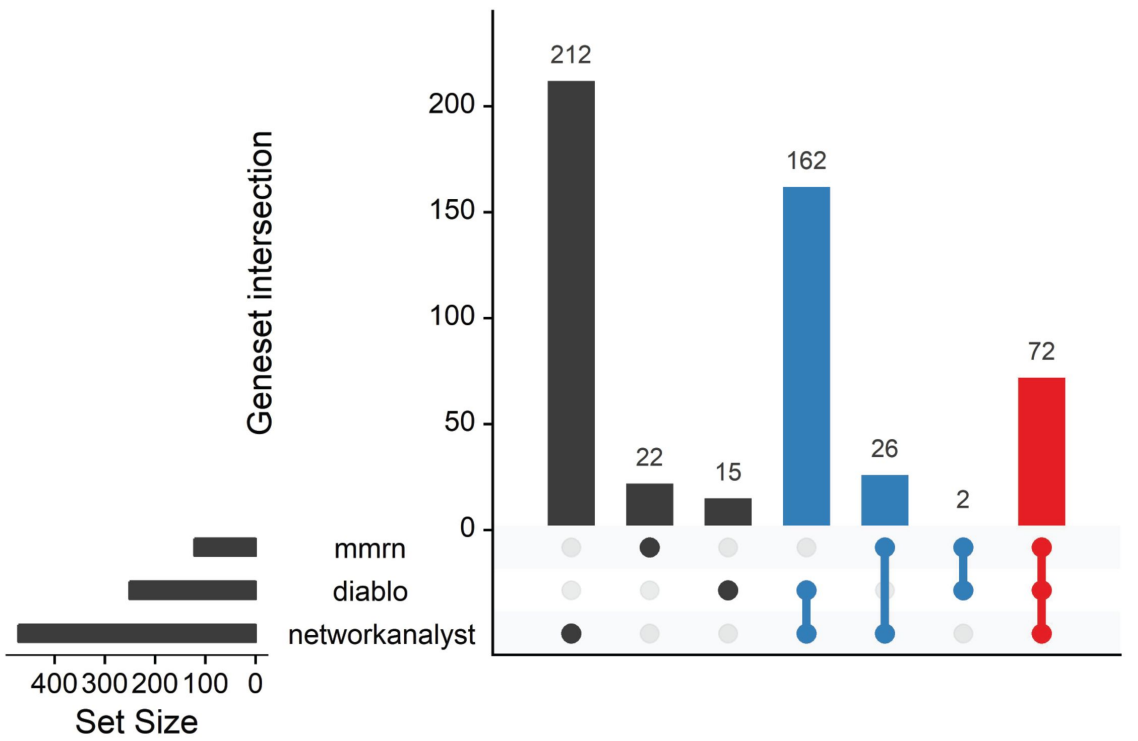
D.


DOL 1/0
DOL 3/0
DOL 7/0
switched.memory.cd38p.bcells
switched.memory.cd38n.bcells
switched.memory.bcells
plasmablasts
pdcs
naive.cd38p.bcells
naive.cd38n.b.cells
naive.bcells
mature.transition.bcells
immature.transition.cd38p.bcells
immature.transition.cd38n.bcells
immature.transition.bcells
cd19pcd20p.bcells
cd19p.igmp.igdp.bcells
cd19p.igmp.igdn.bcells
cd19p.igmn.igdp.bcells
cd19p.bcells
stable network flow cytometry
TLR and inflammatory signaling (M16)
signal transduction, plasma membrane (M82)
regulation of transcription, transcription factors (M213)
regulation of signal transduction (M3)
platelet activation (III) (M42)
olfactory receptors (M228)
myeloid cell cytokines, metallopeptidases and laminins (M78)
Monocyte surface signature (S4)
metabolism of steroids (M225)
lysosome (M209)
interferon alpha response (I) (M158.0)
inositol phosphate metabolism (M129)
innate antiviral response (M150)
heme biosynthesis (II) (M222)
heme biosynthesis (I) (M171)
golgi membrane (I) (M113)
glycerophospholipid metabolism (M114.1)
erythrocyte differentiation (M173)
enriched in neutrophils (I) (M37.1)
enriched in myeloid cells and monocytes (M81)
enriched in monocytes (surface) (M118.1)
enriched in monocytes (IV) (M118.0)
enriched in monocytes (III) (M73)
enriched in membrane proteins (M124)
enriched in dendritic cells (M168)
enriched in activated dendritic cells/monocytes (M64)
enriched in activated dendritic cells (I) (M119)
enriched for ubiquitination (M138)
endoplasmic reticulum (M37.2)
cytoskeletal remodeling (enriched for SRF targets) (M34)
chromosome Y linked (M240)
cell junction (GO) (M4.13)
cell cycle and growth arrest (M31)
cell adhesion, membrane (M133.0)
CCR1, 7 and cell signaling (M59)
stable network transcriptomics
TCA Cycle
Sphingolipid Metabolism
Pyrimidine Metabolism, Uracil containing
Purine Metabolism, Guanine containing
Purine Metabolism, Adenine containing
Polyamine Metabolism
Methionine, Cysteine, SAM and Taurine Metabolism
Lysophospholipid
Lysine Metabolism
Glycine, Serine and Threonine Metabolism
Gamma-glutamyl Amino Acid
Fatty Acid, Monohydroxy
Fatty Acid, Branched
Endocannabinoid
Aminosugar Metabolism
stable network metabolomics
z-scores
-2 -1 0 1 2

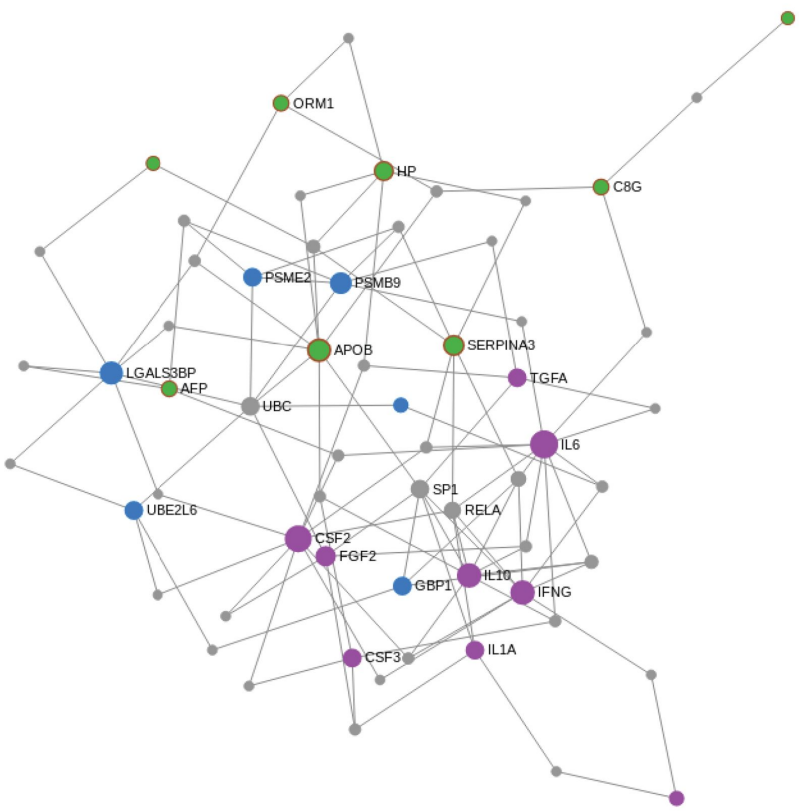
A.



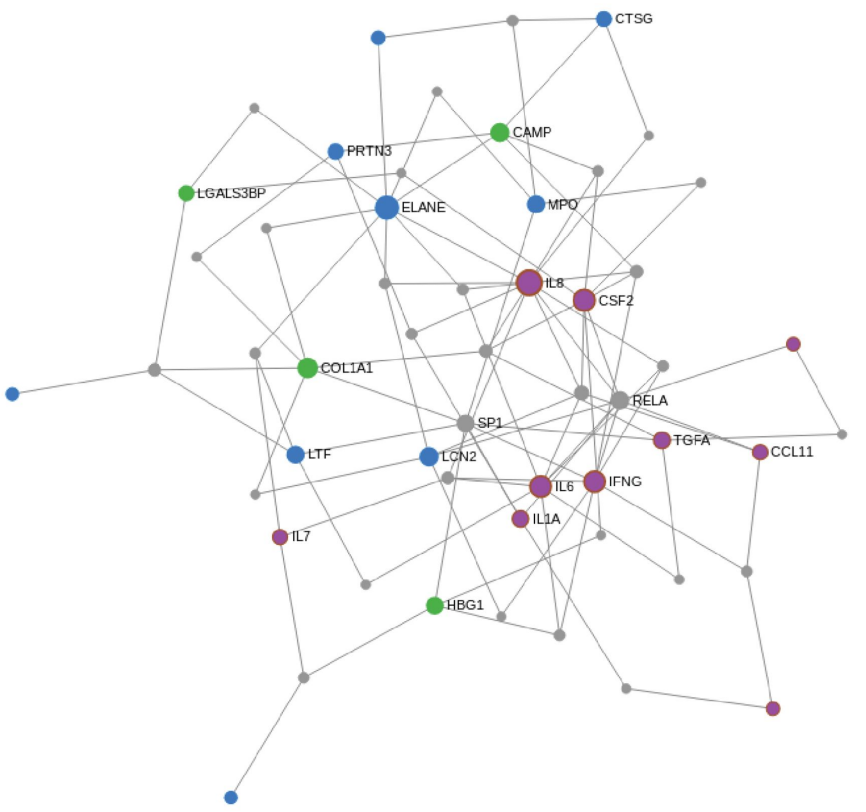
B.



C.



Component 1

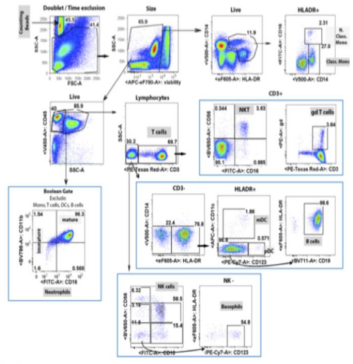


Component 2

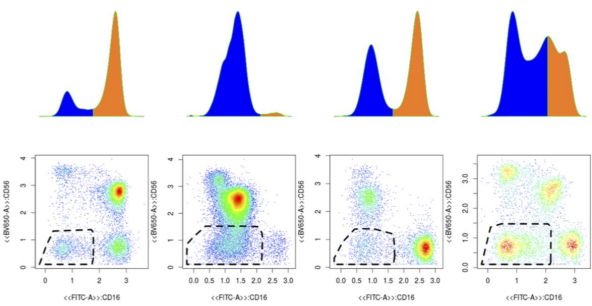
A

Cell population	Anchor Markers
T cells	CD45+/CD3+/CD56-/CD16-
γδ- T cells	CD45+/CD3+/γδ+
B cells	CD45+/CD3-/HLADR+/CD11c-/CD123-/CD19+
Classical Monocytes	CD45+/HLADR+/CD14+/CD16-
Non-classical Monocytes	CD45+/HLADR+/CD14+/CD16+
NKT cells	CD45+/CD3+/CD56+/CD16+
Myeloid Dendritic Cells (mDC)	CD45+/CD3-/HLADR+/CD11c+
Plasmacytoid Dendritic Cells (pDC)	CD45+/CD3-/HLADR+/CD123+/CD11c+
NK-CD56hi	CD45+/CD3-/HLADR-/CD16-/CD56HI
NK-CD56dim	CD45+/CD3-/HLADR-/CD16-/CD56Dim
Mature Neutrophils	Live/CD45+/Boolean gate excluding All other cell types / CD16+/CD11b+
Immature Neutrophils	Live/CD45+/Boolean gate excluding All other cell types / CD16-CD11b$^{+/-}$
Eosinophils	CD45+ hi/CD16-
Basophils	CD45+/CD3-/CD16-/CD56-/CD20-/CD123+

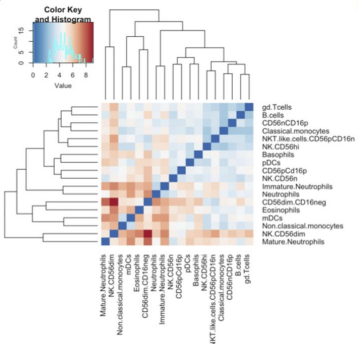
B



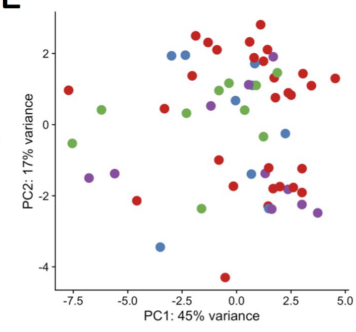
C



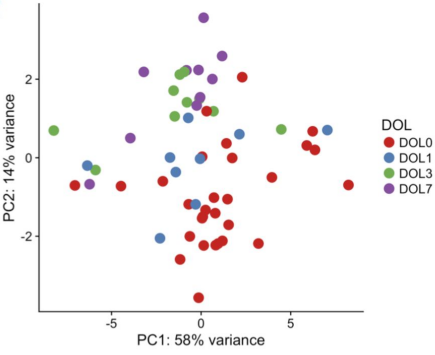
D



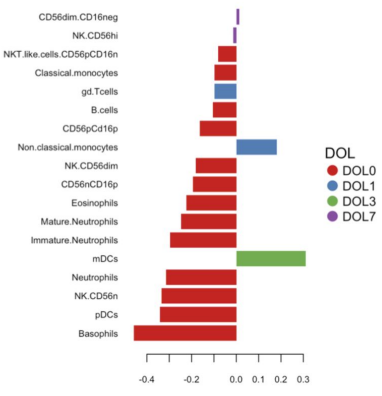
E



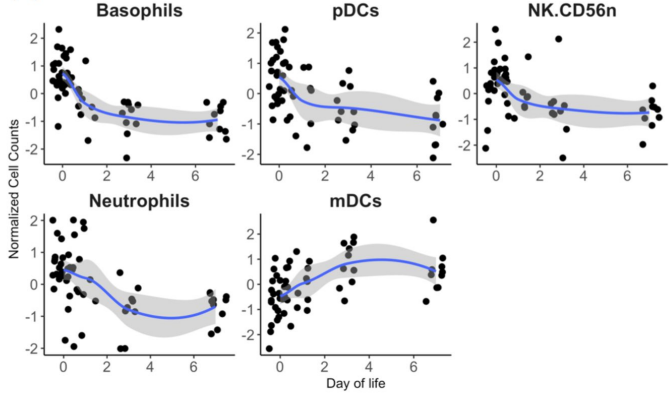
F

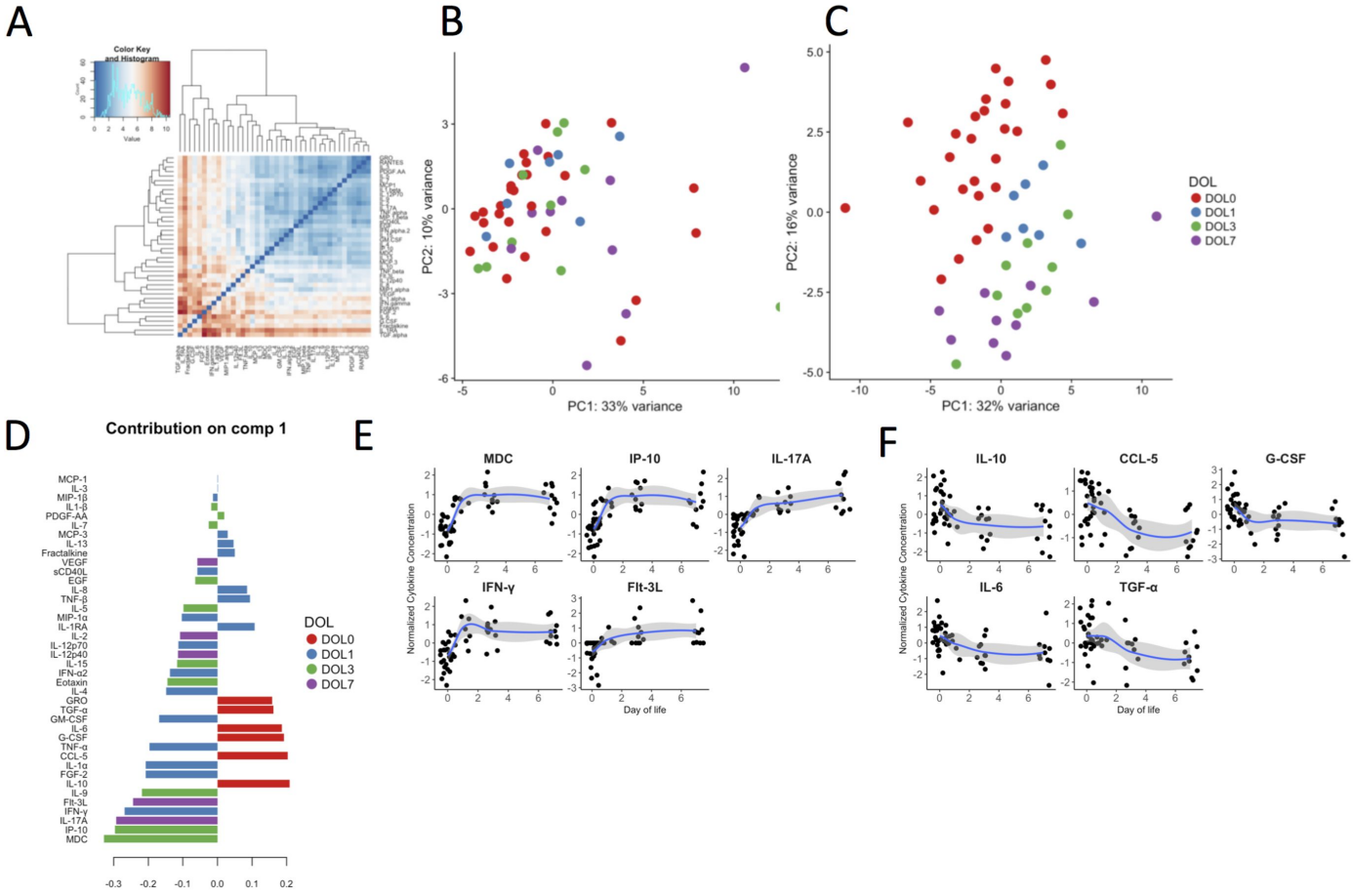


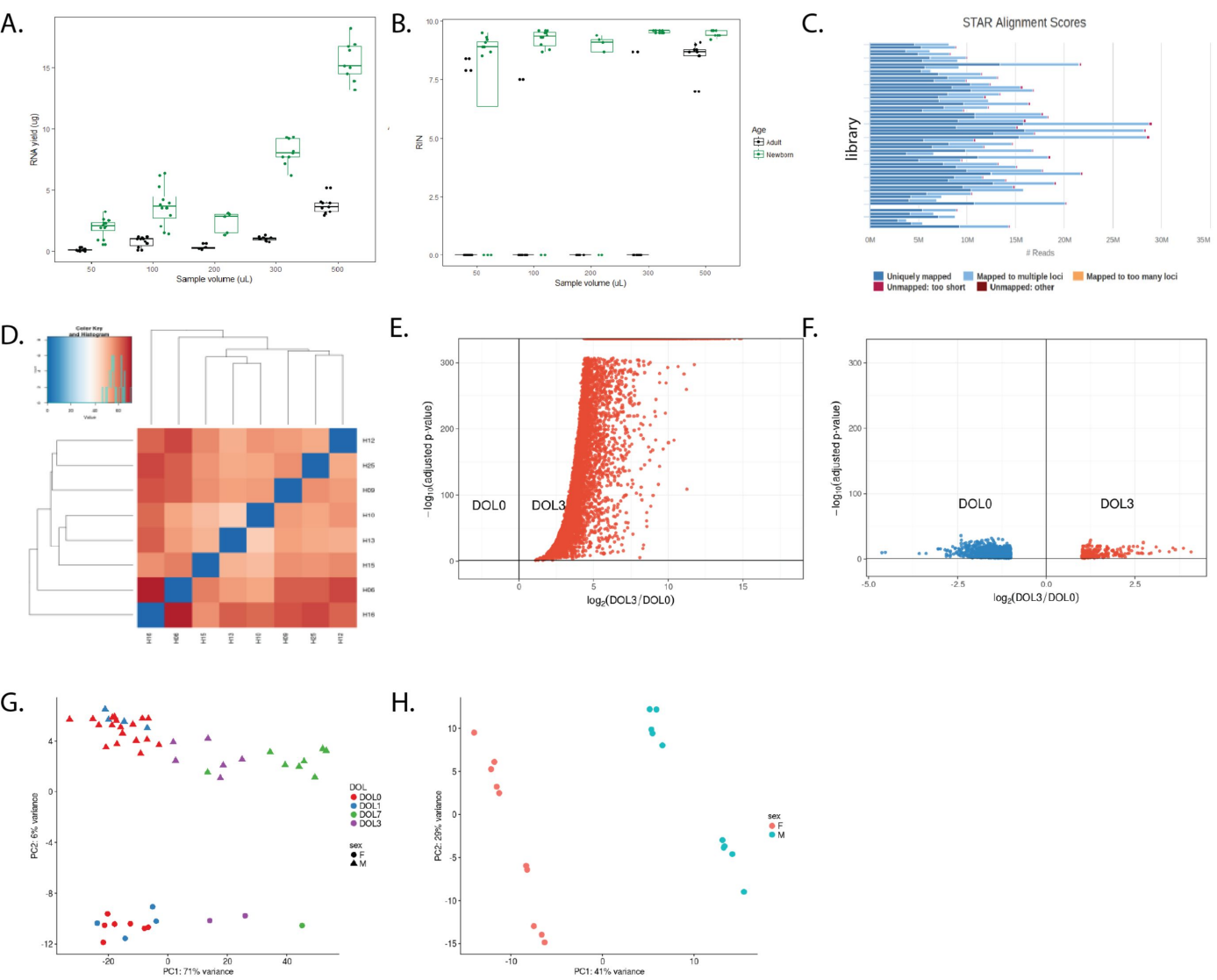
G



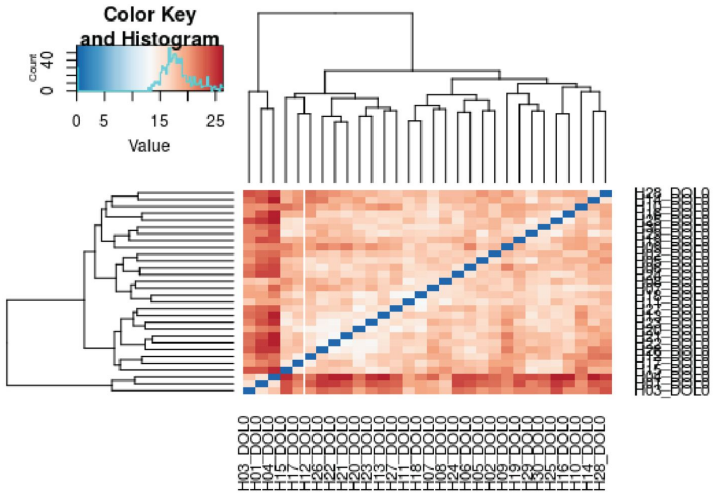
H



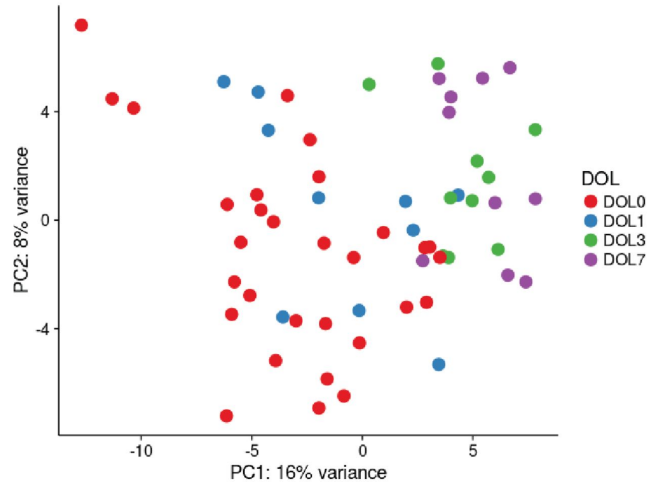

A
Color Key and Histogram
Value
B
PC2: 10% variance
PC1: 33% variance
C
PC2: 16% variance
PC1: 32% variance
DOL
DOL0
DOL1
DOL3
DOL7
D
Contribution on comp 1
MCP-1
IL-3
MIP-1β
IL-1β
PDGF-AA
IL-7
MCP-3
IL-13
Fractalkine
VEGF
sCD40L
EGF
IL-8
TNF-β
IL-5
MIP-1α
IL-1RA
IL-2
IL-12p70
IL-12p40
IL-15
IFN-α2
Eotaxin
IL-4
GRO
TGF-α
GM-CSF
IL-6
G-CSF
TNF-α
CCL-5
IL-1α
FGF-2
IL-10
IL-9
Flt-3L
IFN-γ
IL-17A
IP-10
MDC
-0.3
-0.2
-0.1
0.0
0.1
0.2
DOL
DOL0
DOL1
DOL3
DOL7
E
MDC
IP-10
IL-17A
IFN-γ
Flt-3L
Normalized Cytokine Concentration
Day of life
F
IL-10
CCL-5
G-CSF
IL-6
TGF-α
Normalized Cytokine Concentration
Day of life


A.
RNA yield (ug)
Sample volume (uL)
B.
RIN
Sample volume (uL)
Age
Adult
Newborn
C.
STAR Alignment Scores
library
Reads
Uniquely mapped
Mapped to multiple loci
Mapped to too many loci
Unmapped: too short
Unmapped: other
D.
E.
DOL0
DOL3
log2(DOL3/DOL0)
F.
DOL0
DOL3
log2(DOL3/DOL0)
G.
PC2: 6% variance
PC1: 71% variance
DOL
DOL0
DOL1
DOL7
DOL3
sex
F
M
H.
PC2: 29% variance
PC1: 41% variance
sex
F
M

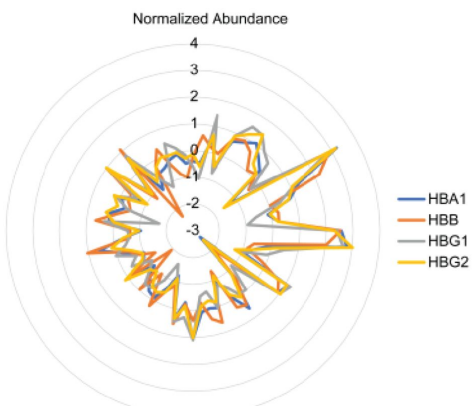
A.



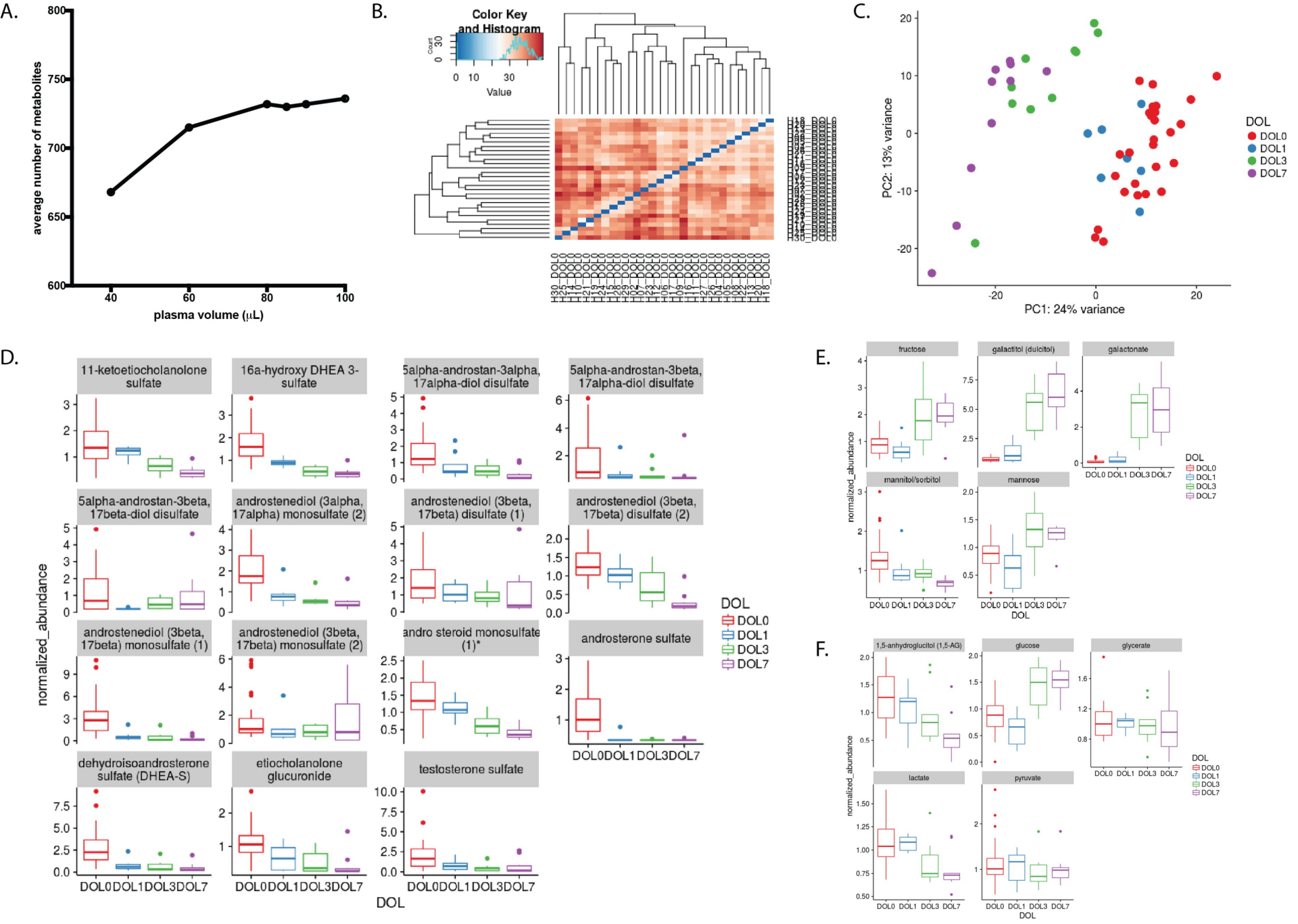
B.

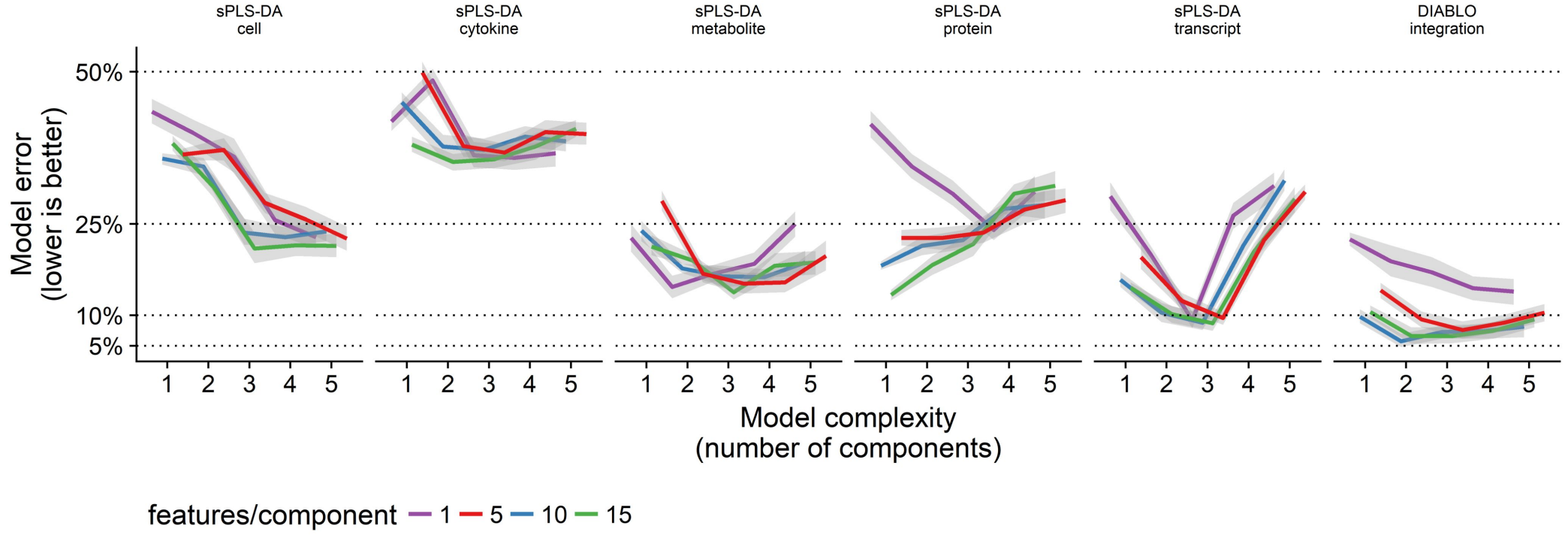


C.



p-value / R	HBA1	HBB	*HBG1*	HBG2
HBA1		5.0E-20	1.9E-15	2.2E-39
HBB	0.8773		1.4E-08	1.6E-15
HBG1	0.8169	0.6577		1.5E-15
HBG2	0.9748	0.8205	0.8178	


A.
average number of metabolites
plasma volume (μL)
B.
Color Key and Histogram
Count
Value
C.
PC2: 13% variance
PC1: 24% variance
DOL
DOL0
DOL1
DOL3
DOL7
D.
11-ketoetiocholanolone sulfate
16a-hydroxy DHEA 3-sulfate
5alpha-androstan-3alpha, 17alpha-diol disulfate
5alpha-androstan-3beta, 17alpha-diol disulfate
5alpha-androstan-3beta, 17beta-diol disulfate
androstenediol (3alpha, 17alpha) monosulfate (2)
androstenediol (3beta, 17beta) disulfate (1)
androstenediol (3beta, 17beta) disulfate (2)
androstenediol (3beta, 17beta) monosulfate (1)
androstenediol (3beta, 17beta) monosulfate (2)
andro steroid monosulfate (1)*
androsterone sulfate
dehydroisoandrosterone sulfate (DHEA-S)
etiocholanolone glucuronide
testosterone sulfate
normalized_abundance
DOL
E.
fructose
galactitol (dulcitol)
galactonate
mannitol/sorbitol
mannose
normalized_abundance
DOL
F.
1,5-anhydroglucitol (1,5-AG)
glucose
glycerate
lactate
pyruvate
normalized_abundance
DOL


sPLS-DA
cell
sPLS-DA
cytokine
sPLS-DA
metabolite
sPLS-DA
protein
sPLS-DA
transcript
DIABLO
integration
Model error
(lower is better)
50%
25%
10%
5%
1
2
3
4
5
Model complexity
(number of components)
features/component
1
5
10
15