Dynamic molecular changes during the first week of human life follow a robust developmental trajectory

Amy H. Lee1, Casey P. Shannon2, Nelly Amenyogbe3,25, Tue B. Bennike4,5,6, Joann Diray-Arce5,6, Olubukola T. Idoko7,8, Erin E. Gill1, Rym Ben-Othman9, William S. Pomat10, Simon D. van Haren5,6, Kim-Anh Lê Cao11, Momoudou Cox7, Alansana Darboe7, Reza Falsafi1, Davide Ferrari11, Daniel J. Harbeson3, Daniel He2, Cai Bing9, Samuel J. Hinshaw1,12, Jorjoh Ndure7, Jainaba Njie-Jobe7, Matthew A. Pettengill5, Peter C. Richmond13,14, Rebecca Ford10, Gerard Saleu10, Geraldine Masiria10, John Paul Matlam10, Wendy Kirarock10, Elishia Roberts7, Mehnoush Malek15, Guzmán Sanchez-Schmitz5,6, Amrit Singh2,16, Asimenia Angelidou5,6,17, Kinga K. Smolen5,6, The EPIC Consortium#, Ryan R. Brinkman15,18, Al Ozonoff5,6,19, Robert E.W. Hancock1, Anita H.J. van den Biggelaar14, Hanno Steen5,6,14, Scott J. Tebbutt2,20,21, Beate Kampmann7,22, Ofer Levy5,6,23 & Tobias R. Kollmann3,9,25

Systems biology can unravel complex biology but has not been extensively applied to human newborns, a group highly vulnerable to a wide range of diseases. We optimized methods to extract transcriptomic, proteomic, metabolomic, cytokine/chemokine, and single cell immune phenotyping data from <1 ml of blood, a volume readily obtained from newborns. Indexing to baseline and applying innovative integrative computational methods reveals dramatic changes along a remarkably stable developmental trajectory over the first week of life. This is most evident in changes of interferon and complement pathways, as well as neutrophil-associated signaling. Validated across two independent cohorts of newborns from West Africa and Australasia, a robust and common trajectory emerges, suggesting a purposeful rather than random developmental path. Systems biology and innovative data integration can provide fresh insights into the molecular ontogeny of the first week of life, a dynamic developmental phase that is key for health and disease.

Correspondence and requests for materials should be addressed to O.L. (email: offer.levy@childrens.harvard.edu) or to T.R.K. (email: tkollm@mac.com).

#A full list of affiliations appears at the end of the paper. These authors contributed equally: Amy H. Lee, Casey P. Shannon, Nelly Amenyogbe, Tue B. Bennike, Joann Diray-Arce, Olubukola Idoko. These authors jointly supervised this work: Anita H.J. van den Biggelaar, Hanno Steen, Scott J. Tebbutt, Beate Kampmann, Ofer Levy, Tobias R. Kollmann.
The first week of life is characterized by heightened susceptibility to infections and is increasingly recognized as a major determinant of overall health for the entire human lifespan. Knowledge of the molecular drivers involved in these processes in newborns (defined as those <28 days of life) is fragmentary. Systems biology approaches, employing high-dimensional molecular and cellular measurements (henceforth referred to as OMICs), along with unbiased analytic approaches, have increased our understanding of basal and altered molecular states in adults and recently in newborns and infants after the first week of life, but such approaches have not been applied systematically to characterize molecular ontogeny over the most critical period, i.e. the first week of life. This is likely due to the analytical challenges posed by the limited amount of biosample that can be obtained and the many rapid physiological changes around birth. The resulting variance in biological measurements has been thought to necessitate a large participant sample size, which would increase complexity and cost.

To overcome these hurdles, we developed a robust experimental and analytical approach feasible with <1 ml of newborn blood. Our data represent the most comprehensive systems biology study yet performed during the first week of human life. Despite substantial between-subject variation, normalizing (indexing) all samples from a given newborn enabled identification of consistent and robust changes over the first week of life across the entire cohort. Furthermore, data integration using independent strategies not only validated signatures across methodologically- and biologically-distinct datasets, but also provided novel findings. The major observations derived from a cohort from West Africa (The Gambia) were validated for an Australasian (Papua New Guinea) cohort. Our results highlight that, contrary to the relatively steady-state biology observed in healthy adults, the first week of human life is highly dynamic. Nevertheless, despite the substantial variability between participants and these dramatic changes, ontogeny followed a robust trajectory common to newborns from very different areas of the world.

Results

Blood processing. One of the objectives of this project was to develop a robust standard operating protocol (SOP) to enable extraction and analysis of data using systems biology (big data) approaches from small blood sample volumes that can readily be obtained for research purposes from newborns within the first week of life (Fig. 1, see also Protocol). Our experimental SOP utilized important sample-sparing modifications whereby we obtained samples for immune phenotyping, transcriptomic, proteomic, and metabonomic analysis from <1 ml of blood (see also Supplementary Methods). We profiled the peripheral blood of each participant twice over their first week of life, i.e. at Day of Life (DOL) 0 (baseline) and additionally at either DOL 1, 3, or 7, and sought to identify variables that differed between the baseline and later time points across all participants. This required indexing either by employing paired statistical tests for univariate analyses or calculating fold changes relative to the DOL0 sample for multivariate analyses, as described in online Methods. The number of samples used in each OMIC platform as well as the analysis stage are shown in Supplementary Table 1 and Supplementary Figure 9.

Immune phenotyping across first week of life. Determining the cellular composition of biological samples is important in systems biology, as relative and absolute cell numbers predict endpoints of interest with high accuracy, e.g. vaccine responses, and enable deconvolution of OMICs data. Analysis of our predefined targeted cell populations revealed substantial between-subject variability in peripheral blood samples obtained over the first week of life (Fig. 2, Supplementary Figure 2, Supplementary Note 1). However, consistent within-subject changes over the first week of life amongst the entire cohort of 30 Gambian newborns emerged when samples were indexed as displayed by principal component analysis (PCA). Univariate analysis identified the following discriminating cellular features over the first week of life: basophils, plasmacytoid dendritic cells (DC), natural killer cells, and neutrophils decreased; in contrast, myeloid DCs increased after DOL0, while many other cell types remained stable. We also detected dramatic but consistent changes in soluble immune markers including plasma cytokines and chemokines over the first week of life (Fig. 1). Based on the relevant univariate analysis, we found that plasma concentrations of C–X–C motif chemokine 10 (CXCL10), interleukin (IL)-17A, macrophage-derived chemokine (MDC), and interferon (IFN)γ increased, while IL-10, Chemokine C–C motif ligand (CCL) 5, granulocyte colony stimulating factor 2 (GCSF), and IL-6 decreased with age over the first week of life; many other soluble immune markers remained unchanged.

Transcriptomic analysis across first week of life. Modern systems biology studies employ gene expression analysis by RNA-Seq. We found that ≥500 μl of adult blood was required to consistently obtain sufficient high-quality RNA, but <100 μl of newborn blood sufficed (Supplementary Figure 3). This likely reflected the relatively high content of white blood cells (WBC) and nucleated red blood cells in newborn blood, which contain abundant globin mRNA. 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) with the former chosen for subsequent studies (Supplementary Methods).

As with the immune phenotyping data, there was substantial between-subject variability in our RNA-Seq data from peripheral blood samples obtained over the first week of life. This was resolved by indexing each participant to their own DOL0 sample, revealing dramatic yet consistent developmental signals that related to age (i.e. ontogeny) to emerge across the entire cohort. We were concerned that the above-mentioned changes in cell composition could provide a basis for the observed differentially expressed (DE) genes, but the inclusion of cell composition data in our model did not affect our results (see Methods (Online) and Supplementary Figure 3 for details) demonstrating that the whole blood transcriptomic signals driving the developmental trajectory observed were not merely a consequence of changes in the underlying cellular composition across DOL.

In comparing DOL1 vs. DOL0, there were few (12) identified DE genes; however, dramatic developmental changes emerged when comparing later days of life to DOL0 (Fig. 3a). 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 Data 5 and Supplementary Methods. In particular, genes with decreased expression across the first week of life are involved in cellular responses to stress, detoxification of reactive oxygen, as well as heme biosynthesis and iron uptake. Conversely genes involved in interferon signaling, Toll-like-receptor (TLR), negative regulation of Retinoic Acid Inducible Gene 1 (RIG-I) and complement activation were upregulated over the first week of life.

Proteomic analysis across first week of life. A total of 684 different proteins were identified across peripheral blood plasma samples (false discovery rate (FDR) < 1%) (Supplementary
Figure 4). Of these, 199 proteins met our criteria, including detection of at least two unique peptides per protein, for further detailed quantification. Substantial between-subject variability was also noted in the plasma proteomic analysis, but again indexing each participant to their baseline enhanced signature detection. This approach indicated a common developmental trajectory over the first week of life, with differences in plasma protein composition compared to DOL0 increasing with increasing age (Fig. 3b). Differentially abundant plasma proteins and their respective pathways are listed in Supplementary Data 6 and Supplementary Methods. At DOL3 vs. DOL0, three pathways were upregulated that center around the complement cascade; at DOL7 vs. DOL0, five additional pathways were upregulated including scavenging heme from plasma and signaling to RAS.

Metabolomic analysis across first week of life. While initial analysis revealed substantial between-subject variation, indexing metabolomic data to DOL0 also revealed a steady but dramatic developmental trajectory in the plasma metabolome (Supplementary Figure 5). Few differences in plasma metabolites were identified comparing DOL1 vs. DOL0, but increasing differences were noted when contrasting DOL3 or DOL7 vs. DOL0 (Fig. 3c). Interestingly, metabolomic differences detected across age involved pathways related to plasma steroids and carbohydrate metabolites, possibly reflecting neurodevelopment, rapid cell proliferation and increased uptake of nutrients in newborns (Supplementary Data 7 and Supplementary Methods)16.

Data integration. Each methodologically- and biologically distinct data type that we examined revealed substantial changes over the first week of life. We thus sought to determine if the observed changes were related to one another across data types, validating consistent age-dependent changes in functional pathways. To minimize the limitations inherent in any single analytical approach, and to detect the most robust signatures, data integration was addressed using three independent strategies. To this end, we employed a novel function-based strategy based on biologically known Molecular Interaction Networks using NetworkAnalyst17; an unbiased data-driven multivariate matrix factorization approach using DIABLO (Data Integration Analysis for Biomarker discovery using Latent cOmponents)18–20; and the multiscale, multifactorial response network (MMRN) approach that estimates correlations across data types21.

NetworkAnalyst enables the creation of networks based on a framework of known protein–protein interactions (PPI) captured in publicly 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 dataset),
as well as first-order interactors that served to connect the seed nodes with each other. To incorporate metabolomics data, synthetic and degradative metabolic enzymes that would determine the levels of differentially detected metabolites were used as seed nodes in network construction. Overall metabolomics, proteomics and transcriptomics data fit well into a single functional network (Fig. 4), 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 across different OMICS datasets (Supplementary
Fig. 2 Indexing cellular and soluble immune markers revealed developmental progression over the first week of life. a, b Principal component analysis was used to plot cellular composition (a) and plasma cytokines/chemokine concentration (b) for each sample; this highlighted the substantial variability between participants and lack of defined clustering by DOL due to higher influence of individual variance over ontogeny. c, d Accounting for repeat measures from the same individual across different sampling days compared to DOL0 (indexing to DOL0) revealed sample clustering by DOL between samples. e, f Normalized cell counts showing developmental trajectories for cell populations that significantly changed (e) or did not change (f) over the first week of life. g, h Normalized plasma cytokine/chemokine concentrations showing developmental trajectories for cytokines/chemokines that significantly changed (g) or did not change (h) over the first week of life. Boxplots display medians with lower and upper hinges representing first and third quartiles. Whiskers reach the highest and lowest values, no more than 1.5× interquartile range from the hinge. ***p ≤ 0.0001, **p ≤ 0.001, *p ≤ 0.01, ’p ≤ 0.05, ns p > 0.05, Kruskal–Wallis test, Benjamini–Hochberg adjusted p values. DOL: day of life

Fig. 3 Transcriptomic, proteomic, and metabolomic analyses identified a robust trajectory of differentially expressed genes, proteins, and metabolites over the first week of life. a Up- and downregulated differentially expressed genes were plotted by DOL (vs. DOL0) and numbers of genes are listed above each point except for downregulated genes at DOL1 vs. DOL0, where the number was zero. b, c Up- and downregulated differentially expressed proteins and metabolites, respectively, plotted by DOL compared to DOL0, with numbers of differentially expressed proteins or metabolites listed above each point. d Zero-order interaction networks for genes differentially expressed at DOL3 vs. DOL0 and DOL7 vs. DOL0. Within networks, upregulated nodes are displayed in red and downregulated nodes in blue. DOL: day of life

Data 8). For example, integrating transcriptomic with proteomic data affirmed the increase in type 1 IFN-related functions and the regulation of complement cascades over the first week of life. This integration also revealed new biological insights not found in any single-data domain analysis, such as changes in cellular replication machinery, creatine metabolism (DOL3), fibrin clotting cascade and signals of increasing adaptive immune and phagosome activity (DOL7).

DIABLO is a multivariate approach to address two of the major concerns faced when integrating multicomponent datasets, namely the complexity of the data, particularly with few samples, each with many observations, and the heterogeneous nature of data measured on different scales and technological platforms18–20. DIABLO constructs components (linear combinations of the original data—cells, cytokines, transcripts, proteins, metabolites) that are maximally correlated across any number of input data types with a specified outcome variable (in this case, DOL), while simultaneously performing marker selection23 to identify a minimal subset of markers associated with this outcome (Supplementary Methods). We created matrices from our five
Using MMRN, we found that associations between data types as an input to DIABLO to identify major ontogeny-related markers (Fig. 5, Supplementary Figure 6 and Supplementary Data 9). The resulting model discriminated well between DOLs, outperforming nonintegrative approaches (Supplementary Figure 6), 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. We next investigated the relationship between features selected by DIABLO across data types and visualized the selected features in an integrative network (Supplementary Methods). We compared this integrative network (Fig. 5a, b) to one derived from features identified using an alternative, nonintegrative sparse discriminant analysis approach (Supplementary Methods). The integrative network was more densely connected (global clustering coefficient = 0.91 vs. 0.68) and composed of fewer, 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 nonintegrative approach favored markers that were not well correlated across data types. The two components of the DIABLO model were composed of distinct sets of features (Fig. 5d, blue bars), representing distinct biology (Fig. 5e, blue bars). The first component reflecting DOL0 was composed of markers consistent with interferon and cytokine signaling, among other immune biology (Fig. 5a, Supplementary Methods). The second component, reflecting progression across DOL1–7 had a distinct granulocyte-flavor, as well as a focus on cytokine signaling in the immune system and cellular response to latent infections (Fig. 5a, Supplementary Methods).

MMRNs are a recently published framework for data integration7. Using MMRN, we found that associations between data types were strongest at DOL1 and decreased across the first week of life (Supplementary Figure 7), confirming the already noted robust trajectory of development. Most of the significantly correlated clusters were transcriptomic (16/21), but we also identified metabolomic (1/21) and flow cytometry-derived (4/21) clusters associated with DOL (Supplementary Data 10). The stable clusters most significantly associated with DOL were composed of blood transcriptomic modules (BTMs) related to DCs and monocytes including cytokine receptors CCR1 and CCR7, TLR and inflammatory signaling, heme biosynthesis, various B-cell subpopulations, as well as metabolic pathways such as purine metabolism7.

**Meta-integration.** To assess the similarities across methodologically distinct integration methods, we analyzed the list of nodes associated with DOLs for each integration approach. We then carried out pathway enrichment to determine which biological processes were identified by each method (Supplementary Data 11 and 12). The selected nodes for Molecular Interaction Networks (NetworkAnalyst) were those nodes of minimum-connected networks when differentially abundant features for DOL3 vs. DOL0 and DOL7 vs. DOL0 were used as seed nodes (3195 features). For DIABLO, a first-order molecular interaction network was constructed using as seed nodes the markers derived from model components 1 and 2 (428 features). For MMRN, selected nodes were those that comprised the stable network shown in Supplementary Figure 7 (675 features).

We observed limited overlap at the individual marker level. However, assessing the above networks using the Reactome pathway annotation system (2208 pathways) revealed that 635, 308 and 84 pathways were significantly over-represented in the NetworkAnalyst, DIABLO and MMRN feature lists, respectively (paired t test DOL0 vs. DOL1, 3, or 7, Benjamini–Hochberg corrected FDR ≤ 0.05; Supplementary Data 11). Importantly, 249 and 25 pathways were identified by NetworkAnalyst and either DIABLO or MMRN, respectively, and 34 pathways were identified by all three approaches, demonstrating a strong congruence between the outputs and conclusions derived from different analytical/OMICS platforms (Fig. 6a). This degree of overlap was unlikely to occur by chance alone as determined by bootstrapping (p value < 0.001; Supplementary Figure 8 and Supplementary Methods). The specific pathways identified by this meta-integration as driving molecular ontology over the first week of life were related to interferon signaling, complement cascade and granulocyte function (Supplementary Data 12). Given the striking convergence observed, we next assessed relevant functional interactions between the pathways identified in silico via our meta-integration. To this end, we used NetworkAnalyst and found that selected common features fitted into minimum-connected networks of experimentally validated interactions.
Interactions (Supplementary Figure 8), implying functional interconnections. This validation confirmed that our meta-integration approach identified functionally important biological interactions central to neonatal ontogeny.

Cross-cohort validation. To validate the generalizability of these integrative models based on data from Gambian (West Africa) newborns, we recruited and characterized an independent newborn cohort (30 participants) from a different region of the world (Papua New Guinea (PNG), Australasia). This validation cohort was processed according to the same methods as described above (Fig. 1 and Supplementary Figure 9). The outputs for the PNG data demonstrated the same basic trajectory as a function of DOL and showed considerable overlap (e.g. in transcriptomics $p < 10^{-138}$ for the Jaccard index/similarity for DOL3 vs. DOL0 and DOL7 vs. DOL0 in the two cohorts). Overall, integrative multivariate modeling using DIABLO predicted the correct DOL well for these independent cohorts (Fig. 6, Supplementary Figure 10).

Fig. 5 DIABLO uncovered biologically relevant features by integrating information across data types. Schematic representation of two contrasting integration approaches using multivariate 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 multivariate 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 and corresponding networks, with dense subgraphs, or network modules, encircled. In particular, the network modules identified in (a) include a number of features selected from all data types. This is not the case in b. 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 DOL0 from all other DOLs (DIABLO component 1) and DOL1, 3, and 7 from each other (DIABLO component 2). Features identified by DIABLO (blue bars) were largely distinct from those identified by more traditional single-OMICS multivariate approaches (red bars; overlaps in gray); 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.

DIABLO: Data Integration Analysis for Biomarker discovery using Latent cOmponents, DOL: day of life.
data, with samples from DOL0 and DOL7 separating from all other DOL and near-perfect AUC values (0.98 and 0.95, respectively) while samples from DOL3 or DOL1 were more challenging to classify (0.84 and 0.60, respectively). A functional zero-order network that included key DIABLO-selected features that predict the DOL highlighted in the PNG cohort pathways related to interferon signaling, complement and neutrophil activity (Fig. 6d), as was observed for the Gambian cohort. Thus, we were able to validate the complex Gambian dataset in an independent and distinct population of newborns.

**Discussion**

Present a holistic suite of complementary methods that addressed key hurdles of applying systems biology to the first week of human life by: (i) overcoming limitations in sample volume via an efficient, field-compatible, sample-sparing SOP to
process peripheral whole blood; (ii) determining the developmental trajectories from each participant’s own baseline (indexing), and (iii) reducing the complexity of the dataset by integrating biologically distinct data types generated via a variety of methods (multi-OMICS integration), which enabled validation of signatures of important pathways. The major conclusions were validated in two independent and distinct cohorts of newborns. This approach revealed a dramatic molecular ontogeny evolving over the first week of human life along a common and robust developmental trajectory. Our observation indicates that development over the first week of human life is not immature and random but follows a precise and age-specific path.

This approach was affirmed by its ability to identify known changes in early life, including those related to the composition of hemoglobin mRNA, increases in complement protein C9 and reductions in plasma steroid metabolites such as pregnanolone/allopregnanolone sulfate and pregnanediol-3-glucuronide (Supplementary Figure 5) that are important for neonatal neurodevelopment, supporting the validity of our findings. Moreover, our approach also identified pathways never previously identified as relevant to ontogeny. For example, many analytical pathways upregulated across the first week of life consistently identified by all of our analytical strategies are relevant for host defense but have not before been identified as central to early human development. These pathways included interleukin 1 signaling, Toll-like receptor signaling, NOTCH signaling, the DHX RNA helicase pathway leading to interferon signaling, as well as regulators of the complement cascade. Upregulation of these pathways across the first week of life may represent a heightened defense function for vulnerable newborns who are especially dependent on innate immunity. Data integration also identified novel and surprising but biologically plausible findings regarding ontogeny. For example, prostaglandin-endoperoxide synthase 2 (PTGS2 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 hematopoietic progenitors and is clinically relevant during premature labor and in necrotizing enterocolitis. Taken together, our findings outlined a developmental process that can serve as a reference, akin to the stable steady state in adults. Deviations from this developmental trajectory could potentially identify at-risk individuals prior to disease onset, indicate potential benefit for interventions, or guide the re-establishment of homeostasis once derailed.

Multivariable analysis of the ROSH data, which included transcriptomics, proteomics, metabolomics, and phosphoproteomics, was performed using NetworkAnalyst to identify pathways that were significantly altered during early life. The analysis identified several key pathways that are central to ontogeny. These pathways centered around and interconnected interferon signaling, the complement cascade, and neutrophil activity. Each of these have previously been recognized as important for host defense but have not before been identified as central to early human development. These pathways included interleukin 1 signaling, Toll-like receptor signaling, NOTCH signaling, the DHX RNA helicase pathway leading to interferon signaling, as well as regulators of the complement cascade. Upregulation of these pathways across the first week of life may represent a heightened defense function for vulnerable newborns who are especially dependent on innate immunity. Data integration also identified novel and surprising but biologically plausible findings regarding ontogeny. For example, prostaglandin-endoperoxide synthase 2 (PTGS2 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 hematopoietic progenitors and is clinically relevant during premature labor and in necrotizing enterocolitis. Taken together, our findings outlined a developmental process that can serve as a reference, akin to the stable steady state in adults. Deviations from this developmental trajectory could potentially identify at-risk individuals prior to disease onset, indicate potential benefit for interventions, or guide the re-establishment of homeostasis once derailed.

Multi-OMICS integration poses an exceptional problem due to the increase in complexity (p data points) relative to the typically achievable sample size (n samples). Here we addressed this problem by reducing data complexity via selection of a subset of variables (MMRN, DIABLO) or, for the NetworkAnalyst approach, imposing functionally relevant molecular interaction information. Furthermore, we applied integrative methods focused on extracting aggregated information from each dataset independently of each other, providing high confidence in results that converged on the same set of key molecular features. Overall, our approach confirmed that integrating across multiple different biological data types dramatically increased robust biological insights. Technical and analytical improvements, such as implementation of our enhanced plasma proteomics workflow, will further increase our understanding of the dynamic molecular changes.

Our integrative approach identified several key pathways as central to ontogeny. These pathways centered around and interconnected interferon signaling, the complement cascade, and neutrophil activity. Each of these have previously been recognized individually as possible contributors to the increase in newborn susceptibility to infection, but their age-dependent change has not previously been known to be centrally important to ontogeny over the first week of human life. Importantly, the three key pathways constituting the core of an early life developmental trajectory were readily validated across two independent cohorts of newborns from very different populations, further confirming that the observed early life trajectory is common and predictable and thus could serve as a baseline reference. The finding of a stable trajectory in the first week of life has profound implications: the newborn immune system is still often viewed as immature, which implies a stochastic, unregulated state. However, the existence of the developmental trajectory we discovered as shared by newborns across very different populations in the world strongly argues that early life immune ontogeny is not random but follows a precise and thus possibly purposeful path. Alternatively, the developmental trajectory we observed over the first week of life could be the result of a limited range of response possibilities in early life, as recently espoused in the hypothesis of stereotypic immune system development. Differentiating between these possibilities (a proactive, purposeful path vs. a reactive, restricted response) will be critical for understanding normal development as well as diagnosing, preventing and treating disease in early life.

In summary, our study presents an integrative analysis across the broadest range of multivariate datasets published to date covering the first week of human life. Our integration over time (indexing data longitudinally for each newborn) and biological space (multi-OMICS integration) enabled exploration of the dynamic molecular and cellular developmental characteristics of early life. In parallel,
we identified signals of potential physiological importance using a practical sample size. Coupled with our field-tested SOP for sample-sparring preanalytical processing, this strategy overcomes key challenges in applying systems biology to neonates. Our data revealed a compelling biological narrative, where out of the apparently noisy age-dependent changes across the first week of life, a consistent developmental trajectory emerged. The interactive contributors to this trajectory relate to cell autonomous as well as innate immune functions, such as interferon-driven signaling, complement cas-

Methods
Overview. A workflow diagram is provided in Supplementary Figure 1 that outlines the precise sample processing steps as well as the statistical and bioinformatic analyses steps applied to the data.

Peripheral blood processing. Thirty healthy, term newborns were enrolled at each of the Medical Research Council (MRC) Unit The Gambia and at the Institute for Medical Research (IMR) in Goroka, Papua New Guinea in accordance with a local Ethics Committee-approved protocol (IMRC SCC 14/16 and IMR B08/2015 and MRAC #16.14). Informed consent was obtained from all human participants. The initial numbers of participants recruited for the Gambian cohort were based on data outputs rather than formal power calculations. This effect size from the Gambian dataset, however, allowed us to estimate the sample size needed for our subsequently employed PNG validation cohort. Specifically, using the approach described by Liu et al., the mean fold-change and standard deviation were estimated in the Gambian transcriptomic data (using the data in Supplementary Table – 12), and, controlling the FDR at 5%, under different assumptions for the proportion of non-DE features (0.8, 0.9), based on what was observed in the Gambian dataset, the number of transcripts DE at DOL7, we estimated that with n = 10, we would have >80% power, even when assuming a very low proportion of DE features (10%).

Following informed consent, mothers were screened for HIV-1 and -2 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-min Apgar score > 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. Venous blood was drawn from infants via venipuncture directly into heparinized collection tubes (Becton Dickinson (BD) Bioscienes; 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 24 h.

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 (catalog 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.: 07272733001, Roche, Basel, Switzerland). All cDNA libraries were pooled and amplified at the sequencing center 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. drought-resistant FASTQ sequences were then used as input for the high-throughput Genome Ensembl (EnsemblGRCh38.86) using STAR v2.5.2 and mapped to Ensembl GRCh38 transcripts. Read-counts were generated using htseq-count (HTSeq 0.6.1p1). All RNA processing and subsequent differential gene expression analyses were performed using R version 3.3.0 and DESeq2 version 1.14.1. Genes with very low counts (with less than ten counts in eight or more samples, or the smallest number of biological replicates within each treatment group) and genes transcripts were prefiltered and removed in silico. Differentially expressed genes were identified using paired analysis with the Wald statistics test and filtering for any genes that showed twofold change and adjusted p value < 0.05 (cut-off at 5% FDR) as the threshold. Functional discovery of pathway enrichment and network enrichment was performed using SIGORA v2.0.1 and NetworkAnalyst, respectively. To test whether changes in cell composition could account for the observed changes in gene expression, we used DESeq2 (with default parameters) to fit two models, one including subject and DOL and the other model including the additional covariate of birth weight. To address the collinearity of these factors, we used PCA, summarizing the cell proportions (flow cytometry) to five principle components (PCs, accounting for 95% of the variance observed). We compared the estimated effect sizes for the DOL term between these two models (Pearson correlation) across all genes and found them to be highly correlated (p < 0.50), indicating that the observed changes in gene expression could not be (fully) explained by changes in the underlying cell composition.

Plastic proteomics. Plasma samples were prepared for proteome analysis using an in-house (Boston Children’s Hospital; Steen laboratory)-developed mStern blotting sample processing and trypsinization protocol, which was adapted for
plasma samples. 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 using 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 (Promega, Madison, WI, USA) at a nominal protease to protein ratio of 1:25 w/w. After incubation for 2 h 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 (Biogнос; Schlieren, Switzerland) were spiked into all samples. Samples were analyzed using a nanoLC system (Eksigent; Dublin, CA) equipped with a LCchip system (chPLC nanoFLX, 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 positive-ion mode with a mass window of 2.0 mDa and a dynamic exclusion time of 15 s. The instrument was operated in data-dependent mode, selecting up to 12 of the most intense precursors for fragmentation from each precursor scan. Label-free protein quantification analysis was performed using MaxQuant 1.5.3.30. Raw-data were downloaded and used to build a matching library and searched against the UniProt Human Reference Proteome database using the settings with the following modifications: Max missed cleavage 3; variable modifications Deamidation (NQ) and Oxidation (M). A de novo search strategy was employed to filter all proteins and peptides to <1% FDR. The list of proteins was further processed in Perseus 1.5.3.5, log2-transformed, and proteins with less than two peptides ( razor) were discarded. The same samples were also included to DOL, and proteins that 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. The R package RmixOmics was used to correct for batch effects for samples sampled with different LC-MS columns. Proteins with a statistically significant change of abundance between different DOL were identified by paired two-sample t test. To correct for multiple hypothesis testing, permutation-based false positive control was applied using standard parameters in Perseus (FDR = 0.05, a=0.1). Significant proteins were further analyzed using SIGORA with the Reactome gene annotation system and DAVID. 

Plasma metabolomics. Plasma samples were run using the nontargeted metabolomics platform of Metabolon Inc. (Durham, NC, USA). Samples were extracted and prepared using Metabolon’s solvent extraction method and run on four independent platforms: reverse-phase/UPLC-MS/MS with positive ion mode electron spray ionization (ESI) retention; positive-phased UPLC-MS/MS with negative mode ESI, HILIC/UPLC-MS/MS with negative ion mode ESI, and a 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 (30,000 @ 1.000). Ionization of metabolites included the following 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. The MS/MS scores were based on a comparison of the ions present in the experimental spectrum to the ions present in the library spectrum. Similarities between these metabolites were based on one of these factors and the use of all three data points can be utilized to distinguish and differentiate biochemistry. The metabolites were confirmed by comparing their mass spectra and chromatographic retention times with more than 3300 commercially available reference standards. All identified metabolites were categorized to level 1 metabolites according to reporting standards set by the Chemical Analysis Working Group of the Metabolomics Standards Initiative and have appropriate orthogonal analytical techniques applied to the metabolite of interest and to a chemical reference standard. These identified metabolites included corresponding accurate mass via MS retention index, chemical and composition ID and accurate mass matched. About 75% of the metabolites identified also had Pubchem ID associated with it. MS/MS fragment ion analysis process was performed using peak-matching algorithms (Waters MassLynx v4.1, Waters corp; Milford, USA) and quantified using an area-under-the-curve in-house algorithm. Metabolite features where all values are missing/undefined or features with a variance of zero and with just a single value in the whole dataset were excluded from the samples. All remaining missing values were imputed with half the minimum value of peak intensity for that feature. Features with interquartile range of zero were excluded to ensure reproducibility. All metabolite features were then log-transformed for normalization, pvalue-scaled to reduce variance in features and for robust size effect comparability across samples. Significant metabolite between DOL were identified using t test between paired samples corrected for multiple hypothesis testing (threshold of adjusted p value <0.05) using MetaboAnalyst 4.0. Hypergeometric test enrichment of metabolite sets were matched against KEGG pathway. The Human Metabolome Database Version 3.6 (HMDB) and Metabolynx Pathway Analysis using the Cytoscape plugin.

Data integration. Data obtained via the immune phenotyping (cellular composition and plasma cytokines), transcriptomics, proteomics and metabolomic 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. In addition, we assessed whether we could derive any novel biological insights from the integration of multiple OMICS data layers that were not revealed in a single dataset alone. To cross-validate results we employed three 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 the following three criteria: retention index (RI) within a narrow RI window of the metabolite interaction networks (Molecular Interaction Network) using Network Analyst that provided context based on annotated molecular interactions; (ii) a new approach to identify the underlying key drivers of ontology 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 Latent cOmponent (DIA-BLO) framework, which is part of the mixOMics R package and the recently published MMRN, querying the informatically derived correlations within a network for statistically significant features.

Molecular Interaction Network. Network Analyst integrates data using protein interactions as a biological framework. Metabolomic data required preprocessing in order to integrate these data 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: metabolites were mapped to their directly interacting enzymes based on their corresponding HMDB via the MetaCyc module in MetaBridge. To construct networks for each OMICS type, the list of proteins derived from the metabolite data processing (above), and lists of differentially expressed genes or proteins identified by transcriptomic and proteomic data were submitted to Network Analyst to produce a zero-order or a minimum-connected network, depending on the size of the dataset. Networks consist of seed nodes (proteins genes) that are used as input to build the network) and edges between the nodes together and are indicative of a molecular interaction between nodes. Minimum-connected integrated networks were constructed using pairwise or all three OMICS lists of differentially present genes/proteins on each later DOL vs. DOLO for transcriptomic, proteomic and metabolomic data. Node lists in each network that were identified as nodes that were not in either set of differentially expressed genes/proteins nor found in minimum-connected networks constructed from each single data type. 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. Transcripts were integrated via their respective encoded gene products. Pathway enrichment analysis was performed using SIGORA R package 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 layer (transcriptomics and metabolomics) a 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 unique data type pathways were subtracted.
gene set libraries, using the hypergeometric test, either separately per component, or after enrichment to include their first-degree neighbors subgraph in the protein—protein interaction data. Obtained p values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure.

Multiscale, multifactorial response network. Examples of systematic integration across multiple large OMICS datasets remain rare. Recently, Li et al. used MMRNs to study the immune response to varicella zoster vaccine across four OMICS datasets. We aimed to compare insights generated by our chosen approaches to theirs and, to that end, created a multiscale integration network as they described, with the following important modifications. First, instead of summarizing the higher dimensional data blocks (transcriptomics, metabolomics) to either gene modules or metabolic pathways (BBM63; Metabolon Inc. annotation, respectively) using a simple average, we used Eigen-gene summarization48, a weighted average based on the first principal component of the data, in order to maximize the variance explained. Second, clusters were identified at each DOL using Euclidean distance and Ward’s method as described44. The optimal number of clusters was determined using the elbow criterion45. Stable clusters were identified by comparing cluster membership using the Symkiewicz–Simpson coefficient46. Finally, cluster association with DOL was assessed using the Correlation Adjusted MÉan RAnk (Camera) gene set test47.

Meta-integration to identify convergence across data integration strategies. We performed meta-integration to address whether: (1) DIABLO-selected features formed a functional network as defined by known protein–protein interactions, and (2) each of the integration methodologies, namely the Molecular Interaction Networks integration, DIABLO and MMRN, would identify similar biological signals across the first week of life. To assess whether DIABLO-selected features formed a functional protein–protein interaction network, 20 features from each transcriptome, proteome, and metabolome data were used to create a minimum-convex network. Metabolites were mapped to the directly interacting enzymes based on their corresponding HMDB IDs via the MetaCyc module in MetaBridge53,54 as above, and used as seed nodes for NetworkAnalyzer. The various data integration approaches resulted in outputs that were not directly comparable. To enable direct comparisons, we simplified their outputs to lists of features of interest. Molecular Interaction Networks integration: list of features that made up a minimum-connected network using nodes from differentially abundant features (transcripts, proteins and metabolites) for DOL5 vs. DOL0 and DOL7 vs. DOL0 (3195 features). DIABLO: a first-order connected network including seed nodes of component 1 and 2 features of the DIABLO model using Network Analyzer (428 features). MMRN: features were selected from the small stable interconnected node-network of transcriptomic and metabolomics features shown in Supplementary Figure 7 (675 features). Pathway enrichment of the small stable interconnected node-network of transcriptomic and metabolomics data by automated density-based cell population identification in the context of acute kidney allograft rejection. PLoS ONE 9, e95224 (2014).



Data availability

All data are publicly available. The transcriptomics data presented in this publication were submitted to the NCBI Gene Expression Omnibus under accession numbers GSE111404 and GSE123070. All other datasets including immune phenotyping, Luminex, metabolomics and proteomics data were archived on ImmPort (https://immport.niaid.nih.gov/home) under accession numbers SDY1256 and SDY1412.

Received: 18 February 2018 Accepted: 24 January 2019

Published online: 12 March 2019

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. Balbus, J. M. et al. Early-life prevention of non-communicable diseases. Lancet 381, 3–4 (2013).
3. Chausabel, D. & Pueltrandur, B. A vision and a prescription for big data-enabled medicine. Nat. Immunol. 16, 435–439 (2015).
4. Olin, A. et al. Stereotypic immune system development in newborn children. Cell 174, 1227–1242 (2018).
5. Jennenwein, M. F., Butler, A. L. & Alter, G. Neonate-omics: charting the unknown immune response in early life. Cell 174, 1051–1053 (2018).
6. Amenyo, L., Levy, O. & Kollmann, T. R. Systems vaccinology: a promise for the young and the poor. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 2040349 (2015).
7. Li, S. et al. Metabolic phenotypes of response to vaccination in humans. Cell 169, 862–877 (2017).
8. Howie, S. R. Blood sample volumes in child health research: review of safe limits. Bull. World Health Organ. 89, 46–53 (2011).
9. Tsang, J. S. Utilizing population variation, vaccination, and systems biology to study human immunology. Trends Immunol. 36, 479–493 (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. 193, 3003–3012 (2014).
12. Tsang, J. S. et al. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. Cell 157, 499–513 (2014).
13. 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).
14. Smith, C. L. et al. Identification of a human neonatal immune-metabolic network associated with bacterial infection. Nat. Commun. 5, 4649 (2014).
15. Henry, E. & Christensen, R. D. Reference intervals in neonatal hematology. Clin. Perinatol. 42, 483–497 (2015).
16. McCallie, R. K. et al. Skin-to-skin contact after birth and the natural course of neurodevelopmental outcomes in healthy term newborns. J. Perinatol. 37, 591–595 (2017).
17. Xia, J., Gill, E. E. & Hancock, R. E. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat. Protoc. 10, 823–844 (2015).
18. Cao, K. A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinform. 12, 253 (2011).
19. Singh, A. et al. DIABLO—an integrative, multi-omics, multivariate method for multi-group classification. Preprint at https://www.biorxiv.org/content/ 10.1101/067611v2. (2018).
20. Rohart, F., Gautier, B., Singh, A. & Le Cao, K. A. mixOmics: an R package for omics feature selection and multiple data integration. PLoS Comput. Biol. 13, e1005752 (2017).
21. Chassabel, D. & Baldwin, N. Democratizing systems immunology with modular transcriptional repertoire analyses. Nat. Rev. Immunol. 14, 271–280 (2014).
22. Breuer, K. et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res. 41, D1228–D1233 (2013).
23. Tibshirani, R. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B (Methodol.) 58, 267–288 (1996).
24. Ismail, A. A., Walker, P. L., Macfuss, R. & Gindal, B. Diagnostic value of serum testosterone measurement in infancy: two case reports. Ann. Clin. Biochem. 26(3 Pt 1), 259–261 (1989).
25. 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).
26. Lugo, R., Ford, H. R. & Grishin, A. Molecular signaling in necrotizing enterocolitis: regulation of intestinal COX-2 expression. J. Pediatr. Surg. 42, 1165–1171 (2007).
27. Reinreit, H. E. et al. Cyclo-oxygenase (COX) inhibitors for treating preterm labour. Cochrane Database Syst. Rev. 2017. https://doi.org/10.1002/14651858. CD001992.pub3 (2015).
28. Fan, J. & Li, J. A selective overview of variable selection in high dimensional feature space. Stat. Sin. 20, 101–148 (2010).
29. Bennike, T. B. et al. A cost-effective high-throughput plasma and serum proteomics workflow enables mapping of the molecular impact of total pancreatostomy with ileal autotransplantation. J. Proteome Res. 17, 1983–1992 (2018).
30. Liu, P. & Hwang, J. T. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. Bioinformatics 23, 739–746 (2007).
31. Westerhuis, J. A., van Velzen, E. J., Hoefloot, H. C. & Smilde, A. K. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metab.- Off. J. Metab. Soc. 6, 119–128 (2017).
32. Lee, J. et al. miFlowCyt: the minimum information about a Flow Cytometry Experiment. Cytometry 73, 926–930 (2008).
33. Hahne, F. et al. bowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinform. 10, 106 (2009).
34. Malek, M. et al. BowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics 31, 606–607 (2015).
35. O'Neill, K., Jalali, A., Aghaeepour, N., Hoos, H. & Brinkman, R. R. Enhanced flowType/RchyOptimy: a BioConductor pipeline for discovery in high-dimensional cytometry data. Bioinformatics 30, 1329–1330 (2014).
36. Ewens, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048 (2016).
Acknowledgements

We would like to thank all the participants and their parents for their time and willingness to support this study. Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health as part of the Hu-man Immunology Project consortium under 5U19AI118608-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. T.R.K.’s laboratory is supported by a Michael Smith Foundation for Health Research Career Investigator Award. O.L.’s laboratory is supported by the following U.S. National Institutes of Health (NIH)/National Institute of Allergy and Infectious Disease Research, the Canadian Institutes of Health Research (CIHR), the Canadian Foundation for AIDS Research, the Canadian Institutes of Health Research, the Canada Research Chair in Systems Biomedicine, and the Canadian Institutes of Health Research (CIHR). The Lundbeck Foundation (R181-2014-3372), the Carlsberg Foundation (CF14-0561), and A.P. Møller Foundation are acknowledged for grants enabling T.B.B.’s work. K.-A.L.C. is supported in part by the National Health and Medical Research Council (NHMRC) Career Development Fellowship (GNT1080345). We gratefully acknowledge the support from Drs. Gary Fleisher, Michael Wessels and Ken Kraft as well as Maria Crenshaw, Mark Liu, Kenny McNaney and Diana Vo (all BCH); Susan Farmer, Manish Sadarangani, Aaron Liu, Gordan Bjornson (all BCH). The Expanded Program on Immunization Consortium (EPIC) contributed collectively to this study. EPIC is an association of academic centers partnering to conduct systems biology studies in newborns and infants, comprised of the investigators listed above at Boston Children’s Hospital (BCH), University of British Columbia (UBC), Medical Research Council Unit The Gambia (MRCG), Univer-site libre de Bruxelles, Telethon Kids Institute and University of Western Australia, and the Papua New Guinea Institute for Medical Research (PNG-IMR).

Author contributions

Project core and lead works were: O.L., T.R.K., H.S., R.E.W.H., S.J.T., A.O., A.H.J.v.d.B., and B. K. Project activities were coordinated by the Administrative Core (O.L., J.D.-A., D.V., K.K.S.; BCH; T.R.K., R.B.-O.; UBC). All infants were enrolled at the MRCG (O.I., B.K.) and PNG-IMR (W.S.P., A.H.J.v.d.B.), where biopsies were collected, processed and shipped (M.C.D., A.D.-A., J.N., J.N., E.R., R.F., G.S.M., J.P.M., W.K.), to the Clinical Core (K.K.S., O.L., S.V.; BCH) for further processing. Data generation and analyses were conducted at centers specializing in RNAseq (A.H.L., E.E.G., R.F., R.E.W.H., UBC), proteomics (T.B.B., H.S.; BCH), cytokine/chemokine analysis (S.X.H., O.L.; BCH; N.A.; D. H.; UBC), flow cytometry (T.R.K., N.A., R.B.-O., C.B., D.I.H.; M.M.; R.R.B.; UBC), with metabolomics data provided to O.L. and J.D.-A. (BCH) on a fee-for-service basis. Investigators listed above at Boston Children’s Hospital (BCH), University of British Columbia (UBC), Medical Research Council Unit The Gambia (MRCG), Université libre de Bruxelles, Telethon Kids Institute and University of Western Australia, and the Papua New Guinea Institute for Medical Research (PNG-IMR).

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-08794-4.
Competing interests: O.L. is a named inventor on patents regarding bactericidal/permeability increasing protein (BPI), including “Therapeutic uses of BPI protein products in BPI-deficient humans” (WO20005931A3) and “BPI and its congeners as radiation mitigators and radiation protectors” (WO201218839A1). R.R.B. has ownership interest in Cytapex Bioinformatics Inc. The remaining authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The EPIC Consortium
Diana Vo5, Ken Kraft5,6, Kerry McEnaney5,18, Sofia Vignolo5 & Arnaud Marchant24