CausER - a framework for inferring causal latent factors using multi-omic human datasets

Xin Bing[1*], Tyler Lovelace[2.3*], Florentina Bunea[1], Marten Wegkamp[1,4], Harinder Singh[5†], Panayiotis V Benos[2†], Jishnu Das[5†]

[1] Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA

[2] Department of Computational & Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA

[3] Joint CMU-Pitt PhD Program in Computational Biology, Pittsburgh, PA, USA

[4] Department of Mathematics, Cornell University, Ithaca, NY, USA

[5] Center for Systems Immunology, Departments of Immunology and Computational & Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA

* Equal contribution/co-first author

[†] Corresponding authors – Jishnu Das (jishnu@pitt.edu), Panayiotis V Benos (benos@pitt.edu), Harinder Singh (harinder@pitt.edu)

Abstract

High-dimensional cellular and molecular profiling of human samples highlights the need for analytical approaches that can integrate multi-omic datasets to generate predictive biomarkers that are in turn accompanied with strong causal inferences. Current methodologies are challenged by the high dimensionality of the combined datasets, the differences in distributions across the datasets, and their integration in a plausible causal framework, beyond merely correlative biomarkers. Here we present CausER, a first-in-class two-step interpretable machine learning approach for high-dimensional multi-omic datasets, that addresses these problems by identifying latent factors and their cause-effect relationships with the system-wide outcome/property of interest. The first step consists of Essential Regression (ER), a novel data-distribution-free regression model that integrates multi-omic datasets and identifies latent factors significantly associated with an outcome. The second involves probabilistic graphical modeling of the significant latent factors to infer plausible causal associations between them and mechanisms that affect outcomes, thereby significantly moving beyond predictive associative markers. By analyzing varied human immunological multi-omic datasets, we demonstrate that CausER significantly outperforms a wide range of state-of-the-art approaches. It generates novel cellular and molecular predictions in a range of contexts, including immunosenescence and sustained immune dysregulation associated with pre-term birth, that are corroborated by biological findings in model systems.

Introduction

Over the last decade, genomic, proteomic, metabolomic and other technologies for generating deep molecular profiles of tissues and cells from model organisms or humans have rapidly expanded[1-4]. However, the explosion in data, especially from a range of such 'omic technologies has not been coupled to a proportional increase in our understanding of the underlying causal chains and mechanisms. Existing analytical approaches have primarily focused on individual "omic" datasets with relatively few attempts at integration of multi-omic datasets. In either case, we[5-9] and others[10-12] have primarily emphasized on uncovering predictive biomarkers (Fig. 1a). A key focus of these efforts is to overcome the "curse of dimensionality" (very large number of variables being measured in relation to a comparatively low number of samples) and the multiplicity of predictive signatures due to multi-collinear data i.e., large correlated sets of variables. While there are several methods for uncovering predictive markers from high dimensional data, none of these analyze cause-effect relationships in relation the outcomes/outputs of interest. This in turn has hampered efforts to undertake perturbative/translational experiments and/or clinical investigations that can test a functionally prioritized set of hypotheses generated by the large datasets.

In addition to the high dimensionality of datasets at any given scale of organization (e.g., cellular, molecular), biological systems, particularly humans, manifest extreme complexity in terms of numbers of molecular components, their interaction rules as well as their hierarchical scales of organization that include macromolecular complexes/condensates, organelles, cells, tissues and organs. Each scale of organization in such a complex system has components and interaction rules that are unique to its level of organization. Thus, predicting changes in properties or behaviors of the system based on measuring components that are operating at different scales of organization represents a formidable challenge.

We propose a novel framework to address these key limitations by focusing on latent factors rather than observables in high dimensional datasets that are significantly associated with a system wide-property or outcome that is of interest. Further, the use of regression on the latent factors rather than the observables comes with rigorous statistical guarantees and provides a major conceptual advance that helps address current limitations. After identifying significant latent factors, we use causal graphical model analyses to examine the connectivity of these factors to the system-wide property or outcome of interest. Our analytical framework, termed CausER, attempts to move beyond biomarkers and derives causal latent factors from thousands of variables from multi-omics datasets across various scales of biological organization, and subsequently identifies potential cause-effect relationships between those factors (Fig. 1a). In so doing CausER generates a prioritized set of latent factors comprised of known observables that are most proximal in the causal graph network to the system property/outcome of interest. By analyzing three human immunological multi-omic datasets, we demonstrate that CausER significantly outperforms a wide range of state-of-the-art approaches in predicting outcomes and provides multi-scale inferences not afforded by the existing methods. The novel predictions are corroborated by biological findings in model systems.

Results

CausER – a framework for inferring causal latent factors

The first step in CausER comprises Essential Regression (ER), a novel data-distribution-free regression model that integrates multi-omic datasets and identifies latent factors that are significantly associated with a system property/outcome (Fig. 1b, Methods). ER involves a latent model approach that we previously described[13, 14] which performs unsupervised identification of latent factors from the input data. Next, ER regresses these latent factors (rather than the original measured variables) to the system property/outcome variable of interest (Fig. 1b, Methods). ER is a paradigm-altering concept in regression analysis in the context of dimensionality reduction with preservation of the underlying information. Existing regression methods use regularization (e.g., L1 regularization – LASSO, L1 + L2 regularization – Elastic Net) or pre-specified group structure (e.g., group LASSO) on the measured features, which perform feature selection that eliminates much of the underlying information (e.g., correlated variables). ER, on the other hand, identifies latent factors in a data-dependent fashion (without the need for pre-specified group structure) and then hones in on specific latent factors significantly associated with the property/outcome of interest. Importantly, each latent factor is summarizing the values of a number of variables that It represents, and therefore preserves the underlying information. Critically, ER makes no assumptions regarding the underlying data generating mechanisms and can be broadly used across multi-omic datasets (Methods). We note that ER enables the further analysis of all observable features within the significant latent factors (Fig. S1). The use of L1-regularization on the significant latent factors identified by ER allows us to identify a sparse set of observables, within these factors, tied to outcome. We term this ER-derivative-approach Composite Regression (CR) (Fig. S1).

The second step in CausER involves causal inference analyses on the ER-identified significant latent factors using directed graphical models[15]. Directed Acyclic Graphs (DAGs) are sometimes referred to as Causal Graphs, because under certain assumptions the learned DAGs from observational data (Markov equivalence classes) asymptotically represent the true data-generating causal graph. Although these algorithms have shown considerable success in analyzing many biological processes and biomedical problems[16-20], including biomarker selection and classification[21-23], scalability limits the datasets that they can be applied.[24, 25] Here, we use the causal learning algorithm for mixed data, CausalMGM,[16, 26] only on the significant latent factors delineated by ER, to overcome the scale limitation. By applying CausalMGM only on the significant latent factors, we greatly reduce the dimensionality of the input dataset while preserving the information of individual (correlated) variables in the latent factors. Thus, CausER (CausalMGM on the significant latent factors from ER) prioritizes further within the significant latent factors (Fig. 1c, Methods) by virtue of their direct connections to the outcome in the graphical model. Furthermore, it predicts potential cause-effect relationships between the latent factors and the property/outcome of interest, which leads to hypotheses generation. The associations of latent factors to outcome revealed this way provide a highly prioritized set of hypotheses which can then be corroborated by prior biological information and subjected to experiment tests or clinical investigations (Fig. 1d, Methods).

Inferring causal factors underlying immunosenescence in a vaccine response

A recent study comprehensively profiled cellular and molecular responses induced by the shingles Zostavax vaccine in a cohort comprising both younger adults and elderly subjects[27]. The high dimensional multi-omic analysis included immune cell frequencies and phenotypes, as well as transcriptomic, metabolomic, cytokine and antibody analyses. The vaccine induced robust antigen-specific antibody titers as well as $CD4^+$ but not $CD8^+$ T cell responses[27]. Using a multiscale, multifactorial response network, the authors identified associations between transcriptomic, metabolomic, cellular phenotypic and cytokine datasets which pointed to immune and metabolic correlates of vaccine immunity[27]. Interestingly, differences in the quality of the vaccine-induced responses by age were also noted[27]. We hypothesized that a method based on latent factors rather than measurables would improve the delineation of components that underlie the quality and magnitude of the vaccine-induced responses. If so, then such a method would be able to leverage the differences in vaccine-induced responses and accurately predict age as the system-wide property of interest. The latent factors identified in this manner could then provide insights into the cellular and molecular basis of age-induced immunosenescence manifested by diminished responses to the Zostavax vaccine.

To explore the above formulation of immunosenescence as a predictor of age, we first applied a suite of state-of-the-art approaches including the least absolute selection and shrinkage operator (LASSO)[28], partial least squares (PLS) regression[29], and principal components/factors regression (PFR)[30] on the entire spectrum of vaccine-induced responses to predict age (Fig. 2a). As most subjects in the cohort were in 2 distinct age groups – adults under 40 and elderly people over 60, we first sought to explore the performance of LASSO, PLS and PFR in predicting the two age groups as binary categorical variables i.e., younger adults and elderly-. The predictive performance of all methods was evaluated in a stringent leave-one-out cross-validation (LOOCV) framework (Methods). We have previously demonstrated that on such multi-omic datasets, cross-validation is a gold standard to evaluate model performance with data held out[5, 6, 8]. In a LOOCV framework, we found that PFR had no predictive power (AUC < 0.5), while LASSO and PLS had weak predictive power in predicting age as a categorical variable (Fig. 2b, AUCs = 0.63 and 0.60 respectively). The ROC curve for LASSO had an interesting shape. It attained a true positive rate of ~0.4 at a false positive rate of ~0.15, but beyond that it was essentially no better than random (Fig. 2b). This observation is consistent with the observation that differences in an age-associated MMRN were driven by only a subset of elderly vaccinees[27]. Thus, a purely predictive modeling approach like LASSO can leverage these relatively straightforward differences to accurately predict age for a subset of the vaccinees, but fails to predict age for others. We then compared these methods to the performance of ER, CR and CausER. In a matched, LOOCV framework, ER and CR were very accurate at predicting age (Fig. 2b, AUCs = 0.79 and 0.77 respectively, $P < 0.01$), while CausER was the best predictor of age as a categorical variable (AUC = 0.86, $P < 0.01$). Together, these results demonstrate that while LASSO, PLS and PFR fail to accurately predict age from Zostavax-induced vaccine responses, ER, CR and CausER can overcome this challenging problem by leveraging non-trivial differences in latent factors comprised of discrete sets of measurables.

Next, we evaluated whether these methods could predict actual age as a continuous variable beyond the categorical classifiers of younger adults and the elderly. As before, performance was measured in a rigorous cross-validation framework (Methods). Using the vaccine-induced responses, PFR was not at all predictive of age (Fig. 2c, Pearson r = -0.71; Fig. S2, Spearman r = -0.82). LASSO and PLS had poor performance in predicting age as a continuous variable (Fig. 2c, Pearson r = 0.29 and 0.13 respectively; Fig. S2, Spearman r = 0.25 and 0.09 respectively). In fact, the predictive power of PLS and PFR were not significantly different from a negative control model built on permuted data (Fig. 2c). However, both ER and CR were significantly predictive of age as a continuous variable (Pearson $r = 0.48$ for both, Spearman r = 0.44 and 0.49 respectively, $P < 0.01$ Fig. 2c, Fig. S2), and as in the previous instance, CausER had the best performance in predicting age as a continuous variable (Pearson $r = 0.61$, Spearman r = 0.59, $P < 0.01$ Fig. 2c, Fig. S2). Together, these results demonstrate that while state-of-the-art methods including LASSO, PLS and PFR fail to predict age either as a categorical or a continuous variable, all three of the new approaches that are based on latent factors – ER, CR and CausER, are able to do so reasonably accurately based on the multi-omic profiles of vaccine-induced responses.

We next explored the likely causal relationships among the latent factors that lead to age-induced immunosenescence and diminished responses to the Zostavax vaccine. CausalMGM was used to construct a causal graph with all latent factors identified in the latent model identification step of ER (Fig. 2d). Notably, majority of the significant latent factors identified by ER were seen to proximal to the outcome variable (age) in the causal graph. Importantly, all 4 latent factors in the Markov blanket generated by CausalMGM were also identified as significant by ER (Fig. 2d). Overall, the significant latent factors revealed by ER had significantly lower network distances (i.e., had stronger cause-effect relationships) from age compared to the non-significant latent factors (Fig. 2e, *P* < 0.05). These results demonstrate that ER and CausalMGM independently converge on the same presumptive cause-effect relationships.

The prioritized CausER hits (Fig. 2d) i.e, significant latent factors identified by ER that are also in the Markov blanket of the outcome variable (age) in the causal graph generated by CausalMGM comprised antigen-specific IgG titers (Z1), a metabolic module (Z19), B cell (Z46) and NK cell frequencies (Z45). CausER provides both prioritized cause-effect relationships and directions of these relationships. While the latter relates to mathematical conditional independence relationships (Methods), the former provides prioritized mechanistic insights.

While the lowering of titers with age is expected and has been previously reported[27], CausER revealed a likely cause-effect relationship between altered B cell and NK cell numbers and immunosenescence. To further dissect the nature of this relationship, we examined correlations between NK cells, B cells and age. We found that NK cells significantly increased, while the numbers of B cells significantly decreased with age (Fig. 2f). More interestingly, there was a significant negative correlation between NK cells and B cells (Fig. 2f), and the correlation remained significant even after correcting for age (Fig. 2f). Our results suggest a novel basis of human immunosenescence in the context of vaccine responses (Fig. 2g). This could involve a previously described mechanistic linkage between NK cells and a weaker germinal center (GC) response in a murine model[31]. NK cells can inhibit CD4 T cell responses including those of T follicular helper cells in a perforin-dependent manner; this leads to a weaker GC response diminished antibody titers and affinity maturation[31,32].

Analyzing latent factors potentially reflective of trained immunity in a vaccine response

Next, we used CausER to analyze the temporal dynamics of transcriptional responses induced by the malaria RTS,S vaccine[33]. RTS,S has a standard regimen of 3 doses separated by a month, and is currently the most advanced malaria vaccine candidate, that has consistently demonstrated 40-80% protective efficacy in malaria-naïve individuals in controlled human challenge studies[5]. There has been intense interest over the last decade at uncovering molecular signatures induced by the RTS,S vaccine and corresponding correlates of protection[5, 34, 35]. In a controlled human infection setting, differential expression of immunoproteasome genes was identified as a pre-challenge correlate of protection[33]. After the third dose, as expected, there was a striking but transitory shift in inflammatory gene expression followed a convergence of the majority of gene signatures back to pre-vaccination levels within 2 weeks after the third dose[33]. We reasoned that aspects of trained immunity induced by the vaccine may be reflected in the transcriptomic signatures that do not converge after 2 weeks. Thus, a sensitive method such as CausER would be able to discriminate between expression profiles at the following time-points – pre-vaccination (G1), the day after the third dose (G2) and 14 days after the third dose (G3) (Fig. 3a) and reveal candidate genes and molecular pathways that could contribute to trained immunity. In this instance, the use of a microarray dataset also afforded the opportunity to explore how CausER performs with noisier but nevertheless valuable datasets generated using older technologies.

As before, the ability of the different methods to discriminate between G1, G2 and G3 transcriptional profiles was measured in a rigorous cross-validation framework (Methods). We found that there were significant differences in the ability of the different methods to discriminate between the three kinds of expression profiles, with CausER and ER having the best performance, significantly better than the other methods (P < 0.01, Figs. 3b, 3c). Next, we chose to focus on the ability of the different methods to specifically distinguish the G3 profile from the other two (Fig. 3d) or just the G1 profile (Fig. S3a). This constituted the most “difficult” discrimination as there are broad differences in the expression profiles between the pre- (G1) and 24-hour-post-vaccination (G2) time-points, but most of these differences disappear by 14 days

(G3)[33]. Consistent with expectation, in this binary classification setting, there was wide variability in the performance of the methods to specifically discriminate the G3 time-point from the G1 and G2 time-points. While PFR and PLS performed poorly, CausER, ER and LASSO had significantly better performance, with CausER being the best performing method ($P < 0.01$, Figs. 3d, Fig S3). In terms of correctly classifying just the true G3 profiles as G3, PLS and PFR had poor performance, while CausER had the best performance, significantly better than other methods ($P< 0.01$, Fig. 3e).

Next, we focused on the CausER hits i.e., the significant latent factors from ER in the Markov blanket of the outcome variable (Fig. 3f). Genes comprising these latent factors were seen to be differentially expressed between the G1 and G3 samples (Fig. 3g, Fig. S3b). Our results suggest that beyond the initial divergence of immunoproteasome genes, there is a sustained divergence (2 weeks post-vaccination) of genes involved in immune-metabolic processes. These results complement recent findings that suggest that targeting immunometabolism is a promising direction in modulating trained immunity[36]. While a vaccine induces a rapid initial divergence in inflammatory signatures reflecting the activation of innate immune cells and their engagement with adaptive B and T cells, it may also induce alterations in the innate immune compartment that are discernible at later time points and contribute to a distinct form of immune memory[36].

Uncovering latent factors that distinguish immune system states of term and pre-term infants

Finally, we focused on a multi-omic longitudinal cohort that analyzed immune cell populations and plasma proteins in 100 newborn children during their first 3 months of life[37] (Fig. 4a). Striking differences were observed in immune parameters between preterm and term children at birth. However, the immune trajectories appeared to achieve a stereotypic convergence within the first 3 months of life[37] (Fig. 4a). We hypothesized that CausER might be able to uncover latent factors that distinguish immune system states of term and pre-term infants after 3 months of life and therefore reveal features that could impact later life (Fig. 4a). As expected, based on the striking differences at birth between term and pre-term children, all methods (LASSO, PLS, PFR, ER, CR and CausER) were be able to discriminate between these 2 groups using immune parameters measured in the first week of life (Fig. S4). All model performances were measured in a rigorous cross-validation framework (Methods). However, given the stereotypic convergence in the first 3 months (12 weeks) of life[37], we found that PLS and PFR were unable to accurately discriminate between term and pre-term children using immune parameters measured at 12 weeks of life (Figs. 4b-4c). However, LASSO was able to accurately distinguish between term and pre-term births using the 12-week profiles (Figs. 4b-4c), suggesting that despite broad convergence, a small subset of immune parameters still remain different term and pre-term infants between at 3 months of life. More importantly, ER and CR were able to accurately discriminate between term and pre-term births using immune profiles at 3 months of life, significantly better than other methods (Figs. 4b-4c, $P < 0.01$). ER identified only 2 significant latent factors, and based on CausalMGM analyses, one of these 2 significant latent factors was in the Markov blanket i.e., for this dataset, this single latent factor was the sole CausER hit (Fig. 4d).

We visualized the immune cell populations and plasma proteins in this CausER hit (Fig. 4d). These profiles had clearly remained divergent even at 3 months of life (Fig. 4d) despite the broad stereotypic convergence of most other immune parameters. At 3 months of life, term infants had an anti-inflammatory milieu including high IL-10, while pre-term infants had a pro-inflammatory milieu including elevated IL-6 and IL-8 (Fig. 4d). These findings agree with a previous study that IL-10 is highly expressed in the uterus and placenta and has a key role in controlling inflammation-induced pre-term labor in a murine model[38]. Furthermore, regulatory B cells are a key source of IL-10 and appear to be important in sustaining pregnancy till term[39-41]. It is also known that modulation of pro-vs-anti-inflammatory environments by relevant cytokines and chemokines at the maternal-fetal interface (decidua) is a critical component of the bifurcation between term and pre-term births[39]. Thus, our analyses of immune system states of term and pre-term infants at 3 months of life revealed that pre-term infants had a pro-inflammatory state, while term infants had an anti-inflammatory state (Fig. 4e). These findings could have long-term implications for the health of pre-term infants.

Discussion

Over the last two decades, while there have been rapid advances in high-throughput experimental technologies to generate deep molecular profiles, computational analyses of these high-dimensional datasets have primarily focused on biomarker discovery[42]. This is because rigorous statistical approaches for analyzing high-dimensional datasets, such as regularized regression and bootstrap aggregated classification, are focused on uncovering predictive biomarkers which may simply be correlative surrogates of outcome or system-wide property but unrelated to the underlying causal factors. Incorrect extrapolation of insights derived from biomarker-based approaches can lead to perturbation experiments with low success. Alternatively, efforts to move beyond biomarkers to mechanistic insights often use biological priors, which may be incomplete or suffer from sampling/study biases[43]. Further, while there have been advances in causal modeling[44], existing approaches are difficult to apply to high-dimensional datasets due to the computational intractability of applying these approaches on[15] and the multi-collinearity of the data. The methods presented in this manuscript address this fundamental limitation in systems biology. ER and CausER are first-in-class machine learning methods that can both handle high-dimensional multi-omic datasets with co-linear variables and prioritize cause-effect relationships between the input features and the outcome of interest.

The CausER framework pushes the envelope on multiple key challenges in systems biology. First, it establishes a rigorous framework with provable statistical guarantees that explores a large space of higher-order relationships from high-dimensional features and uncovers latent factors tied to the outcome variable via directed cause-effect relationships. Second, unlike existing causal reasoning approaches that are constrained by the size of the input data, CausER can be applied to modern high-dimensional datasets. The time complexities of the different steps are essentially quadratic and not exponential like some other causal reasoning approaches. Third, ER makes no assumptions regarding data-generating mechanisms and CausER can integrate multi-omic datasets to capture the interplay across a plethora of biological processes at multiple scales of organization of the system. A key innovation within the framework is the sequential use of two orthogonal methods for statistical inference, ER and CausalMGM. These methods have different theoretical bases and assumptions and yet converge on common causal latent factors, underscoring the robustness of our approach.

Here, we applied ER and CausER to three biologically diverse contexts. In each case, we leveraged an existing study that had generated high-dimensional omic profiles to address key questions that had not been the focus of the original studies, in part because of limitations of methods used. Such questions could now be addressed by the methodological advances of ER and CausER over state-of-the-art approaches. We demonstrated that ER and CausER significantly outperform PFR and PLS across contexts, and either outperform or match LASSO in terms of predictive performance. While we used three examples to illustrate the superior performance of ER and CausER, these methods come with broad theoretical guarantees to outperform PLS, PFR and LASSO across contexts (Methods, Fig. S5). Further, while the existing methods simply identify correlates, ER and CausER provide mechanistic insights, some of which are consistent with prior knowledge, while others are novel. Our findings have broad implications across domains in systems biology and are likely to transform both computational workflows used to analyze multi-omic datasets and downstream experiments designed based on the insights gleaned via these analyses.

Online Methods

Detailed descriptions of the theoretical underpinnings and associated proofs of ER, CR and CausER are provided in Supplementary File 1. Supplementary File 1 also describes details of the application of ER, CR, CausER, LASSO, PLS and PFR to the different datasets of interest.

Code Availability

Detailed code and documentation for ER and CR and CausER are available at https://github.com/bingx1990/Application-of-ER-and-CausalMGM.git

Author Contributions

J.D. designed the study, and oversaw all aspects of it. X.B., F.B. and M.W. jointly conceived the ER framework. J.D., P.B. and H.S. jointly conceived the CausER framework. X.B. and T.L. implemented the ER and CausER frameworks, and carried out all computational analyses. J.D. and H.S. interpreted the results. J.D., H.S. and P.B. wrote the main text. X.B., T.L., F.B. and M.W. wrote the supplementary methods including formal proofs.

Figure Legends

Figure 1 – An overview of Essential Regression and CausER

a) Schematic illustrating the different kinds of multi-omic datasets typically used in systems analyses and the key advantages of the methods introduced in this study (ER and CausER) over existing approaches.

b) Schematic summarizing the steps in ER.

c) Schematic summarizing the steps in CausER.

d) Conceptual overview of key advances afforded by ER and CausER.

Figure 2 – Identifying causal signatures of age-induced immunosenescent responses to the Zostavax vaccine

a) Schematic summarizing the input data and the problem of interest.

b) ROC curves for the different methods at discriminating between elderly and younger adults in a LOOCV framework.

c) Pearson correlations of the different methods at predicting age as a continuous variable, as measured in a LOOCV cross-validation framework.

d) CausER graph – CausalMGM on all Z's. Markov blanket highlighted with a blue border and bolder fonts. A directed edge X --> Y indicates X is a cause of Y, while a bidirected edge X <-> Y indicates the presence of a latent confounder that is a common cause of X and Y. A partially oriented edge X o-> Y indicates that Y is not a cause of X, but either X or a latent confounder causes Y. Unoriented edge indicates directionality couldn't be inferred for that edge.

e) Network distances in the causal graph generated by CausalMGM of the significant and non-significant Z's (identified by ER) from the outcome variable of interest.

f) Mechanistic insights obtained from CausER.

Figure 3 – Identifying differences in vaccine-induced transcriptomic profiles over time

a) Schematic summarizing the input data and the problem of interest.

b) Ternary classification accuracy of the different methods at discriminating among G1, G2 and G3 in a replicated *k*-fold cross-validation framework.

c) Confusion matrix summarizing the performance of the different methods at discriminating among G1, G2 and G3 in a LOOCV framework.

d) ROC curves for the different methods at discriminating between G3 and (G1 & G2) combined in a LOOCV framework.

e) Fraction of true G3 correctly classified as G3 (as measured in a LOOCV framework).

f) CausER graph – CausalMGM on the significant Z's from ER. Markov blanket highlighted with a blue border and bolder fonts. A directed edge X --> Y indicates X is a cause of Y, while a bidirected edge X <-> Y indicates the presence of a latent confounder that is a common cause of X and Y. A partially oriented edge X o-> Y indicates that Y is not a cause of X, but either X or a latent confounder causes Y. Unoriented edge indicates directionality couldn't be inferred for that edge.

g) Heatmap of genes in CausER hits (significant Z's in the Markov blanket) for G1 and G3 samples.

Figure 4 – Uncovering specific immune parameters from term and pre-term infants that do not achieve stereotypic convergence

a) Schematic summarizing the input data and the problem of interest.

b) Classification accuracy of the different methods at discriminating between term and pre-term births using immune profiles at 3 months after birth, measured in a replicated *k*-fold cross validation framework.

c) ROC curves for the different methods at discriminating between term and pre-term births as measured in a LOOCV framework.

d) Heatmap of features (plasma proteins and immune cells) in the single CausER hit (significant Z in the Markov blanket).

e) Mechanistic insights obtained from CausER.

Supplementary Figures

Fig. S1 (accompanying Fig. 1) – Schematic of CR.

Fig. S2 (accompanying Fig. 2) – Spearman correlations of the different methods at predicting age as a continuous variable, as measured in a LOOCV cross-validation framework.

Fig. S3 (accompanying Fig. 3)
a) ROC curves for the different methods at discriminating between G3 and G1 in a LOOCV framework.
b) Heatmap of genes in CausER hits (significant Z's in the Markov blanket) for G1, G2 and G3 samples

Fig. S4 (accompanying Fig. 4)
a) Classification accuracy of the different methods at discriminating between term and pre-term births using immune profiles at 1 week after birth, measured in a replicated k-fold cross validation framework
b) ROC curves for the different methods at discriminating between term and pre-term births as measured in a LOOCV framework.

Fig. S5 – Predictive performance of PLS, PFR, LASSO and ER on simulated datasets.

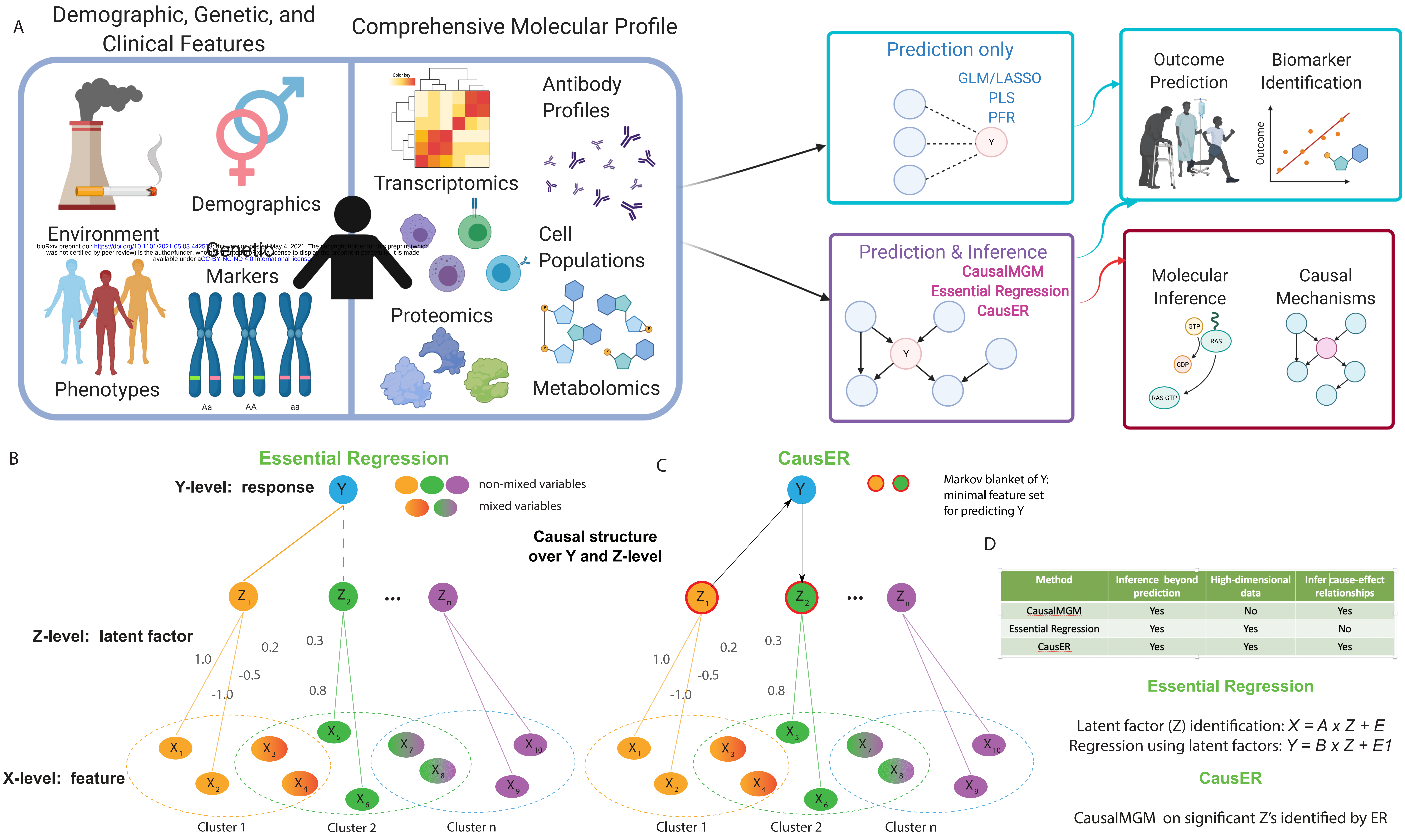
Supplementary File 1

Detailed descriptions of the theoretical underpinnings, associated proofs of ER, CR and CausER. The file also describes details of the applications of ER, CR, CausER, LASSO, PLS and PFR to the different datasets of interest.

References

1. Hagan, T. & Pulendran, B. Will Systems Biology Deliver Its Promise and Contribute to the Development of New or Improved Vaccines? From Data to Understanding through Systems Biology. *Cold Spring Harb Perspect Biol* **10** (2018).
2. Pulendran, B., Li, S. & Nakaya, H.I. Systems vaccinology. *Immunity* **33**, 516-529 (2010).
3. Davis, M.M., Tato, C.M. & Furman, D. Systems immunology: just getting started. *Nat Immunol* **18**, 725-732 (2017).
4. Villani, A.C., Sarkizova, S. & Hacohen, N. Systems Immunology: Learning the Rules of the Immune System. *Annu Rev Immunol* **36**, 813-842 (2018).
5. Suscovich, T.J. et al. Mapping functional humoral correlates of protection against malaria challenge following RTS,S/AS01 vaccination. *Sci Transl Med* **12** (2020).
6. Das, J. et al. Mining for humoral correlates of HIV control and latent reservoir size. *PLoS Pathog* **16**, e1008868 (2020).
7. Goetghebuer, T. et al. Initiation of Antiretroviral Therapy Before Pregnancy Reduces the Risk of Infection-related Hospitalization in Human Immunodeficiency Virus-exposed Uninfected Infants Born in a High-income Country. *Clin Infect Dis* **68**, 1193-1203 (2019).
8. Ackerman, M.E. et al. Route of immunization defines multiple mechanisms of vaccine-mediated protection against SIV. *Nat Med* **24**, 1590-1598 (2018).
9. Sadanand, S. et al. Temporal variation in HIV-specific IgG subclass antibodies during acute infection differentiates spontaneous controllers from chronic progressors. *AIDS* **32**, 443-450 (2018).
10. Vafaee, F. et al. A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. *NPJ Syst Biol Appl* **4**, 20 (2018).
11. Li, S. et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat Immunol* **15**, 195-204 (2014).
12. Nakaya, H.I. et al. Systems biology of vaccination for seasonal influenza in humans. *Nat Immunol* **12**, 786-795 (2011).
13. Bing, X., Bunea, F., Royer, M. & Das, J. Latent Model-Based Clustering for Biological Discovery. *iScience* **14**, 125-135 (2019).
14. Bing, X., Bunea, F., Ning, Y. & Wegkamp, M. Adaptive estimation in structured factor models with applications to overlapping clustering. *The Annals of Statistics* **48**, 2055-2081, 2027 (2020).
15. Ge, X., Raghu, V.K., Chrysanthis, P.K. & Benos, P.V. CausalMGM: an interactive web-based causal discovery tool. *Nucleic Acids Res* **48**, W597-W602 (2020).
16. Sedgewick, A.J. et al. Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis. *Bioinformatics* **35**, 1204-1212 (2019).
17. Schadt, E.E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**, 710-717 (2005).
18. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. & Nolan, G.P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523-529 (2005).
19. Manatakis, D.V., Raghu, V.K. & Benos, P.V. piMGM: incorporating multi-source priors in mixed graphical models for learning disease networks. *Bioinformatics* **34**, i848-i856 (2018).
20. Kitsios, G.D. et al. Respiratory Microbiome Profiling for Etiologic Diagnosis of Pneumonia in Mechanically Ventilated Patients. *Front Microbiol* **9**, 1413 (2018).
21. Abecassis, I. et al. PARP1 rs1805407 Increases Sensitivity to PARP1 Inhibitors in Cancer Cells Suggesting an Improved Therapeutic Strategy. *Sci Rep* **9**, 3309 (2019).

22. Raghu, V.K. et al. Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models. *Thorax* **74**, 643-649 (2019).
23. Raghu, V.K. et al. Biomarker identification for statin sensitivity of cancer cell lines. *Biochem Biophys Res Commun* **495**, 659-665 (2018).
24. Raghu, V.K. et al. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *Int J Data Sci Anal* **6**, 33-45 (2018).
25. Raghu, V.K., Poon, A. & Benos, P.V. in Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery, Vol. 92 48--65 (PMLR, Proceedings of Machine Learning Research; 2018).
26. Sedgewick, A.J., Shi, I., Donovan, R.M. & Benos, P.V. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics* **17 Suppl 5**, 175 (2016).
27. Li, S. et al. Metabolic Phenotypes of Response to Vaccination in Humans. *Cell* **169**, 862-877 e817 (2017).
28. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).
29. Boulesteix, A.L. & Strimmer, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform* **8**, 32-44 (2007).
30. Bair, E., Hastie, T., Paul, D. & Tibshirani, R. Prediction by Supervised Principal Components. *Journal of the American Statistical Association* **101**, 119-137 (2006).
31. Rydyznski, C. et al. Generation of cellular immune memory and B-cell immunity is impaired by natural killer cells. *Nat Commun* **6**, 6375 (2015).
32. Rydyznski, C.E. et al. Affinity Maturation Is Impaired by Natural Killer Cell Suppression of Germinal Centers. *Cell Rep* **24**, 3367-3373 e3364 (2018).
33. Vahey, M.T. et al. Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine. *J Infect Dis* **201**, 580-589 (2010).
34. Kazmin, D. et al. Systems analysis of protective immune responses to RTS,S malaria vaccination in humans. *Proc Natl Acad Sci U S A* **114**, 2425-2430 (2017).
35. Neafsey, D.E. et al. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. *N Engl J Med* **373**, 2025-2037 (2015).
36. Arts, R.J., Joosten, L.A. & Netea, M.G. Immunometabolic circuits in trained immunity. *Semin Immunol* **28**, 425-430 (2016).
37. Olin, A. et al. Stereotypic Immune System Development in Newborn Children. *Cell* **174**, 1277-1292 e1214 (2018).
38. Robertson, S.A., Skinner, R.J. & Care, A.S. Essential role for IL-10 in resistance to lipopolysaccharide-induced preterm labor in mice. *J Immunol* **177**, 4888-4896 (2006).
39. Gomez-Lopez, N., StLouis, D., Lehr, M.A., Sanchez-Rodriguez, E.N. & Arenas-Hernandez, M. Immune cells in term and preterm labor. *Cell Mol Immunol* **11**, 571-581 (2014).
40. Rolle, L. et al. Cutting edge: IL-10-producing regulatory B cells in early human pregnancy. *Am J Reprod Immunol* **70**, 448-453 (2013).
41. Jensen, F., Muzzio, D., Soldati, R., Fest, S. & Zenclussen, A.C. Regulatory B10 cells restore pregnancy tolerance in a mouse model. *Biol Reprod* **89**, 90 (2013).
42. Libbrecht, M.W. & Noble, W.S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321-332 (2015).
43. Cusick, M.E. et al. Literature-curated protein interaction datasets. *Nat Methods* **6**, 39-46 (2009).
44. Pearl, J. An introduction to causal inference. *Int J Biostat* **6**, Article 7 (2010).



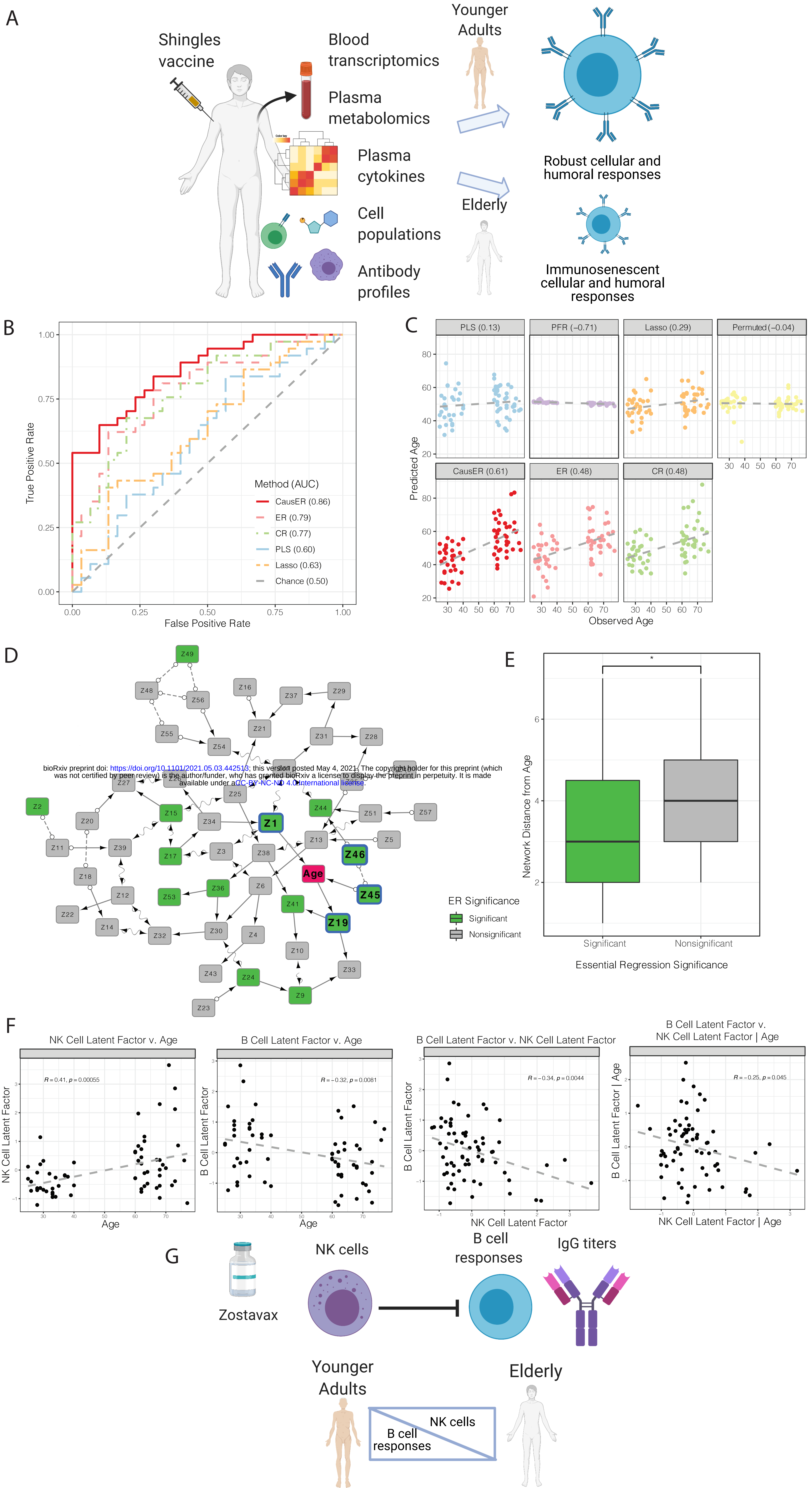
Method	Inference beyond prediction	High-dimensional data	Infer cause-effect relationships
CausalMGM	Yes	No	Yes
Essential Regression	Yes	Yes	No
CausER	Yes	Yes	Yes

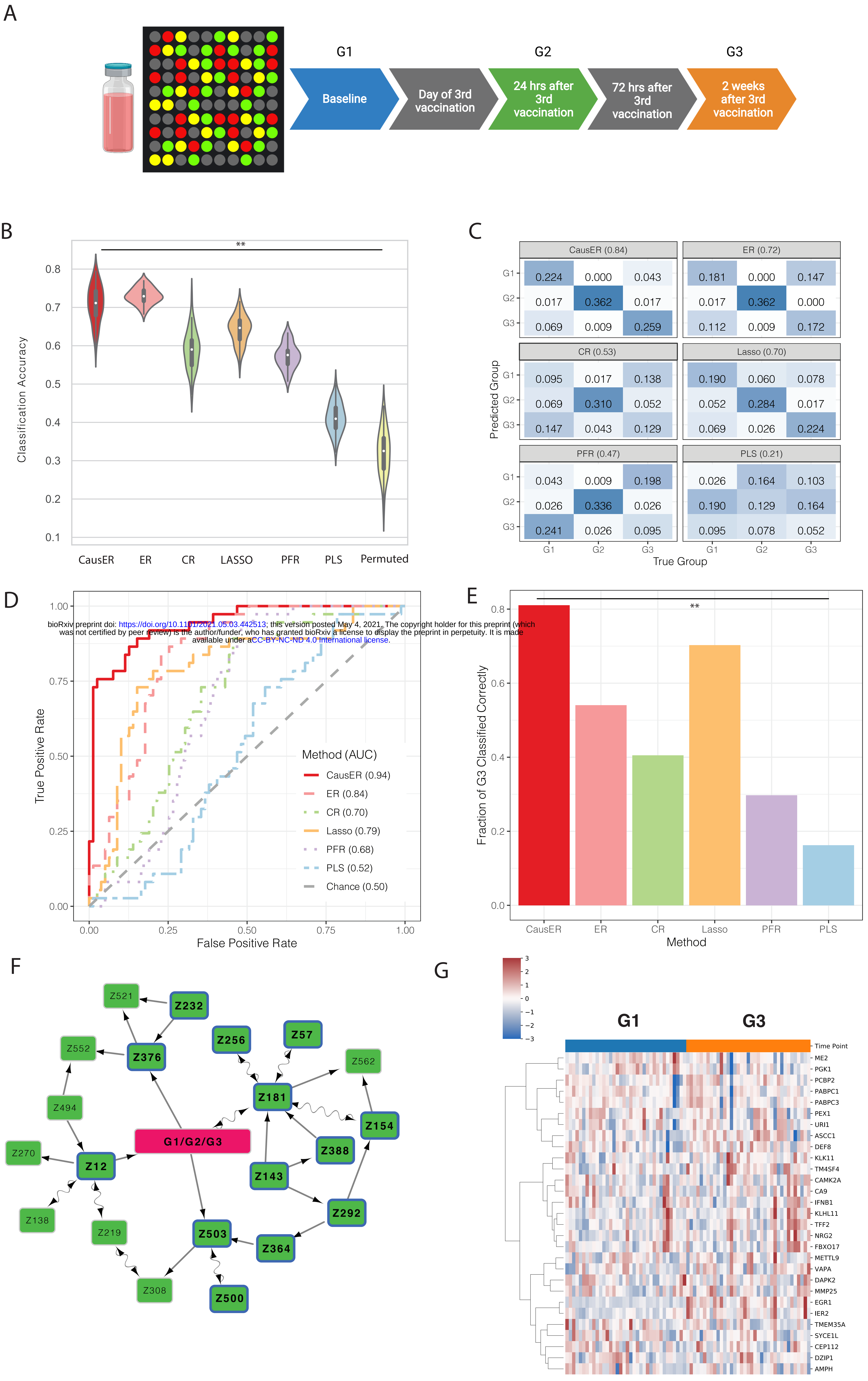
Essential Regression

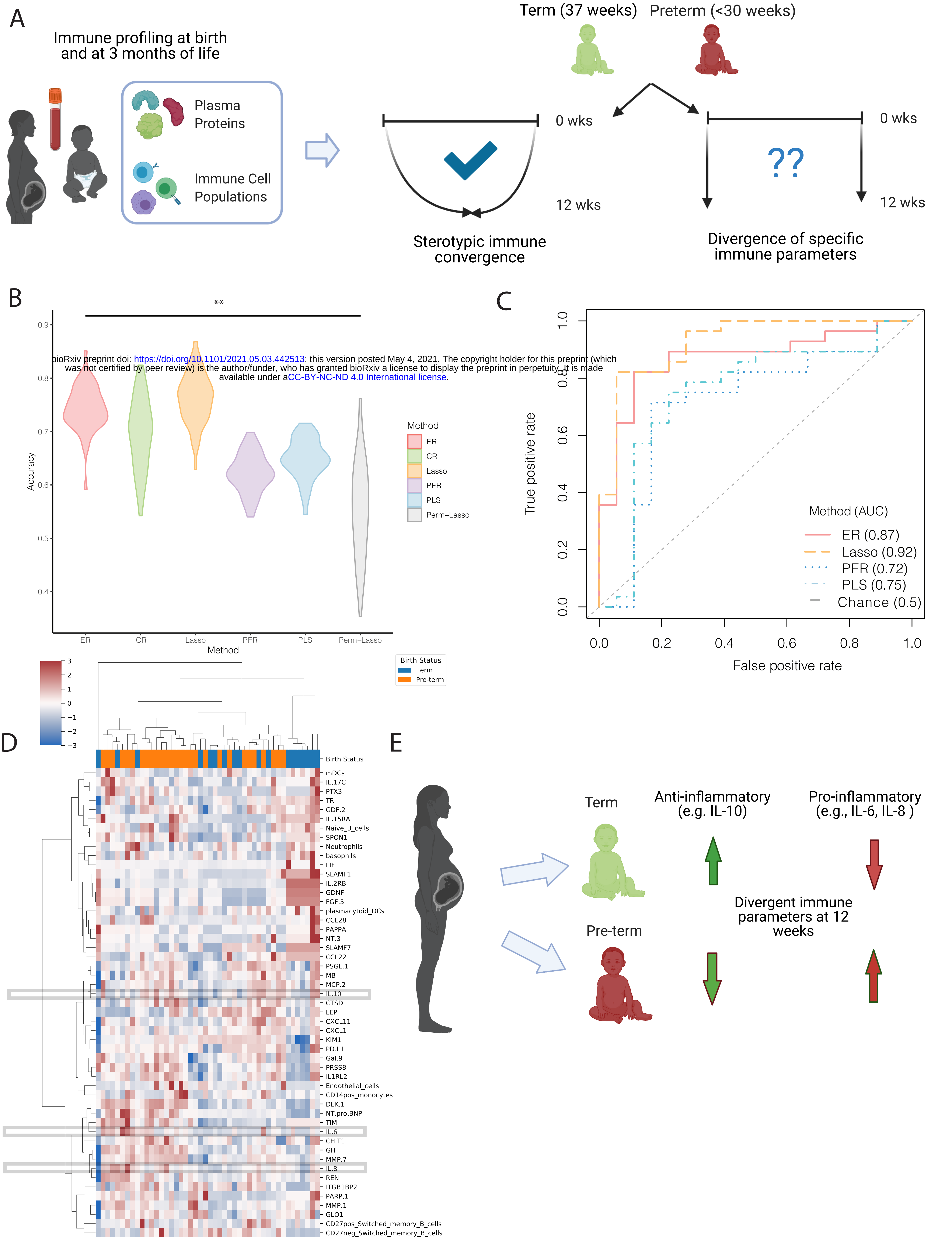
Latent factor (Z) identification: $X = A \times Z + E$
Regression using latent factors: $Y = B \times Z + E1$

CausER

CausalMGM on significant Z's identified by ER


A
Shingles vaccine
Blood transcriptomics
Plasma metabolomics
Plasma cytokines
Cell populations
Antibody profiles
Younger Adults
Elderly
Robust cellular and humoral responses
Immunosenescent cellular and humoral responses
B
True Positive Rate
False Positive Rate
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Method (AUC)
CausER (0.86)
ER (0.79)
CR (0.77)
PLS (0.60)
Lasso (0.63)
Chance (0.50)
C
PLS (0.13)
PFR (−0.71)
Lasso (0.29)
Permuted (−0.04)
CausER (0.61)
ER (0.48)
CR (0.48)
Predicted Age
Observed Age
80
60
40
20
30 40 50 60 70
D
Z49
Z48
Z56
Z55
Z54
Z16
Z21
Z37
Z29
Z31
Z28
Z27
Z26
Z25
Z2
Z20
Z15
Z34
Z1
Z44
Z51
Z57
Z11
Z39
Z17
Z3
Z38
Z13
Z5
Z46
Z18
Age
Z45
Z12
Z53
Z36
Z6
Z41
Z22
Z19
Z14
Z32
Z30
Z4
Z10
Z33
Z43
Z24
Z9
Z23
ER Significance
🟩 Significant
⬜ Nonsignificant
E
*
Network Distance from Age
6
4
2
Significant
Nonsignificant
Essential Regression Significance
F
NK Cell Latent Factor v. Age
R = 0.41, p = 0.00055
NK Cell Latent Factor
Age
B Cell Latent Factor v. Age
R = −0.32, p = 0.0081
B Cell Latent Factor
Age
B Cell Latent Factor v. NK Cell Latent Factor
R = −0.34, p = 0.0044
B Cell Latent Factor
NK Cell Latent Factor
B Cell Latent Factor v. NK Cell Latent Factor | Age
R = −0.25, p = 0.045
B Cell Latent Factor | Age
NK Cell Latent Factor | Age
G
Zostavax
NK cells
B cell responses
IgG titers
Younger Adults
Elderly
NK cells
B cell responses


A
G1
G2
G3
Baseline
Day of 3rd vaccination
24 hrs after 3rd vaccination
72 hrs after 3rd vaccination
2 weeks after 3rd vaccination
B
Classification Accuracy
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
**
CausER
ER
CR
LASSO
PFR
PLS
Permuted
C
CausER (0.84)
0.224 0.000 0.043
0.017 0.362 0.017
0.069 0.009 0.259
ER (0.72)
0.181 0.000 0.147
0.017 0.362 0.000
0.112 0.009 0.172
CR (0.53)
0.095 0.017 0.138
0.069 0.310 0.052
0.147 0.043 0.129
Lasso (0.70)
0.190 0.060 0.078
0.052 0.284 0.017
0.069 0.026 0.224
PFR (0.47)
0.043 0.009 0.198
0.026 0.336 0.026
0.241 0.026 0.095
PLS (0.21)
0.026 0.164 0.103
0.190 0.129 0.164
0.095 0.078 0.052
Predicted Group
G1
G2
G3
True Group
G1
G2
G3
D
True Positive Rate
1.00
0.75
0.50
0.25
0.00
False Positive Rate
0.00
0.25
0.50
0.75
1.00
Method (AUC)
CausER (0.94)
ER (0.84)
CR (0.70)
Lasso (0.79)
PFR (0.68)
PLS (0.52)
Chance (0.50)
E
Fraction of G3 Classified Correctly
0.8
0.6
0.4
0.2
0.0
**
CausER
ER
CR
Lasso
PFR
PLS
Method
F
Z521
Z232
Z552
Z376
Z256
Z57
Z562
Z494
Z181
Z154
G1/G2/G3
Z270
Z12
Z388
Z143
Z138
Z219
Z292
Z503
Z364
Z308
Z500
G
3
2
1
0
−1
−2
−3
G1
G3
Time Point
ME2
PGK1
PCBP2
PABPC1
PABPC3
PEX1
URI1
ASCC1
DEF8
KLK11
TM4SF4
CAMK2A
CA9
IFNB1
KLHL11
TFF2
NRG2
FBXO17
METTL9
VAPA
DAPK2
MMP25
EGR1
IER2
TMEM35A
SYCE1L
CEP112
DZIP1
AMPH


A
Immune profiling at birth and at 3 months of life
Plasma Proteins
Immune Cell Populations
Term (37 weeks)
Preterm (<30 weeks)
0 wks
12 wks
Sterotypic immune convergence
0 wks
??
12 wks
Divergence of specific immune parameters
B
**
0.9
0.8
0.7
0.6
0.5
0.4
Accuracy
ER
CR
Lasso
PFR
PLS
Perm–Lasso
Method
Method
ER
CR
Lasso
PFR
PLS
Perm–Lasso
C
1.0
0.8
0.6
0.4
0.2
0.0
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Method (AUC)
ER (0.87)
Lasso (0.92)
PFR (0.72)
PLS (0.75)
Chance (0.5)
D
3
2
1
0
−1
−2
−3
Birth Status
Term
Pre-term
Birth Status
mDCs
IL.17C
PTX3
TR
GDF.2
IL.15RA
Naive_B_cells
SPON1
Neutrophils
basophils
LIF
SLAMF1
IL.2RB
GDNF
FGF.5
plasmacytoid_DCs
CCL28
PAPPA
NT.3
SLAMF7
CCL22
PSGL.1
MB
MCP.2
IL.10
CTSD
LEP
CXCL11
CXCL1
KIM1
PD.L1
Gal.9
PRSS8
IL1RL2
Endothelial_cells
CD14pos_monocytes
DLK.1
NT.pro.BNP
TIM
IL.6
CHIT1
GH
MMP.7
IL.8
REN
ITGB1BP2
PARP.1
MMP.1
GLO1
CD27pos_Switched_memory_B_cells
CD27neg_Switched_memory_B_cells
E
Term
Anti-inflammatory (e.g. IL-10)
Pro-inflammatory (e.g., IL-6, IL-8)
Divergent immune parameters at 12 weeks
Pre-term