BaySyn: Bayesian Evidence Synthesis for Multi-system Multiomic Integration

Rupam Bhattacharyya[†1], Nicholas Henderson[1] and Veerabhadran Baladandayuthapani[1]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor Michigan 48109, USA*

[†]*Corresponding Author, EMail: rupamb@umich.edu*

The discovery of cancer drivers and drug targets are often limited to the biological systems - from cancer model systems to patients. While multiomic patient databases have sparse drug response data, cancer model systems databases, despite covering a broad range of pharmacogenomic platforms, provide lower lineage-specific sample sizes, resulting in reduced statistical power to detect both functional driver genes and their associations with drug sensitivity profiles. Hence, integrating evidence across model systems, taking into account the pros and cons of each system, in addition to multiomic integration, can more efficiently deconvolve cellular mechanisms of cancer as well as learn therapeutic associations. To this end, we propose *BaySyn* - a hierarchical Bayesian evidence synthesis framework for multi-system multiomic integration. BaySyn detects functionally relevant driver genes based on their associations with upstream regulators using additive Gaussian process models and uses this evidence to calibrate Bayesian variable selection models in the (drug) outcome layer. We apply BaySyn to multiomic cancer cell line and patient datasets from the Cancer Cell Line Encyclopedia and The Cancer Genome Atlas, respectively, across pan-gynecological cancers. Our mechanistic models implicate several relevant functional genes across cancers such as PTPN6 and ERBB2 in the KEGG adherens junction gene set. Furthermore, our outcome model is able to make higher number of discoveries in drug response models than its uncalibrated counterparts under the same thresholds of Type I error control, including detection of known lineage-specific biomarker associations such as BCL11A in breast and FGFRL1 in ovarian cancers. All our results and implementation codes are freely available via an interactive R Shiny dashboard at tinyurl.com/BaySynApp. The supplementary materials are available online at tinyurl.com/BaySynSup.

Keywords: Additive Gaussian processes, cancer driver genes, gene-drug associations, hierarchical Bayesian variable selection, KEGG gene sets, spike-and-slab priors.

1. Introduction

With the advent of sophisticated techniques and platforms, large-scale datasets covering multiple layers of cellular omics are becoming increasingly available.[1,2] Consistent advancements have been made in the last few years towards adding more dimensions to these high-throughput datasets, namely (1) additional to patient-level disease databases, model systems such as cell lines, patient-derived xenografts and organoids are being studied extensively in context of cancer and other diseases;[3,4] (2) assessing clinical information and therapeutic response with omics data to make pharmacogenomic discoveries is becoming increasingly common.[5,6] Multiple challenges arise during investigations of such datasets, including but not limited to computational inefficiency, complex nature of associations among the omic variables considered, and the biological interpretability and clinical implications of the results.[7] Specifically in context of cancer, the necessity to not only detect biomarker associations

with drug/treatment regimens but also to assess the functional relevance and mechanism of such associations is paramount, potentially guiding future therapeutic advances. Thus, novel algorithms that integrate multi-omics patient and model systems profiles can potentially reveal novel biomarkers, drug targets and predictive models in cancer.

Multi-dimensional data integration in cancer To address the wide range of complexity and variability in both detection and management of cancer, a number of multi-omics approaches have been able to uncover intricate molecular mechanisms and discover prognostic candidates.[8] Data integration approaches have proven particularly useful - both vertical (multiple experiments on a common cohort of samples)[9,10] and horizontal (meta-analysis of different cohorts)[11,12] integration methods have been developed.[13] To simultaneously identify pharmacogenomic associations and corresponding functional mechanisms, singular usage of either of these dimensions is insufficient due to the richness of the currently available omics databases. Multi-omics patient databases of cancer such as The Cancer Genome Atlas (TCGA),[14] while rich in transcriptomic, proteomic and other levels of omics profiles, do not typically provide comprehensive and systematic drug response on the same cohort of patients, restricting utilization of these profiles directly in pharmacogenomic contexts. Model systems databases such as the Cancer Cell Line Encyclopedia (CCLE)[15] and Genomics of Drug Sensitivity in Cancer (GDSC)[16] provide both molecular profiles and drug sensitivity information on the same set of models, but the cancer- or lineage-specific sample sizes of such databases are lower than their patient counterparts and association models built solely on them may suffer from the lack of sufficient statistical power to detect all the true signals. In this work, we propose a solution to this, based on a multi-stage hierarchical Bayesian framework that synthesizes information from both patient and model system databases across multiomic levels to improve the identification of novel cancer driver genes and association with drug responses.

A Bayesian evidence synthesis procedure Our integrative framework is called BaySyn: a multi-stage hierarchical Bayesian evidence synthesis pipeline for analysis of multi-system multiomic data. The first stage identifies cancer driver genes by detecting transcriptomic associations with upstream changes, which are then utilized to inform biomarker association models in the second stage to improve selection. Specifically, the first stage uses additive Gaussian process regression models to detect potential nonlinear associations of gene expression data with corresponding copy number and methylation profiles for both cell line cancer lineages and patient cancer types. To tackle the issue of lower sample size in cell line data, we propose multi-lineage versions of these mechanistic models that can deconvolve lineage and upstream main effects as well as any potential interactions, in addition to single-lineage versions of the same. Evidence synthesized across a common pool of genes from the two sources is then used in a calibrated Bayesian variable selection procedure in the second stage to identify genes having high association with an outcome variable of interest, such as drug response data. Specifically, the evidence quantifications from the mechanistic models are used in these outcome models to upweight the prior probability of selection of different biomarkers in a spike-and-slab prior setting. A conceptual schematic of the procedure is presented in Figure 1, providing a high-level summary of the multi-model system evidence synthesis through the mechanistic models and calibrated biomarker selection via the outcome models. We apply our framework to multiomic CCLE and TCGA datasets from pan-gynecological cancers (breast, ovary, and uterus lineages). Our mechanistic models provide cancer-specific and cross-lineage evidence that implicate

several relevant functional genes such as PTPN6 and ERBB2 in the KEGG adherens junction gene set. Furthermore, our outcome model is able to make higher number of discoveries in drug response models than its uncalibrated counterparts under the same thresholds of type I error control, including detection of known lineage-specific biomarker associations such as BCL11A in breast and FGFRL1 in ovarian cancers.

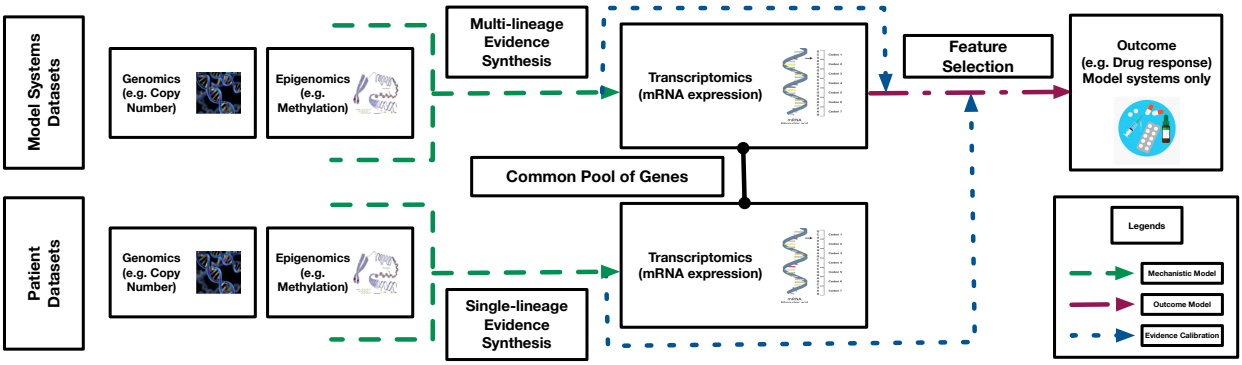

Fig. 1: Conceptual schema of the *BaySyn* framework.

The rest of the paper is organized as follows. Section 2 summarizes the multi-stage data integration framework. Section 3 describes the CCLE and TCGA data processing and analysis procedures, along with summarization of interesting results. We finish with a brief discussion of our proposed procedure and findings in Section 4. All the processed datasets, R codes for the pipeline, and the complete set of real data results are available for access via an interactive R Shiny dashboard at tinyurl.com/BaySynApp. The supplementary materials are available online at tinyurl.com/BaySynSup.

2. Methods

Multi-stage integration pipeline Following Figure 1, for a given set of samples (patients/model systems), we build gene-specific mechanistic models to infer functional relevance of the genes in the samples of interest based on the association of the gene's expression pattern with its upstream covariates such as copy number changes or DNA methylation. Particularly, in case of model systems, certain cancer lineages may contain a low number of samples and the mechanistic models may suffer from a lack of sufficient statistical power to identify true associations with upstream factors. Therefore, we build two versions of the mechanistic models depending on the sample size scenarios - a multi-lineage model that can borrow strength across samples from different lineages (used in this work for modeling the cell line samples; Section 2.1.1), and a single-lineage version that can be applied to a set of samples from a single cancer lineage/type (used in this work in context of the patient samples; Section 2.1.2). Based on statistical summaries of significance of the upstream factors for each gene from these models, we then build the outcome-specific Bayesian hierarchical variable selection models (outcome models, in short; Section 2.2) that can incorporate such prior information and borrow strength to improve selection of genes. The pseudocode for the complete framework is available at Supplementary Notes Section S1.1. The specifics of each type of model are described in full detail in the rest of this section.

2.1. *Mechanistic Models*

For the mechanistic models, we investigate a gene of interest specifically in relation with its upstream factors to detect whether it is a functional driver, and repeat the procedure across the complete pool of genes included in the analyses. This approach offers a highly parallelizable framework, and the efficiency only depends on the computational resources used by each individual model. Further, the class of genomic associations with upstream factors that we are interested in may be highly nonlinear, as has been indicated in past cancer literature.[17,18] Therefore, we intend to equip our models with sufficiently flexible specifications that can identify a broad range of association patterns. Keeping these useful features in mind, we describe the mathematical details of the multi- and single-lineage mechanistic models below.

2.1.1. *Multi-lineage Mechanistic Models*

Notations We begin with setting up some notations. Let M denote the number of lineages across which we intend to borrow strength in a single mechanistic model, and let $\{n_1, \ldots, n_M\}$ denote the lineage-specific sample sizes, with $n = \sum_{c=1}^{M} n_c$ being the total sample size. Across a total of $j \in \{1, \ldots, q\}$ genes, let G_{ij} denote the (continuous) normalized expression data for the j^{th} gene in the i^{th} sample. Let L_i denote the lineage (tissue/cancer type) of the i^{th} sample, and let $\mathbf{U}_{ij} = (U_{ij1}, \ldots, U_{ijp_j})^T$ denote the $p_j \times 1$ vector of upstream information from sample i matched to gene j. Our mechanistic models are gene-specific, allowing different sample sizes for each gene. However, for simplicity of notations, we describe the models assuming a fixed n.

Model structure For the j^{th} gene, we build an additive multi-lineage mechanistic model containing separable components for the main effects of lineage and each upstream covariate, along with any possible interactions of lineage with the upstream factors. Assuming the G_{ij}s to be mean-centered, the general mathematical form of such a model is presented in the following equation.

$$G_{ij} = \underbrace{f_{1j}(L_i)}_{\text{Lineage main effect}} + \underbrace{\sum_{v=1}^{p_j} f_{2jv}(U_{ijv})}_{\text{Upstream main effects}} + \underbrace{\sum_{v=1}^{p_j} f_{3jv}(L_i, U_{ijv})}_{\text{Interaction effects}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}, \forall i \in \{1, \ldots n\}. \quad (1)$$

The simplest choice is to specify each component $f_{\bullet}$ as a linear model. Such models have been explored in context of cancer omics.[19] Although they are computationally simple, they may not be fully able to capture the general range of cellular association patterns. An obvious nonlinear extension is to use splines to construct piece-wise linear mean profiles. Such approaches have also been explored in this context.[20] However, there are multifold challenges – including specifying the number of knots (hence the degree of adaptable nonlinearity) and increasing computational intensity with increasing number of covariates. To build a general class of additive association models while maintaining a reasonable extent of computational efficiency, we use Gaussian process (GP) models.

To build an additive GP model with interaction effects, we adapt an existing approach proposed in context of longitudinal data.[21] In a repeated measures setting, this approach provides a way to incorporate sample-level baseline effects and treatment effects in a nonlinear fashion. We extend this idea to our scenario to include lineage-level baseline effects (treating the experiments on cell lines from the same lineage akin to a repeated experiment setting) and changes in the effects of upstream covariates across different lineages. While samples belonging to cancers sharing some larger group-specific commonalities (e.g. all gynecological cancers) may share patterns of mechanistic impacts

of upstream platforms on gene expressions, there may still be cancer-specific differences in the exact effects. Briefly, we use a GP equipped with a zero-sum (zs) kernel for the main effect of the categorical lineage variable, one with an exponentiated quadratic (eq) kernel for the main effects of the continuous upstream variables, and a product of the zs and eq kernels for their interactions, following existing approaches.[21,22] The specifics of the GP model along with the prior choices are described in detail in Supplementary Notes Section S1.2.

Model fitting and hypothesis testing The interest now is in building mechanistic models and testing for different main and interaction effects of interest. We use a dynamic Hamiltonian Monte Carlo (HMC) sampler to obtain draws from the posterior distributions of the parameters. Since we are interested in evaluating the roles of lineage, upstream factors, and any possible interactions in explaining the variability in gene expressions, we are interested in testing the following hypotheses.

(1) **Lineage main effect:** $H_{0Lj} : f_{1j} = \text{constant}$.
(2) **Upstream main effects:** $H_{0Uj} : f_{2jv} = \text{constant}, \forall v \in \{1, \ldots, p_j\}$.
(3) **All upstream effects:** $H_{0UIj} : f_{2jv}, f_{3jv} = \text{constant}, \forall v \in \{1, \ldots, p_j\}$.

To perform these tests, we use model comparison procedures using HMC-based draws of the joint log-posterior function of the parameters in a model. For a model M containing all or some of the components in Equation (1), let $H_{0\bullet}$ be the test of interest and $\text{M}_\bullet$ be the null model, which is a submodel of M not containing the components set to constant under $H_{0\bullet}$. For example, if we are interested in testing the lineage main effect in a main effects-only model M, $\text{M}_\bullet$ would be an upstream-only model. We define *pseudo-Bayes factors* ($\text{pBF}_{\bullet j}$s) as scalar summaries of component significance, defined to be the mean difference of the log-posteriors evaluated across the MCMC draws between the two models being compared. The pBFs for the three hypotheses above and for the j^{th} gene are denoted respectively by pBF_{Lj}, pBF_{Uj}, and pBF_{UIj}. Note that these quantities are approximations for the traditional log-Bayes factors (lBFs) for comparing Bayesian models under equal model priors. To compute an lBF, one has to compute the expected posteriors for each model, followed by their log-ratio. Here, we are computing an empirical average of the difference of log-posteriors of the model parameters. The exact expressions of these quantities for a given HMC sample of the parameters are derived in Supplementary Notes Section S1.3. We use standard cut-offs for significance used for lBFs at a $\log_{10}(\bullet)$-scale: < 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), and > 2 (decisive).[23] From now on, by pBF we always mean a quantity already in this scale.

Sequential evidence detection To identify driver genes, we quantify evidence of any upstream effect on gene expression untangled from any possible lineage effect. To this end, mimicking classical approaches in regression settings, we follow a sequential scheme as described in Supplementary Figure S1.

(1) Test for any lineage main effect using pBF_{Lj}. If $\text{pBF}_{Lj} \leq 1$, go to Step 2. Else go to Step 3.
(2) Test H_{0Uj} using pBF_{Uj}. Set mechanistic evidence $\mathcal{E}_{j1} = \text{pBF}_{Uj}$.
(3) Test H_{0UIj} using pBF_{UIj}. Set mechanistic evidence $\mathcal{E}_{j1} = \text{pBF}_{UIj}$.

2.1.2. *Single-lineage Mechanistic Models*

These models do not include any lineage main or interaction effects. Thus, from Equation (1), the full models reduce to the following for the j^{th} gene, using same notations as before.

$$G_{ij} = \underbrace{\sum_{v=1}^{p_j} f_{jv}(U_{ijv})}_{\text{Upstream main effects}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}, \forall i \in \{1, \dots n\}. \tag{2}$$

We use the same eq kernel parametrization for the GP priors on each $f_{\bullet}$ as we used for the $f_{2\bullet}$ components in the multi-lineage models. We now test $H_{0j} : f_{jv} = \text{constant}, \forall v \in \{1, \dots, p_j\}$ for each gene. We compare the full model in Equation (2) with a noise-only null model. The derivation of the corresponding pBF_j is described in Supplementary Notes Section S1.4. We assign the evidence $\mathcal{E}_{j2} = \text{pBF}_j$, as described in Supplementary Figure S1.

2.2. *Outcome Model*

For a given pool of genes, it is possible to compute multiple lines of evidence ($\boldsymbol{\mathcal{E}}_j = (\mathcal{E}_{j1}, \dots, \mathcal{E}_{jE})^T$ for gene j). For example, for a given gene j, we may compute one pBF from a multi-lineage model built on cell line samples, and another pBF from a single-lineage model built on patient samples ($E = 2$). With interest in some disease- or therapy-related phenotype/outcome Y and the selection of biomarkers associated with it, the goal is to inform the outcome model about any level of evidence captured in these $\mathcal{E}_{je}$s in a covariate-specific way to possibly improve selection.

(1) Sufficiently strong evidence in favor of a covariate $\implies$ higher prior probability of inclusion.

(2) Otherwise, a uniform prior is placed on selection/non-selection for that particular covariate.

We utilize a hierarchical Bayesian setting with calibrated spike-and-slab priors, described below. Let Y_i be the mean-centered continuous outcome for the i^{th} sample. Simple extensions to categorical/censored outcomes are possible, but in this work we only focus on continuous outcomes. The mathematical form of the calibrated Bayesian variable selection (cBVS) model is then the following.

$$Y_i = \sum_{j=1}^{q} \underbrace{\beta_j}_{\text{Gene expression coefficients}} G_{ij} + \underbrace{\eta_i}_{\text{Error}}, i \in \{1, \dots, n\}. \tag{3}$$

Model and prior specifications The errors η_i are iid $\text{N}(0, \tau^2), \forall i \in \{1, ..., n\}$. A standard conjugate prior is used for $\tau^2 \sim \text{Inverse-Gamma}(\frac{\nu}{2}, \frac{\nu\lambda}{2})$. Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$ denote the q-dimensional vector of regression coefficients. We place a calibrated hierarchical spike-and-slab prior on $\boldsymbol{\beta}$.

$$\begin{aligned} \boldsymbol{\beta}|\boldsymbol{\delta}, \tau &\sim \mathbf{N}_q(\mathbf{0}, \mathbf{D}_{\boldsymbol{\delta},\tau}), \\ \delta_j|\theta_j &\sim \text{Bernoulli}(\theta_j), \quad \forall j \in \{1, ..., q\}, \\ \theta_j &\sim \text{Beta}\Big(\mathcal{F}(\boldsymbol{\mathcal{E}}_j), \frac{1}{\mathcal{F}(\boldsymbol{\mathcal{E}}_j)}\Big), \quad \forall j \in \{1, ..., q\}. \end{aligned} \tag{4}$$

Here $\mathbf{D}_{\boldsymbol{\delta},\tau} = \tau^2 \mathbf{A}_{\boldsymbol{\delta}}$, where $\mathbf{A}_{\boldsymbol{\delta}}$ is the $q \times q$ diagonal matrix $\mathbf{A}_{\boldsymbol{\delta}} = \text{diag}\{\delta_1 v_1 + (1-\delta_1)v_0, \dots, \delta_q v_1 + (1-\delta_q)v_0\}$ and $v_1 \geq v_0 > 0$ are respectively the slab and spike variances. The binary latent variables δ_j are variable inclusion indicators with $\delta_j = 1$ meaning that the j^{th} variable is included in the model. $\mathcal{F}$ is a calibration function mapping the evidence vector $\boldsymbol{\mathcal{E}}_j = (\mathcal{E}_{j1}, \dots, \mathcal{E}_{jE})^T$ to the prior covariate

inclusion probability θ_j. The advantages of the hierarchical formulation (Equation (4)) coupled with the evidence calibration function $\mathcal{F}$ are multifold. First, by adapting $\mathcal{F}$, our framework allows the user to incorporate other significance quantities (such as p-values) into the final outcome model. Any external upstream information, including categorical and continuous covariates, can be used in the mechanistic layer to compute such summary statistics. Finally, by tuning $\mathcal{F}$ appropriately, our framework allows the user to control the impact of the prior information on selection, as we show below. We discuss all these in more detail in Section 4.

Choice of evidence calibration function We use a calibration function $\mathcal{F}$ on $\mathbb{R}^E \to [0,1]$ to aggregate multi-dimensional prior evidence into a scalar prior probability. To this end, we use a four-parameter logistic map reflecting the maximal evidence across all sources on a continuous and non-decreasing spectrum of evidence strength. The exact mathematical form and the motivation behind this choice are described in Supplementary Notes Section S1.5. Using this function, the calibrated prior means of θ_j (representative values of maximal evidence at the pBF/ ln(10) scale in parentheses) are as follows: 0.502 (0.25), 0.543 (0.75), 0.726 (1.5), 0.962 (3). As illustrated in Supplementary Figure S2, the corresponding prior distributions of θ_j shift from an uniform prior to one concentrated close to one with increase in prior evidence strength.

Variable selection Inference is centered around the posterior $\wp(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\theta}, \tau | \boldsymbol{Y}, \boldsymbol{G}, \boldsymbol{\mathcal{E}}, \nu, \lambda, \upsilon_0, \upsilon_1)$, where $\boldsymbol{\beta}, \boldsymbol{\delta}$, and $\boldsymbol{\theta}$ are the $q \times 1$ vectors of all β_js, δ_js, and θ_js respectively, $\boldsymbol{Y}_{n\times 1}$ is the outcome vector, $\boldsymbol{G}_{n\times q}$ is the design matrix, and $\boldsymbol{\mathcal{E}}_{q\times E}$ is the matrix of the $\mathcal{E}_{je}$s. We approximate this using a Gibbs sampler implemented via the *rjags R package*.[24] We obtain posterior estimates of the parameters (i.e., $\hat{\beta}_j$s, $\hat{\theta}_j$s, and $\hat{\tau}$) as their corresponding empirical posterior means. Model selection is performed using the collection of $1 - \hat{\theta}_j$ as p-value type quantities and applying a false discovery rate (FDR) control procedure,[25] described in Supplementary Notes Section S1.6.

3. Multi-system and Multi-platform Integrative Analyses of Pan-Gynecological Cancers

We perform an integrative analysis of cancer cell lines data from CCLE and patient samples from TCGA.[14,15] Using multi-lineage mechanistic models for cell line samples and single-lineage mechanistic models for patient samples, we quantify gene-specific associations of expression with corresponding copy number and methylation data. We then use the pBFs from these two sources to inform and build cBVS models of drug response on gene expression based on the cell line samples. Specifically, our multi-lineage mechanistic models on the cell line samples borrow strength by combining data across three gynecological lineages - breast, ovary, and uterus. The single-lineage mechanistic models on the patient samples are built separately for each of the three corresponding TCGA cancer types by tissue - breast invasive carcinoma (BRCA), ovarian serous cystadenocarcinoma (OV), and uterine carcinosarcoma (UCS). The outcome models on the cell line samples are built in a lineage-specific way for a collection of drugs of interest in gynecological cancers. Our investigations are aimed broadly at answering two sets of questions.

(1) We assess within-system and between-system patterns of functional evidence garnered by the mechanistic models (i.e., a gene may have strong mechanistic evidence of association with the upstream factors for the cell lines only, the patients only, both, or none).
(2) We identify panels of genes whose expressions are associated with responses to specific drugs in the cell line samples, potentially offering novel introspection into treatment selection and the cellular mechanisms/targets of such drugs.

3.1. *Data Processing and Analysis Pipeline*

Multi-omics cell line and patient data Gene expression, copy number, and DNA methylation data on cancer cell lines from CCLE, drug response data from GDSC, along with annotation information to match genes to upstream information, are downloaded from the depmap portal.[26] Gene expression, copy number, and DNA methylation data on TCGA patient samples, along with annotation information matching genes to upstream covariates, are downloaded from the Xena browser.[27] Sample size and other filtering requirements result in a pool of 5,792 genes and 65 drugs to be included in all further analyses, as described in Supplementary Notes Section S1.7. Summary information on each dataset are available in Supplementary Table S1 and Supplementary Figures S3-S8.

BaySyn analysis of gynecological cancers For each gene, a multi-lineage mechanistic model with $M = 3$ (breast, ovary, uterus) is constructed (termed the CL model hereafter) and hypothesis tests are performed as described in Supplementary Figure S1. Further, for each gene, three single-lineage mechanistic models (one for each cancer type – BRCA, OV, UCS) are built on the patient samples and upstream effects are quantified following Supplementary Figure S1. As a post-model fitting investigation, we perform gene set enrichment analyses (GSEA)[28] using these four sets of evidence (CL, BRCA, UCS, OV) for the Kyoto Encyclopedia of Genes and Genomes (KEGG)[29] and gene ontology (GO) gene sets.[30,31] For our analyses, we use the gene set enrichment (GAGE) procedure implemented in the *gage R package* due to the reason that our pBFs are on a different scale than typical expression levels or fold-change summaries.[32] The gene set-specific hypothesis that we test is whether the set in question exhibits significantly higher level of activity as summarized by the evidence statistics compared to the genes outside the gene set, due to the unidirectional nature of the pBFs. For each lineage, drug-specific response association models are built using the cBVS procedure, and variable selection is performed using a 10% FDR control threshold. Illustrative examples of annotated and integrated datasets for each stage of modeling are presented in Supplementary Notes Section S1.8 and Supplementary Figures S9-S11.

3.2. *Results*

Utility of borrowing strength to detect mechanistic evidence Figure 2a summarizes the number of genes inferred to be at the decisive level of evidence (in favor of associations with corresponding upstream covariates) across the three single-lineage models for each TCGA patient cancer type and the multi-lineage model for the cell lines data. The connected dots at the bottom indicate the intersection of the mechanistic models for which the number of genes summarized by the bar height are decisive. The top three combinations of models in terms of detecting decisive evidence all belong to some combination of the TCGA data sets (BRCA only, BRCA and OV, BRCA and UCS - in decreasing order). However, except for the BRCA dataset which utilizes > 750 samples for all genes to build the mechanistic models, the cell lines mechanistic models borrowing strength across three lineages detect more unique signals (4^{th} in the ranking) than the other TCGA datasets. This further validates the utility of building joint nonlinear association models with main and interaction components that can identify shared patterns of association across smaller datasets which would potentially be missed in dataset-specific models. The list of genes uniquely identified by the cell lines mechanistic model is available in Supplementary Table S2.

KEGG gene set enrichment analyses illustrate utility of mechanistic evidences To assess the utility of the mechanistic evidence quantities and to validate their use in future detection of novel

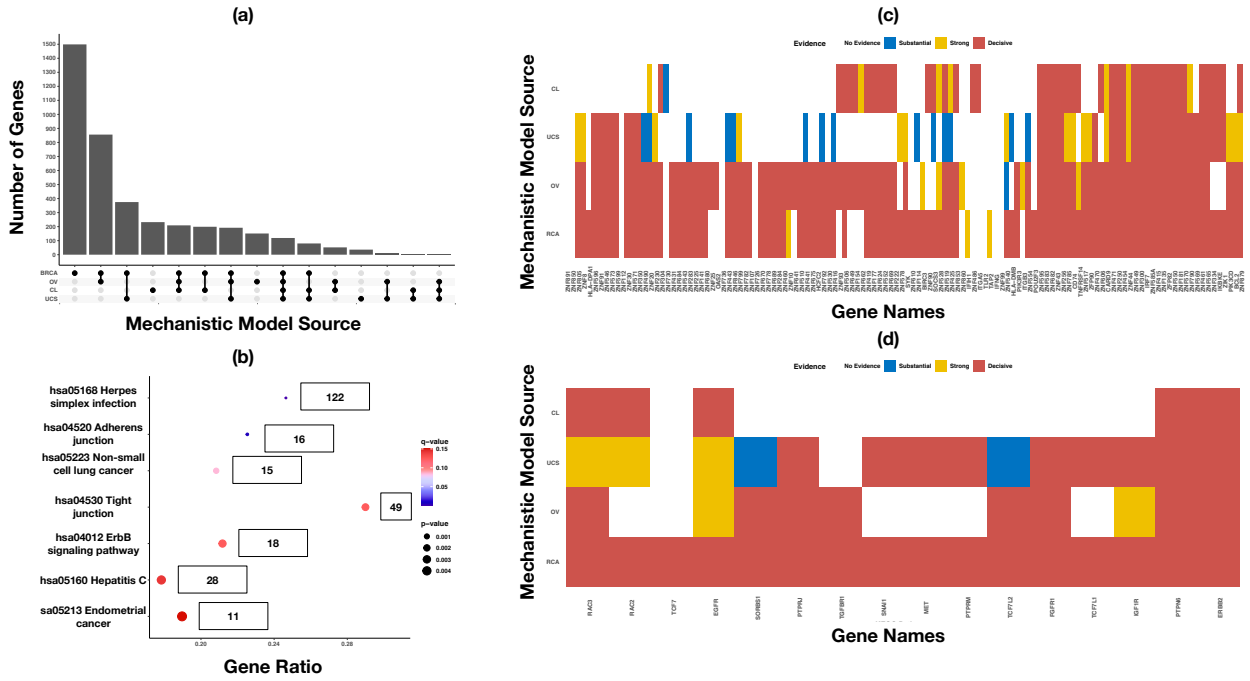

Fig. 2: Mechanistic evidence summary and gene set enrichment results. Panel (a) presents an upset plot of the number of genes at the decisive level of evidence based on the mechanistic models for different intersections of the patient and cell line datasets. Panel (b) presents a dotplot summarizing significance levels for KEGG gene sets. The gene sets are ordered from top to bottom in decreasing order of q-values ($\leq$ 0.2 included). The labels beside the dots indicate set sizes in our analyses. Panels (c) and (d) present heatmaps summarizing levels of mechanistic evidence for the genes in KEGG herpes simplex infection and adherens junction gene sets respectively. Genes in the rows are ordered based on clusters resulting from the evidence statistics.

functional drivers, we perform GSEA using the four evidence sources and the KEGG and GO gene sets. Due to space limitations we only discuss the KEGG results here. The GO results are presented in Supplementary Figures S17-S32. Several KEGG gene sets have been implicated to have significant roles generally in cancer[33,34] and specifically in gynecological cancers.[35–38] The results from our KEGG GSEA are summarized in Figure 2b, exhibiting the seven gene sets with FDR-controlled q-value < 0.2. The gene set-specific mechanistic evidences are summarized in Figure 2c-d for the top two KEGG gene sets; the rest are presented in Supplementary Figures S12-S16. The top gene set identified in the KEGG analyses is the herpes simplex infection pathway (p-value $= 3.88 \times 10^{-16}$) (Figure 2b). This gene set contains a large cluster of genes exhibiting decisive evidence across majority of the mechanistic models, as can be seen in Figure 2c. Following these genes are two major clusters - one containing genes at the decisive level for the BRCA, OV, and CL mechanistic models, and one containing genes at the decisive level for all three TCGA cancers. The consistent nature of functional evidence across this gene set is in agreement with findings from past investigations - multiple studies have indicated the prognostic value of members of this pathway in gynecological cancers - including breast,[39] ovarian,[40] and endometrial[41] cancer. The second-highest gene set in the KEGG analyses is the adherens junction gene set (p-value $= 5.52 \times 10^{-5}$) (Figure 2b). The genes PTPN6 and ERBB2 exhibit decisive levels of mechanistic evidence in all four models (Figure 2d).

Different upstream mechanisms of the ERBB2 gene have been implicated in different gynecological cancers, such as copy number changes in ovarian tumors[42] and somatic mutations in breast cancer.[43] The EGFR gene has also shown promise as a potential therapeutic target in multiple gynecological cancers,[44,45] which is in alignment with our findings of some signal in all the TCGA and cell line models (Figure 2d).

Calibrated drug response models identify high-association lineage-specific biomarkers We build calibrated hierarchical Bayesian variable selection-based drug response models for each lineage × drug combination across all 65 drugs and all three cell line lineages. Figure 3a presents a wordcloud where each gene is weighted by the total number of times it is selected in a drug response model at the 10% FDR-controlled cutoff. The genes BAHCC1, ALOX12P2, and SYCP2 emerge as the top candidates, with selection in 14, 12, and 12 models respectively. While this summary allows us to identify general candidates for future pharmacogenomic investigations, it does not indicate any potential lineage-specific utility of these genes. To this end, Figure 3b summarizes the number of times the top genes across all drug response models are selected in each lineage. For breast, genes BAHCC1, BCL11A, and SYCP2 are at the top, with respectively eight, eight, and six detected drug associations. The role of BCL11A in triple-negative breast cancer (TNBC) stemness is well known, and it is considered to be one of the first utilizable targets for treatment of TNBCs.[46] A similar confirmation can be obtained for SYCP2, which has recently been identified as a prognostic biomarker in breast cancer.[47] However, to the best of our knowledge, BAHCC1 has not so far been identified to have breast cancer-specific functional roles, which renders it as a novel detection that deserves deeper investigations. Top genes in the two other lineages also include both novel and known functional drivers - such as ALOX12P2 (nine selections, novel) and FGFRL1 (eight selections, known)[48] for ovary and FBXO17 (seven selections, novel) for uterus.

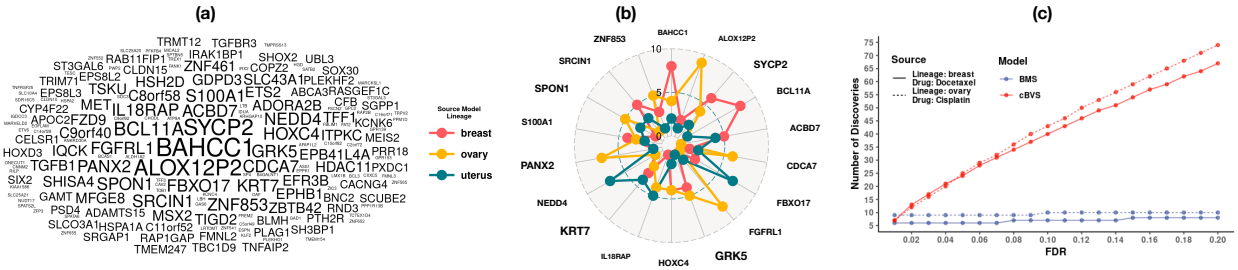

Fig. 3: Drug response model summaries. Panel (a) presents a wordcloud of top genes across all the drug response models (three lineages × 65 drugs). The sizes of the words are proportional to the total number of times across all models that a gene is selected based on a 10% FDR-controlled threshold. Panel (b) presents a radar chart of the top 18 genes (selected in at least nine drug response models) according to the three lineages. Panel (c) presents a discovery plot across increasing FDR control thresholds for the drug docetaxel in lineage breast and the drug cisplatin in lineage ovary. BMS refers to an uncalibrated Bayesian variable selection model based on the Bayesian model averaging procedure (see Supplementary Notes Section S1.9).

Calibration improves statistical power to detect gene-drug associations To assess the discoveries for specific lineage × drug combinations, we focus on two drugs with known use in specific cancer lineages - docetaxel for breast and cisplatin for ovary. The number of discoveries across different FDR

thresholds for these are presented in Figure 3c-d and the corresponding discoveries are summarized in Supplementary Tables S3-S4. Similar plots and tables for all other models are available in our R Shiny dashboard at tinyurl.com/BaySynApp. Evidently, compared to an uncalibrated Bayesian variable selection procedure implemented via the BMS R package (see Supplementary Notes Section S1.9), cBVS models make more discoveries at the same level of error control, allowing a continuum of assessment for top candidates emerging across increasing control thresholds. This indicates the utility of synthesizing mechanistic evidence and calibrating the outcome models with such evidences. Several examples of cell lines-based discoveries guided by evidences discovered in patient data emerge. For example, the model for docetaxel response in breast cell lines identify an association with the gene GRK5 at 10% FDR control. Cell lines overexpressing GRK5 have previously been observed to demonstrate an increase in resistance to docetaxel in male gynecological cancers,[49] and our finding suggests that it deserves further investigations in female gynecological cancers as well. Another top discovery at the same FDR threshold is the gene CD83, expression of which is known to be enhanced by docetaxel in metastatic breast cancers.[50] For the response model of cisplatin in the ovarian lineage, multiple solute-carrier family (SLC) genes are selected at the 10% FDR threshold. These genes are known potential biomarkers of ovarian cancer and are under investigation for prognostic utility.[51] Another interesting discovery is that of the CDCA7 gene from the cell division cycle pathway, silencing of which has recently been shown to downregulate cisplatin resistance in lung cancer subtypes, making it a potential therapeutic target.[52] Our finding seems to indicate similar scope in ovarian cancer, demanding further investigation. Notably, all four of these discussed findings had no cell lines-based mechanistic evidence, but had decisive evidence from at least one TCGA source – which further underscores the importance of synthesizing evidence across model systems.

4. Summary and Discussion

We propose BaySyn, a hierarchical multi-stage Bayesian evidence synthesis procedure for multi-system multiomic integration. BaySyn detects functionally relevant driver genes based on their associations with upstream regulators and uses this information to guide variable selection in outcome association models. We apply our framework to multiomic cancer cell line and patient datasets for pan-gynecological cancers. pBFs from the mechanistic layer of BaySyn exhibit high enrichment in previously known KEGG gene sets and detect driver genes known to have functional roles in the cancers studied. Calibrated outcome models for drug responses identify several confirmatory and novel lineage-drug-gene combinations providing further evidence on the profitability of our approach towards future precision oncology endeavors.

Several features of our framework makes it readily adaptable to more general settings and richer datasets. The calibrated spike-and-slab prior can be generalized to include any number (more upstream platforms such as miRNA or mutation) and form (other evidence metrics such as p-values) of prior information by tuning the calibration function accordingly. The outcome model can easily be extended to include other biomarkers such as proteomics. While we use cell lines data to illustrate the integrative approach across model systems, it is straightforward to apply our pipeline to datasets from cancer model systems with higher fidelity to human tumors[53] - such as organoids[54] or patient-derived xenografts[55] - as such databases become increasingly comprehensive and available. Further, both the stages of our framework are highly parallelizable and individual runs are quite efficient - a single gene-specific multi-lineage mechanistic model with interactions takes approximately 20 minutes on

average to complete, while a single lineage-drug specific outcome model takes approximately 12 minutes on average (both based on runs on a single core of a 2015 Macbook Air with 8 GB memory and Intel i5 processor). Thus, extending our analyses to include larger gene-drug panels with similar sample sizes is straightforward with existing parallel computing resources.

Limitations and Future Work Certain improvements are of interest given the biological context of our work. First, although we assess mechanistic relevance at a gene-by-gene basis, at a molecular level, genes interact in functional pathways to result in downstream modifications. This motivates joint models for driver genes in a multivariable setting accounting for underlying gene-gene interactions. Second, the relatively low lineage-specific sample sizes in cell lines data make fully Bayesian exploration of the posteriors feasible in the outcome models. Higher data dimensions would result in increased computation times; where-in approximate Bayesian computation schemes such as the E-M based variable selection[56] or variational Bayes[57] would need to be employed. Third, while our framework allows integration of covariate-specific prior information in a variable selection framework, more granular information (both sample- and covariate-specific) may be available, allowing improved learning of the molecular functions driving the changes in an outcome of interest. For example, sample-specific data on tumor heterogeneity may be available, and such data may need to be incorporated in the outcome models driving changes in the covariate effects. Finally, as outlined in Supplementary Notes Section S1.5, in the presence of multiple lines of evidence, how best to aggregate them depends heavily on the context - while multiple possible approaches exist, a case-specific decision must be made to ensure best utilization of the evidences. A data-driven procedure of choosing evidence weights would eliminate this requirement. We leave these tasks for future exploration.

Acknowledgments RB and VB were partially supported by NIH grant R01CA244845-01A1 and VB by P30 CA-046592.

References

1. I. Subramanian et al. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.
2. A. Conesa and S. Beck. *Scientific data*, 6(1):1–4, 2019.
3. B. A. Ruggeri et al. *Biochemical pharmacology*, 87(1):150–161, 2014.
4. J. Kim et al. *Nature Reviews Molecular Cell Biology*, 21(10):571–584, 2020.
5. M. V. Relling and W. E. Evans. *Nature*, 526(7573):343–350, 2015.
6. D. M. Roden et al. *Annals of internal medicine*, 145(10):749–757, 2006.
7. S. Tarazona et al. *Nature Computational Science*, 1(6):395–402, 2021.
8. S. Chakraborty et al. *BioMed research international*, 2018, 2018.
9. A. Kaplan and E. F. Lock. *Cancer informatics*, 16:1176935117718517, 2017.
10. C. Cheng et al. *Integrating Omics Data*, pp. 380, 2015.
11. P. W. Angel et al. *PLoS computational biology*, 16(9):e1008219, 2020.
12. Z. Tu et al. *Integrating Omics Data*, 88:88–109, 2015.
13. G. Tseng et al. *Integrating omics data*. Cambridge University Press, 2015.
14. J. N. Weinstein et al. *Nature genetics*, 45(10):1113–1120, 2013.
15. J. Barretina et al. *Nature*, 483(7391):603–607, 2012.
16. W. Yang et al. *Nucleic acids research*, 41(D1):D955–D961, 2012.
17. H. K. Solvang et al. *BMC bioinformatics*, 12(1):1–12, 2011.
18. K. Litovkin et al. *Journal of cancer research and clinical oncology*, 140(11):1849–1861, 2014.
19. W. Wang et al. *Bioinformatics*, 29(2):149–159, 2013.
20. E. J. McGuffey. *Statistical methods for integrating genomics data*. Texas A&M University, 2015.
21. J. Timonen et al. *Bioinformatics*, 37(13):1860–1867, 2021.
22. C. G. Kaufman and S. R. Sain. *Bayesian Analysis*, 5(1):123–149, 2010.
23. R. E. Kass and A. E. Raftery. *Journal of the american statistical association*, 90(430):773–795, 1995.
24. M. Plummer et al. *Vienna, Austria*, 2016.
25. V. Baladandayuthapani et al. *Journal of the american statistical association*, 105(492):1358–1375, 2010.
26. A. Tsherniak et al. *Cell*, 170(3):564–576, 2017.
27. M. J. Goldman et al. *Nature biotechnology*, 38(6):675–678, 2020.
28. A. Subramanian et al. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.
29. M. Kanehisa and S. Goto. *Nucleic Acids Research*, 28(1):27–30, Jan 2000.
30. M. Ashburner et al. *Nature genetics*, 25(1):25–29, 2000.
31. *Nucleic acids research*, 49(D1):D325–D334, 2021.
32. W. Luo et al. *BMC bioinformatics*, 10(1):1–17, 2009.
33. L. Chen et al. *Artificial Intelligence in Medicine*, 76:27–36, 2017.
34. F. Yuan et al. *Mathematical Biosciences*, 304:1–8, 2018.
35. A. D. Campos-Parra et al. *Gynecologic oncology*, 143(2):406–413, 2016.
36. X. Yang et al. *OncoTargets and therapy*, 11:1457, 2018.
37. T. Zhang et al. *PLoS One*, 13(5):e0196351, 2018.
38. J. Chen et al. *Medicine*, 99(18), 2020.
39. S. M. Ghouse et al. *Frontiers in oncology*, 10:384, 2020.
40. M. Nakamori et al. *Clinical cancer research*, 9(7):2727–2733, 2003.
41. X.-Y. Zhou et al. *Journal of Clinical Laboratory Analysis*, 36(4):e24315, 2022.
42. I. Dimova et al. *International Journal of Gynecologic Cancer*, 16(1), 2006.
43. J. Y. Hou et al. *Gynecologic Oncology Reports*, 32, 2020.
44. H. D. Reyes et al. *Molecular diagnosis & therapy*, 18(2):137–151, 2014.
45. K. K. Kim et al. *Scientific reports*, 5(1):1–11, 2015.
46. A. Errico. *Nature Reviews Clinical Oncology*, 12(3):127–127, 2015.
47. C. Wu and Y. Tuo. *Future Oncology*, 15(8):817–826, 2019.
48. H. Tai et al. *Journal of immunology research*, 2018, 2018.
49. J. B. Black et al. *Cancer Research*, 78(13_Supplement):LB–312, 2018.
50. M. Buoncervello et al. *Neoplasia*, 14(9):855–IN19, 2012.
51. H. Chen et al. *Annals of Translational Medicine*, 9(15), 2021.
52. W. Zeng et al. 2021.
53. A. Goodspeed et al. *Molecular Cancer Research*, 14(1):3–13, 2016.
54. J. Drost and H. Clevers. *Nature Reviews Cancer*, 18(7):407–418, 2018.
55. F. Invrea et al. *Current opinion in biotechnology*, 63:151–156, 2020.
56. V. Ročková and E. I. George. *Journal of the American Statistical Association*, 109(506):828–846, 2014.
57. C. W. Fox and S. J. Roberts. *Artificial intelligence review*, 38(2):85–95, 2012.