Diagram

Description automatically generated

**Figure S1**. Detailed analysis pipeline of BioVLAB-Cancer-Pharmacogenomics. Black lines denote the data flow through the pipeline and grey boxes denote the corresponding file formats. The input multi-omics data for a tumor consists of somatic copy number alterations (SCNAs), single nucleotide variations (SNVs) in variant call format (VCF) file, gene expression quantification of tumor and matched normal samples, and DNA methylation profile. The server-side area is colored in grey, and shows the pipeline of data being processed. ABSOULTE takes SCNAs and SNVs in mutation annotation format (MAF) created from VCF file as input, and outputs tumor purity and cell ploidy. PyClone takes the multiplicities of SNVs and tumor purity created by ABSOLUTE as input, and outputs the information about subclones. nJSD takes gene expression of both the tumor and its matched normal as input and outputs transcriptomic intratumor heterogeneity (ITH). EDec takes gene expression and DNA methylation as input, and outputs cell type-specific gene expressions and their proportions. Based on the output of four component tools, we prepare a report for ITH and pharmacogenomics (PG) reports as final output of the system. The ITH reports comprise of matched TCGA clinical notes and ITH measured in genomic/transcriptomic/copy number level. The PG reports comprise of tumor cell type decomposition, matched CCLE cell line, and its drug response. TSV, Tab-separated values.

Diagram

Description automatically generated with medium confidence

**Figure S2**. Epigenomic deconvolution to match tumor to cell line. Tumors consist of heterogeneous cell types, e.g., cancer epithelial, normal epithelial, stromal, and immune cells. Epigenomic Deconvolution (EDec) decomposes tumors into cell types in a matrix factorization fashion. **A1** DNA methylation profile of samples is given. **A2** Cell type-specific methylation pattern is retrieved from external source. Epithelial cells have distinct pattern of methylation biomarkers between cancer and normal. **A3** Calculate cell type-wise proportions. **A4** Gene expression profile of samples is given. **A5** Derive cell type-specific gene expression values. **B** Cancer epithelial cells can have distinct population by intratumoral heterogeneity. We used multiple cancer epithelial cell types to capture cancerous basis for each sample. Reconstructing gene expression using only these cancer epithelial cells gives each tumor sample its pure cancerous gene expression profile.

Graphical user interface, application, website

Description automatically generated

**Figure S3**. User-level walkthrough of BioVLAB-Cancer-Pharmacogenomics. Although the system architecture is rather complicated, what users have to do is very simple, since interaction with users is done through a web GUI. (A) User is provided with graphical illustration and text-based manual from the website, and can click buttons to get started. (B) User needs to provide AWS credentials, email, and multi-omics data of tumor. The email is used for notifying the user once the reports are ready. (C) Upon submitting data, the analysis process takes place in background and user can close the window. User can monitor the progress of workflow before the analysis is done. (D) When analysis is finished and reports are ready, the user gets notified via email with a link to the final report webpage.

A screenshot of a video game

Description automatically generated with medium confidence

**Figure S4**. Purifying cancer gene expression profiles using epigenomic deconvolution facilitate the comparison between patient-derived tumor samples and cancer cell lines. (A) The distribution of the cosine distance between the gene expression profiles of TCGA-BRCA samples and CCLE cell lines. (B) Schematic diagram of the proposed effect of epigenomic deconvolution in tumor-cell line comparison. (C) The proportions of subtype-matching cell lines in K-nearest neighbor cell lines. Shaded area denotes the standard error. (D) PCA analysis of gene expression profiles before (left panel) and after (right panel) epigenomic deconvolution. Samples are colored according to their corresponding PAM50 subtypes. (E) Silhouette scores of PC embeddings with regards to the PAM50 subtypes. (F) The distribution of subtype-matching and subtype-mismatching TCGA-BRCA-CCLE pairs before (left panel) and after (right panel) the epigenomic deconvolution.

Chart, bar chart

Description automatically generated

**Figure S5**. Retrospective analysis for the incomplete remission rates of TCGA-BRCA patients. Patients were divided into two groups: (1) patients who had been treated with at least one of the pharmacogenomically suggested drugs and (2) patients whose treatment involved none of the suggested drugs. Patient whose response was measured as Partial Respons or Clinical Progressive Disease was classified as incomplete remission. Fisher's exact p=0.063.

Chart, line chart

Description automatically generated

**Figure S6**. Kaplan-Meier plots for overall survivals of TCGA-BRCA patients stratified by various ITH measures. TCGA-BRCA patients with clinical information were divided into two groups with equal size according to transcriptomic ITH (tITH, left panel), genomic ITH (gITH, middle panel) and multi-omics ITH (moITH, right panel). Cox proportional hazards regression p-values are shown.

**Supplementary Note S1.**

Amazon hosts publicly available datasets as open S3 buckets. With the AWS Open Data, many valuable genomic databases, such as TCGA, Therapeutically Applicable Research To Generate Effective Treatments (TARGET), CCLE, Clinical Proteomic Tumor Analysis Consortium (CPTAC), and International Cancer Genome Consortium (ICGC), can be accessed freely. BioVLAB framework allows users without expert knowledge on computing system to utilize the AWS cloud platform. The user needs to create an AWS account and provide credentials such as AWS access key and AWS secret key. Then, the system automatically performs the aforementioned analyses of the bulk cancer multi-omics data provided by the user. By copying the machine image, the complete execution environment for computational tools and databases is automatically copied to the private computing space of the user. By executing the machine image, the whole analysis pipeline is performed on the isolated computing instance of the user and the analysis result is notified to the user. All computations are done in an EC2 instance owned by the user and the results are saved to an S3 bucket owned by the user. This way, our system provides users to exploit the AWS system with data privacy fully ensured.

**Supplementary Note S2.**

Analyzing molecular data of a user in terms of very large databases, e.g., TCGA and CCLE, is a challenging task, especially for small research labs. We have been developing BioVLAB system on the cloud so that users can perform data analysis without worrying about the computer infrastructure, installing and pipelining multiple computational tools, and summarizing and navigating the analysis results. BioVLAB-Microarray is the earliest implementation of the BioVLAB framework. It helps the user analyze microarray data on AWS by providing flexible and configurable graphical user interface (GUI) workflow. The user can explore similar genes and perform analysis such as subset extraction, component analysis, clustering of genes and network analysis from the input microarray data. BioVLAB-MMIA is a web-based system for the integrative analysis of microRNA and mRNA expression. Since it is widely known that microRNA regulates the abundances of mRNA, integrated analysis of such two-omics can help discover new knowledges. Later, the system was further developed to BioVLAB-MMIA-NGS to allow the use of RNA-Seq data for the quantification of mRNA expression. A recent implementation of BioVLAB framework is BioVLAB-mCpG-SNP-EXPRESS. BioVLAB-mCpG-SNP-EXPRESS integrates DNA methylation, genomic sequence, and gene expression profiles to characterize the mechanisms of gene expression regulation through both genomic and epigenomic features. Deploying machine learning tools in analysis pipelines on the cloud is extensively surveyed and discussed in.

**Supplementary Note S3.**

In order to trace genetic mutations in cancer, the system uses PyClone. PyClone is a tool that statistically infer the subclonal population of cancer using Bayesian inference. It clusters positions of mutations in terms of allele frequencies. Using Bayesian hierarchical clustering, PyClone groups mutations from input by their allelic prevalence and infer ITH at the mutation level. BioVLAB-Cancer-Pharmacogenomics uses nJSD for transcriptomic ITH, or tITH. In nJSD, each gene was represented as a probability distribution proportionate to its neighborhood genes' expression values, which is entropy of the gene. Neighborhood genes are determined in the protein-protein interaction network. Then, gene-level entropy is summed and averaged to come up with a network level entropy. To define distance, Jensen-Shannon Divergence is used. Copy number alterations are typically computed as relative ratio of reads mapped to reference variant to those mapped to alternative variant. Cancers often go through gene duplications, or sometimes whole genome duplications. Therefore, it is important to understand the exact copy number of genes per cancer cell. ABSOLUTE estimates tumor purity and malignant cell ploidy with copy number segmentation and mutation allele frequency.

**Supplementary Note S4.**

To estimate the relative abundance of constituent cell types and their expression profiles, EDec executes three consecutive steps, all of which can be automatically performed on First, EDec exploits a collection of cell type-specific DNA methylation profiles to extract `informative' CpG loci that show statistically significant difference in their DNA methylation levels across cell types. The rationale behind selecting the informative CpG loci for the estimation of cell type proportions is that the DNA methylation at a specific CpG locus is a binary event, which are usually robust to the cell-to-cell variation within a cell type and dramatically different across cell types. Then, using the set of informative CpG loci, EDec tries to simultaneously estimate the methylation profiles and the relative abundances for K cell types, where the number of estimated cell types K is given as a predefined parameter. Formally, this step can be formulated as a non-negative matrix factorization of user-given methylation profile into the product of cell type-specific methylation profile matrix and relative abundance matrix. Finally, the cell-type specific expression profiles are estimated with user-given bulk expression profile and the proportions of estimated cell types. In BioVLAB-Cancer-Pharmacogenomics, we assume that a tumor consists of one or more subpopulations of cancerous cells with different molecular characteristics. Moreover, to avoid the linear dependencies across the cancer-specific expression profiles of tumor samples, K is intentionally set to a sufficiently large number.