HHS Public Access

Author manuscript

Semin Cancer Biol. Author manuscript; available in PMC 2021 April 01.

Published in final edited form as:

Semin Cancer Biol. 2020 April ; 61: 1–10. doi:10.1016/j.semcancer.2019.08.021.

Statistical genomics in rare cancer

Farnoosh Abbas-Aghababazadeh, **Qianxing Mo**, **Brooke L. Fridley***

Department of Biostatistics & Bioinformatics, Moffitt Cancer Center, Tampa, FL, 33612, USA

Abstract

Rare cancers make of more than 20% of cancer cases. Due to the rare nature, less research has been conducted on rare cancers resulting in worse outcomes for patients with rare cancers compared to common cancers. The ability to study rare cancers is impaired by the ability to collect a large enough set of patients to complete an adequately powered genomic study. In this manuscript we outline analytical approaches and public genomic datasets that have been used in genomic studies of rare cancers. These statistical analysis approaches and study designs include: gene set / pathway analyses, pedigree and consortium studies, meta-analysis or horizontal integration, and integration of multiple types of genomic information or vertical integration. We also discuss some of the publicly available resources that can be leveraged in rare cancer genomic studies.

Keywords

Data integration; Pathway analysis; Meta-analysis; Consortium; Pedigree studies; Heterogeneity

1. Introduction

Rare cancers are roughly defined as cancers with fewer than 15 new diagnoses in 100,000 people a year; however there is no universally adopted definition. Collectively rare cancer account for more than 20% of all cancer diagnoses in a year [1]. In the United States, pediatric and adolescence young adults are disproportionally impacted by rare cancers, with more than 2/3 of cancers in pediatric, adolescent or young adult individuals (age < 20 years) being rare [2]. Due to the rare nature, less research has been conducted on rare cancers resulting in worse outcomes for patients with rare cancers compared to common cancers. In a study conducted by the Surveillance of Rare Cancers in Europe (RARECARE) Project, they found the five-year survival rate to be 47% for rare cancers compared to 65% for common cancers [3]. Additionally, not much is known regarding ways to prevent and adequately diagnoses many rare cancers.

*Corresponding author at: Department of Biostatistics & Bioinformatics, Moffitt Cancer Center, 12902 Magnolia Drive, MRC-Bio2, Tampa, FL, 33612-9416, USA. Brooke.Fridley@moffitt.org (B.L. Fridley).

Declaration of Competing Interest

None declared.

Rare cancers are often defined in terms of their uncommon site of origin (heart), uncommon cell type of origin (small cell cancer of the cervix), unique molecular feature (*RYB6-NTRK3* fusion in breast cancer), or uncommon host (male breast cancer). Recently, with the advent of high-throughput genomics, researchers have been able to further stratify common cancers into unique subtypes based on a set of molecular features, where often the number of patients that fall into these specific subtypes becomes small and meets the general definition of a rare cancer. The ability to study these rare cancers and subtypes is impaired by the ability to collect a large enough set of patients to complete an adequately powered genomic study. As illustrated in Fig. 1, the power to detect a moderate effect size (Cohen's d = 0.50) for differentially expressed genes with a type I error rate of 0.00001 (Bonferroni adjustment for testing 5000 genes) improves greatly as the sample size increase, where 226 subjects are needed in each group to be able to detect the moderately differentially expressed genes with 80% power. In the following sections, we outline some analytical approaches and public genomic datasets that have been used in genomic studies of rare cancers.

2. Public data resources

Large genomic repositories have been created for use by cancer researchers. These repositories can be used to: (1) combine with other studies of a given rare cancer to increase sample size (e.g. meta-analysis or horizontal integration); (2) determine novel hypothesis to be tested in future prospective studies; (3) replication or validation of findings; or (4) compare / contrast between cancers (pan-cancer). Below we outline three large public data resources that can be leveraged in the study of rare cancers.

2.1. Gene expression omnibus (GEO)

GEO is a public repository for high-throughput transcriptomic datasets generated by array- and sequence-based technologies [4,5]. Such datasets hold great value for knowledge discovery, particularly when integrated. The GEO repository is publicly accessible at (https://www.ncbi.nlm.nih.gov/geo) and includes hundreds of studies on a variety of rare cancers. For example, GEO contains 5420 studies on Osteosarcoma, 550 on Ewing's sarcoma, 160 on male breast cancer, 1480 on Gastrointestinal stromal tumors (GISTs), 426 on chondrosarcoma, 1331 on mesothelioma, and 1547 on ependymoma.

2.2. The Cancer genome atlas (TCGA)

The TCGA was a multi-institutional collaborative project aimed to comprehensively catalogue genomic alterations of a variety of cancers through high-throughput genomic and bioinformatic analyses [6]. TCGA has characterized 33 cancer types, including 10 rare cancers: adrenocortical carcinoma (ACC) [7], cholangiocarcinoma (CHOL) [8], kidney chromophobe carcinoma (KICH) [9], mesothelioma (MESO) [10], pheochromocytoma and paraganglioma (PCPG) [11], sarcoma (SARC) [12], testicular germ cell tumor (TGCT) [13], thymoma (THYM) [9], uterine carcinosarcoma (UCS) [14], and uveal melanoma (UVM) [15] (https://portal.gdc.cancer.gov). TCGA has generated integrative multi-omics data including large-scale genomic, epigenomic, transcriptomic and proteomic datasets along with slide images for histopathology and details on patient's information which have become a great resource for cancer research.

2.3. Therapeutically applicable research to generate effective treatment (TARGET)

The TARGET initiative is employing genomic data to accelerate molecular discoveries and drug development for difficult to treat childhood cancers, such as, acute lymphoblastic leukemia (ALL) [16], acute myeloid leukemia (AML) [17], neuroblastoma (NBL) [18], osteosarcoma (OS), and Wilms' tumor (WT) [19]. Pediatric ALL was the first disease to be piloted for the TARGET initiative, which is jointly managed by the NCI Office of Cancer Genomics (OCG) and Cancer Therapy Evaluation Program (CTEP). TARGET datasets include large-scale genomic data including gene-expression, copy number variation, epigenetics, along with annotated clinical information for a selected set of pediatric cancers (https://ocg.cancer.gov/programs/target/data-matrix).

3. Consortium and collaborative networks

One major limitation in the study of rare cancers is the lack of sufficiently size cohort of cancer patients by any one given research group. Hence, consortium and networks have been created to pool resources in the study of rare cancers [20]. This approach has also been employed in studies of common cancers as a means to increase power to detect relative small effects, particular in the context of genome-wide genetic association studies (GWAS) [21–23]. One of the largest initiative of this kind for studying rare cancers is the International Rare Cancer Initiative (IRCI)1 and the International Cancer Genome Consortium (ICGC) [24,25]. Common cancers are the primary cancers being studied in the ICGC. However, some studies are underway involving rare cancers, such as Ewing's sarcoma, osteosarcoma, chondrosarcoma and medulloblastoma, which all primary effect children or adolescent young adults. The pooling of familial data for rare genetic based rare cancer has often been done to localize disease loci for rare cancers. This approach has been successfully in the setting of Li-Fraumeni syndrome (LFS) [26,27]. LFS rare inherited syndrome that can lead to the development of a number of cancers, including sarcoma (such as osteosarcoma and soft-tissue sarcomas), leukemia, brain (central nervous system) cancers, cancer of the adrenal cortex and breast cancer.

In completing statistical analyses in the context of consortium or networks, additional care is needed to insure adequacy of results. In particular, assessment of batch effects in the genomic data between studies is a major concern and needs to be considered when completing statistical analyses. Batch effects in large datasets can often been visualized using data reduction method, such as plotting the first two principal components from a principal component analysis (PCA), as illustrated in Fig. 2. If study or batch effects are observed, adjusting for study/batch in the model or normalization of the data using more sophisticated methods (e.g., COMBAT [28]) are warrented [29]. In the setting of GWAS adjustment for population stratification is also required when completing the statistical analysis [30].

4. Analysis of cancer "families" and meta-analysis

Since the etiology, diagnosis, and treatment of some rare cancers are similar to more common cancers, often researchers look at biologically similarly common cancers to the rare cancer of interest to extrapolated or compare findings. As an example, male breast cancer is

a rare disease accounting for approximately <1% of all breast cancer diagnoses worldwide [31–34] and shares many similarities with female breast cancer [31,35]. Another way to overcome some of the challenges in studying rare cancers is to group rare cancers into "families". In addition, genetic information is increasingly used to group cancers according to a tumor's molecular subtype. However, a limitation in grouping rare cancers together in "families" is that this can add heterogeneity. In the following section we discuss various statistical approaches for meta-analysis for the analysis of cancer families. Many of these methods allow researcher to assess the degree of heterogeneity between the cancers being pooled together.

4.1. Meta-analysis methods

Due to the large number of features and limited sample size in rare cancer study, multi-study genomic data integration called "horizontal integration" has been considered to improve the discovery of new biological insights and reach more general and reliable conclusions along with increasing statistical power [36–39]. Meta-analysis methods are based on summary statistics from the analysis of each individual study (e.g., effect sizes, p-values), whereby these study specific summary statistics are aggregated together to get a combined level of association for the entire set of studies or cancer "family" [40–43]. Several meta-analysis methods have been suggested to genomic applications such as combining p-values, effect sizes, and rankings. The strengths and limitations of meta-analysis methods are evaluated particularly with respect to their ability to assess variation across independent genomic studies (e.g., platform variability, inconsistent annotation, various methods for data processing, and patient heterogeneity) beyond within-study variation.

4.1.1. Methods combining p-values—Combining the p-values from multiple independent studies has a benefit of its simplicity and extensibility to different kinds of outcome variables [36,42,44]. When the outcome variable is not binary (e.g. multi-class or censored survival), association p-values can be computed, while effect sizes may not be well defined. Fisher's method [45] and Stouffer's method [46] are widely used to combine results from different studies [47]. Other p-values based meta-analysis methods have been applied such as taking the minimum and maximum p-values [48,49], or a weighted modification to Fisher's method [50]. Major limitation of such traditional methods of combining p-values is that they can be performed parametrically under the assumption that p-values are uniformly distributed under the null hypothesis [51,52]. In addition, the existing traditional combining p-values typically do not account the data heterogeneity and do not take into account direction of effect.

4.1.2. Methods combining effect sizes—Most popular statistical methods to combine effect sizes (e.g., standardized mean differences, correlation coefficients, odds ratios, etc.) are based on fixed- and random-effects model [53]. Under the fixed-effects model, estimated effect sizes are assumed to be homogeneous across studies and all differences in observed effects are due to sampling error or within-study variability, while in practice such assumption is questionable. In contrast, the random-effects model incorporates the variability of the effect size across studies in addition to the within-study variability using two-stage hierarchical process [44,54–56]. One of the most troublesome aspects of a

meta-analysis is the determination of whether there is true heterogeneity (i.e., across study variability), as it can influence the choice of the statistical method to combine effect sizes. Cochran (1954) [57] proposed a ***Q*** test to determine the heterogeneity across studies; however its statistical power depends on the number of studies [58,59]. Another approach is to assume a random-effects model that consists of estimating the across study variance (τ^2). To address the limitations of ***Q*** test and the between-study variance methods, Higgins et al. (2002, 2003) proposed the I^2 statistic that is the percentage of the total variability in a set of effect sizes due to true heterogeneity [60,61].

To illustrate these methods for assessing heterogeneity, we used the data from the TCGA sarcoma (SARC) study, were multiple types of sarcomas were represented. Using RNA-seq data collected on 58 dedifferentiated liposarcoma (DDLPS), 104 leiomyosarcoma (LMS), 25 myxofibrosarcoma (MFS), and 50 undifferentiated pleomorphic sarcoma (UPS) we applied fixed- and random-effects meta-analysis for to determine the association of gene expression with overall survival using hazard ratios (HZ) from Cox PH models. Fig. 3 presents forest plots for *TRPV6* (I^2 = 85%, p-value < 0.01) and *KIF21A* (I^2 = 62%, p-value = 0.05), both showing substantial heterogeneity across 4 studies.

4.1.3. Methods combining ranks—Proposed ranking-based meta-analysis methods combine robust rank statistics instead of p-values or effect sizes to address issues regarding outliers and heterogeneity in genomic studies [62,63]. The product, mean or sum of ranks [64–66] from genomic studies is calculated as the test statistics along with assessing the statistical significance using permutation testing [67–69].

5. Pedigree and population based genetic studies

Pedigree or family studies have been the backbone for studying rare Mendelian or inherited cancers and cancer syndromes. Mendelian traits are traits that are controlled by a single locus in which one inherits the disease predisposing allele from either their mother or father. An example is *BRCA1* related breast cancer, ovarian cancer, and Lynch Syndrome, as illustrated in two example pedigrees from the National Cancer Institute (Fig. 4). Often in the study of these cancers, pedigrees are recruited based on the proband, the person serving as the starting point for the genetic study of a family and usually the first person enrolled in the family. These study design are often used in genetic epidemiology studies [70] involving familial aggregation [71], segregation [72,73], and linkage [74–76] studies of both rare and common cancers. Pedigrees studies have been successfully applied to the study of Li-Fraumeni Syndrome (LFS). LFS is a rare inherited syndrome that leads to the development of many cancers, often at a younger age of onset. LFS families are predisposed to both rare cancers (e.g., osteosarcoma, Wilms tumors) and common cancers (e.g., colon and breast cancer) at a higher rate than the general population [77,78]. Using 39 pedigrees, researchers were able to determine that the most common cause of LFS is inherited mutations in the tumor suppressor gene *TP53* [26].

With the advent of microarray technology and the ability to genotype thousands of single-nucleotide polymorphism, as opposed to genotyping hundreds of micro-satellites or restriction fragment-length polymorphisms (RFLPs), population based genetic association

studies involving families or unrelated individuals became common place as a means to study the genetic basis of cancer risk [79]. However, these approaches often failed in the study of rare cancers due to the limited sample size to detect the small effect size, as the majority of cancer susceptibility loci discovered in genome-wide association studies (GWAS) have low effect sizes (OR < 1.3) [80]. However, GWAS have been successfully applied in the setting of some rare cancers, such as neuroblastoma, where GWAS have been able to determine 8 neuroblastoma risk loci including *BARD1* [81] (OR = 1.4) and a loci within 6p22 [82] (OR = 1.4) (see GWAS catalog at https://www.ebi.ac.uk/gwas/home for entire list of known risk loci).

6. Analysis of pathways and gene sets

Due to the limited power in the study of rare cancers due to relatively small sample size, researchers often focus on particular pathways of interest, thus reduced the multiple testing burden. As defined by the NHGRI, “a biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in the cell. It can trigger the assembly of new molecules, such as a fat or protein, turn genes on and off, or spur a cell to move”. However, it should be noted that assigning genes to a pathway is somewhat artificial, as pathways are not entities in of themselves, but rather inter-related dynamic groups of genes. Often researchers use pathways as defined in KEGG: Kyoto encyclopedia of gene and genomes [83,84]. Other possible pathway or gene sets often use are those defined in Gene Ontology (GO) [85] (sets of genes that are biological related) and those genes related to a particular transcription factor defined by ChIP-Seq studies [86].

There are two approaches for using pathway information, a “hypothesis-driven” approach or an “agnostic” approach. The “hypothesis-driven” approach is one in which a researcher *a priori* looks only at the biological factors within a particular pathway of interest, as illustrated in the paper by Mezzapelle et al [87]. In this research, mesothelioma researchers focused on *EGRF* and downstream signaling pathways in a set of 77 malignant pleural mesothelioma tumors. In the “agnostic” approach, a researcher is looking at aggregation of signal or enrichment of signal in a large set of pathways, such as investigation of all KEGG pathways. Aggregation of the association signals for a set of genes within a pathway may be beneficial as it incorporates biological knowledge, reduces the multiple-testing burden, and may increase the association signal, thus increasing the power to detect biologically meaningful results. In agnostic pathway analysis, there are two types of statistical methods for assessing pathways: competitive/enrichment methods or self-contained methods [88,89]. These two approaches are based on two different null hypotheses. In the competitive or enrichment methods one is testing the null hypothesis that genes within the pathway are more associated with the phenotype of interest (i.e., survival) then genes outside the pathway. That is, one is looking to determining if there is more “signal” within the pathway than expected or if the pathway is “enriched for signal”. Commonly used pathway analysis approaches that are competitive in nature are *GSEA: gene set enrichment analysis* [90] or methods in the flavor of Fisher Exact Test, such as those implemented in commonly used online tool *EnrichR* [91,92]], and *DAVID* [93,94], where users upload a list of genes found to be associated with the phenotype of interest (i.e., genes differentially expressed between two treatment conditions). Additionally, commercially available software tools, such as

Ingenuity Pathway Analysis (IPA) and MetaCore software analysis tool (GeneGo), have implementations of enrichment or competitive pathway analysis methods. In using the competitive or enrichment methods, experiments and analyses need to be genome-wide (i.e., transcriptomic studies conducted using high-throughput sequencing or microarray technologies). These agnostic enrichment approaches have been successfully used in Ewing's sarcoma genomic studies [95,96].

In contrast, the self-contained pathway or gene set approaches are assessing if the genes in the pathway are associated with the phenotype of interest. These methods can be used in both candidate pathway studies (i.e., targeted or custom panels) and genome-wide studies. Examples of methods in this framework are modeling based methods like the *global test* Bioconductor package which uses a random effects model [97] or summarization methods which aggregate the results for the genes in the pathway to the pathway level, such as Fisher's Method for meta-analysis [45] and variations on Fisher's Method, such as the Gamma Method [98,99]. A study of acute megakaryoblastic leukemia (AMKL), a rare subtype of acute myeloid leukemia (AML), researchers combined two studies of pediatric cases (N = 14 and N = 79) and completed high-throughput sequencing. The self-contained global testing method was used successfully to assess the genomic features in a gene set with determine subtypes of AMKL [100].

7. Network analysis and module construction

In contrast to gene set or pathway analysis, where the analysis is focused on *a prior* defined set of genes, network analysis is focused on determining sets of molecular features that are co-expressed or related and form a tight network or modules which aid researchers in understanding of gene regulation (i.e., co-regulated genes) or signaling networks. Additionally, many network analysis methods are used as the basis for determining modules of related features (i.e., co-expressed genes) [101]. In the study of rare cancers, reducing the dimensionally from individual features to a module of related features can reduce the multiple testing burden and therefore increase the power to detect biological relevant associations. Methods for determining networks primarily falling into three different paradigms: relevance networks, Gaussian graphical models (GMMs) and Bayesian belief networks [102,103]. Relevance networks [104] are based on a pairwise distance measures, often the correlation coefficient, where genes are connected if the absolute value of the measure is greater than some threshold. Since originally proposed in 2000, this approach has been modified extensive, with correlation based networks the basis of many module detection algorithms, including weighted gene co-expression network analysis (WGCNA) [105–107] which has been successfully applied to studies of Wilms tumors [108], Adrenocortical carcinoma [109] and Osteosarcoma [110]. However, a drawback of using correlation as the pairwise measure of relatedness captures both the direct and indirect relationship between features. To overcome this limitation, GMMs based on the partial correlation coefficient (i.e., inverse covariance matrix) only incorporate the direct relationship between features are often used to construct networks [111]. This approach has been used extensive in analysis of genomic data, with a recent publication applying GMMs to molecular data collected on 15 cancers in the TCGA [112]. Finally, Bayesian belief networks (BBNs) [113,114], or probabilistic directed acyclic graphical models are based on

probability statements and infer direction of causation. BBNs have been applied extensively in the study of protein signaling networks [115,116], and gene expression networks [113].

One challenge in determining networks from high-dimensional molecular data is the enormous number of possible networks due to the large number of possible nodes (genes) and edges (connections). To overcome this challenge, many methods impose a sparsity or penalty term to estimate a robust correlation or partial correlation matrix for which the network is derived [117–121]. For example, Schafer and Strimmer (2005) proposed a shrinkage approach for estimating the covariance matrix needed for determining the partial correlation coefficients involving thousands of genes [122].

8. Integrative analyses

Due to advances in our ability to assess genomic, transcriptomic, epigenomic and proteomic features in cancer, many studies are assessing multiple layers of information on the same set of samples/subjects. In the context of rare cancers, the accumulation of multiple types of data on the same set of subjects can aid in the detection of biologically relevant findings, even when the sample size is relative small. In these multiple-omic studies, the data can be arranged vertically, where each of the data sets is represented by a data matrix with columns corresponding to samples and rows corresponding to genomic features (Fig. 5A). How to integrate these high-dimensional multi-omics data has become a great challenge for biostatisticians and bioinformaticians. In general, multi-omics data analysis can be classified as supervised or unsupervised integrative analysis. We will use examples to illustrate typical supervised and unsupervised analyses in the following sections. For comprehensive review of integrative analysis, we refer the readers to Kristensen et al. [123], Richardson et al. [41], and Wu et al. [124].

In supervised integrative analyses (e.g., classification or signature development), a clinical phenotype of the samples (e.g., overall survival, progression free survival) is usually include the dependent variable (Y) and the multi-omics features are included as exploratory variables (X’s) for statistical modeling (Fig. 5C & E). The supervised analyses could reveal genomic predictors of diseases and lead to identification of drug-targetable molecular alterations. In genomic application, supervised analysis is to identify omics features that are associated with clinical outcomes such as patient survival in a vertical integrative analysis (Fig. 5C) and tumor pathological stages (or treatment responses) in a horizontal integrative analysis or meta-analysis (Fig. 5E).

For example, in a vertical supervised integrative analysis, Jiang et al. developed prognostic Cox models for cutaneous melanoma that included clinical variables, mutation, mRNA gene expression, methylation and copy number data into the model [125]. To reduce the high dimensional omics variables into low dimensional variables that can be modeled, they used variable selection and dimension reduction approaches. For the variable selection approach, first, they fit a penalized Cox model with the clinical variables to identify 5 most important variables. Next, they fitted a penalized Cox model with single-omics data to identify the 10 most informative variables for each molecular data type. Lastly, they fit Cox models with the selected clinical and omics variables. For the dimension reduction approach, they used

sparse principal component (SPCA) and sparse partial least squares (SPLS) to reduce the single-omics data to 10 SPCA and 10 SPLS components, respectively; then they fitted Cox models with various combination of the selected clinical variables and SPCA (or SPLS) components. As a result, they found that inclusion of multi-omics variables led to prognostic models with improved prediction performances. Zhao et al. used a similar approach to integrate multi-omics data to develop prognostic models for breast invasive carcinoma, glioblastoma multiforme (GBM), acute myeloid leukemia (AML) and lung squamous cell carcinoma (LUSC) [126].

For horizontal integrative analysis, if omics data from different cancer types are combined for analysis, it often refers to as a “pancancer” analysis (Fig. 5D). The major goal of pan-cancer analysis is to define commonalities and differences of genomic alterations across cancer types [6]. For example, Kandoth et al. [127] analyzed the mutation landscape of 12 major cancer types and found that *TP53* and *PIK3CA* were the most commonly mutated genes, *ARID1A* were frequently mutated in bladder urothelial carcinoma (BLCA), uterine corpus endometrial carcinoma (UCEC), lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), and *EGFR* were frequently mutated in GBM and LUAD. In contrast, *VHL* and *PBRM1* mutations were exclusive to kidney renal clear cell carcinoma (KIRC), and *NPM1* and *FLT3* mutations were exclusive to AML. Numerous pan-cancer analyses have been performed, including pan-cancer analyses of copy number alteration, enhancer expression, oncogenic signaling pathways, and transcriptional metabolic dysregulation [128–131]. If omics data from different cohorts of the same study are combined for analysis, horizontal integrative analysis often refers to meta-analysis (Fig. 5E). Meta-analysis is often used to increase statistical power and achieve a consensus conclusion (see section above on meta-analysis).

In unsupervised analysis, clinical variables are not directly used in the analysis and the major goal is to understand the underlying structure of omics data. Integrative clustering is an important unsupervised method that has been widely used to characterize cancers by grouping cancer samples and genomic features into meaningful subgroups (Fig. 5B & D). To characterize a specific cancer, it is often involved in vertical integrative analysis of multi-omics data sets. Traditionally, multi-omics data sets were clustered separately and with the resulting clusters manually integrated for characterization of the cancer. This single-omics clustering approach was used to cluster TCGA rare cancer studies involving Pheochromocytoma and Paraganglioma (PCPG) [11], uterine carcinosarcoma (USC) [14], uveal melanoma (UVM) [15] and testicular germ cell tumors (TGCT) [13].

An obvious disadvantage of the single-omics clustering analysis was that the inherent correlations among the multi-omics data sets were not taken into account during the analysis. As a result, it was not easy to identify molecular patterns across multiple platforms. Realizing the weakness of the single-omics clustering approach, researchers used cluster-of-clusters-analysis (COCA) in order to achieve integrative clustering assignments for breast cancer [132]. This approach is a step-wise integration approach. At the first step, single-omics clustering analysis was performed and a sample cluster assignment matrix was generated. At the second step, clustering analysis was performed on the sample cluster assignment matrix to generate joint cluster assignments for the samples. The COCA

approach has been used to perform clustering analysis of rare cancers of adrenocortical carcinoma (ACC) [7], cholangiocarcinoma (CHOL) [8] and thymic epithelial tumors (THYM) [133]. It should be noted that the resulting integrative clusters were not directly driven by the genomic patterns embedded in the multi-omic data sets because the COCA approach does not directly use the multi-omics data.

Unlike the step-wise clustering approach or COCA, the integrative clustering method iCluster developed by Shen et al. [134] directly models multi-omic data sets to obtain clusters. The iCluster model is a Gaussian latent variable model in which a latent variable is used to capture the correlative structure of multi-omics data. A penalized expectation-maximization algorithm using a lasso-like penalty is used to obtain optimized solutions. Integrative sample cluster assignments are obtained by performing k-means clustering on the latent variable matrix. In an effort to accommodate different natures of multi-omics data, Shen et al. further extended the iCluster method by making use of lasso, elastic net and fused lasso penalty functions to make feature selection more flexible [135]. A limitation of the iCluster method is that it can only model continuous multi-omics data. Besides continuous data, omics data can be in the forms of binary (e.g., gene mutation status: yes, or no), multi-category (e.g., copy number states: gain, normal, loss) and counts (e.g., gene expression measurement by RNA-seq). The iClusterPlus method developed by Mo et al. [136] is a significant enhancement of the iCluster method, which uses linear regression to model continuous data, Poisson regression to model count data, logistic regression to model binary data, and multi-logit regression to model multi-categorical data. Recently, Mo et al. [137] developed the iClusterBayes method, a fully Bayesian latent variable model under the iClusterPlus framework. This method uses Bayesian variable selection techniques to identify informative features that contribute to sample clustering. An advantage of the iClusterBayes method is that it provides posterior probability estimation for each omics feature, which can be used as a criterion for feature selection. The iCluster methods have been used by TCGA and other research groups to characterize common and rare cancers, including mesothelioma (MESO) [10] and adult soft tissue sarcoma (SARC) [12].

Integrative non-negative matrix factorization (intNMF) is another powerful method for integrative clustering analysis of multi-omics data. NMF was initially applied to microarray gene expression data to identify cancer subtypes [138–140]. Usually, analysis is completed for a range of number of clusters (e.g., 2–10) with the selection of number of clustered determined by a number of diagnostic measures (e.g., consensus matrix [141], cophenetic correlation coefficient [138], dispersion coefficient [140]). Zhang et al. [142] extended NMF to multi-omics data to identify multi-dimensional modules (patterns). The joint NMF framework of Zhang et al. was effective in detection coordinated patterns across multiple data sets, but it was sensitive to random noise and confounding effects [142]. To remedy that, Yang and Michailidis (2016) extended the joint NMF framework by using a new factorization algorithm that was more robust to heterogeneous effects in the multi-omics data [143]. Chalise and Fridley (2017) further extended the joint NMF framework with a focus on integrative clustering analysis [144]. By analyzing the TCGA breast cancer (BRCA) and glioblastoma (GBM) multi-omics data sets, Chalise and Fridley found that the clusters identified by intNMF and iCluster largely overlapped, demonstrating its capability in identification of cancer subtypes inherent in the data.

To illustrate the integrative clustering analysis, we analyzed 241 sarcoma tumor samples from the TCGA SARC study that had somatic mutation, copy number, methylation and mRNA expression data using the recently developed iClusterBayes software [137]. Fig. 6A shows the three iClusters of SARC along with the driver omics features that made major contribution to the sample clustering. These three iClusters highly overlapped with the five iClusters reported by TCGA that were based on the iCluster analysis of copy number, methylation, mRNA, and miRNA expression data [12]. Specifically, iCluster 1 contained only leiomyosarcoma (LMS) tumors; iCluster 2 is dominated by dedifferentiated liposarcoma (DDLPS), leiomyosarcoma (LMS) and synovial sarcoma (SS); iCluster 3 is dominated by undifferentiated pleomorphic sarcoma (UPS), DDLPS, myxofibrosarcoma (MFS) and LMS (Fig. 6B). Fig. 6C shows the samples clustering on the two-dimensional latent variable spaces. The iCluster 1 was characterized by relatively low mutation rates of the driver genes except for RB1 and TP53, normality of chr12:58018979-70771592, trend to loss of chr13:49062990-50471179, and cluster-specific methylation and mRNA expression patterns (Fig. 6A). In contrast, the iCluster 2 and 3 were characterized by relatively high mutation rates of the driver genes, trend to amplification of chr12:58018979-70771592, normality of chr13:49062990-50471179, and cluster-specific methylation and mRNA expression patterns (Fig. 6A). In terms of overall survival, the iCluster 1 was the best, followed by the iCluster 2 the second, and the iCluster 3 the worst ($p = 0.009$, Fig. 6D).

9. Conclusions

There are numerous challenges in completing genomic studies of rare cancers. In terms of statistical analysis of genomic studies, the primary challenge is the limited sample size for an adequately powered study. This review paper outlines and illustrates various approaches that can be leveraged to increase the power to detect biologically relevant genomic factors related to rare cancers. Many of these approaches involve some level of integration or aggregation, whether it involves combination of association signal across a set of studies (i.e. meta-analysis or horizontal integration), integration of information across multiple types of omic data (i.e., vertical integration), or aggregations of association signal in a pathway (i.e., pathway or gene set analysis). In completing horizontal integration or meta-analysis, consortiums are often developed to bring together multiple research groups to increase the sample size; however, care is needed in the analysis of consortium data to assess degree of heterogeneity and potential batch effects which can impact the interpretability of the statistical results. Most genomic studies involve unrelated individuals, but family-based studies that leverage the pedigree structure have been successfully applied to rare cancer genetic epidemiology studies. Lastly, leveraging all possible publicly available resources and information is an important consideration when planning and completing genomic studies of rare cancers.

References

[1]. Keat N, Law K, Seymour M, Welch J, Trimble T, Lascombe D, Negrouk A, International rare cancers initiative, Lancet Oncol. 14 (2013) 109–110. [PubMed: 23369681]

[2]. DeSantis CE, Kramer JL, Jemal A, The burden of rare cancers in the United States, CA Cancer J. Clin 67 (2017) 261–272. [PubMed: 28542893]

[3]. Gatta G, van der Zwan JM, Casali PG, Siesling S, Dei Tos AP, Kunkler I, Otter R, Licitra L, Mallone S, Tavilla A, et al., Rare cancers are not so rare: the rare cancer burden in Europe, Eur. J. Cancer 47 (2011) 2493–2511. [PubMed: 22033323]

[4]. Edgar R, Domrachev M, Lash AE, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, Nucleic Acids Res. 30 (2002) 207–210. [PubMed: 11752295]

[5]. Barrett T. Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al., NCBI GEO: archive for functional genomics data sets–update, Nucleic Acids Res. 41 (2013) D991–995. [PubMed: 23193258]

[6]. N. Cancer Genome Atlas Research, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, The cancer genome atlas pan-cancer analysis project, Nat. Genet 45 (2013) 1113–1120. [PubMed: 24071849]

[7]. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, Lerario AM, Else T, Knijnenburg TA, Ciriello G, et al., Comprehensive pan-genomic characterization of adrenocortical carcinoma, Cancer Cell 29 (2016) 723–736. [PubMed: 27165744]

[8]. Farshidfar F, Zheng S, Gingras MC, Newton Y, Shih J, Robertson AG, Hinoue T, Hoadley KA, Gibb EA, Roszik J, et al., Integrative genomic analysis of cholangiocarcinoma identifies distinct IDH-Mutant molecular profiles, Cell Rep. 18 (2017) 2780–2794. [PubMed: 28297679]

[9]. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al., Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of Cancer, Cell 173 (2018) 291–304 e296. [PubMed: 29625048]

[10]. Hmeljak J, Sanchez-Vega F, Hoadley KA, Shih J, Stewart C, Heiman D, Tarpey P, Danilova L, Drill E, Gibb EA, et al., Integrative molecular characterization of malignant pleural mesothelioma, Cancer Discov. 8 (2018) 1548–1565. [PubMed: 30322867]

[11]. Fishbein L, Leshchiner I, Walter V, Danilova L, Robertson AG, Johnson AR, Lichtenberg TM, Murray BA, Ghayee HK, Else T, et al., Comprehensive molecular characterization of pheochromocytoma and paraganglioma, Cancer Cell 31 (2017) 181–193. [PubMed: 28162975]

[12]. Cancer Genome Atlas Research Network, edsc. Electronic address, N. Cancer Genome Atlas Research, Comprehensive and integrated genomic characterization of adult soft tissue sarcomas, Cell 171 (950-965) (2017) e928.

[13]. Shen H, Shih J, Hollern DP, Wang L, Bowlby R, Tickoo SK, Thorsson V, Mungall AJ, Newton Y, Hegde AM, et al., Integrated molecular characterization of testicular germ cell tumors, Cell Rep. 23 (2018) 3392–3406. [PubMed: 29898407]

[14]. Cherniack AD, Shen H, Walter V, Stewart C, Murray BA, Bowlby R, Hu X, Ling S, Soslow RA, Broaddus RR, et al., Integrated molecular characterization of uterine carcinosarcoma, Cancer Cell 31 (2017) 411–423. [PubMed: 28292439]

[15]. Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, Hess JM, Uzunangelov V, Walter V, Danilova L, et al., Integrative analysis identifies four molecular and clinical subsets in uveal melanoma, Cancer Cell 32 (204-220) (2017) e215.

[16]. Liu Y, Easton J, Shao Y, Maciaszek J, Wang Z, Wilkinson MR, McCastlain K, Edmonson M, Pounds SB, Shi L, et al., The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia, Nat. Genet 49 (2017) 1211–1218. [PubMed: 28671688]

[17]. Bolouri H, Farrar JE, Triche T Jr, Ries RE, Lim EL, Alonzo TA, Ma Y, Moore R, Mungall AJ, Marra MA, et al., The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions, Nat. Med 24 (2018) 103–112. [PubMed: 29227476]

[18]. Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, Carter SL, Cibulskis K, Hanna M, Kiezun A, et al., The genetic landscape of high-risk neuroblastoma, Nat. Genet 45 (2013) 279–284. [PubMed: 23334666]

[19]. Armstrong AE, Gadd S, Huff V, Gerhard DS, Dome JS, Perlman EJ, A unique subset of low-risk Wilms tumors is characterized by loss of function of TRIM28 (KAP1), a gene critical in early renal development: a children's oncology group study, PLoS One 13 (2018) e0208936. [PubMed: 30543698]

[20]. Blay JY, Coindre JM, Ducimetiere F, Ray-Coquard I, The value of research collaborations and consortia in rare cancers, Lancet Oncol. 17 (2016) e62–e69. [PubMed: 26868355]

[21]. B.C.A.C. Ovarian Cancer Association Consortium, B. Consortium of Modifiers of, Brea, Hollestelle A, van der Baan FH, Berchuck A, Johnatty SE, Aben KK, Agnarsson BA, Aittomaki K, et al., No clinical utility of KRAS variant rs61764370 for ovarian or breast cancer, Gynecol. Oncol (2015).

[22]. Phelan CM, Kuchenbaecker KB, Tyrer JP, Kar SP, Lawrenson K, Winham SJ, Dennis J, Pirie A, Riggan MJ, Chornokur G, et al., Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer, Nat. Genet 49 (2017) 680–691. [PubMed: 28346442]

[23]. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, et al., Genome-wide association study identifies novel breast cancer susceptibility loci, Nature 447 (2007) 1087–1093. [PubMed: 17529967]

[24]. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, et al., International cancer genome consortium data portal–a one-stop shop for cancer genomics data, Database 2011 (2011) bar026. [PubMed: 21930502]

[25]. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, Stein LD, Ferretti V, The international cancer genome consortium data portal, Nat. Biotechnol 37 (2019) 367–369. [PubMed: 30877282]

[26]. Varley JM, McGown G, Thorncroft M, Santibanez-Koref MF, Kelsey AM, Tricker KJ, Evans DG, Birch JM, Germ-line mutations of TP53 in Li-Fraumeni families: an extended study of 39 families, Cancer Res. 57 (1997) 3245–3252. [PubMed: 9242456]

[27]. Eng C, Schneider K, Fraumeni JF Jr, Li FP, Third international workshop on collaborative interdisciplinary studies of p53 and other predisposing genes in Li-Fraumeni syndrome, Cancer Epidemiol. Biomarkers Prev 6 (1997) 379–383. [PubMed: 9149899]

[28]. Johnson WE, Li C, Rabinovic A, Adjusting batch effects in microarray expression data using empirical Bayes methods, Biostatistics 8 (2007) 118–127. [PubMed: 16632515]

[29]. Abbas-Aghababazadeh F, Li Q, Fridley BL, Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing, PLoS One 13 (2018) e0206312. [PubMed: 30379879]

[30]. Price AL, Zaitlen NA, Reich D, Patterson N, New approaches to population stratification in genome-wide association studies, Nature reviews 11 (2010) 459–463.

[31]. Deb S, Wong SQ, Li J, Do H, Weiss J, Byrne D, Chakrabarti A, Bosma T, kConFab I, Fellowes A, et al., Mutational profiling of familial male breast cancers reveals similarities with luminal A female breast cancer with rare TP53 mutations, Br. J. Cancer 111 (2014) 2351–2360. [PubMed: 25490678]

[32]. Weiss JR, Moysich KB, Swede H, Epidemiology of male breast cancer, Cancer Epidemiol. Biomarkers Prev 14 (2005) 20–26. [PubMed: 15668471]

[33]. Korde LA, Zujewski JA, Kamin L, Giordano S, Domchek S, Anderson WF, Bartlett JM, Gelmon K, Nahleh Z, Bergh J, et al., Multidisciplinary meeting on male breast cancer: summary and research recommendations, J. Clin. Oncol 28 (2010) 2114–2122. [PubMed: 20308661]

[34]. Harlan LC, Zujewski JA, Goodman MT, Stevens JL, Breast cancer in men in the United States: a population-based study of diagnosis, treatment, and survival, Cancer 116 (2010) 3558–3568. [PubMed: 20564105]

[35]. Giordano SH, A review of the diagnosis and management of male breast cancer, Oncologist 10 (2005) 471–479. [PubMed: 16079314]

[36]. Chang LC, Lin HM, Sibille E, Tseng GC, Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline, BMC Bioinformatics 14 (2013) 368. [PubMed: 24359104]

[37]. Wang X, Chua HX, Chen P, Ong RT, Sim X, Zhang W, Takeuchi F, Liu X. Khor CC, Tay WT, et al., Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies, Hum. Mol. Genet 22 (2013) 2303–2311. [PubMed: 23406875]

[38]. Ramasamy A, Mondry A, Holmes CC, Altman DG, Key issues in conducting a meta-analysis of gene expression microarray datasets, PLoS Med. 5 (2008) e184. [PubMed: 18767902]

[39]. Thompson JR, Attia J, Minelli C, The meta-analysis of genome-wide association studies, Brief Bioinform 12 (2011) 259–269. [PubMed: 21546449]

[40]. Mo Q, Nikolos F, Chen F, Tramel Z, Lee YC, Hayashi K, Xiao J, Shen J, Chan KS, Prognostic power of a tumor differentiation gene signature for bladder urothelial carcinomas, J. Natl. Cancer Inst 110 (2018) 448–459. [PubMed: 29342309]

[41]. Richardson S, Tseng GC, Sun W, Statistical methods in integrative genomics, Annu. Rev. Stat. Appl 3 (2016) 181–209. [PubMed: 27482531]

[42]. Tseng GC, Ghosh D, Feingold E, Comprehensive literature review and statistical considerations for microarray meta-analysis, Nucleic Acids Res. 40 (2012) 3785–3799. [PubMed: 22262733]

[43]. Borenstein M, Hedges LV, Higgins J, Rothstein, Introduction to Meta-Analysis, (Chichester, UK), (2009).

[44]. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM, Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer, Cancer Res. 62 (2002) 4427–4433. [PubMed: 12154050]

[45]. Fisher RA, Statistical Methods for Research Workers, Oliver and Boyd, London, 1932.

[46]. Stouffer SA, Suchman EA, DeVinney LC, Star SA, Williams RMJ, The American soldier, Adjustment During Army Life Vol 1 Princeton University Press, Princeton, 1949.

[47]. van Zwet WR, Oosterhoff J, On the combination of independent test statistics, Ann. Math. Stat 38 (1967) 659–680.

[48]. Won S, Morris N, Lu Q, Elston RC, Choosing an optimal method to combine P-values, Stat. Med 28 (2009) 1537–1553. [PubMed: 19266501]

[49]. Tippett LHC, The Methods of Statistics; An Introduction Mainly for Workers in the Biological Sciences, Williams & Norgate ltd., London, 1931.

[50]. Li J, Tseng GC, An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies, Ann. Appl. Stat 5 (2011) 994–1019.

[51]. Barton SJ. Crozier SR, Lillycrop KA, Godfrey KM, Inskip HM, Correction of unexpected distributions of P values from analysis of whole genome arrays by rectifying violation of statistical assumptions, BMC Genomics 14 (2013) 161. [PubMed: 23496791]

[52]. Fodor AA, Tickle TL, Richardson C, Towards the uniform distribution of null P values on Affymetrix microarrays, Genome Biol. 8 (2007) R69. [PubMed: 17472745]

[53]. Borenstein M, Hedges LV, Higgins JP, Rothstein HR, A basic introduction to fixed-effect and random-effects models for meta-analysis, Res. Synth. Methods 1 (2010) 97–111. [PubMed: 26061376]

[54]. Brockwell SE, Gordon IR, A comparison of statistical methods for meta-analysis, Stat. Med 20 (2001) 825–840. [PubMed: 11252006]

[55]. Goldstein H, Multilevel Statistical Models, Wiley, Chichester, West Sussex, 2011.

[56]. Viechtbauer W, Bias and efficiency of meta-analytic variance estimators in the random-effects model, J. Educ. Behav. Stat 30 (2005) 261–293.

[57]. Cochran WG, The combination of estimates from different experiments, Biometrics 10 (1954) 101–129.

[58]. Paul SR, Donner A, Small sample performance of tests of homogeneity of odds ratios in K 2 x 2 tables, Stat. Med 11 (1992) 159–165. [PubMed: 1579755]

[59]. Hardy RJ, Thompson SG, Detecting and describing heterogeneity in meta-analysis, Stat. Med 17 (1998) 841–856. [PubMed: 9595615]

[60]. Higgins JP, Thompson SG, Deeks JJ, Altman DG, Measuring inconsistency in meta-analyses, Bmj 327 (2003) 557–560. [PubMed: 12958120]

[61]. Higgins JP, Thompson SG, Quantifying heterogeneity in a meta-analysis, Stat. Med 21 (2002) 1539–1558. [PubMed: 12111919]

[62]. Lin S, Ding J, Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA Studies, Biometrics 65 (2009) 9–18. [PubMed: 18479487]

[63]. Deng K, Han SM, Li KJ, Liu JS, Bayesian aggregation of order-based rank data, J. Am. Stat. Assoc 109 (2014) 1023–1039.

[64]. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J, RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis, Bioinformatics 22 (2006) 2825–2827. [PubMed: 16982708]

[65]. Dreyfuss JM, Johnson MD, Park PJ, Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers, Mol. Cancer 8 (2009) 71. [PubMed: 19732454]

[66]. Zintzaras E, Ioannidis JP, Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays, Comput. Biol. Chem 32 (2008) 38–46. [PubMed: 17988949]

[67]. DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R, Combining Results of Microarray Experiments: A Rank Aggregation Approach. Statistical Applications in Genetics and Molecular Biology 5, Article15, (2006).

[68]. Hong F, Breitling R, A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments, Bioinformatics 24 (2008) 374–382. [PubMed: 18204063]

[69]. Li X, Wang XL, Xiao GH, A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications, Brief. Bioinformatics 20 (2019) 178–189. [PubMed: 28968705]

[70]. Balding DJ, Bishop MJ, Cannings C, Handbook of Statistical Genetics, John Wiley & Sons, Chichester, England; Hoboken, NJ, 2007.

[71]. Liang KY, Beaty TH, Statistical designs for familial aggregation, Stat. Methods Med. Res 9 (2000) 543–562. [PubMed: 11308070]

[72]. Jarvik GP, Complex segregation analyses: uses and limitations, Am. J. Hum. Genet 63 (1998) 942–946. [PubMed: 9758633]

[73]. Genetic Approaches to Familial Aggregation. II. Segregation Analysis. In Fundamentals of Genetic Epidemiology, pp 233–283.

[74]. Elston RC, Methods of linkage analysis–and the assumptions underlying them [see comment], Am. J. Hum. Genet 63 (1998) 931–934. [PubMed: 9758631]

[75]. Teare MD, Barrett JH, Genetic genetic linkage, Lancet 366 (9490) (2005) 1036–1044. [PubMed: 16168786]

[76]. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES, Parametric and nonparametric linkage analysis: a unified multipoint approach, Am. J. Hum. Genet 58 (1996) 1347–1363. [PubMed: 8651312]

[77]. Malkin D, Li-fraumeni syndrome, Genes Cancer 2 (2011) 475–484. [PubMed: 21779515]

[78]. Varley JM, Evans DG, Birch JM, Li-Fraumeni syndrome–a molecular and clinical review, Br. J. Cancer 76 (1997) 1–14.

[79]. Balding DJ, A tutorial on statistical methods for population association studies, Nature reviews 7 (2006) 781–791.

[80]. Chung CC, Magalhaes WC, Gonzalez-Bosquet J, Chanock SJ, Genome-wide association studies in cancer–current and future directions, Carcinogenesis 31 (2010) 111–120. [PubMed: 19906782]

[81]. Capasso M, Devoto M, Hou C, Asgharzadeh S, Glessner JT, Attiyeh EF, Mosse YP, Kim C, Diskin SJ, Cole KA, et al., Common variations in BARD1 influence susceptibility to high-risk neuroblastoma, Nat. Genet 41 (2009) 718–723. [PubMed: 19412175]

[82]. Maris JM, Mosse YP, Bradfield JP, Hou C, Monni S, Scott RH, Asgharzadeh S, Attiyeh EF, Diskin SJ, Laudenslager M, et al., Chromosome 6p22 locus associated with clinically aggressive neuroblastoma, N. Engl. J. Med 358 (2008) 2585–2593. [PubMed: 18463370]

[83]. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res. 45 (2017) D353–D361. [PubMed: 27899662]

[84]. Kanehisa M, Goto S, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30. [PubMed: 10592173]

[85]. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet 25 (2000) 25–29. [PubMed: 10802651]

[86]. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A, ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments, Bioinformatics 26 (2010) 2438–2444. [PubMed: 20709693]

[87]. Mezzapelle R, Miglio U, Rena O, Paganotti A, Allegrini S, Antona J, Molinari F, Frattini M, Monga G, Alabiso O, et al., Mutation analysis of the EGFR gene and downstream signalling pathway in histologic samples of malignant pleural mesothelioma, Br. J. Cancer 108 (2013) 1743–1749. [PubMed: 23558893]

[88]. Goeman JJ, Buhlmann P, Analyzing gene expression data in terms of gene sets: methodological issues, Bioinformatics 23 (2007) 980–987. [PubMed: 17303618]

[89]. Fridley BL, Biernacka JM, Gene set analysis of SNP data: benefits, challenges, and future directions, Eur. J. Hum. Genet 19 (2011) 837–843. [PubMed: 21487444]

[90]. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U. S. A 102 (2005) 15545–15550. [PubMed: 16199517]

[91]. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinformatics 14 (2013) 128. [PubMed: 23586463]

[92]. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, Nucleic Acids Res. 44 (2016) W90–97. [PubMed: 27141961]

[93]. Huang da W, Sherman BT, Lempicki RA, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc 4 (2009) 44–57. [PubMed: 19131956]

[94]. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA, DAVID: database for annotation, visualization, and integrated discovery, Genome Biol. 4 (2003) P3. [PubMed: 12734009]

[95]. Ferreira BI, Alonso J, Carrillo J, Acquadro F, Largo C, Suela J, Teixeira MR, Cerveira N, Molares A, Gomez-Lopez G, et al., Array CGH and gene-expression profiling reveals distinct genomic instability patterns associated with DNA repair and cell-cycle checkpoint pathways in Ewing's sarcoma, Oncogene 27 (2008) 2084–2090. [PubMed: 17952124]

[96]. Kikuta K, Tochigi N, Shimoda T, Yabe H, Morioka H, Toyama Y, Hosono A, Beppu Y, Kawai A, Hirohashi S, et al., Nucleophosmin as a candidate prognostic biomarker of Ewing's sarcoma revealed by proteomics, Clin. Cancer Res 15 (2009) 2885–2894. [PubMed: 19351769]

[97]. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC, A global test for groups of genes: testing association with a clinical outcome, Bioinformatics 20 (2004) 93–99. [PubMed: 14693814]

[98]. Biernacka JM, Jenkins GD, Wang L, Moyer AM, Fridley BL, Use of the gamma method for self-contained gene-set analysis of SNP data, Eur. J. Hum. Genet 20 (2012) 565–571. [PubMed: 22166939]

[99]. Fridley BL, Jenkins GD, Grill DE, Kennedy RB, Poland GA, Oberg AL, Soft truncation thresholding for gene set analysis of RNA-seq data: application to a vaccine study, Sci. Rep 3 (2013) 2898. [PubMed: 24104466]

[100]. de Rooij JD, Branstetter C, Ma J, Li Y, Walsh MP, Cheng J, Obulkasim A, Dang J, Easton J, Verboon LJ, et al., Pediatric non-Down syndrome acute megakaryoblastic leukemia is characterized by distinct genomic subsets with varying outcomes, Nat. Genet 49 (2017) 451–456. [PubMed: 28112737]

[101]. Saelens W, Cannoodt R, Saeys Y, A comprehensive evaluation of module detection methods for gene expression data, Nat. Commun 9 (2018) 1090. [PubMed: 29545622]

[102]. Werhli AV, Grzegorczyk M, Husmeier D, Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks, Bioinformatics 22 (2006) 2523–2531. [PubMed: 16844710]

[103]. Grzegorczyk M, Extracting protein regulatory networks with graphical models, Proteomics 1 (2007) 51–59.

[104]. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS, Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, Proc. Natl. Acad. Sci. U. S. A 97 (2000) 12182–12186. [PubMed: 11027309]

[105]. Langfelder P, Horvath S, WGCNA: an R package for weighted correlation network analysis, BMC Bioinformatics 9 (2008) 559. [PubMed: 19114008]

[106]. Zhang B, Horvath S, A general framework for weighted gene co-expression network analysis, Stat. Appl. Genet. Mol. Biol 4 (2005) Article17.

[107]. Yip AM, Horvath S, Gene network interconnectedness and the generalized topological overlap measure, BMC Bioinformatics 8 (2007) 22. [PubMed: 17250769]

[108]. Wang X, Song P, Huang C, Yuan N, Zhao X, Xu C, Weighted gene coexpression network analysis for identifying hub genes in association with prognosis in Wilms tumor, Mol. Med. Rep 19 (2019) 2041–2050. [PubMed: 30664180]

[109]. Yuan L, Qian G, Chen L, Wu CL, Dan HC, Xiao Y, Wang X, Co-expression network analysis of biomarkers for adrenocortical carcinoma, Front. Genet 9 (2018) 328. [PubMed: 30158955]

[110]. Zhang H, Guo L, Zhang Z, Sun Y, Kang H, Song C, Liu H, Lei Z, Wang J, Mi B, et al., Co-expression network analysis identified gene signatures in Osteosarcoma as a predictive tool for lung metastasis and survival, J. Cancer 10 (2019) 3706–3716. [PubMed: 31333788]

[111]. Schafer J, Strimmer K, An empirical Bayes approach to inferring large-scale gene association networks, Bioinformatics 21 (2005) 754–764. [PubMed: 15479708]

[112]. Zhao H, Duan ZH, Cancer genetic network inference using gaussian graphical models. Bioinform. Biol. Insights 13 (2019) 1177932219839402. [PubMed: 31007526]

[113]. Friedman N, Linial M, Nachman I, Pe'er D, Using bayesian networks to analyze expression data, J. Comput. Biol 7 (2000) 601–620. [PubMed: 11108481]

[114]. Ni Y, Muller P, Wei L, Ji Y, Bayesian graphical models for computational network biology, BMC Bioinformatics 19 (2018) 63. [PubMed: 29589555]

[115]. Bulashevska S, Bulashevska A, Eils R, Bayesian statistical modelling of human protein interaction network incorporating protein disorder information, BMC Bioinformatics 11 (2010) 46. [PubMed: 20100321]

[116]. Hill SM, Lu Y, Molina J, Heiser LM, Spellman PT, Speed TP, Gray JW, Mills GB, Mukherjee S, Bayesian inference of signaling network topology in a cancer cell line, Bioinformatics 28 (2012) 2804–2810. [PubMed: 22923301]

[117]. Kramer N, Schafer J, Boulesteix AL, Regularized estimation of large-scale gene association networks using graphical Gaussian models, BMC Bioinformatics 10 (2009) 384. [PubMed: 19930695]

[118]. Yin J, Li H, A sparse conditional gaussian graphical model for analysis of genetical genomics data, Ann. Appl. Stat 5 (2011) 2630–2650. [PubMed: 22905077]

[119]. Chun H, Zhang X, Zhao H, Gene regulation network inference with joint sparse Gaussian graphical models, J. Comput. Graph. Stat 24 (2015) 954–974. [PubMed: 26858518]

[120]. Blum Y, Houee-Bigot M, Causeur D, Sparse factor model for co-expression networks with an application using prior biological knowledge, Stat. Appl. Genet. Mol. Biol 15 (2016) 253–272. [PubMed: 27166726]

[121]. Serra A, Coretto P, Fratello M, Tagliaferri R, Stegle O, Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data, Bioinformatics 34 (2018) 625–634. [PubMed: 29040390]

[122]. Schafer J, Strimmer K, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, Stat. Appl. Genet. Mol. Biol 4 (2005) Article32.

[123]. Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HK, Frigessi A, Borresen-Dale AL, Principles and methods of integrative genomic analyses in cancer, Nat. Rev. Cancer 14 (2014) 299–313. [PubMed: 24759209]

[124]. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S, A selective review of multi-level omics data integration using variable selection, High Throughput (2019) 8.

[125]. Jiang Y, Shi X, Zhao Q, Krauthammer M, Rothberg BE, Ma S, Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis, Genomics 107 (2016) 223–230. [PubMed: 27141884]

[126]. Zhao Q, Shi X, Xie Y, Huang J, Shia B, Ma S, Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA, Brief Bioinform 16 (2015) 291–303. [PubMed: 24632304]

[127]. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al., Mutational landscape and significance across 12 major cancer types, Nature 502 (2013) 333–339. [PubMed: 24132290]

[128]. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhsng CZ, Wala J, Mermel CH, et al., Pan-cancer patterns of somatic copy number alteration, Nat. Genet 45 (2013) 1134–1140. [PubMed: 24071852]

[129]. Chen H, Li C, Peng X, Zhou Z, Weinstein JN, Cancer N Genome Atlas Research, H. Liang, A pan-cancer analysis of enhancer expression in nearly 9000 patient samples, Cell 173 (2018) 386–399 e312. [PubMed: 29625054]

[130]. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al., Oncogenic signaling pathways in the Cancer genome atlas, Cell 173 (321-337) (2018) e310.

[131]. Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ, Pan-cancer analysis of transcriptional metabolic dysregulation using the cancer genome atlas, Nat. Commun 9 (2018) 5330. [PubMed: 30552315]

[132]. Network CGA, Comprehensive molecular portraits of human breast tumours, Nature 490 (2012) 61–70. [PubMed: 23000897]

[133]. Radovich M, Pickering CR, Felau I, Ha G, Zhang H, Jo H, Hoadley KA, Anur P, Zhang J, McLellan M, et al., The integrated genomic landscape of thymic epithelial tumors, Cancer Cell 33 (244-258) (2018) e210.

[134]. Shen R, Olshen AB, Ladanyi M, Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis, Bioinformatics 25 (2009) 2906–2912. [PubMed: 19759197]

[135]. Shen R, Wang S, Mo Q, Sparse integrative clustering of multiple omics data sets, Ann. Appl. Stat 7 (2013) 269–294. [PubMed: 24587839]

[136]. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R, Pattern discovery and cancer gene identification in integrated cancer genomic data, Proc. Natl. Acad. Sci. U. S. A 110 (2013) 4245–4250. [PubMed: 23431203]

[137]. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG, A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data, Biostatistics 19 (2018) 71–86. [PubMed: 28541380]

[138]. Brunet JP, Tamayo P, Golub TR, Mesirov JP, Metagenes and molecular pattern discovery using matrix factorization, Proc. Natl. Acad. Sci. U. S. A 101 (2004) 4164–4169. [PubMed: 15016911]

[139]. Gao Y, Church G, Improving molecular cancer class discovery through sparse non-negative matrix factorization, Bioinformatics 21 (2005) 3970–3975. [PubMed: 16244221]

[140]. Kim H, Park H, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, Bioinformatics 23 (2007) 1495–1502. [PubMed: 17483501]

[141]. Monti S, Tamayo P, Mesirov J, Golub T, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, Mach. Learn. 52 (2003) 91–118.

[142]. Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ, Discovery of multi-dimensional modules by integrative analysis of cancer genomic data, Nucleic Acids Res. 40 (2012) 9379–9391. [PubMed: 22879375]

[143]. Yang Z, Michailidis G, A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data, Bioinformatics 32 (2016) 1–8. [PubMed: 26377073]

[144]. Chalise P, Fridley BL, Integrative clustering of multi-level' omic data based on non-negative matrix factorization algorithm, PLoS One 12 (2017) e0176278. [PubMed: 28459819]

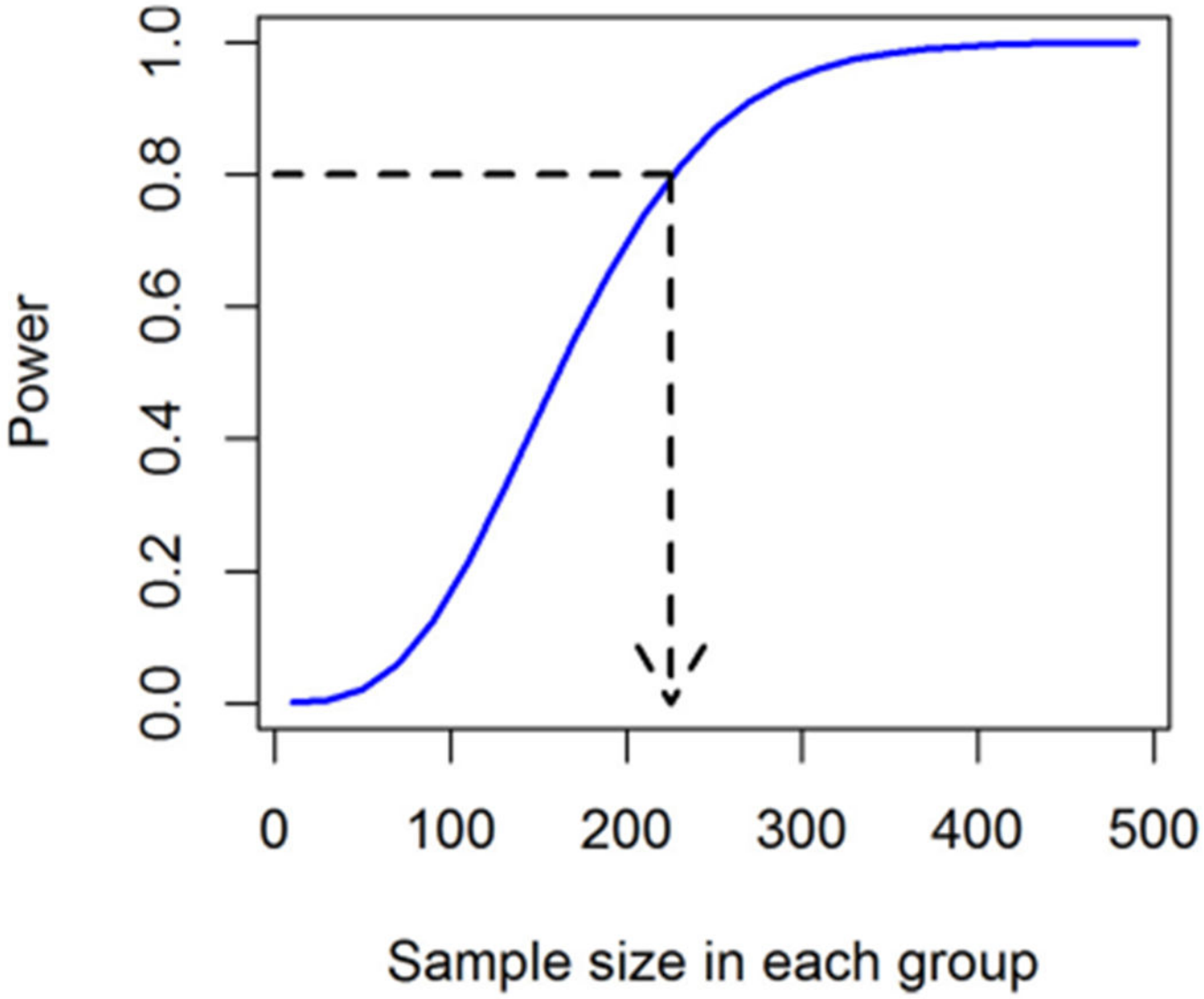

Fig. 1.

Power to detect a moderate effect size (0.50) as a function of power and sample size, with a significance level set to 0.00001 to account for the testing of multiple genes.

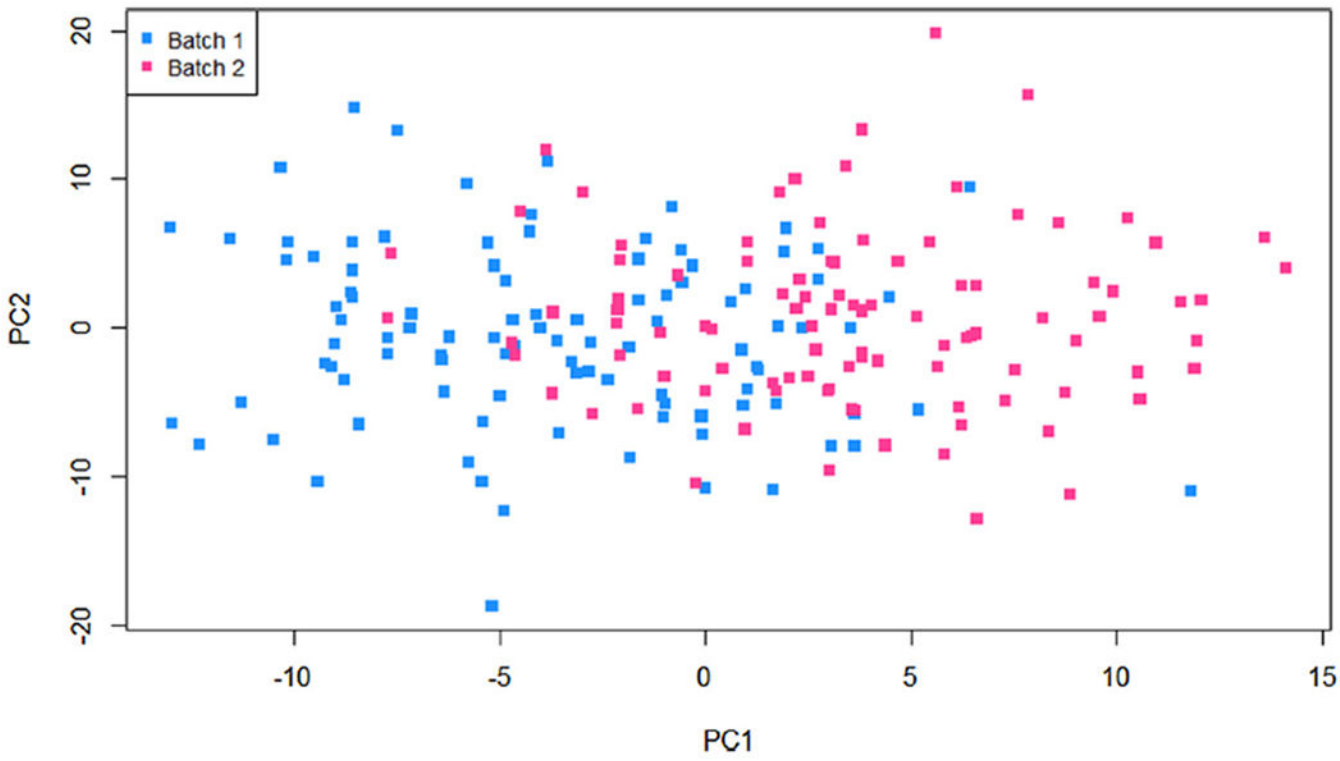

Fig. 2.

Plot of first and second principal components (PCs) from principal component analysis (PCA) for visualization of batch effects.

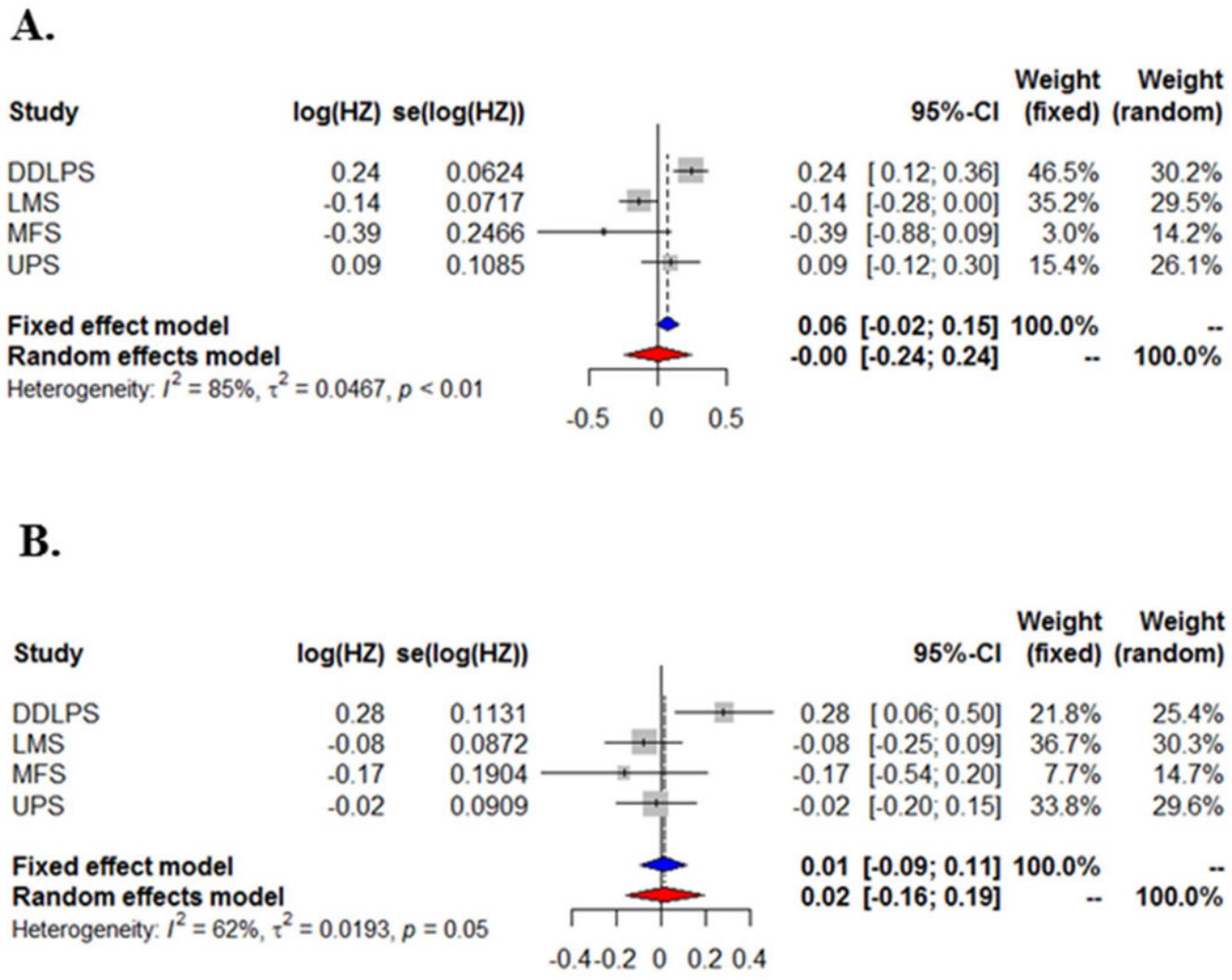

Fig. 3.

Forest plots for TCGA SARC study to assess true heterogeneity across 4 types of sarcoma (DDLPS, LMS, MFS, and UPS) using fixed- and random-effects models for overall survival with gene expression for (A) *TRPV6* and (B) *KIF21A*.

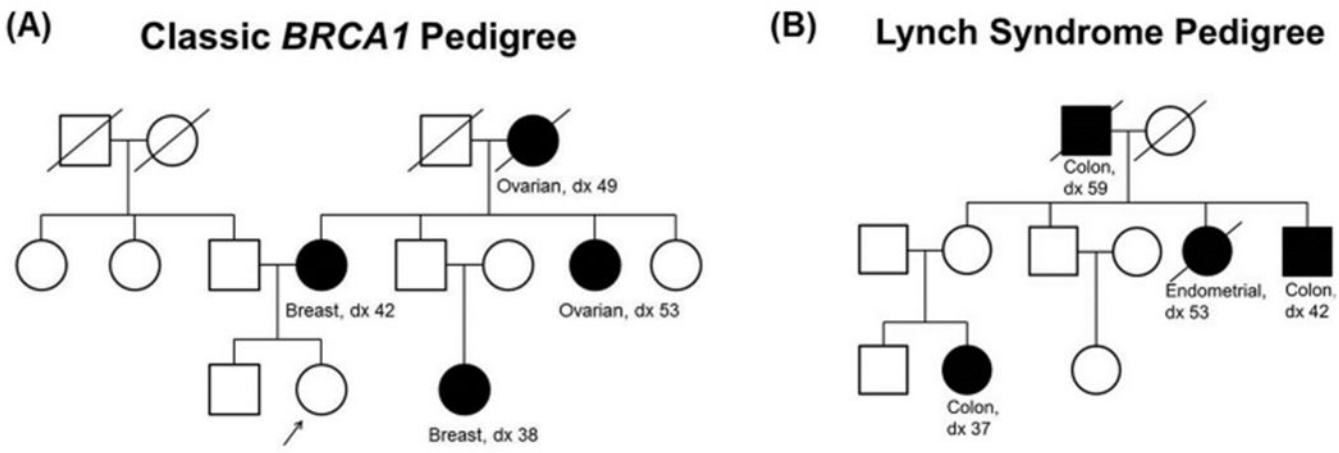

Fig. 4.

Example pedigrees of for two inherited cancers. (A) *BRCA1* related breast and ovarian cancer and (B) Lynch Syndrome.

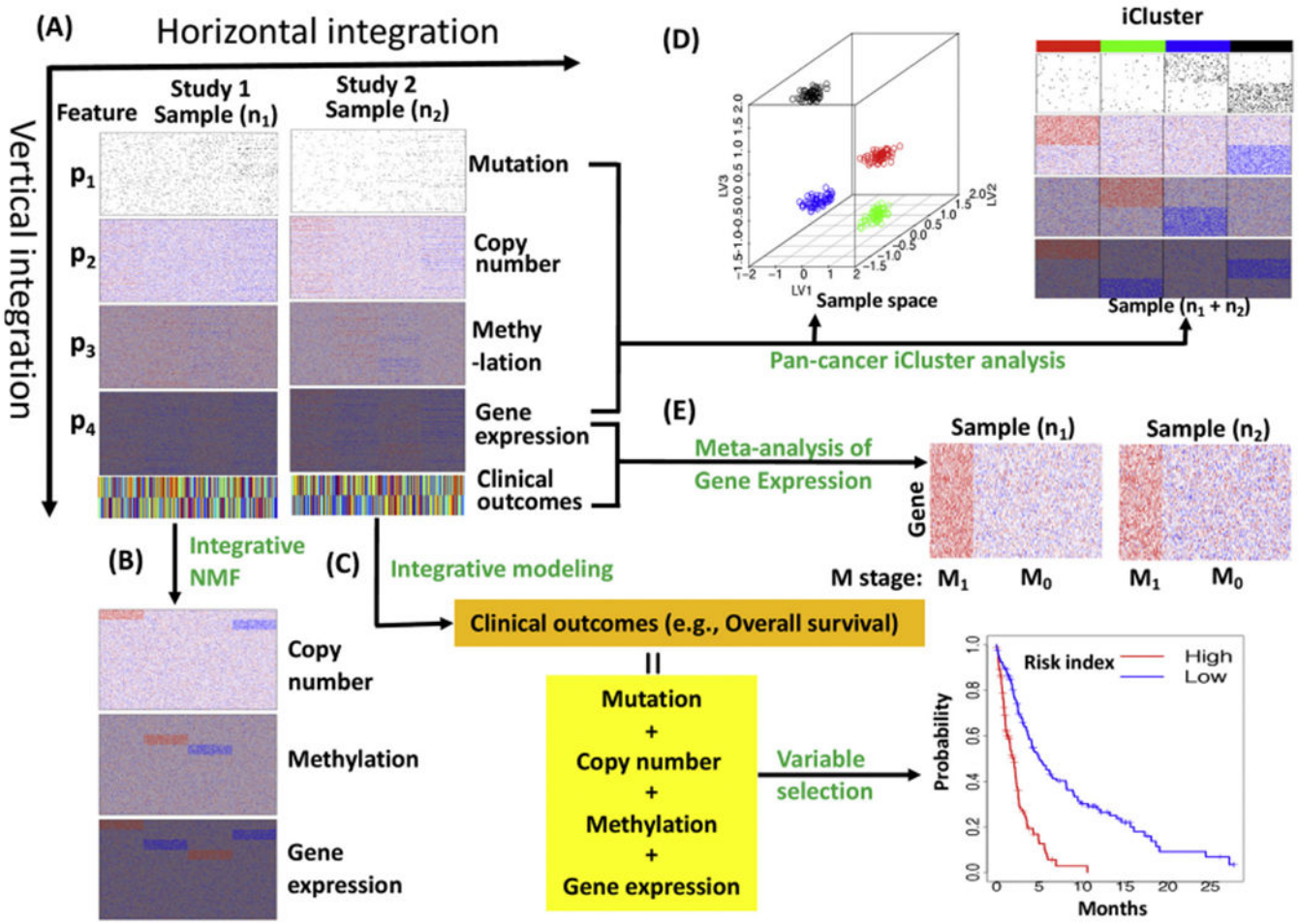

Fig. 5.

A scheme of integrative analyses of multi-omics data. (A) multi-omics data including mutation, DNA copy number, methylation and gene expression for two TCGA-like studies. (B) Unsupervised vertical integration analysis of copy number, methylation and gene expression using integrative NMF. (C) Supervised vertical integration analysis of mutation, copy number, methylation and gene expression. (D) Unsupervised pan-cancer iCluster analysis (vertical and horizontal integration). (E) Supervised horizontal integration analysis (Meta-analysis) of gene expression for pathological metastasis (M) stages. M1: metastasis, M0: no metastasis.

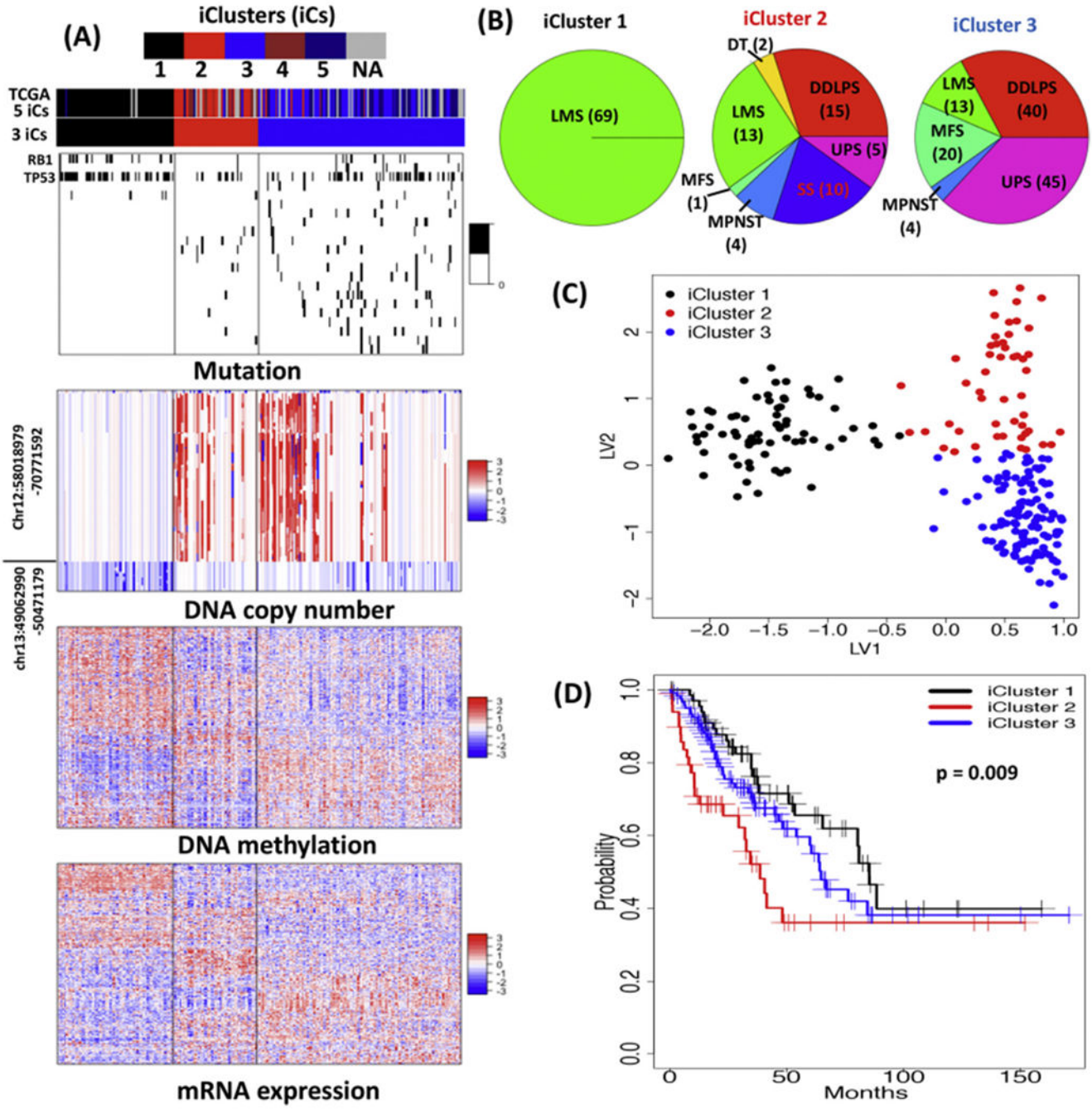

Fig. 6.

SARC iClusters. (A) Heatmaps of the driver omics features. The color bars on the heatmaps indicate the 3 iClusters and TCGA 5 iClusters12. Heatmaps from the top to bottom are for the mutation (black: mutated; white: normal), copy number (red: trend to amplification; white: normal; blue: trend to deletion), methylation (red: hyper-methylated; blue: hypo-methylated) and gene expression (red: high expression; blue: low expression), respectively. The driver genes from top to bottom on the mutation heatmap are *RB1, RYR2, TP53, CAC1F, PTEN, MKI67, MYH7, MGAM, TRDN, KRTAP5-5, FAT3, CFTR, ZNF831,*

MYOCD, WNK2, DCC, PEG3, DSCAM, ASTN2, PAPPA, DH6, MYHB. (B) Numbers of sarcoma distributed in the 3 iClusters. DT: desmoid tumor; MPNST: malignant peripheral nerve sheath tumor. (C) Sample clusters on the two-dimensional latent spaces. LV: latent variable. (D) Kaplan-Meir survival curves of the 3 iClusters and Log-rank test p-value.