Using Single Nucleotide Variations in Cancer Single-Cell RNA-Seq Data for Subpopulation Identification and Genotype-phenotype Linkage Analysis

Olivier Poirion[1], Xun Zhu[1, 2], Travers Ching[1, 2], Lana Garmire[1, 2 *]

[1] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA.

[2] Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI 96822, USA.

[*] To whom correspondence should be addressed. Email address: lgarmire@cc.hawaii.edu

Abstract

Characterization of subpopulations is a key challenge in the emerging field of single-cell RNA-seq (scRNA-seq). In scRNA-seq data, gene expression has been used as feature to detect subpopulations, however, this data type is subject to significant amount of noise. Unconventionally, we propose to use filtered, effective and expressed nucleotide variations (eeSNVs) information from scRNA-seq data as improved predictive features for subpopulation identification. We developed a linear modeling framework called SSrGE (Sparse SNV inference to reflect Gene Expression) to detect eeSNVs that are associated with gene expression profiles. In all the datasets tested, these eeSNVs show better accuracy than gene expression for retrieving cell subpopulations. Moreover, bipartite graphs of cells in combination with eeSNVs have better visual representation of the different cell subpopulations than the other methods that use gene expression data. We ranked genes, according to their eeSNVs in the cancer scRNA-seq data, and found that genes in antigen processing and presentation pathway have top ranked eeSNVs. They include HLA-A, HLA-B, HLA-C and HLA-DRA in Human Leukocyte Antigen (HLA) complex, and B2M in histocompatibility complex MHC. Moreover, previously experimentally validated cancer relevant genes, such as KRAS and SPARC, are highly ranked for eeSNVs too. In summary, we emphasize that SNV features hidden in scRNA-seq data have merits for both subpopulation identification and linkage of genotype-phenotype relationship. The computational method is freely available at https://github.com/lanagarmire/SSrGE.

Introduction

Characterization of phenotypic diversity is a key challenge in the emerging field of single-cell RNA-sequencing (scRNA-seq). In scRNA-seq data, patterns of gene expression (GE) are conventionally used as features to explore the heterogeneity among single cells (Harris et al., 2015; Pierson & Yau, 2015; Vallejos, Marioni, &

Richardson, 2015). However, GE features are subject to significant amount of noises (Kolodziejczyk, Kim, Svensson, Marioni, & Teichmann, 2015). One issue of GE is the batch effect, where results obtained from two different runs of experiments may present substantial variations (Stegle, Teichmann, & Marioni, 2015), even when the input materials are identical. Additionally, the expression of certain genes vary with cell cycle (Buettner et al., 2015), increasing the heterogeneity observed in single cells. To cope with these sources of variations, normalization of GE is usually a mandatory step before downstream functional analysis (except those done with Unique Molecular Identifiers). Even with these procedures, other sources of biases still exist, e.g. dependent on read depth, cell capture efficiency and experimental protocols etc.

Single nucleotide variations (SNVs) are small genetic alterations occurring in specific cells as compared to the population background. SNVs may manifest their effects on gene expression per *cis* and/or *trans* effect (Bryois et al., 2014; Hu, Lan, Xu, Beyene, & Greenwood, 2007). It is regarded that cancer evolution involves the disruption of the genetic stability e.g. increasing number of new SNVs (Berdasco & Esteller, 2010; Navin et al., 2011). A cell may become the precursor of a subpopulation (clone) upon gaining a set of SNVs. Large heterogeneity exists not only between tumors but also within the same tumor (Almendro, Marusyk, & Polyak, 2013; Burrell, McGranahan, Bartek, & Swanton, 2013). Therefore, investigating the patterns of SNVs provides means to understand tumor heterogeneity.

In single cells, SNVs are conventionally obtained from the single-cell exome-sequencing approach (Zafar, Wang, Nakhleh, Navin, & Chen, 2016). Previously, the resulting SNVs were used to infer cancer cell subpopulations (Ross & Markowetz, 2016; Welch, Hartemink, & Prins, 2016). In this study, we propose to obtain useful SNV-based genetic information from scRNA-seq data, in addition to the GE information. Rather than being considered the “by-products” of scRNA-seq, the SNVs not only have the potential to improve the accuracy of identifying subpopulations compared to GE, but also offer unique opportunities to study the genetic events (genotype) that are associated with gene expression (phenotype) (Gamazon et al., 2015; Pineda et al., 2015). Moreover, when the coupled DNA- and RNA- based single-cell sequencing techniques become mature, the computational methodology proposed in this report can be easily adopted as well.

Here we first built a computational pipeline to identify SNVs from scRNA-seq raw reads directly. We then

constructed a linear modeling framework to obtain filtered, effective and expressed SNVs (eeSNVs) associated with gene expression profiles. In all the datasets tested, these eeSNVs show better accuracies at retrieving cell subpopulation identities, compared to those from gene expression (GE). Moreover, when combined with cell entities into bipartite graphs, they demonstrate improved visual representation of the cell subpopulations. We ranked eeSNVs and genes according to their overall significance in the linear models, and discovered that several top-ranked genes (e.g. HLA genes) appear commonly in all cancer scRNA-seq data. In summary, we emphasize the SNV approach that were previously understudied in scRNA-seq analysis, which can successfully identify subpopulation complexities and highlight genotype-phenotype relationships.

Results

eeSNV detection from scRNA-seq data

We implemented a pipeline to identify SNVs directly from FASTQ files of scRNA-seq data, following the SNV guideline of GATK (Suppl. Figure S1). We applied this pipeline to four scRNA-seq cancer datasets (Kim, Ting, Miyamoto and Patel, see Methods), and tested the efficiency of SNV features on retrieving single cell groups of interest. These datasets vary in tissue types, origins (Mouse or Human), read lengths and map-ability (Table 1). They all have pre-defined cell types (subclasses), providing good references to assess the performance of a variety of clustering methods used in this study.

To link the relationship between SNV and GE, we developed a method called “Sparse SNV inference to reflect Gene Expression” (SSrGE), as detailed in Materials and Methods. This method uses SNVs as predictors to fit a linear model for gene expression, under LASSO regularization and feature selection (Tibshirani, 1996). The output is a subset of effective, expressed SNVs (eeSNVs) selected by LASSO, which serve as refined descriptive features for subsequent subpopulation identification (Suppl. Figure S2). To directly pinpoint the contributions of SNVs relevant to protein coding genes, we used the SNVs residing between transcription starting and ending sites of genes as the inputs. In SSrGE, the value of the regularization parameter a is the only tuning variable, controlling the sparsity of the linear models and influences the number of eeSNVs.

eeSNVs are better features than gene expression to identify subpopulations

We measured the performance of SNVs and gene expression (GE) in the four datasets with five clustering approaches. These clustering approaches include two dimension reduction methods, namely Principal Component Analysis (PCA) (I T, 2002) and Factor Analysis (FA) (Cattell, 1952), followed by either K-Means or the hierarchical agglomerative method (agglo) with WARD linkage (Joe H Ward, 1963). We also used a recent algorithm SIMLR specifically designed for scRNA-seq data clustering and visualization (Wang et al., 2016). To evaluate the accuracy of obtained subpopulations in each dataset, we used the metric of Adjusted Mutual Information (AMI) over 30 bootstrap runs, from the optimal *a* parameters (Supplementary Table S1). Even though the numbers are much reduced from the original SNVs, eeSNVs are still better features to retrieve cancer cell subpopulations compared to GE, independent of the clustering methods used (Figure 1). Among the clustering algorithms, SIMLR is a better choice in general using eeSNV features. In addition, we also computed Adjusted Rand index (ARI) (Vinh, Epps, & Bailey, 2010) and V-measure (Rosenberg & Hirschberg, 2007), two other metrics for modularity measurements and obtained similar trends (Suppl. Figure S3).

Visualization of subpopulations with bipartite graphs

Bipartite graphs are useful to represent binary relations between two different classes of objects. We next represented the binary eeSNVs features and the single cells with bipartite graphs using ForceAtlas2 algorithm (Bastian, Heymann, & Jacomy, 2009). We drew an edge (link) between a cell node and a given eeSNV node whenever an eeSNV is detected. The results show that bipartite graph is a robust and more discriminative alternative (Figure 2), comparing to PCA plots (using GE and eeSNVs) as well as SIMLR (using GE). For Kim dataset, bipartite graph separates the three classes perfectly. However, gene based visualization approaches using either PCA or SIMLR have misclassifications. For Ting data, eeSNVs-cell bipartite graph gives clear visualization of all six different subgroups of single cells. Other three approaches have more exaggerated separations among the same mouse circulating tumor cells (CTC) subgroup MP (orange color), but mix some other subpopulations (e.g. GM, MP and TuGMP groups). Miyamoto dataset is the most difficult one to visualize among the four datasets, due to its high number (24) of reference classes and heterogeneity among CTCs. Bipartite graph is not only able to condense the whole populations, but also separate subpopulations (e.g. the orange colored PC subpopulation) much better than the other three methods.

Characteristics of eeSNVs

Since the selection of eeSNVs is dependent on regularization parameter *a*, we next explored their relationship. For every dataset, increasing the value of *a* decreases the number of selected eeSNVs overall (Figure 3A), as well as the average number of eeSNVs associated with every expressed gene (Figure 3B). The optimal *a* depends on the clustering algorithm and the dataset used (Suppl. Table S1 and Suppl. Figure S4). Increasing the value of *a* increases the proportion of eeSNVs that have annotations in human dbSNP138 database, indicating that these eeSNVs are biologically valid (Figure 3C). Finally, increasing *a* generally increases the average number of cells sharing the same eeSNVs, supporting the hypothesis that cancer cells differentiate with a growing number of genetic mutations over time (Figure 3D). Note the slight drop of the average number of cells sharing the same eeSNVs in Kim data when $a > 0.6$, this is due to over-penalization (eg. $a = 0.8$ yields only 34 eeSNVs).

Cancer relevance of eeSNVs

To further explore the biological functions, we ranked the different eeSNVs and the genes harboring them, using eeSNVs' coefficients from SSrGE models (Suppl. Tables S2). We found that eeSNVs from multiple genes in Human Leukocyte Antigen (HLA) complex, such as HLA-A, HLA-B, HLA-C and HLA-DRA, are top ranked in all three human datasets (Table 2 and Suppl. Tables S2). HLA is a family encoding the major histocompatibility complex (MHC) proteins in human. Beta-2-microglobulin (B2M), on the other hand, is ranked 7th and 45th in Ting and Patel datasets, respectively (Table 2). Unlike HLA that is present in human only, B2M encodes a serum protein involved in the histocompatibility complex MHC that is also present in mice. Other previously identified tumor driver genes are also ranked top by SSrGE, demonstrating the significance of mutations on cis-gene expression (Table 2 and Suppl. Tables S2). Notably, *KRAS*, previously linked to tumor heterogeneity by the original scRNA-Seq study (Kim et al., 2015), is ranked 13th among all eeSNV containing genes (Suppl. Tables S2). *AR* and *KLK3*, two genes reported to show genomic heterogeneity in tumor development in the original study (Miyamoto et al., 2015), are ranked 6th and 19th, respectively. *EGFR*, the therapeutic target in Patel study with an important oncogenic variant EGFRvIII (Patel et al., 2014), is ranked 88th out of 4,225 genes (Supplementary File 1). Therefore, genes top-ranked by their eeSNVs are empirically validated.

Next we conducted more systematic investigation to identify KEGG pathways enriched in each dataset, using

these genes as the input for DAVID annotation tool (Huang, Sherman, & Lempicki, 2009) (Figure 4). The pathway-gene bipartite graph illustrates the relationships between these genes and enriched pathways (Figure 5). As expected, Antigen processing and presentation pathway stands out as the most enriched pathway, with the sum -log10 (p-value) of 9.22. “Phagosome” is the second most enriched pathway in all four data sets (Figure 4), largely due its members in HLA families (Figure 5). Additionally, pathways related to cell junctions and adhesion (focal adhesion, tight junction, cell adhesion molecules CAMs), protein processing (protein processing in endoplasmic reticulum and proteasome), and PI3K-AKT signaling pathway are also highly enriched with eeSNVs (Figure 5).

Discussion

Using GE to accurately analyze scRNA-seq data has many challenges, including technological biases such as the choice of the sequencing platforms, the experimental protocols and conditions. These biases may lead to various confounding factors in interpreting GE data (Stegle et al., 2015). SNVs, on the other hand, are less prone to these issues given their binary nature. In this report, we demonstrate that eeSNVs extracted from scRNA-seq data are ideal features to characterize cell subpopulations. Moreover, they provide a means to examine the relationship between eeSNVs and gene expression in the same scRNA-Seq sample.

eeSNVs have improved accuracy on identifying tumor single-cell subpopulations

The process of selecting eeSNVs linked to GE allows us to identify representative genotype markers for cell subpopulations. We speculate the following reasons attributed to the better accuracies of eeSNVs compared to GE. First, eeSNVs are binary features rather than continuous features like GE, thus eeSNVs are more robust at separating subpopulations. We have noticed that SNVs are less affected by batch effects (Suppl. Figure S5). Secondly, LASSO penalization works as a feature selection method and minimizes the spurious SNVs (false positive) from the filtered set of eeSNVs. Thirdly, since eeSNVs are obtained from the same samples as scRNA-seq data, they are more likely to have biological impact, and this is supported the observation that they have high prevalence of dbSNP annotations.

eeSNVs have improved efficiency at identifying tumor single-cell subpopulations

A small number of eeSNVs can be used to discriminate distinct single-cell subpopulations, as compared to

thousands of genes that are normally used for scRNA-seq analyses. Taking advantage of the eeSNV-GE relationship, a very small number of top eeSNVs still can clearly separate cell subpopulations of the different datasets (eg. 8 eeSNV features have decent accuracy for Kim dataset). Moreover, our SSrGE package can be easily parallelized and process each gene independently. It has the potential to scale up to (very) large datasets, well-poised for the new wave of scRNA-Seq technologies that can generate thousands of cells at one time (Tirosh et al., 2016). One can also easily rank the eeSNVs and the genes harboring them, for the purpose of identifying robust eeSNVs as genetic markers for a variety of cancers.

eeSNVs highlights genes linked to cancer phenotypes

SSrGE uses an accumulative ranking approach to select eeSNVs linked to the expression of a particular gene. Particularly, HLA class I genes (HLA-A, HLA-B and HLA-C) are top-ranked for the three human datasets, and they contribute to "antigen processing and presentation pathway", the most enriched pathways of the four datasets. HLA has amongst the highest polymorphic genes of the human genome (de Bakker et al., 2006), and the somatic mutations of genes in this family were reported in the development and progression of various cancers (Network & others, 2014; Shukla et al., 2015). HLA genes with eeSNVs could be used as fingerprints to characterize the cellular state of the cancer cells. *B2M*, another gene with top-scored eeSNVs in Ting and Patel datasets, is also known to be a mutational hotspot (Chang, Campoli, Restifo, Wang, & Ferrone, 2005). It is directly linked to immune response as tumor cell proliferation (Chang et al., 2005; Network & others, 2014). Many other top-ranked genes, such as *KRAS* and *SPARC*, were reported to be driver genes in the original studies of the different dataset. Thus, it is reasonable to speculate that SSrGE is capable of identifying some driver genes. However, SSrGE may miss some driver mutations, since its primary goal is to identify a minimal set of eeSNV features by LASSO penalization and LASSO may select one of those highly correlated SNV features that correspond to GE.

Advantages of using bipartite graphs to represent scRNA-seq data

Bipartite graphs are a natural way to visualize eeSNV-cell relationships. We have used force-directed graph drawing algorithms involve spring-like attractive forces and electrical repulsions between nodes that are connected by edges. This approach has the advantage to reveal "outlier" single cells, with a small set of eeSNVs, compared to those distance-based approaches. Moreover, the bipartite representation also reveals directly the

relationship between single cells and the eeSNV features. Contrary to dimension reduction approaches such as PCA that requires linear transformation of features into principle components, bipartite graphs preserve all the binary information between cell and eeSNV. Graph analysis software such as Gephi (Bastian et al., 2009) or Cytoscape (Shannon et al., 2003) can be utilized to explore the bipartite relationships in an interactive manner.

Conclusion

We demonstrated the efficiency of using eeSNVs for cell subpopulation identification over multiple datasets. eeSNVs are excellent genetic markers for intra-tumor heterogeneity and may serve as genetic candidates of new treatment options. We also have developed SSrGE, a linear model framework that correlates genotype (eeSNV) and phenotype (GE) information in scRNA-seq data. This method has the potential to be routinely integrated in future scRNA-seq analyses. Moreover, SSrGE can also be used to analyze coupled DNA- and RNA-based single-cell sequencing techniques on the horizon (Dey, Kester, Spanjaard, Bienko, & van Oudenaarden, 2015; Kim et al., 2015).

Materials and Methods

scRNA-seq datasets

All four datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) portal (Barrett et al., 2013).

Kim dataset (accession GSE73121): contains three cell populations from matched primary and metastasis tumor from the same patient (Kim et al., 2015). Patient Derivated Xenographs (PDX) were constructed using cells from the primary Clear Cell Renal Cell Carcinoma (PDX-pRCC) tumor and from the lung metastasic tumor (PDX-mRCC). Also, metastasic cells from the patient (Pt-mRCC) were sequenced.

Patel dataset (accession GSE57872): contains five glioblastoma cell populations isolated from 5 individual tumors from different patients (MGH26, MGH28, MGH29 MGH30 and MGH31) and two gliomasphere cell lines, CSC6 and CSC8, used as control (Patel et al., 2014).

Miyamoto dataset (accession GSE67980): contains 122 CTCs from Prostate cancer from 18 patients, 30 single cells derivated from 4 different cancer cell lines: VCaP, LNCaP, PC3 and DU145, and 5 leukocyte cells from a

healthy patient (HD1) (Miyamoto et al., 2015). A total of 23 classes (18 CTC classes + 4 cancer cell lines + 1 healthy leukocyte cell lines) was obtained.

Ting dataset (subset of accession GSE51372): contains 75 CTCs from Pancreatic cancer from 5 different KPC mice (MP2, MP3, MP4, MP6, MP7), 18 CTCs from two GFP-lineage traced mice (GMP1 and GMP2), 20 single cells from one GFP-lineage traced mouse (TuGMP3), 12 single cells from a mouse embryonic fibroblast cell line (MEF), 12 single cells from mouse white blood (WBC) and 16 single cells from the nb508 mouse pancreatic cell line (nb508) (Ting et al., 2014). KPC mice have uniform genetic cancer drivers (Tp53, Kras). Due to their shared genotype, we merged all the KPC CTCs into one single reference class. CTCs from GMP1 did not pass the QC test and were dismissed. CTCs from GMP2 mice were labeled as GMP. Finally, 6 reference classes were used: MP, nb508, GMP, TuGMP, MEF and WBC.

SNV detection using scRNA-Seq data

The SNV detection pipeline using scRNA-Seq data follows the guidelines of GATK (http://gatkforums.broadinstitute.org/wdl/discussion/3891/calling-variants-in-rnaseq). It includes four steps: alignment of spliced transcripts to the reference genome (hg19 or mm10), BAM file preprocessing, read realignment and recalibration, and variant calling and filtering (Suppl. Figure S1) (Auwera et al., 2013).

Specifically, FASTQ files were first aligned using STAR aligner (Dobin & Gingeras, 2015), using mm10 and hg19 as reference genomes for mouse and human datasets, respectively. The BAM file quality check was done by FastQC (Andrews & others, 2010), and samples with lower than 50% of unique sequences were removed (default of FastQC). Also, samples with more than 20% of the duplicated reads were removed by STAR. Finally, samples with insufficient reads were also removed, if their reads were below the mean minus two times the standard deviation of the entire single-cell population. Raw gene counts X_j were estimated using featureCounts (Liao, Smyth, & Shi, 2013), and normalized using the logarithmic transformation:

$$f(X_j) = \log_2(1 + X_j \cdot \frac{10^9}{(G_j.R)})$$

where X_j is the raw expression of gene j, R is the total number of reads and G_j is the length of the gene j. Bam files were pre-processed and reordered using Picard Tools (http://broadinstitute.github.io/picard/), before subject to realignment and recalibration using GATK tools (Guidot et al., 2007). SNVs are then calculated and filtered

using GATK tools using default parameters.

SNV annotation

To annotate human SNV datasets, dbSNP138 from the NCBI Single Nucleotide Polymorphism database (Sherry et al., 2001) and reference INDELs from 1000 genomes (1000_phase1 as Mills_and_1000G_gold_standard) (Consortium & others, 2015) were used. To annotate the mouse SNV dataset, dbSNPv137 for SNPs and INDELs were downloaded from the Mouse Genomes Project of the Sanger Institute, using the following link: ftp://ftp-mouse.sanger.ac.uk/REL-1303-SNPs_Indels-GRCm38/ (Keane et al., 2011). The mouse SNP databases were sorted using SortVcf command of Picard Tools, in order to be properly used by Picard Tools and GATK.

SSrGE package to calculate eeSNVs

For each dataset, we denote M_{SNV} and M_{GE} as the SNV and gene expression matrices, respectively. M_{SNV} is binary and $M_{SNV\,c,s} \in \{0, 1\}$ designates the presence/absence of SNV s in cell c. $M_{GE\,c,s}$ is the log transformed gene expression value of the gene g in cell c. A gene and its associated SNVs were only considered when the gene was expressed in at least one sample. Sparse linear regression using LASSO was then applied to identify W_g , the linear coefficients associated to the SNVs. The objective function (to minimize) is:

$$min_{W_g} \frac{1}{2n} ||M_{SNV} . W^T{}_g - M_{GE_{*,g}}||_2^2 + \alpha . ||W_g||_1$$

where α is the regularization parameter.

An SNV was considered as eeSNV when $W_g(s) \neq 0$. To derive sensible eeSNVs, the linear regression was only done on a particular gene, when at least 10 cells in the population expressed it.

Ranking of eeSNVs and genes

SSrGE generates coefficients of eeSNVs for each gene, as a metric for their contributions to the gene expression. The score of an eeSNV is given by the sum of its weights over all genes:

$$score_{eeSNV} = \sum_g |W^T{}_g(s)|$$

Each gene also receives a score according to its associated eeSNVs:

$$score_{gene_g} = \sum_{eeSNV \in gene_g} score_{eeSNV}$$

In practice, we first obtained eeSNVs using a minimum filtering of α=0.1, before using these two scores above to rank eeSNVs and the genes.

Subpopulation clustering algorithms

We combined two dimension reduction algorithms: Principal Component Analysis (PCA) (I T, 2002) and Factor Analysis (FA) (Cattell, 1952) with two popular clustering approaches: the K-Means algorithm (MacQueen & others, 1967) and agglomerative hierarchical clustering (agglo) with WARD linkage (Joe H Ward, 1963). We also used SIMLR, a recent algorithm specifically tailored to cluster and visualize scRNA-seq data, which learns the similarity matrix from subpopulations (Wang, Zhu, Pierson, & Batzoglou, 2016). Similar to the original SIMLR study, we used the embedding of the cells produced by the algorithm to apply K-Means algorithm.

PCA and FA were performed using their corresponding implementation in Scikit-Learn (*sklearn*) (Pedregosa, Weiss, & Brucher, 2011). For PCA, FA and SIMLR, we used various input dimensions *D* [2, 3, 5, 10, 15, 20, 25, 30] to project the data. To cluster the data with K-Means or the hierarchical agglomerative procedure, we used a different cluster numbers *N* (2 to 80) to obtain the best clustering results from each dataset. We computed accuracy metrics for each (*D*, *N)* pair and chose the combination that gives the overall best score. Between the two clustering methods, K-Means was the implementation of *sklearn* package with the default parameter, and hierarchical clustering was done by the *AgglomerativeClustering* implementation of *sklearn*, using WARD linkage.

Validation metrics

To assess the accuracy of the obtained clusters, we used three metrics: Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI) and V-measure (Rosenberg & Hirschberg, 2007; Vinh et al., 2010). These metrics compare the obtained clusters *C* to some reference classes *K* and generate scores between 0 and 1 for AMI and V-measure, and between -1 and 1 for ARI. A score of 1 means perfect match between the obtained clusters and the reference classes. For ARI, a score below 0 indicates a random clustering.

Rand Index (RI) was computed by: $RI = \frac{a+b}{C_2^{n_{sample}}}$, where a is the number of con-concordant sample pairs in

obtained clusters C and reference classes K, where as b is the number of dis-concordant samples. As an improvement, ARI normalizes RI against random chances: $ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}$ (Vinh et al., 2010).

AMI, similarly to ARI, normalizes Mutual Information (MI) against chances (Vinh et al., 2010). The Mutual Information between two sets of classes C and K is equal to: $\mathrm{MI}\ (C, K) = \sum_i^{|C|} \sum_j^{|K|} P(i, j) \log\left(\frac{P(i,j)}{P(i)P'(j)}\right)$, where $P(i)$ is the probability that an object from C belongs to the class i, $P'(j)$ is the probability that an object from K belongs to class j, and $P(i, j)$ is the probability that an object are in both class i and j. AMI is equal to: $AMI(C, K) = \frac{MI(C,K) - E(MI(C,K))}{\max(\{H(C), H(K)\}) - E(MI(C,K))}$, where $H(C)$ and $H(K)$ designates the entropy of C and K.

V-measure, similar to F-measure, calculates the harmonic mean between homogeneity and completeness. Homogeneity is defined as $1 - \frac{H(C|K)}{H(C)}$, where $H(C|K)$ is the conditional entropy of C given K. Completeness is the symmetrical of homogeneity: $1 - \frac{H(K|C)}{H(K)}$.

Graph visualization

The different datasets were transformed into GraphML files with Python scripts using iGraph library (Csardi & Nepusz, 2006). Graphs were visualized using GePhi software (Bastian et al., 2009) and spatialized using ForceAtlas2 (Jacomy, Venturini, & Bastian, 2011), a specific graph layout implemented into the GePhi software.

Pathway enrichment analysis

We used the KEGG pathway database to identify pathways related to specific genes (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2015). For each dataset, we selected the top 100 ranked genes by SSrGE and used the online interface of DAVID 6.8 functional annotation tool to identify significant pathways (Huang et al., 2009). We used the default significance value (adjusted p-value threshold of 0.10).

Code availability

The SNV calling pipeline and SSrGE are available through the following GitHub project: https://github.com/lanagarmire/SSrGE.

Acknowledgements

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by NIH/NIGMS, R01 LM012373 awarded by NLM, R01 HD084633 awarded by NICHD and Hawaii Community Foundation Medical Research Grant 14ADVC-64566 to L.X. Garmire. We acknowledge K. Chaudhary for manuscript proofreading.

Author contributions

LG envisioned this project. OP implemented the project and conducted genomics analysis, XZ and TC helped on implementation. OP and LG wrote the manuscript. All authors have read and agreed on the manuscript.

Competing financial interests

The author(s) declare no competing financial interests.

Bibliography

Almendro, V., Marusyk, A., & Polyak, K. (2013). Cellular heterogeneity and molecular evolution in cancer. *Annual Review of Pathology: Mechanisms of Disease*, *8*, 277–302. article.

Andrews, S., & others. (2010). FastQC: A quality control tool for high throughput sequence data. *Reference Source*. article.

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … others. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 10–11. article.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … others. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991--D995. article.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewPDFInterstitial/154Forum/1009

Berdasco, M., & Esteller, M. (2010). Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Developmental Cell*, *19*(5), 698–711. article.

Bryois, J., Buil, A., Evans, D. M., Kemp, J. P., Montgomery, S. B., Conrad, D. F., … others. (2014). Cis and trans effects of human genomic variants on gene expression. *PLoS Genet*, *10*(7), e1004461. article.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., … Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden

subpopulations of cells. *Nature Biotechnology*, *33*(2), 155–160. article.

Burrell, R. A., McGranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, *501*(7467), 338–345. article.

Cattell, R. B. (1952). Factor analysis: an introduction and manual for the psychologist and social scientist. article.

Chang, C.-C., Campoli, M., Restifo, N. P., Wang, X., & Ferrone, S. (2005). Immune selection of hot-spot $β$2-microglobulin gene mutations, HLA-A2 allospecificity loss, and antigen-processing machinery component down-regulation in melanoma cells derived from recurrent metastases following immunotherapy. *The Journal of Immunology*, *174*(3), 1462–1471. article.

Consortium, 1000 Genomes Project, & others. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. article.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, *Complex Sy*(1695), 1695. Retrieved from http://igraph.sf.net

de Bakker, P. I. W., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., … others. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nature Genetics*, *38*(10), 1166–1172. article.

Dey, S. S., Kester, L., Spanjaard, B., Bienko, M., & van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, *33*(3), 285–289. article.

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics*, 11–14. article.

Gamazon, E. R., Wheeler, H. E., Shah, K., Mozaffari, S. V, Aquino-Michaels, K., Carroll, R. J., … others. (2015). PrediXcan: Trait Mapping Using Human Transcriptome Regulation. *bioRxiv*, 20164. article.

Guidot, A., Prior, P., Schoenfeld, J., Carrère, S., Genin, S., & Boucher, C. (2007). Genomic structure and phylogeny of the plant pathogen Ralstonia solanacearum inferred from gene distribution analysis. *Journal of Bacteriology*, *189*(2), 377–87. http://doi.org/10.1128/JB.00999-06

Harris, K., Magno, L., Katona, L., Lönnerberg, P., Manchado, A. B. M., Somogyi, P., … Hjerling-Leffler, J. (2015). Molecular organization of CA1 interneuron classes. *bioRxiv*, 34595. article.

Hu, P., Lan, H., Xu, W., Beyene, J., & Greenwood, C. M. T. (2007). Identifying cis-and trans-acting single-nucleotide polymorphisms controlling lymphocyte gene expression in humans. In *BMC proceedings* (Vol. 1, p. 1). inproceedings.

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. article.

I T, J. (2002). *"Principal Component Analysis,2nd ed." Journal of the American Statistical Association* (Vol. 98). Springer Series in Statistics. Retrieved from http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95442-4

Jacomy, M., Venturini, T., & Bastian, M. (2011). ForceAtlas2, A Graph Layout Algorithm for Handy Network Visualization, 1–21.

Joe H Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *48*, 236–244. Retrieved from http://www.jstor.org/pss/2282967?searchUrl=/action/doAdvancedSearch?q0=Ward&f0=all&c1=AND&q1

=&f1=all&wc=on&Search=Search&sd=1963&ed=1963&la=&jo=&jc.Statistics_JournaloftheAmericanStatisticalAssociation=j100549&Search=yes

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, gkv1070. article.

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., … others. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, *477*(7364), 289–294. article.

Kim, K.-T., Lee, H. W., Lee, H.-O., Kim, S. C., Seo, Y. J., Chung, W., … others. (2015). Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol*, *16*(1), 127. article.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular Cell*, *58*(4), 610–620. article.

Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, btt656. article.

MacQueen, J., & others. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). inproceedings.

Miyamoto, D. T., Zheng, Y., Wittner, B. S., Lee, R. J., Zhu, H., Broderick, K. T., … others. (2015). RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science*, *349*(6254), 1351–1356. article.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., … others. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, *472*(7341), 90–94. article.

Network, C. G. A. R., & others. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202–209. article.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., … others. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, *344*(6190), 1396–1401. article.

Pedregosa, F., Weiss, R., & Brucher, M. (2011). Scikit-learn□: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(October), 2825–2830. Retrieved from http://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pierson, E., & Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, *16*(1), 1–10. article.

Pineda, S., Real, F. X., Kogevinas, M., Carrato, A., Chanock, S. J., Malats, N., & Van Steen, K. (2015). Integration analysis of three omics data using penalized regression methods: An application to bladder cancer. *PLoS Genet*, *11*(12), e1005689. article.

Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A conditional entropy-based external cluster evaluation measure. *Computational Linguistics*, (June), 410–420.

Ross, E. M., & Markowetz, F. (2016). OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, *17*(1), 1. article.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*,

13(11), 2498–2504. article.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. article.

Shukla, S. A., Rooney, M. S., Rajasagi, M., Tiao, G., Dixon, P. M., Lawrence, M. S., … others. (2015). Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nature Biotechnology*, *33*(11), 1152–1158. article.

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, *16*(3), 133–145. article.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. article.

Ting, D. T., Wittner, B. S., Ligorio, M., Jordan, N. V., Shah, A. M., Miyamoto, D. T., … others. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Reports*, *8*(6), 1905–1918. article.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., … others. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, *352*(6282), 189–196. article.

Vallejos, C. A., Marioni, J. C., & Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol*, *11*(6), e1004333. article.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*(Oct), 2837–2854. article.

Wang, B., Zhu, J., Pierson, E., & Batzoglou, S. (2016). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *bioRxiv*, 52225. article.

Welch, J. D., Hartemink, A. J., & Prins, J. F. (2016). SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology*, *17*(1), 1. article.

Zafar, H., Wang, Y., Nakhleh, L., Navin, N., & Chen, K. (2016). Monovar: single-nucleotide variant detection in single cells. *Nature Methods*. article.

Figure legends

Figure 1: Comparison of clustering accuracy using eeSNV and gene expression (GE) features.

(A) Bar plot comparing the clustering performance using eeSNV vs. gene expression (GE) as features, over four datasets and five different clustering strategies. Y-axis is the adjusted mutual information (AMI) obtained across 30 bootstrap runs (mean ± s.d.). *: P<0.05, ** P<0.01 and *** P<0.001. (B) Heatmap of the rankings among different methods and datasets as shown in (A).

Figure 2: Comparison of clustering visualization using eeSNV and gene expression (GE) features.

(A) bipartite graphs using eeSNVs and cell representations. (B) Principle Component Analysis (PCA) results using gene expression. (C) PCA results using eeSNVs. (D) SIMILR results using gene expression.

Figure 3: Characteristics of the eeSNVs.

X-axis: the regularization parameter *a* values. And the Y-axes are: (A) Log10 transformation of the number of eeSNVs. (B) The average number of eeSNVs per gene. (C) The proportion of SNVs with dbSNP138 annotations (human datasets). (D) The average number of cells sharing eeSNVs. Insert: Patel dataset.

Figure 4: KEGG pathways enriched with eeSNVs in the four datasets of this study. Pathways are sorted by the sum of the -log10(p-value) of each dataset, in the descending order.

Figure 5: Bipartite graph for KEGG pathways and genes enriched with eeSNVs from each dataset in this study.

Tables

Table 1: Summary of scRNA-seq datasets used in this study.

Data	Description	Type	Sub-classes	Cell counts	Reads Per cell	Map-ability	Read length	Expressed genes
Kim dataset (Kim et al., 2015)	Renal carcinoma cancer cell from patient and PDX	Human	3	91	4,1M	82 %	100	18,288
Ting dataset (Ting et al., 2014)	Pancreas Circulating Tumor cells (CTC) Cancer	Mouse	6	116	13,7M	39 %	50	15,868
Miyamoto data (Miyamoto et al., 2015)	Prostate CTCs Cancer	Human	24	133	2,0M	44 %	50	18,224

Patel data (Patel et al., 2014)	Glioblastoma tumor cells	Human	7	593	3,2M	63 %	25	25,053

Table 2: A list of interested genes highly ranked. Ranks with '*' designate cancer driver genes reported in the original studies.

Dataset	Kim	Patel	Miyamoto	Ting (mouse)
HLA-A	**32**	**8**	**2**	-
HLA-B	**3**	**105**	**1**	-
HLA-C	**1**	**98**	**4**	-
HLA-DRA	**71**	771	200	-
B2M	1617	**45**	301	**7**
KRAS	**13***	2101	2254	235*
TRP53	*NA*	*NA*	*NA*	365*
SPARC	**22**	**37**	567	**79**
EGFR	2231	**88***	*NA*	*NA*
AR	*NA*	*NA*	**6***	*NA*
KLK3	*NA*	*NA*	**19***	*NA*

Supplemental Materials

Supplementary Figure S1: the SNV calling pipeline. The follows GATK's "Best Practice" workflow for SNP and INDEL calling, with four steps. Step 1: alignment. Step 2: preprocessing of BAM files. Step 3: read realignment and recalibration. Step 4: variant calling.

Supplementary Figure S2: Sketch of Sparse SNV inference to Reflect Gene Expression (SSrGE) linear models. The SNVs calculated from the SNV calling pipeline (Supplementary Figure S1) are transformed into a predictor matrix M_{SNV}. Gene expression is the response matrix M_{GE}. For each gene, a LASSO regression is fitted to identify non-null coefficient matrix W. The output of the models is a set of filtered eeSNVs and a set of corresponding genes in which eeSNVs are found.

Supplementary Figure S3: Bar plot comparing the clustering performance using eeSNV vs. gene expression (GE) as features, over four datasets and five different clustering strategies. The metrics used are (A): Adjusted Rand Index (ARI), and (B): V-measure.

Supplementary Figure S4: Relationship between the best accuracy metrics and the LASSO regularization

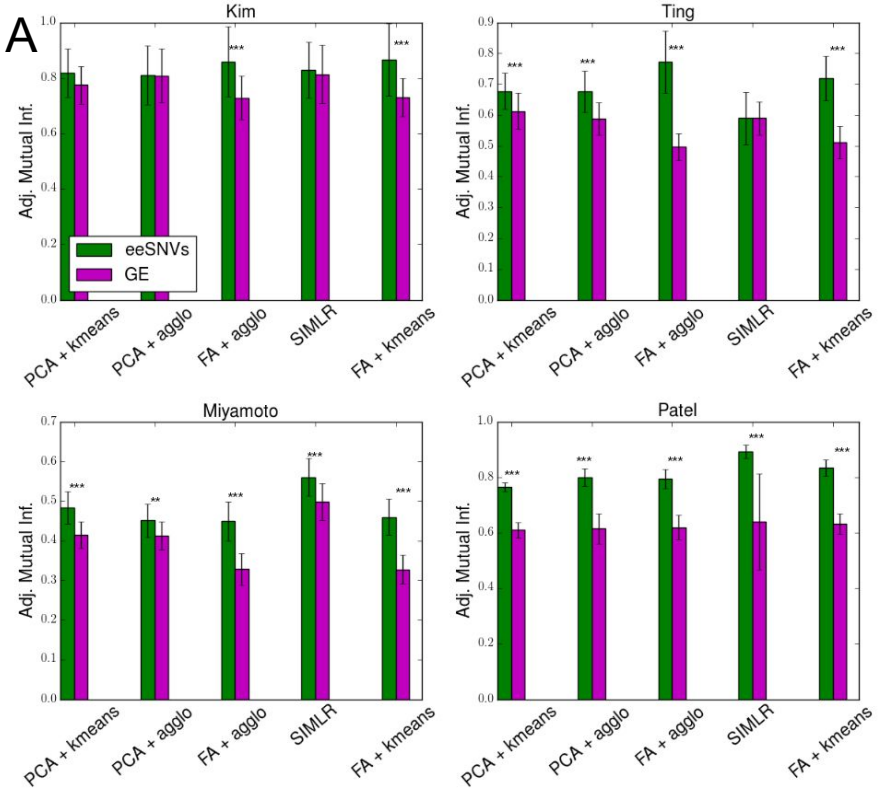
parameter *a*, over the four datasets and five different clustering approaches. The accuracy metrics are: (A) Adjusted Mutual Information (AMI), B: Adjusted Rand Index (ARI), and (C): V-measure.

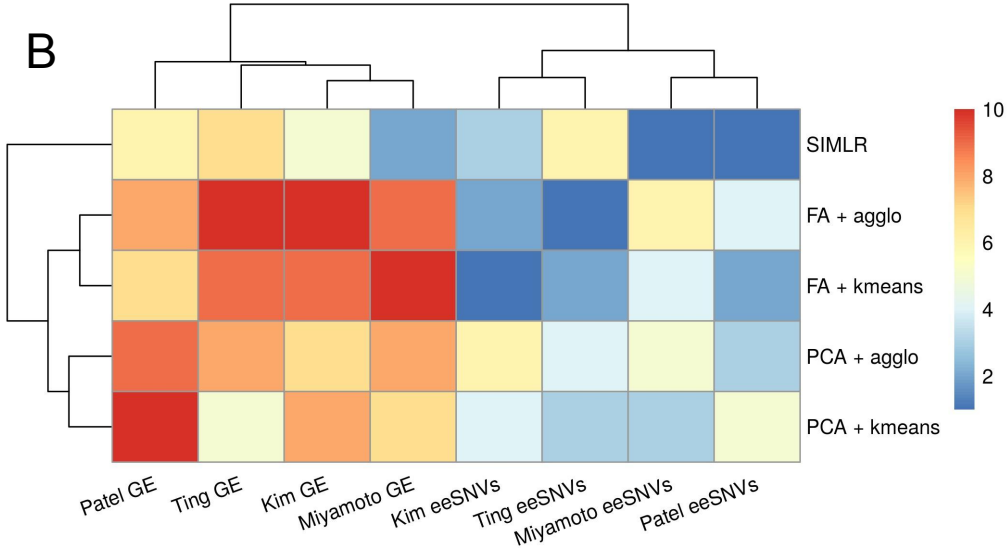
Supplementary Figure S5: Comparison of the batch-effect on SNVs and gene expression, using scRNA-Seq data from glioblastoma patient MGH26.

Supplementary Table S1: Regularization values (*a*) used for the clustering procedures along with the number of eeSNVs features.

Supplementary Tables S2: Ranked eeSNVs and genes for each dataset (with minimum regularization filtering *a*=0.1).

Supplementary File 1: Detailed description on top ranked genes and eeSNVs.





Clustering results with eeSNVs selected according to alpha

Figure 1

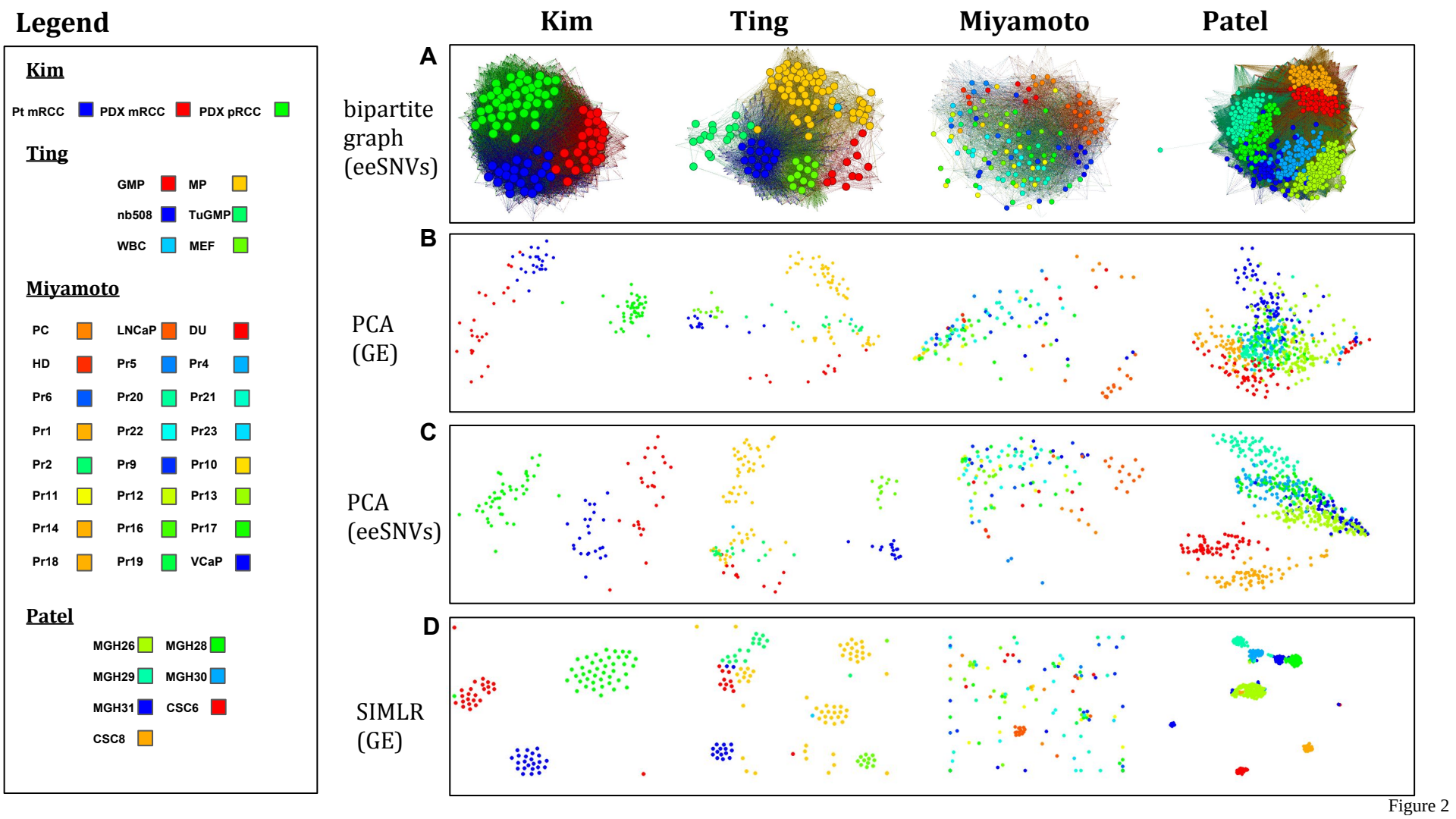

Legend
Kim
Pt mRCC
PDX mRCC
PDX pRCC
Ting
GMP
MP
nb508
TuGMP
WBC
MEF
Miyamoto
PC
LNCaP
DU
HD
Pr5
Pr4
Pr6
Pr20
Pr21
Pr1
Pr22
Pr23
Pr2
Pr9
Pr10
Pr11
Pr12
Pr13
Pr14
Pr16
Pr17
Pr18
Pr19
VCaP
Patel
MGH26
MGH28
MGH29
MGH30
MGH31
CSC6
CSC8
Kim
Ting
Miyamoto
Patel
A
bipartite graph (eeSNVs)
B
PCA (GE)
C
PCA (eeSNVs)
D
SIMLR (GE)

Figure 2

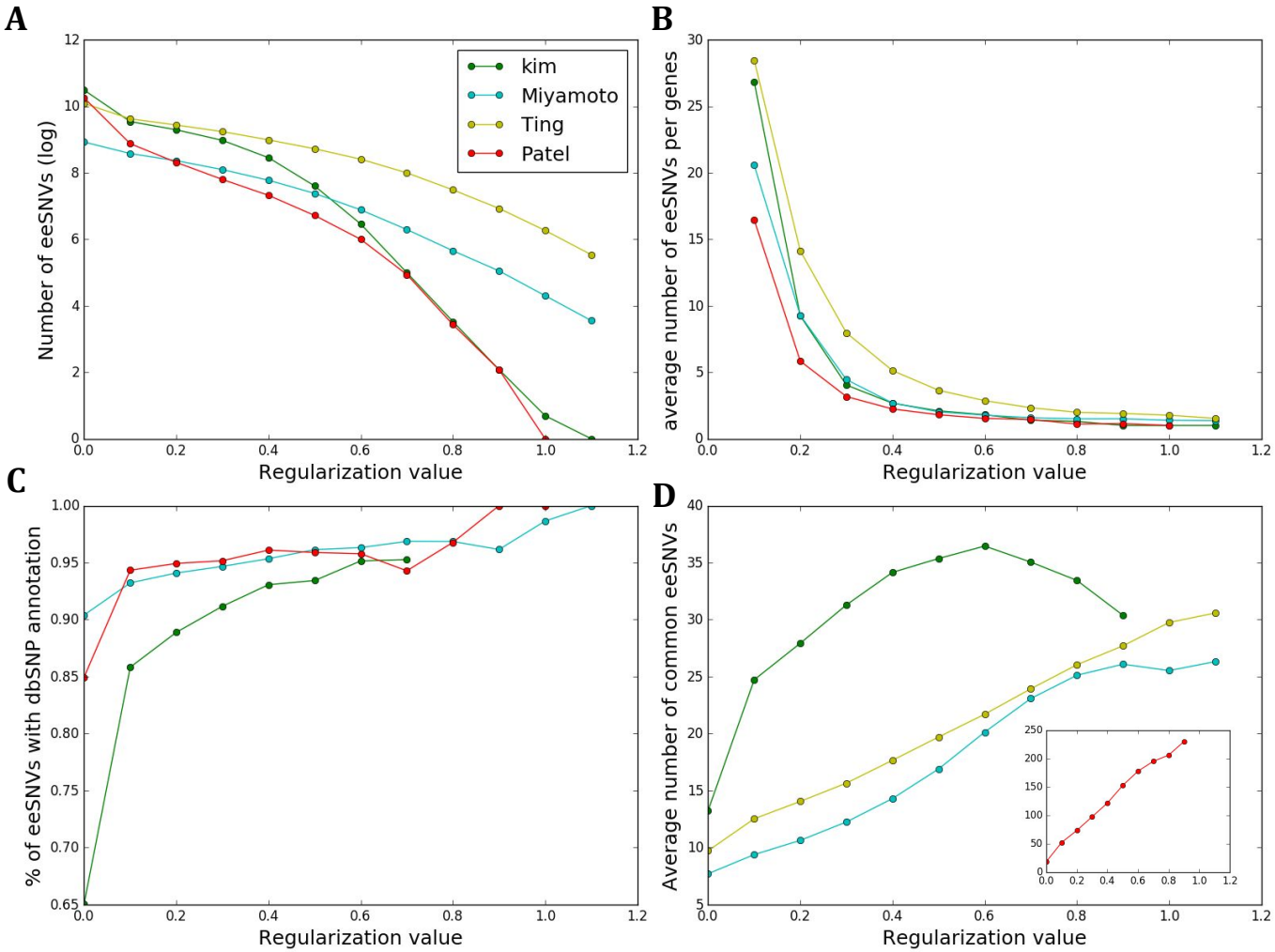

A
B
C
D
kim
Miyamoto
Ting
Patel
Number of eeSNVs (log)
average number of eeSNVs per genes
% of eeSNVs with dbSNP annotation
Average number of common eeSNVs
Regularization value

Figure 3

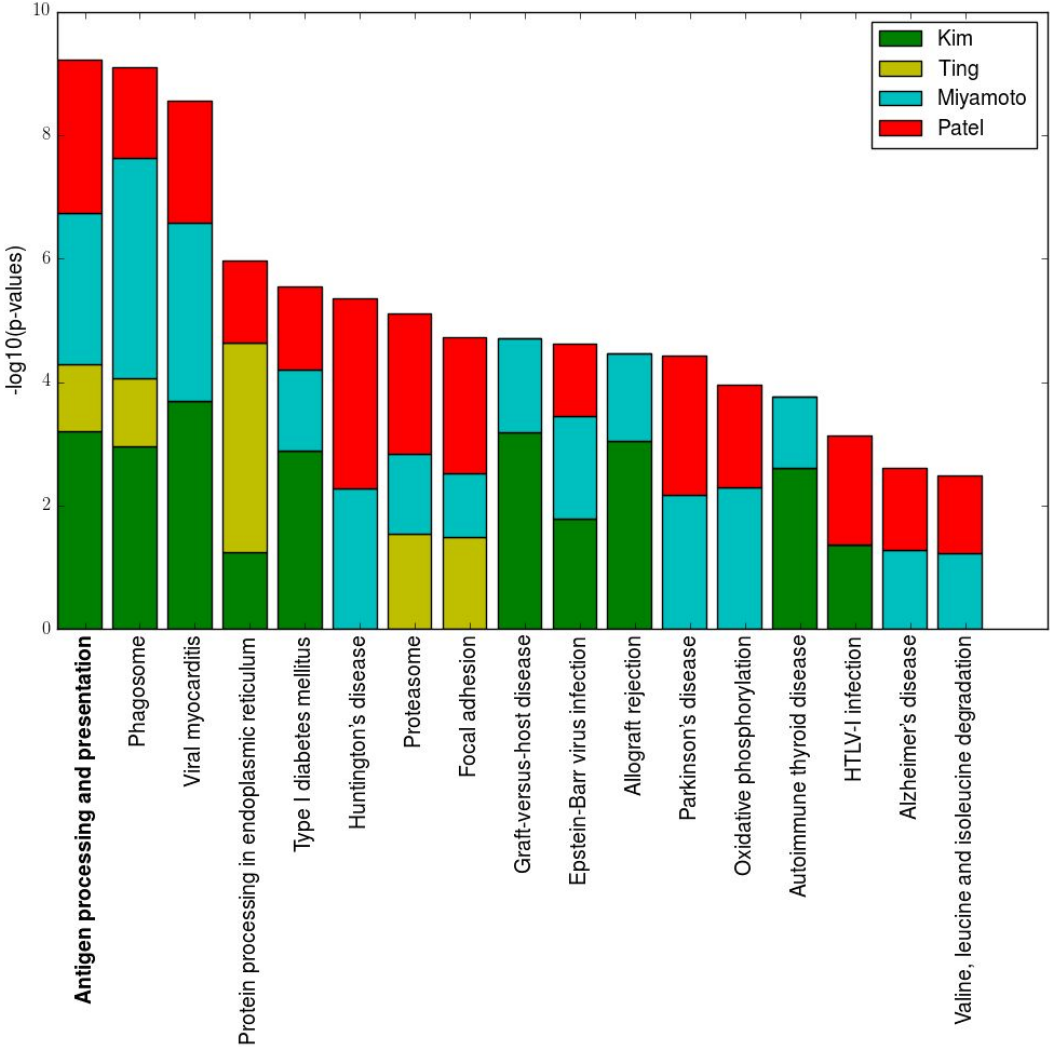

-log10(p-values)
0
2
4
6
8
10
Kim
Ting
Miyamoto
Patel
Antigen processing and presentation
Phagosome
Viral myocarditis
Protein processing in endoplasmic reticulum
Type I diabetes mellitus
Huntington's disease
Proteasome
Focal adhesion
Graft-versus-host disease
Epstein-Barr virus infection
Allograft rejection
Parkinson's disease
Oxidative phosphorylation
Autoimmune thyroid disease
HTLV-I infection
Alzheimer's disease
Valine, leucine and isoleucine degradation

Figure 4

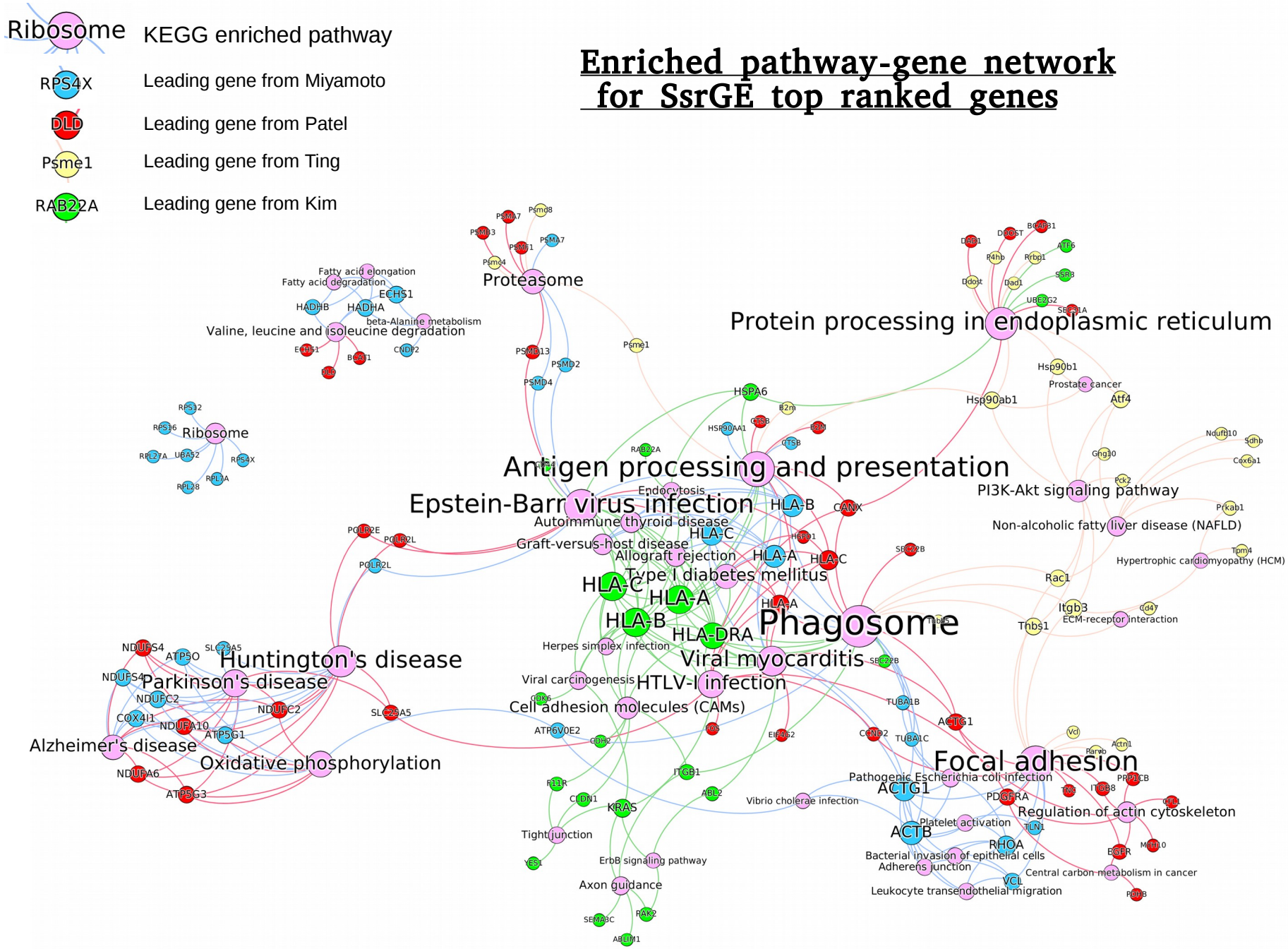

Enriched pathway-gene network for SsrGE top ranked genes
Ribosome
KEGG enriched pathway
RPS4X
Leading gene from Miyamoto
DLD
Leading gene from Patel
Psme1
Leading gene from Ting
RAB22A
Leading gene from Kim
Proteasome
Protein processing in endoplasmic reticulum
Antigen processing and presentation
Epstein-Barr virus infection
Phagosome
Huntington's disease
Parkinson's disease
Alzheimer's disease
Oxidative phosphorylation
Focal adhesion
Regulation of actin cytoskeleton
Viral myocarditis
HTLV-I infection
Cell adhesion molecules (CAMs)
Type I diabetes mellitus
Allograft rejection
Graft-versus-host disease
Autoimmune thyroid disease
Herpes simplex infection
Viral carcinogenesis
Vibrio cholerae infection
Pathogenic Escherichia coli infection
Platelet activation
Bacterial invasion of epithelial cells
Adherens junction
Leukocyte transendothelial migration
Central carbon metabolism in cancer
Tight junction
ErbB signaling pathway
Axon guidance
Fatty acid elongation
Fatty acid degradation
Valine, leucine and isoleucine degradation
beta-Alanine metabolism
Prostate cancer
PI3K-Akt signaling pathway
Non-alcoholic fatty liver disease (NAFLD)
Hypertrophic cardiomyopathy (HCM)
ECM-receptor interaction
HLA-A
HLA-B
HLA-C
HLA-DRA
HLA-DRB1
HLA-DRB5
HLA-E
HLA-F
HLA-G
CANX
HSPA6
HSP90AA1
Hsp90ab1
Hsp90b1
Atf4
Rac1
Itgb3
Thbs1
ACTG1
ACTB
KRAS
RHOA
VCL
EGFR
PDGFRA
ITGB1
ABL2
CLDN1
JAM2
TUBA1B
TUBA1C
CNDP2
ECHS1
HADHA
HADHB
CNDP2
PSMD13
PSMD2
PSMD4
PSMA7
PSMB3
PSMG1
Psme1
RPS12
RPS16
RPL27A
UBA52
RPS4X
RPL7A
RPL28
POLR2E
POLR2L
SLC25A5
ATP5O
NDUFS4
NDUFC2
COX4I1
NDUFA10
ATP5G1
NDUFA6
ATP5G3
ATP6V0E2
CDK6
CDH2
ITGB1
ABL2
CLDN1
KRAS
YES1
SEMA3C
PAK2
ABLIM1
B2M
CTSB
UTSB
RAB22A
Endocytosis
Ndufb10
Sdhb
Cox6a1
Ghg10
Pck2
Prkab1
Tpm4
Cd47
Vcl
Parvb
Actn1
PRP1CB
ITGB8
CFL1
TLN1
MCH10
PDHB
Ddost
Dad1
UBE2G2
SSR3
ATP6
Rrbp1
R4hp
DAD1
DDOST
BCAP31
SEC22B
TMC6
Gnb1
Prkcb
Ndufb10
Sdhb
Cox6a1
Ghg10
Pck2
Prkab1
Tpm4
Cd47
Vcl
Parvb
Actn1
PRP1CB
ITGB8
CFL1
TLN1
MCH10
PDHB

Figure 5