Using Single Nucleotide Variations in Single-Cell RNA-Seq to Identify Tumor Subpopulations and Genotype-phenotype Linkage

Olivier Poirion[1], Xun Zhu[1, 2], Travers Ching[1, 2], Lana X. 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

Despite its popularity, characterization of subpopulations with transcript abundance is subject to significant amount of noise. We propose to use effective and expressed nucleotide variations (eeSNVs) from scRNA-seq as alternative features for tumor subpopulation identification. We developed a linear modeling framework SSrGE to link eeSNVs associated with gene expression. In all the cancer datasets tested, eeSNVs achieve better accuracies and more complexity than gene expression for identifying subpopulations. Previously validated cancer relevant genes are also highly ranked, confirming the significance of the method. Moreover, SSrGE is capable of analyzing coupled DNA-seq and RNA-seq data from the same single cells, demonstrating its power over the cutting-edge single-cell genomics techniques. In summary, SNV features from scRNA-seq data have merits for both subpopulation identification and linkage of genotype-phenotype relationship. SSrGE method is available at https://github.com/lanagarmire/SSrGE.

Keywords:

Single cell; RNA-Seq; single nucleotide variation; cancer; heterogeneity; subpopulations; genotype; linear model; visualization

Background

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 [1–3]. However, GE features are subject to significant amount of noises [4]. One issue of GE is the batch effect, where results obtained from two different runs of experiments may present substantial variations [5], even when the input materials are identical. Additionally, the expression of certain genes vary with cell cycle [6], 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 [7,8]. It is regarded that cancer evolution involves the disruption of the genetic stability e.g. increasing number of new SNVs [9,10]. 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 [11,12]. 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 [13]. Previously, the resulting SNVs were used to infer cancer cell subpopulations [14,15]. 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) [16,17]. 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 was 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 [18]. The output is a subset of 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) [19] and Factor Analysis (FA) [20], followed by either K-Means or the hierarchical agglomerative method (agglo) with WARD linkage [21]. 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 (Suppl. 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) [22] and V-measure [23], 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 [24]. 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). Finally, we also compared the CIS effect of the eeSNVs, i.e. the ability of a given eeSNV to predict the expression of its own gene, (See methods). The vast majority of the eeSNVs, and all top ranked eeSNVs, have a low CIS effect, indicating that eeSNVs are mostly used by the method as predictor to infer the expression of other genes (Suppl. Figure S5).

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. 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 [25] (Figure 4A). The pathway-gene bipartite graph illustrates the relationships between these genes and enriched pathways (Figure 4B). As expected, Antigen processing and presentation pathway stands out as the most enriched pathway, with the sum -log10 (p-value) of 9.22 (Figure 4A). "Phagosome" is the second most enriched pathway in all four data sets, largely due its members in HLA families (Figure 4B). 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

4A).

Heterogeneity markers using eeSNVs

We exemplified the potential of eeSNV as heterogeneity markers using Kim dataset. First, we reconstructed the pseudo-time ordering of the single-cells entirely using eeSNVs, rather than GE. We built a Minimum Spanning Tree, similarly to the Monocle algorithm [26], to reconstruct the pseudo-time ordering of the single-cells. The graph beautifully captures the continuity among cells, from the primary to metastasized tumors (Figure 5A). Moreover, it highlights ramifications inside each of the subgroups, demonstrating the intra-group heterogeneity. On the contrary, pseudo-time reconstruction using GE showed much less complexity and more singularity (Supplementary Figure S6). Next, we used our method to identify eeSNVs specific to each single-cell subgroup and ranked the genes according to these eeSNVs. We compared the characteristics of the metastasis cells to primary tumor cells. Two top ranked genes identified by the method, CD44 (1th) and LPP (2th), are known to promote cancer cell dissemination and metastasis growth after genomic alteration [27–30] (Suppl. Tables S3). Other top ranked genes related to metastasis are also identified, including LAMPC2 (7th), HSP90B1 (14th), MET (44th) and FN1 (52th). As expected, “Pathways in Cancer” are the top ranked pathway enriched with mutations (Figure 5B). Additionally, “Focal Adhesion”, “Endocytosis” pathways are among the other significantly mutated pathways, providing new insights on the mechanistic difference between primary and metastasized RCC tumors (Figure 5B).

Integrating DNA- and RNA-Seq data measured in the same single cells

Coupled DNA-Seq and RNA-Seq measurements from the same single cell are the new horizon of single-cell genomics. To demonstrate the power of SSrGE in integrating DNA and RNA data, we downloaded the only accessible public single cells data, which have DNA methylation and RNA-Seq records from the same hepatocellular carcinoma (HCC) single cells (Hou dataset) [31]. We then inferred SNVs from the aligned reduced representation bisulfite sequencing (RRBS) reads (See Methods). Using our methods, we identified eeSNVs, which perfectly separate normal hepatocellular cells from cancer cells and highlight the two cancer subtypes identified in the original study (Figure 6A). Pseudo-time ordering shows not only an early divergence between the two previously assumed subtypes, but also unveils significant ramifications amongst subtype type II, indicating potential new subgroups (Figure 6B).

We postulated that a considerable part of bisulfite reads was aligned with methylation islands associated with gene promoter regions. We thus annotated eeSNVs within 1500bp upstream of the transcription starting codon, and obtained genes with these eeSNVs, which are significantly prevalent in certain groups. When comparing HCC vs. normal control cells, two genes PRMT2, SULF2 show statistically significant mutations in HCC cells (P-values < 0.05). Downregultion of PRMT2 was previously associated with breast cancer [32], SULF2 was known to be upregulated in HCC and promotes HCC growth [33]. When comparing HCC subgroup I vs. II, CTBP2 is significantly mutated (P-value = 0.01) in subgroup I. CTBP2 is a transcriptional co-repressor that promotes cancer cell migration and invasion by inhibiting tumor suppressor genes, and was previously associated with worse prognostic in HCC [34].

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 [5]. 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 S7). 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 impacts, and this is supported the observation that they have high prevalence of dbSNP annotations.

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 [35]. 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 highlight 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 [36], and the somatic mutations of genes in this family were reported in the development and progression of various cancers [37,38]. 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 [39]. It is directly linked to immune response as tumor cell proliferation [37,39]. 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.

eeSNVs reveal higher degree of single-cell heterogeneity than gene expression

We have showed with strong evidence that eeSNVs unveil inter- and intra- tumor cells heterogeneities better than gene expression count data obtained from the same RNA-Seq reads. Reconstructing the pseudo-time ordering of cancer cells from the same tumor (Kim dataset) displays branching even inside primary tumor and metastasis subgroups, which gene expression data are unable to do. We identified genes enriched with SNVs specific to the metastasis, which were not reported in the original HCC single cell study [31]. Most interestingly, we showed that eeSNVs can also be retrieved from RRBS reads in a multi-omics single-cell HCC dataset, a twist from their original purpose of single-cell DNA methylation. Again, genes ranked by eeSNVs from RRBS reads only differentiate normal from cancer cells but also the different cancer subtypes. We identified several genes that are significant in either HCC or HCC subgroup, whose promoters are highly impacted by eeSNVs. Thus, we have demonstrated that our method is on the fore-front to analyze data generated by new single-cell technologies extracting multi-omics from the same cells [31,40].

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 [24] or Cytoscape [41] 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. Moreover, we have showed the capacity of SSrGE in analyzing multi-omics data from the same single cells, obtained from the most cutting-edge genomics techniques [42,43]. Our method has the great promise as part of routine scRNA-seq analyses, as well as multi-omics single-cell integration projects.

Materials and Methods

scRNA-seq datasets

All four datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) portal [44].

Kim dataset (accession GSE73121): contains three cell populations from matched primary and metastasis tumor from the same patient [45]. 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, metastatic 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 [46].

Miyamoto dataset (accession GSE67980): contains 122 CTCs from Prostate cancer from 18 patients, 30 single cells derived from 4 different cancer cell lines: VCaP, LNCaP, PC3 and DU145, and 5 leukocyte cells from a healthy patient (HD1) [47]. 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) [48]. 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.

Hou dataset (accession GSE65364): contains 25 hepatocellular carcinoma single-cells (Ca) extracted from the same patient and 6 normal liver cells (HepG2) obtained from the adjacent normal tissue of another HCC patient [31]. The 32 cells were sequenced using scTrio-seq in order to obtain reads from both RNA-seq and reduced representation bisulfate sequencing (RRBS). The authors highlighted that one of the Ca cells (Ca_26) was likely to be a normal cell, based on CNV measurements, and thus we discarded this cell. We used only the RRBS reads and as Gene expression measurements. As controls, we also used the bulk genome of all the RNA-Seq and RRBS reads of the HepG2 group.

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) [49].

Specifically, FASTQ files were first aligned using STAR aligner [50], using mm10 and hg19 as reference genomes for mouse and human datasets, respectively. The BAM file quality check was done by FastQC [51], 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 [52], 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 [53]. SNVs are then calculated and filtered using GATK tools using default parameters.

SNV detection using RRBS data

We first aligned the RRBS reads on the hg19 reference genome using the Bismark software [54]. We then processed the bam files using all the preprocessing steps as described in "SNV detection using scRNA-seq data" section (i.e. Picard Preprocessing, Order reads, Split reads and Realignments), except the base recalibration step. Finally, we called the SNVs using the BS-SNPer software [55]. We only considered the SNVs with the status "PASS".

SNV annotation

To annotate human SNV datasets, dbSNP138 from the NCBI Single Nucleotide Polymorphism database [56] and reference INDELs from 1000 genomes (1000_phase1 as Mills_and_1000G_gold_standard) [57] 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/ [58]. 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.

Ranking of eeSNVs and genes for a subpopulation

For a given single-cell subpopulation *p*, a eeSNV is specific to the subpopulation *p* if only it is significantly more present for cells of *p*. For each eeSNV we took only the subset of cells expressing the gene *g* associated with the eeSNV. We then computed the Fisher's exact test to compare the presence of the eeSNV between single-cells inside and outside *p*. We considered a eeSNV as significant for p-value < 0.05. *p'* designates the subset of cells from *p* expressing *g*. The score of a eeSNV for *p* is given by:

$$score^{p}_{eeSNV} = \frac{|\{cell\ with\ eeSNV | cell\ \in\ p'\}|}{|p'|}\ score_{eeSNV}$$

The score of a given gene *g* for *p* is thus given by:

$$score^{p}_{gene_g} = \sum_{eeSNV \in gene_g} score^{p}_{eeSNV}$$

To rank eeSNVs from the promoter regions of the RRBS reads in Hou dataset, we applied a similar methodology: we annotated the eeSNVs within 1500bp upstream of genes' starting codon regions.

CIS scores of eeSNVs

The CIS score of an eeSNV *s* is the fraction of contribution of an eeSNV to the expression of the gene *g* that it resides in among the total score:

$$CISscore_{eeSNV} = \frac{|W^{T}{}_{g}(s)|}{score_{eeSNV}}$$

Pseudo-time ordering reconstruction

To estimate the trajectory of cell evolvement, we adopted the following procedure, motivated by the method described earlier [26]. We first constructed the following distance matrix to reflect the correlation between each pair of cells:

$$\mathrm{D}_{\mathrm{i,j}} = 1 - \mathrm{Correlation}\left(\mathrm{Sample}_{\mathrm{i}}, \mathrm{Sample}_{\mathrm{j}}\right)$$

Using this distance matrix as the adjacency matrix, we constructed a weighted undirected complete graph with each node representing a cell. We then find the minimum spanning tree of this complete graph. Finally, we plotted the graph and mapped the original labels as colors of the nodes.

Subpopulation clustering algorithms

We combined two dimension reduction algorithms: Principal Component Analysis (PCA) [19] and Factor Analysis (FA) [20] with two popular clustering approaches: the K-Means algorithm [59] and agglomerative hierarchical clustering (agglo) with WARD linkage [21]. We also used SIMLR, a recent algorithm specifically tailored to cluster and visualize scRNA-seq data, which learns the similarity matrix from subpopulations [60]. 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*) [61]. 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 [22,23]. 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, whereas b is the number of dis-concordant samples. As an improvement, ARI normalizes RI against random chances: $ARI = \frac{RI-E(RI)}{\max(RI)-E(RI)}$ [22].

AMI, similarly to ARI, normalizes Mutual Information (MI) against chances [22]. The Mutual Information between two sets of classes C and K is equal to: 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 [62]. Graphs were visualized using GePhi software [24] and spatialized using ForceAtlas2 [63], 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 [64]. We selected genes scored with significant eeSNVs for the metastasis cells from Kim dataset and for the CTCs for the Ting dataset. We then used DAVID 6.8 functional annotation tool to identify significant pathways amongst these genes [25]. We used the default significance value (adjusted p-value threshold of 0.10). Significant pathways are then represented as a bipartite graph using Gephi: Nodes are either genes or pathway and the size of each nodes represent the score of the genes or, in the case of pathways, the sum of the scores of the genes linked to the pathways. We used the same methodology to infer significant pathways of cancer cells, compared to normal cells, from Hou dataset. However, we used all the genes ranked rather than only the significant genes, since only few genes are found to be significant for cancer cells.

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

1. Harris K, Magno L, Katona L, Lönnerberg P, Manchado ABM, Somogyi P, et al. Molecular organization of CA1 interneuron classes. bioRxiv. Cold Spring Harbor Labs Journals; 2015;34595.

2. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. Springer; 2015;16:1–10.

3. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput Biol. Public Library of Science; 2015;11:e1004333.

4. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol. Cell. Elsevier; 2015;58:610–20.

5. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. Nature Publishing Group; 2015;16:133–45.

6. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. Nature Publishing Group; 2015;33:155–60.

7. Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, et al. Cis and trans effects of human genomic variants on gene expression. PLoS Genet. Public Library of Science; 2014;10:e1004461.

8. Hu P, Lan H, Xu W, Beyene J, Greenwood CMT. Identifying cis-and trans-acting single-nucleotide polymorphisms controlling lymphocyte gene expression in humans. BMC Proc. 2007. p. 1.

9. Berdasco M, Esteller M. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. Dev. Cell. Elsevier; 2010;19:698–711.

10. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. Nature Publishing Group; 2011;472:90–4.

11. Almendro V, Marusyk A, Polyak K. Cellular heterogeneity and molecular evolution in cancer. Annu. Rev. Pathol. Mech. Dis. Annual Reviews; 2013;8:277–302.

12. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature. Nature Publishing Group; 2013;501:338–45.

13. Zafar H, Wang Y, Nakhleh L, Navin N, Chen K. Monovar: single-nucleotide variant detection in single cells. Nat. Methods. 2016;

14. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol. BioMed Central; 2016;17:1.

15. Welch JD, Hartemink AJ, Prins JF. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. Genome Biol. BioMed Central; 2016;17:1.

16. Gamazon ER, Wheeler HE, Shah K, Mozaffari S V, Aquino-Michaels K, Carroll RJ, et al. PrediXcan: Trait Mapping Using Human Transcriptome Regulation. bioRxiv. Cold Spring Harbor Labs Journals; 2015;20164.

17. Pineda S, Real FX, Kogevinas M, Carrato A, Chanock SJ, Malats N, et al. Integration analysis of three omics data using penalized regression methods: An application to bladder cancer. PLoS Genet. Public Library of Science; 2015;11:e1005689.

18. Tibshirani R. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B. JSTOR; 1996;267–88.

19. I T J. “Principal Component Analysis,2nd ed” [Internet]. J. Am. Stat. Assoc. Springer Series in Statistics; 2002. Available from: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95442-4

20. Cattell RB. Factor analysis: an introduction and manual for the psychologist and social scientist. Harper; 1952;

21. Joe H Ward J. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. [Internet]. American Statistical Association; 1963;48:236–44. Available 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_JournaloftheAmericanStatisticalAss

ociation=j100549&Search=yes

22. Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. 2010;11:2837–54.

23. Rosenberg A, Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure. Comput. Linguist. 2007;410–20.

24. Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks. 2009 [cited 2013 Mar 6]; Available from: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewPDFInterstitial/154Forum/1009

25. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. Nature Publishing Group; 2009;4:44–57.

26. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. Nat. Biotechnol. NIH Public Access; 2014;32:381.

27. Kuriyama S, Yoshida M, Yano S, Aiba N, Kohno T, Minamiya Y, et al. LPP inhibits collective cell migration during lung cancer dissemination. Oncogene. Nature Publishing Group; 2016;35:952–64.

28. Fedele M, Battista S, Manfioletti G, Croce CM, Giancotti V, Fusco A. Role of the high mobility group A proteins in human lipomas. Carcinogenesis. Oxford Univ Press; 2001;22:1583–91.

29. Godar S, Ince TA, Bell GW, Feldser D, Donaher JL, Bergh J, et al. Growth-inhibitory and tumor-suppressive functions of p53 depend on its repression of CD44 expression. Cell. Elsevier; 2008;134:62–73.

30. Wielenga VJM, Heider K-H, Johan G, Offerhaus A, Adolf GR, van den Berg FM, et al. Expression of CD44 variant proteins in human colorectal cancer is related to tumor progression. Cancer Res. AACR; 1993;53:4754–6.

31. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. Nature Publishing Group; 2016;26:304–19.

32. Oh TG, Bailey P, Dray E, Smith AG, Goode J, Eriksson N, et al. PRMT2 and ROR$γ$ expression are associated with breast cancer survival outcomes. Mol. Endocrinol. Endocrine Society Chevy Chase, MD; 2014;28:1166–85.

33. Lai J-P, Sandhu DS, Yu C, Moser CD, Hu C, Shire AM, et al. Sulfatase 2 protects hepatocellular carcinoma cells against apoptosis induced by the PI3K inhibitor LY294002 and ERK and JNK kinase inhibitors. Liver Int. Wiley Online Library; 2010;30:1522–8.

34. Zheng X, Song T, Dou C, Jia Y, Liu Q. CtBP2 is an independent prognostic marker that promotes GLI1 induced epithelial-mesenchymal transition in hepatocellular carcinoma. Oncotarget. Impact Journals, LLC; 2015;6:3752.

35. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (80-.). American Association for the Advancement of Science; 2016;352:189–96.

36. de Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. Nature Publishing Group; 2006;38:1166–72.

37. Network CGAR, others. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. Nature Publishing Group; 2014;513:202–9.

38. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat. Biotechnol. Nature Publishing Group; 2015;33:1152–8.

39. Chang C-C, Campoli M, Restifo NP, Wang X, Ferrone S. 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. J. Immunol. Am Assoc Immnol; 2005;174:1462–71.

40. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods. Nature Publishing Group; 2015;12:519–22.

41. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. Cold Spring Harbor Lab; 2003;13:2498–504.

42. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat. Biotechnol. Nature Publishing Group; 2015;33:285–9.

43. Kim K-T, Lee HW, Lee H-O, Kim SC, Seo YJ, Chung W, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 2015;16:127.

44. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. Oxford Univ Press; 2013;41:D991--D995.

45. Kim K-T, Lee HW, Lee H-O, Song HJ, Shin S, Kim H, et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome Biol. BioMed Central; 2016;17:80.

46. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science (80-.). American Association for the Advancement of Science; 2014;344:1396–401.

47. Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, Broderick KT, et al. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science (80-.). American Association for the Advancement of Science; 2015;349:1351–6.

48. Ting DT, Wittner BS, Ligorio M, Jordan NV, Shah AM, Miyamoto DT, et al. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep. Elsevier; 2014;8:1905–18.

49. Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinforma. Wiley Online Library; 2013;10–1.

50. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr. Protoc. Bioinforma. Wiley Online Library; 2015;11–4.

51. Andrews S, others. FastQC: A quality control tool for high throughput sequence data. Ref. Source. 2010;

52. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. Oxford Univ Press; 2013;btt656.

53. Guidot A, Prior P, Schoenfeld J, Carrère S, Genin S, Boucher C. Genomic structure and phylogeny of the plant pathogen Ralstonia solanacearum inferred from gene distribution analysis. J. Bacteriol. [Internet]. 2007 [cited 2012 Nov 11];189:377–87. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1797399&tool=pmcentrez&rendertype=abstract

54. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. bioinformatics. Oxford Univ Press; 2011;27:1571–2.

55. Gao S, Zou D, Mao L, Liu H, Song P, Chen Y, et al. BS-SNPer: SNP calling in bisulfite-seq data. Bioinformatics. Oxford Univ Press; 2015;btv507.

56. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. Oxford Univ Press; 2001;29:308–11.

57. Consortium 1000 Genomes Project, others. A global reference for human genetic variation. Nature. Nature Publishing Group; 2015;526:68–74.

58. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. Nature Publishing Group; 2011;477:289–94.

59. MacQueen J, others. Some methods for classification and analysis of multivariate observations. Proc. fifth Berkeley Symp. Math. Stat. Probab. 1967. p. 281–97.

60. Wang B, Zhu J, Pierson E, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. bioRxiv. Cold Spring Harbor Labs Journals; 2016;52225.

61. Pedregosa F, Weiss R, Brucher M. Scikit-learn□: Machine Learning in Python. J. Mach. Learn. Res. [Internet]. 2011;12:2825–30. Available from: http://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf

62. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. [Internet]. InterJournal, Complex Systems 1695.; 2006;Complex Sy:1695. Available from: http://igraph.sf.net

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

64. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. Oxford Univ Press; 2015;gkv1070.

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: Gene and KEGG pathways enriched with eeSNVs in the four scRNA-seq datasets.

(A) KEGG pathways enriched with genes containing eeSNVs in the four datasets. Pathways are sorted by the sum of the -log10(p-value) of each dataset, in the descending order. (B) Bipartite graph for KEGG pathways and genes enriched with eeSNVs. Pathways and genes in each dataset are colored as shown in the gragh. The size of the nodes is proportional to the normed gene scores, according to the eeSNV scores, and to the sum of the normed gene scores for the pathway nodes.

Figure 5: Heterogeneity revealed by Kim dataset.

(A) Pseudo-time ordering reconstruction of the different subgroups. (B) Bipartite graph for KEGG pathways and genes enriched with eeSNVs. The size of the nodes is proportional to the gene scores, according to the eeSNV scores, and to the sum of the gene scores for the pathway nodes. Also, lighter green indicates genes with a lower rank

Figure 6: **Heterogeneity revealed by eeSNVs from multi-omics single cell HCC (Hou) dataset.**

(A) Bipartite-graph representation of the single cells using eeSNVs from RRBS reads. (B) Pseudo-time ordering reconstruction of the HCC cells**.**

Tables

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

Data	**Description**	**Type**	**Organism**	**Sub-class**	**Cell count**	**Reads Per cell**	**Map-ability**	**Read length**	**Expres gene**
Kim dataset [45]	Renal carcinoma cancer cell from patient and PDX	RNA-seq	Human	3	91	4.1M	82 %	100	1828
Ting dataset [48]	Pancreas Circulating Tumor cells (CTC) Cancer	RNA-seq	Mouse	6	116	13.7M	39 %	50	1586
Miyamoto dataset [47]	Prostate CTCs Cancer	RNA-seq	Human	24	133	2.0M	44 %	50	1822
Patel dataset [46]	Glioblastoma tumor cells	RNA-seq	Human	7	593	3.2M	63 %	25	2505

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. It 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: CIS score of the eeSNVs for the four datasets. (I) CIS effect for all the dataset and (II) CIS score for the top 100 eeSNVs.

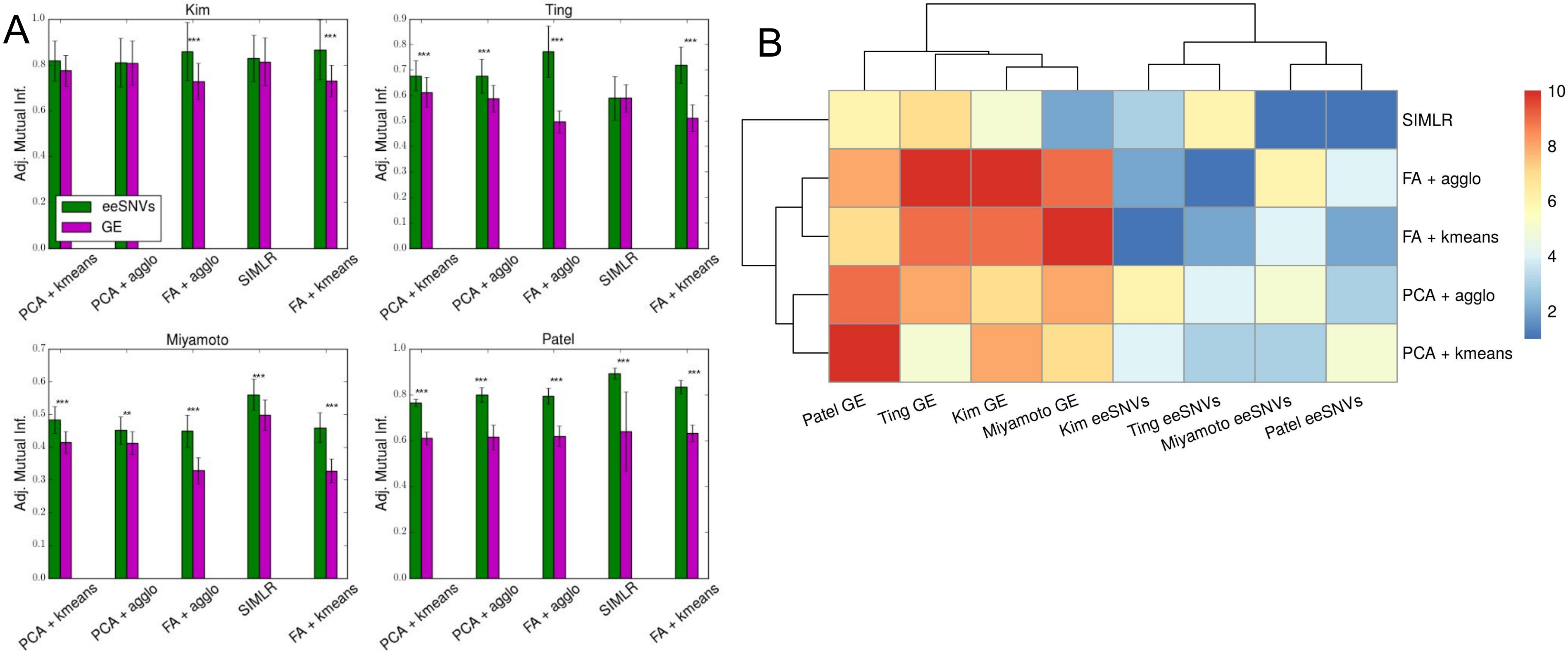
Supplementary Figure S6: Pseudo-time reconstruction using the Monocle algorithm with gene expression from genes having eeSNVs as features.

Supplementary Figure S7: Comparison of the batch-effect on SNVs and gene expression, using scRNA-seq data from glioblastoma patient MGH26.

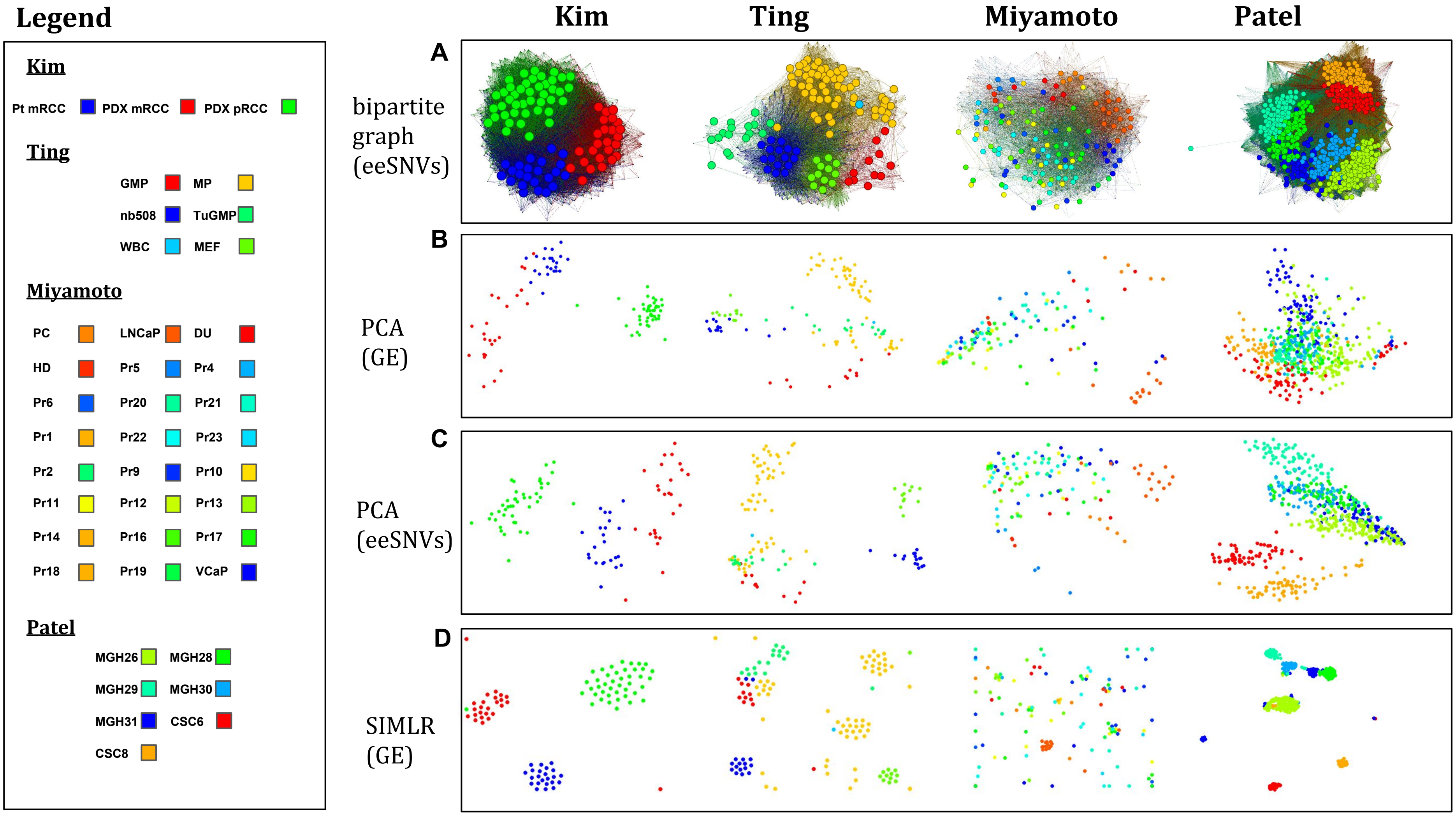
Supplementary Table S1: Regularization values (a) used for the clustering procedures along with the number of eeSNVs features.

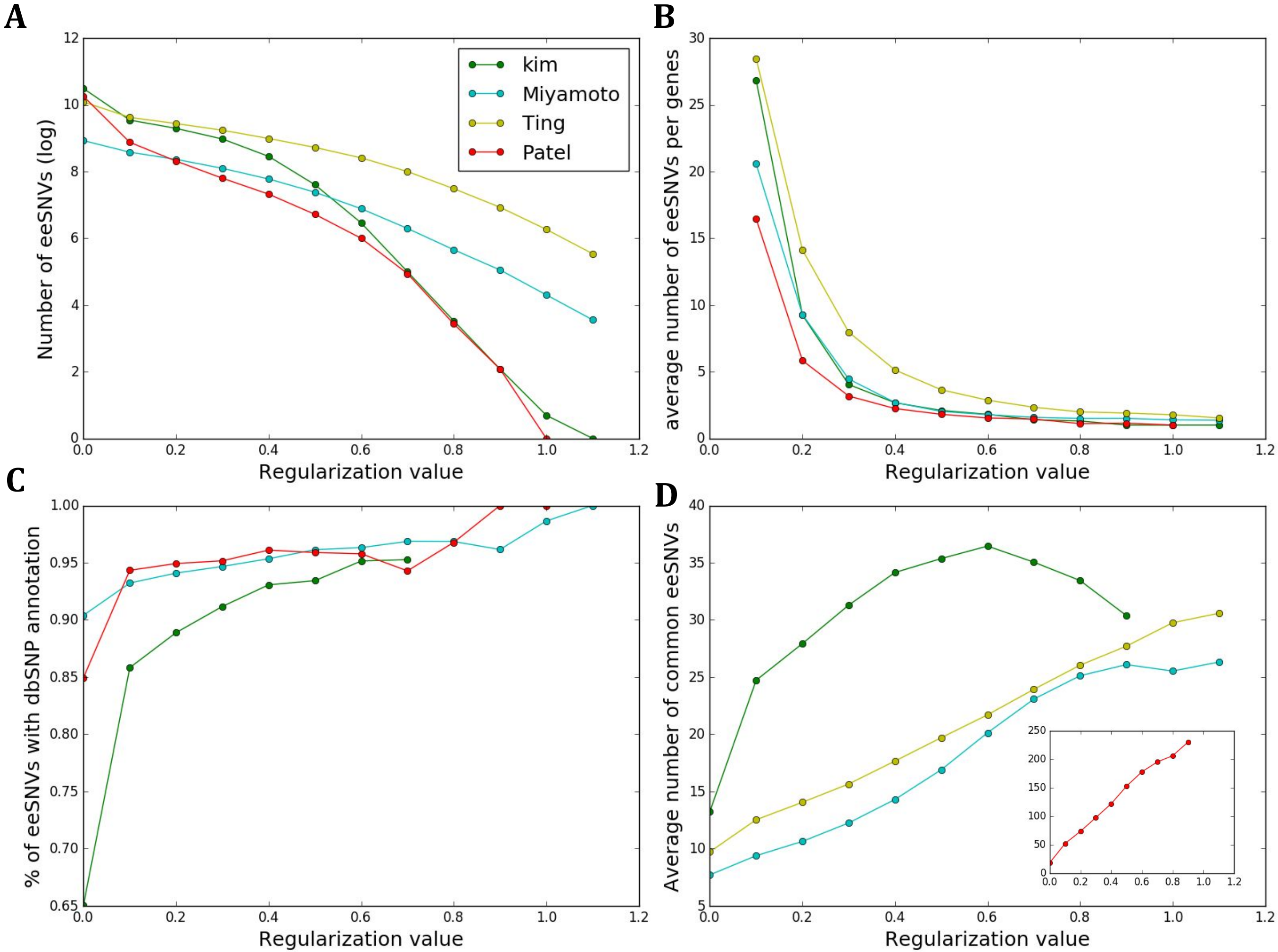
Supplementary Tables S2: Ranked eeSNVs and genes for each dataset (with minimum regularization filtering a=0.1).

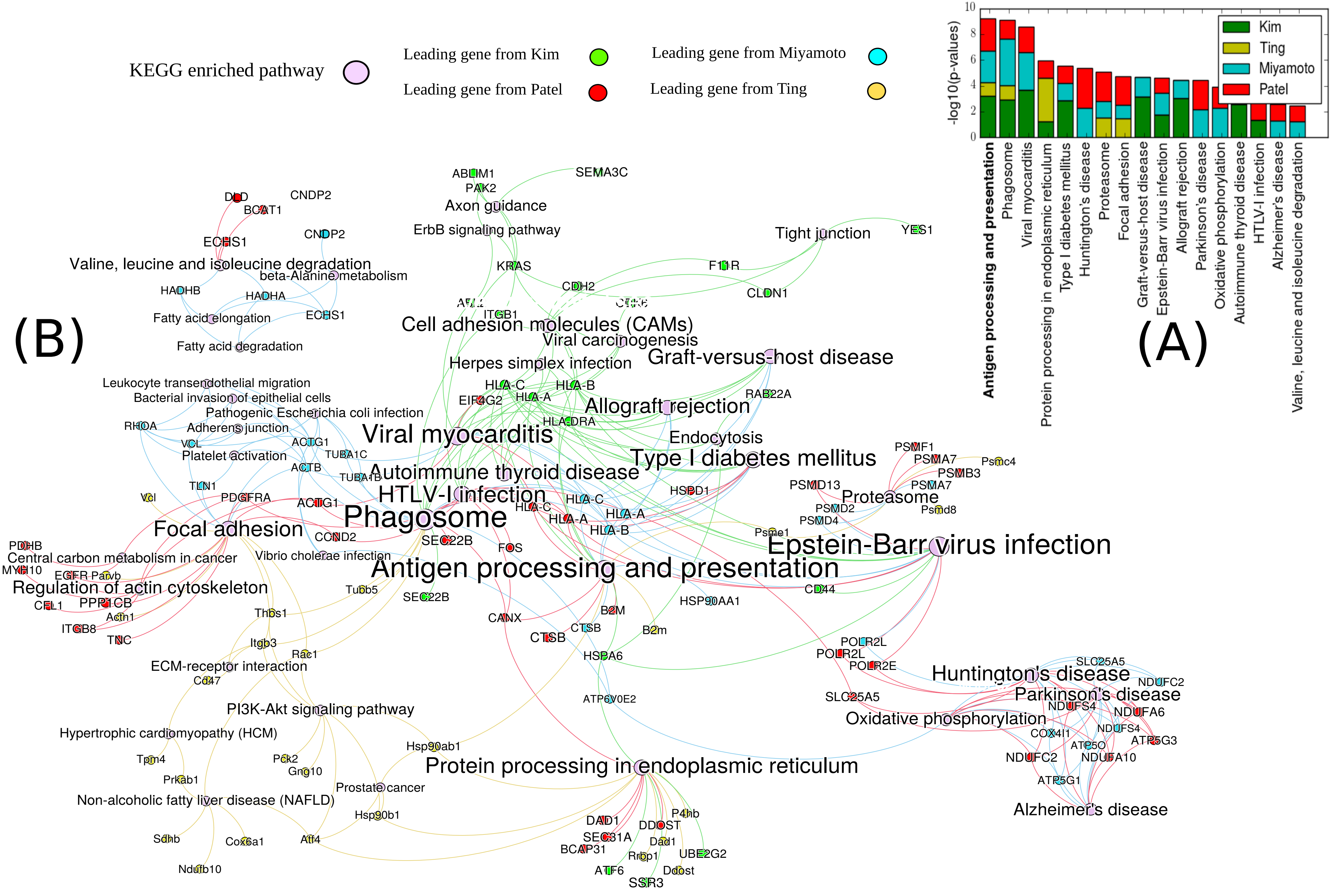
Supplementary Tables S3: Ranked genes for the metastasis single-cells from the Kim dataset (mRCC).

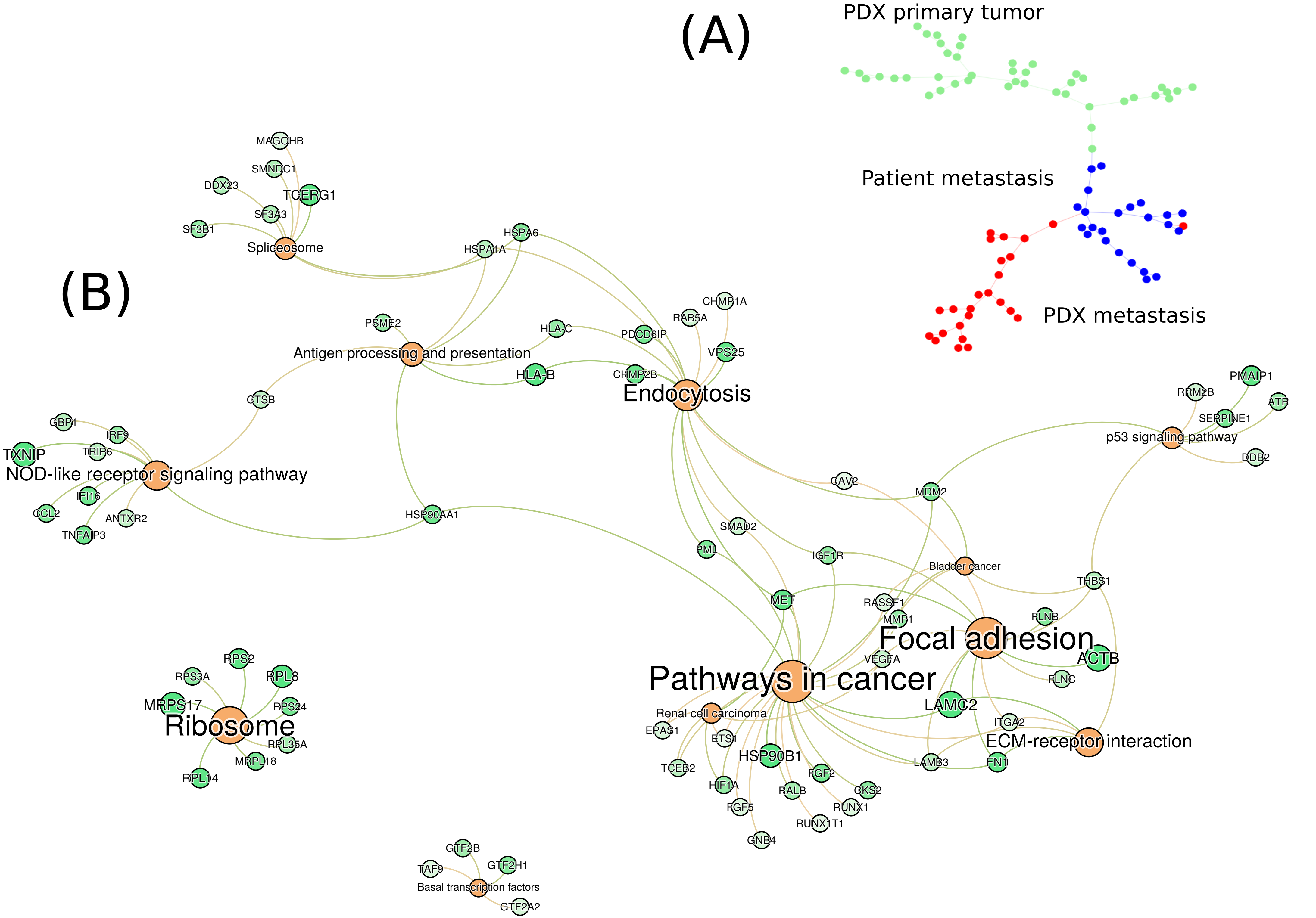


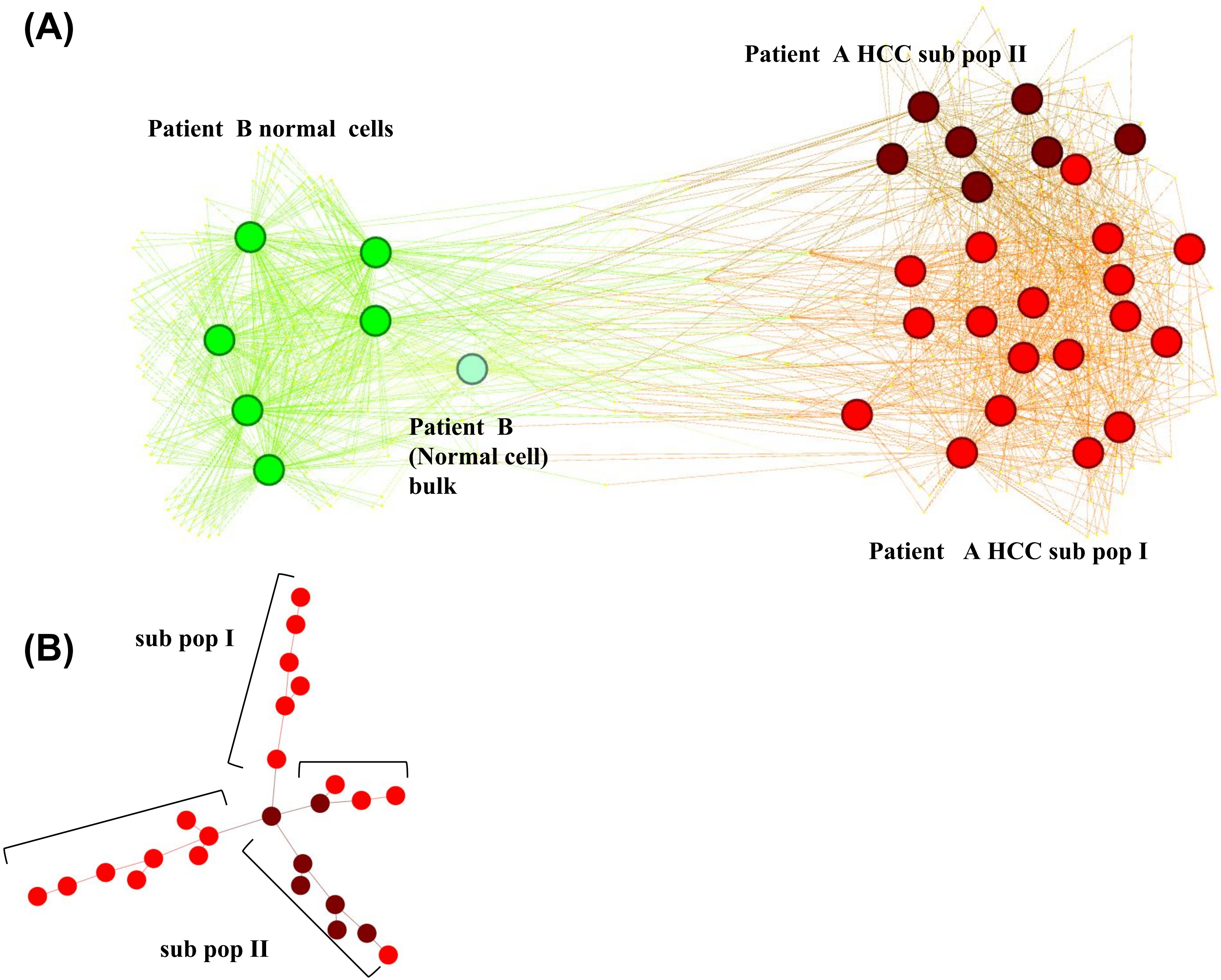
Clustering results with eeSNVs selected according to alpha


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)


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
Regularization value
Regularization value
Regularization value
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
10
12
5
10
15
20
25
30
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
35
40
50
100
150
200
250


KEGG enriched pathway
Leading gene from Kim
Leading gene from Patel
Leading gene from Miyamoto
Leading gene from Ting
(A)
-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
(B)
DLD
CNDP2
BCAT1
ECHS1
CNDP2
Valine, leucine and isoleucine degradation
beta-Alanine metabolism
HADHB
HADHA
ECHS1
Fatty acid elongation
Fatty acid degradation
ABLIM1
SEMA3C
PAK2
Axon guidance
ErbB signaling pathway
Tight junction
YES1
KRAS
F11R
CDH2
CLDN1
ABL2
ITGB1
CTNNB
Cell adhesion molecules (CAMs)
Viral carcinogenesis
Herpes simplex infection
Graft-versus-host disease
Leukocyte transendothelial migration
Bacterial invasion of epithelial cells
Pathogenic Escherichia coli infection
Adherens junction
RHOA
VCL
ACTG1
Platelet activation
TUBA1C
ACTB
TUBA1B
TLN1
Vcl
PDGFRA
ACTG1
HLA-C
HLA-B
EIF4G2
HLA-A
Allograft rejection
RAB22A
HLA-DRA
Viral myocarditis
Endocytosis
Type I diabetes mellitus
Autoimmune thyroid disease
HTLV-I infection
Phagosome
HLA-C
HLA-C
HLA-A
HLA-A
HLA-B
HSPD1
PSMF1
PSMA7
Psmc4
PSMB3
PSMD13
PSMA7
Proteasome
PSMD2
Psmd8
PSMD4
Psme1
Focal adhesion
CCND2
SEC22B
FOS
Epstein-Barr virus infection
PDHB
Central carbon metabolism in cancer
Vibrio cholerae infection
Antigen processing and presentation
MYH10
EGFR Parvb
Regulation of actin cytoskeleton
Tubb5
CD44
SEC22B
HSP90AA1
CFL1
PPP1CB
Actn1
Thbs1
CANX
B2M
ITGB8
TNC
Itgb3
CTSB
CTSB
B2m
Rac1
HSPA6
POLR2L
POLR2L
POLR2E
ECM-receptor interaction
Cd47
Huntington's disease
SLC25A5
NDUFC2
Parkinson's disease
ATP6V0E2
SLC25A5
NDUFS4
NDUFA6
PI3K-Akt signaling pathway
Oxidative phosphorylation
NDUFS4
Hypertrophic cardiomyopathy (HCM)
COX4I1
ATP5G3
ATP5O
Tpm4
Hsp90ab1
Pck2
NDUFC2
NDUFA10
Gng10
Protein processing in endoplasmic reticulum
Prkab1
Prostate cancer
ATP5G1
Non-alcoholic fatty liver disease (NAFLD)
Alzheimer's disease
Hsp90b1
P4hb
DAD1
DDOST
SEC61A
Sdhb
Cox6a1
Atf4
Dad1
BCAP31
Rrbp1
UBE2G2
Ndufb10
ATF6
Ddost
SSR3


(A)
PDX primary tumor
Patient metastasis
PDX metastasis
(B)
MAGOHB
SMNDC1
DDX23
TCERG1
SF3A3
SF3B1
Spliceosome
HSPA6
HSPA1A
PSME2
HLA-C
Antigen processing and presentation
HLA-B
CHMP1A
RAB5A
PDCD6IP
VPS25
CHMP2B
Endocytosis
CTSB
GBP1
IRF9
TRIP6
TXNIP
NOD-like receptor signaling pathway
IFI16
CCL2
ANTXR2
TNFAIP3
HSP90AA1
PMAIP1
RRM2B
ATR
SERPINE1
p53 signaling pathway
DDB2
CAV2
MDM2
SMAD2
PML
IGF1R
Bladder cancer
THBS1
MET
RASSF1
MMP1
FLNB
Focal adhesion
VEGFA
ACTB
FLNC
Pathways in cancer
LAMC2
ITGA2
ECM-receptor interaction
Renal cell carcinoma
EPAS1
ETS1
TCEB2
HSP90B1
HIF1A
FGF2
CKS2
LAMB3
FN1
RALB
RUNX1
FGF5
RUNX1T1
GNB4
RPS2
RPS3A
RPL8
MRPS17
RPS24
Ribosome
RPL35A
MRPL18
RPL14
GTF2B
TAF9
GTF2H1
Basal transcription factors
GTF2A2


(A)
Patient A HCC sub pop II
Patient B normal cells
Patient B
(Normal cell)
bulk
Patient A HCC sub pop I
(B)
sub pop I
sub pop II