TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

Blavatnik School of Computer Science

Methods for the Analysis of Multi-omic and Single-cell Data

THESIS SUBMITTED FOR THE DEGREE OF
"DOCTOR OF PHILOSOPHY"

by

Nimrod Rappoport

The work on this thesis has been carried out
under the supervision of

Prof. Ron Shamir and Prof. Amos Tanay

Submitted to the Senate of Tel-Aviv University
June 2023

Acknowledgments

This dissertation summarizes my research from the last several years. I am grateful for the opportunity to perform research that has the potential to discover unknown truths and improve human life.

I would like to thank my advisors, Prof. Ron Shamir and Prof. Amos Tanay for their guidance, advice and support. This dual mentorship exposed me to diverse scientific fields, and to diverse styles of guidance, which I think benefitted me greatly.

I would also like to thank the members of the Shamir and Tanay labs with whom I spent these last few years. Gilit Zohar-Oren who was always willing to help me, even when I asked for help on the same subject for the fifth time. Liran Shlush and his lab, and specifically Nili Saar-Furer, with whom I collaborated on a recent research project. All my other collaborators, from Japan, through Israel, and all the way to Florida. And to Elad Chomsky, who was also my collaborator and passed away during my studies.

Last but not least, I would like to thank my family for the support during these years.

Declarations

The results published here are based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

This study was supported by grant 2016694 of the United States–Israel Binational Science Foundation (BSF) and the United States National Science Foundation (NSF), by the Israel Science Foundation (grant 1339/18 and grant 3165/19 within the Israel Precision Medicine Partnership program), by the German-Israeli Project DFG RE 4193/1-1, by the Bella Walter Memorial Fund of the Israel Cancer Association, by Len Blavatnik and the Blavatnik Family foundation, by the NIH 4DN nucleomic tools program, and by the European Research Council grant (scAssembly). N.R. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University, and by the Planning and Budgeting Committee (PBC) fellowship for excellent PhD students in Data Sciences. N.R. was also supported by awards from the Herczeg Institute on Aging, and from the Tel Aviv University Healthy Longevity Research Center.

The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Preface

This thesis is based on the following papers that were published throughout the PhD period in scientific journals:

1. Nimrod Rappoport, Ron Shamir, Multi-omic and multi-view clustering algorithms: review and cancer benchmark, Nucleic Acids Research, Volume 46, Issue 20, 16 November 2018, Pages 10546–10562 [1]. Available in: https://doi.org/10.1093/nar/gky889

2. Nimrod Rappoport, Ron Shamir, NEMO: cancer subtyping by integration of partial multi-omic data, Bioinformatics, Volume 35, Issue 18, September 2019, Pages 3348–3356 [2]. Available in: https://doi.org/10.1093/bioinformatics/btz058

3. Nimrod Rappoport, Ron Shamir, Inaccuracy of the log-rank approximation in cancer data analysis, Mol Syst Biol. (2019) 15: e8754 [3]. Available in: https://doi.org/10.15252/msb.20188754

4. Nimrod Rappoport, Roy Safra, Ron Shamir, MONET: Multi-omic module discovery by omic selection, PLOS Computational Biology 16(9): e1008182 [4]. Available in: https://doi.org/10.1371/journal.pcbi.1008182

5. Nimrod Rappoport*, Elad Chomsky*, Takashi Nagano, Charlie Seibert, Yaniv Lubling, Yael Baran, Aviezer Lifshitz, Wing Leung, Zohar Mukamel, Ron Shamir, Peter Fraser, Amos Tanay, Single cell Hi-C identifies plastic chromosome conformations underlying the gastrulation enhancer landscape, accepted to Nature Communications.
   * denotes equal contribution.

The following work done during the Ph.D. is not covered in the thesis:

6. Yaniv Harari, Yoav Ram, Nimrod Rappoport, Lilach Hadany, Martin Kupiec, Spontaneous changes in ploidy are common in yeast, Current Biology, Volume 28, Issue 6, 2018, Pages 825-835.e4, ISSN 0960-9822. Available in: https://doi.org/10.1016/j.cub.2018.01.062

Abstract

Two major trends in biomedical data generation have become prominent in recent years. First, experimental methods can now measure several different types of molecular parameters for biological tissues. Each such type is called an omic, and multi-omic datasets - where multiple omics are measured for each sample - are becoming more common. Second, novel experiments can now measure omic data at single cell resolution, rather than measure averages across all cells in a tissue.

While multi-omic and single cell datasets are increasing in availability, algorithms for their analysis are still lacking. Existing algorithms do not address several characteristics of multi-omic datasets, such as the presence of partial data, and the different structure of the data in different omics. For single cell data, extant methods do not provide parametric models that take full advantage of the fine-tuned single-cell measurements to extract new biological knowledge.

In this work we developed methods to analyze multi-omic and single-cell datasets. For multi-omic data, these methods focused on better characterization of cancer subtypes. For single-cell data, they focused on understanding epigenetic characteristics of cells, and specifically the genome organization using single-cell Hi-C data.

Contents

Chapter 1 - Introduction

1.1. Omics data

1.1.1. Omic technologies and their analysis

Looking at the life sciences from a bird's-eye view, there is a general trend of using tools with increasingly higher resolution. While early biological research focused mainly on taxonomy, the advent of the microscope allowed investigation of very small entities [5]. Further developments originating in physics and chemistry allowed research at the molecular level, for example by using crystallography to probe the structure of specific proteins [6]. Molecular research identified the main agents in the biology of the cell as it is currently understood – proteins, DNA and RNA molecules [7]. These agents were shown to not only teach us about cellular function, but also on how disease develops [8].

In the second half of the 20th century several techniques were developed that could probe and quantify to some extent the aforementioned molecular agents. These include southern blots for DNA [9], northern blots [10] and later real-time PCR for RNA [11], and western blots for protein [12]. Additionally, methods to sequence DNA (that is – read the string of nucleotides a DNA molecule is made of) were developed [13]. These methods were limited in that they could only measure a small number of molecular agents at a time, which should have been predefined when the experiment was planned.

A major breakthrough occurred with the invention of microarrays, which could quantify RNA molecules from hundreds or thousands of different genes [14]. This high number of measured genes allowed an unbiased approach to biological research – to first measure expression levels of all the genes, and then deduce based on these data which genes are relevant. Experiments that measure a high number of molecular parameters are called "high-throughput".

Shortly after microarrays appeared, new techniques to perform DNA sequencing were discovered (Figure 1) [15]. These led to a dramatic decrease in the cost of sequencing, and to generation of methods that use sequencing as a "subroutine" to measure other cellular characteristics. For example, RNA sequencing can be performed by reverse-transcribing RNA into DNA and then sequencing the DNA. Molecular data generated by high-throughput experiments are usually accompanied by the suffix "omic" – genomic, transcriptomic (for gene expression), methylomic (for DNA methylation), and the fields of studying these data are called genomics,

transcriptomics and methylomics accordingly. The term “omics” refers to all these types of data and their fields of study.

There are several commonly used omic data types. First, DNA mutations measure changes that occurred to the sequence of the DNA molecule. These include points mutations, and small-scale insertions and deletions. Additionally, larger changes to the DNA include copy number aberrations and translocations. Gene expression, which refers to transcribed mRNA molecules, measured first by microarrays and then using RNA-sequencing (RNA-seq), quantify the presence of messenger RNA molecules [14], [16]. Other types of RNA molecules, such as micro-RNA and long non-coding RNA, can also be quantified [17], [18]. DNA methylation measures the frequency of methyl groups, which affect gene expression, in specific DNA nucleotides [19]. Other omics that are involved in gene expression regulation include ChIP-seq [20], which is often used to measure histone modifications, and DNase-seq, MNase-seq and ATAC-seq, which determine how open a DNA region is [21]. There are omics that measure additional molecular entities, such as proteomics (proteins) [22], metabolomics (metabolites) [23], and glycomics (sugars) [24].

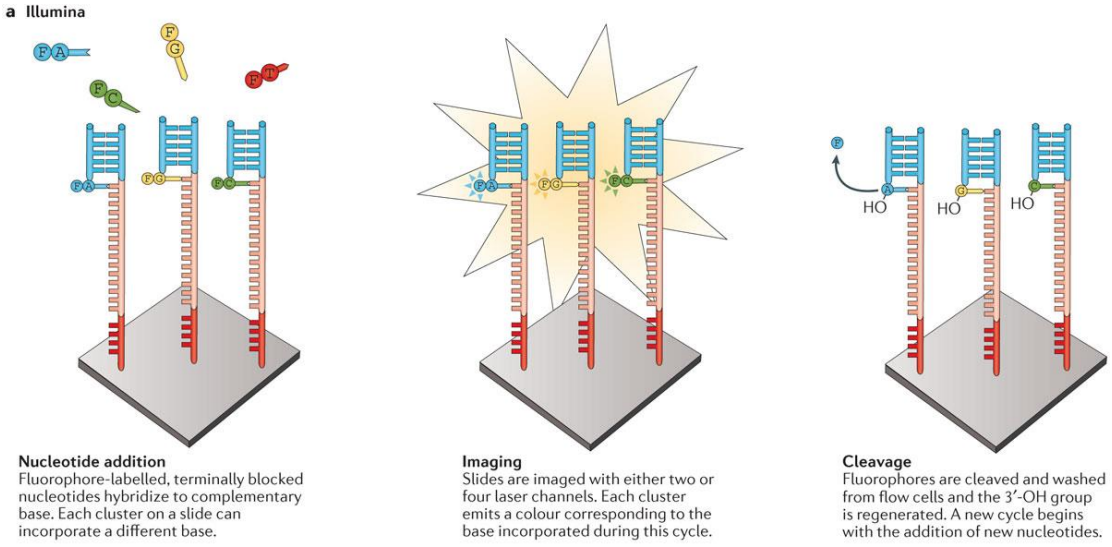


Figure 1: Illumina’s sequencing method. Terminally-blocked nucleotides are added, such that only one nucleotide is added to the end of each DNA molecule. The nucleotides are fluorophore-labeled, such that imaging allows recognition of the added nucleotide. The fluorophore is then cleaved and washed, and the next cycle beings. Source: [25].

At the same years that molecular biology techniques were developed, the field of computer science underwent major developments. In addition to breakthroughs in the theory of computation and algorithms, computation power and storage increased drastically [26]. This allowed large amounts of data to be digitally stored and rapidly accessed, and created the need for algorithms to analyze the collected data. These developments gave birth to branches of computer science that deal specifically with analyzing large amounts of data, such as machine

and statistical learning [27], [28]. Moreover, different computational communities showed increased interest in “big data” problems, as demonstrated for example by the work of the algorithms’ community on clustering [29].

An important consequence of the emergence of high-throughput experiments and omic data is that it necessitates computational analysis. Classical biological experiments that only measure a small number of parameters can be analyzed either manually, or using standard statistical software to run statistical tests. In contrast, analysis of high-throughput data cannot be done manually, and requires sophisticated methods that must do much more than test a fixed set of hypotheses.

Computer scientists were interested in biology before the emergence of omics data. Notable examples of the use of algorithms for biological research can be found in the field of evolution [30], and in comparing genetic sequences using string alignment approaches [31]. The convergence of the large-scale biological data and mathematical analysis led to the creation of the discipline of bioinformatics [32].

An important point to note about analysis of multi-omic data is that of dimensionality. Biological data has a high dimension (tens of thousands for gene expression, hundreds of thousands for DNA methylation), while the number of samples is usually relatively small due to the still high cost of obtaining omics data. These characteristics present several challenges for multi-omic data analysis. First, robust statistical analysis usually requires that the number of samples be much higher than the number of features measured for these samples [33]. Second, even if the number of samples is large, high dimensional data is challenging to analyze due to the “curse of dimensionality” – an umbrella term for non-intuitive mathematical properties of high dimensions, such as the behavior of distance metrics [33]. To handle the high dimension and the high ratio of features to samples, methods usually take advantage of the fact that biological data can often be well represented in a lower dimension.

Studying omics data is now a ubiquitous paradigm in biological research [34]. Its uses span through many biological fields, organisms and computational tasks, and we will name a small number of examples that illustrate this breadth. Spellman et al. identified genes regulated by the cell cycle in the baking yeast using microarrays [35]. Hannum et al. and Horvath used regression to predict the age of human individuals based on methylation data from their blood [36], [37]. Geiger et al. used mass spectrometry data to cluster mouse tissues based on their protein expression [38].

1.1.2. Multi-omics data

Different omic technologies measure different molecular features. Each one of these data types can be used to characterize biological samples, but joint analysis of multiple omics data types has the potential to reveal more comprehensive, systems-level insights (Figure 2) [34], [39]. This notion is what stands at the root of multi-omic analysis. In order to realize this idea, two factors are required: multi-omic datasets and algorithms to analyze them.

Even with the decreasing cost of sequencing experiments, omic studies are expensive relative to other biological experiments. Collecting datasets with multiple omics measured per sample therefore remains an effort that can mostly be carried by large research consortia. A prominent multi-omic dataset was collected by The Cancer Genome Atlas (TCGA) consortium [40]. The goal of TCGA was to provide a comprehensive, multi-omic resource with data from human cancer patients. The project, which started as a joint effort between the USA's National Cancer Institute and National Human Genome Research Institute, began in 2006 and collected data for over a dozen years. The data spans 33 different cancer types, and more than 20,000 individuals. It includes multiple omics measured from each tumor, including DNA mutations, DNA copy number, gene expression, DNA methylation, miRNA expression, and quantification of selected proteins (Figure 3). While not all omics data were collected for all tumors, TCGA still serves as the largest multi-omic data to date, to the best of our knowledge. In addition to the omic data, the project collected clinical information about the participating patients and cancers, including classical pathologic characteristics of the tumors, as well as the age, sex and survival of the patients. Importantly, the clinical and omic data of TCGA are publicly available, and only the raw DNA reads, which could be possibly used to identify the patients, have restricted access.

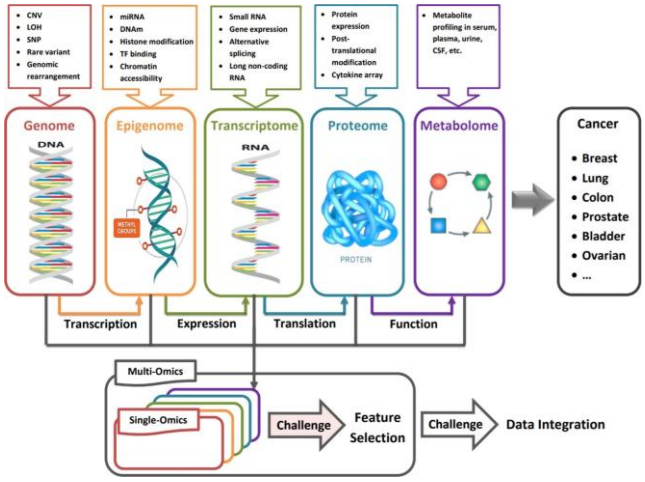


Figure 2: Omics data and integration. The figure lists omic data type families (e.g. Genome) and specific data types (e.g CNV, SNP). Source: [41].

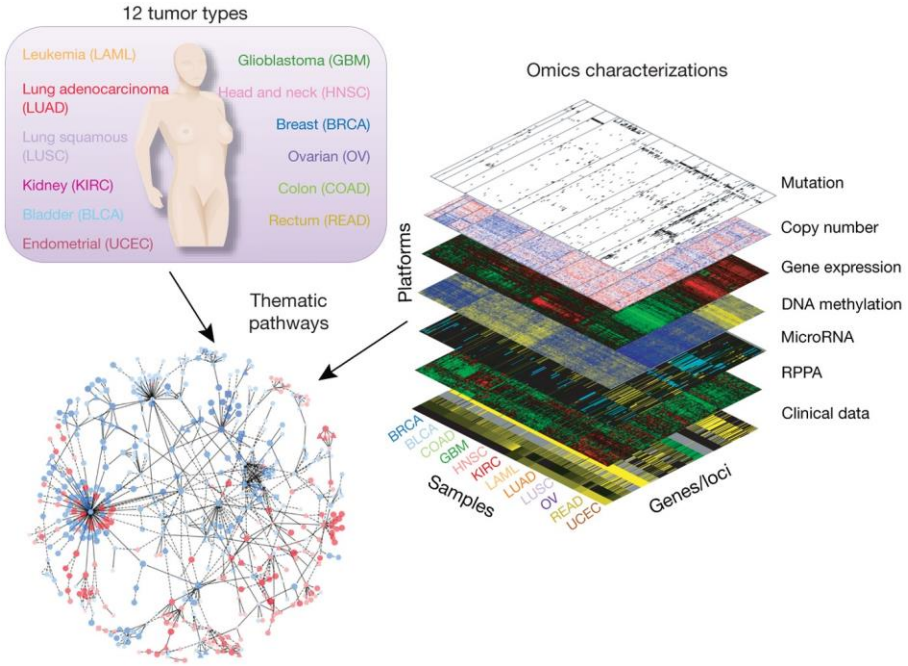


Figure 3: TCGA multi-omic data. The figure lists some of the cancer types and omic data types available in TCGA. Source: [42].

From a computational standpoint, there are several tasks that can be addressed with multi-omic data. These tasks can be either specific to the input omics, or more general. An example of a task that is omic-specific is that of finding expression Quantitative Trait Loci (eQTL) [43], [44]. In this problem the input is made of the genotypes of many individuals, and these individuals' expression of each gene in some cell type. The goal is to find genetic variants that are associated with gene expression. Different versions of the problem exist where gene expression is replaced with another omic, e.g. by DNA methylation (meQTL) [45], but the first omic must remain the genotype. Other multi-omic tasks are more general, and can be solved using different omics. These tasks include multi-omic clustering, dimension reduction, and classification [39].

Multi-omic analysis is challenging for several reasons. First, the high dimensionality of the data that we described when discussing omic data is further exacerbated in multi-omics. Not only does the number of features increase (as it is now the sum of the number of features in each omic separately), but the cost obtaining each sample increases as well, so the number of samples will tend to be lower in multi-omic datasets. Another major challenge in multi-omic analysis is in integrating the different types of data, which have diverse distributions. For example, DNA genotype data is commonly represented as a vector of integers over {0, 1, 2} (since each locus usually has only two possible alleles in the population, this representation counts the number of appearances of the minor allele), RNA-sequencing data counts RNA molecules and can have any integer, and DNA methylation data is provided as continuous numbers between 0 and 1. Integrative analyses of these different spaces has to find ways to make them comparable.

1.1.3. Single-cell data

The omic experiments that we discussed so far are performed on a tissue or on a large group of cells, and therefore the measured features are only averages across these cells. For example, an RNA-seq experiment will list the number of RNA molecules it detected from a specific gene aggregated across all the cells in the sample. It cannot be deduced whether the cells in the experiment all express this gene to the same extent, or whether there is variability among them. It is possible that this gene is actually expressed by only a subset of the cells, or that it is expressed to a varying degree in each cell depending on the cell's state. Furthermore, if the sample contains numerous cell types, we cannot distinguish between features that vary across samples because of actual differences in the measured feature, or because of differences in cell type frequency.

A partial way to overcome this limitation is to perform cell sorting using FACS [46]. This technique allows to partition biological cells into different groups based on the presence of predefined proteins. RNA-seq can then be applied to each group separately. This approach is limited in that it still cannot investigate variability within a cell type. Furthermore, it requires the prior selection of proteins that allegedly separate between cell types, and thus deviates from the unbiased approach that is a major advantage of omic studies.

Aware to these problems, biologists have developed single-cell omic approaches [47]. In these experiments, instead of performing measurements that are averages across many cells, features are measured in each cell separately. Experiments that measure averages are called *bulk* experiments, as opposed to *single-cell* experiments. While in bulk experiments the output is a matrix of samples by features, in single-cell experiments the output is a cells by features matrix for each sample. The number of cells for each biological sample can vary.

Single-cell omic experiments currently mostly use DNA sequencing as a readout. The underlying idea behind these techniques is to add a short sequence of DNA (or RNA), called a *barcode*, to all DNA (or RNA) molecules from a cell, such that all molecules from the same cell have the same barcode, but different cells have different barcodes (Figure 4) [48]. This way, when the molecules are sequenced, the barcode can be used to associate each molecule with the cell from which it came. Additionally, some techniques add a second barcode, that differs between the molecules in the same cell. Because DNA is duplicated many times prior to sequencing, the second barcode, which is called the *Unique Molecular Identifier* (UMI), allows to tell whether two sequenced molecules are duplications of the same original molecule, or come from different molecules. This enables better quantification of the data.

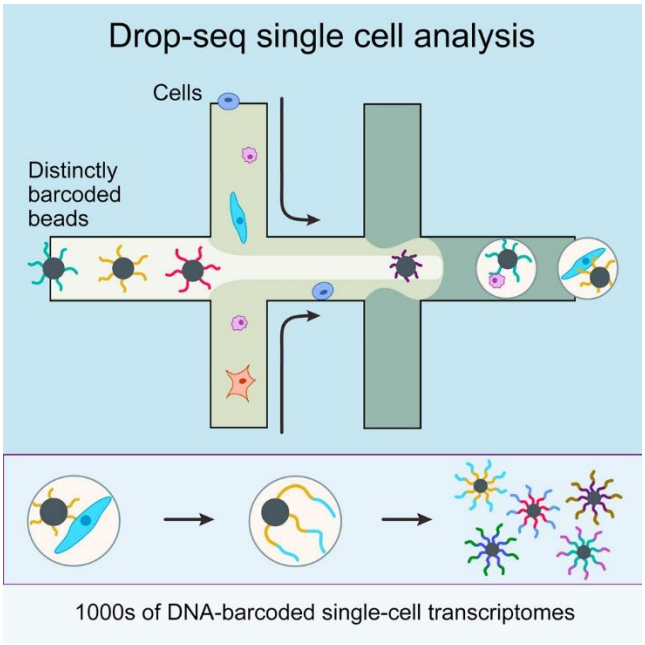


Figure 4: Microfluidics-based single-cell RNA-sequencing. Cells and beads with distinct barcodes are inserted together into aqueous droplets in an oil medium, such that each droplet usually contains at most one cell. Within each droplet the cell is lysed, and its mRNA molecules are hybridized to the distinct barcodes. Source: [49].

The first single-cell methods that were developed and gained popularity were for RNA sequencing [48]. Contributing to this popularity is the entry of commercial companies, such as 10x Genomics, into this space [50]. Such companies provide a standardized toolkit that reduces the technical expertise required to perform single-cell experiments. In addition to single-cell RNA-seq, methods to quantify other omic data at single-cell resolution were developed. These include single-cell ATAC, methylation, and Hi-C, which we will cover later [51]–[53].

The main computational challenge of single-cell data is its sampling depth, which results in sparsity and high variance. The number of sequenced reads from a cell is generally very low compared to bulk experiments. For example, single-cell RNA-seq (scRNA-seq) data usually have a few thousands of UMIs per cell, across about twenty thousand genes. The expected value of the cell by gene matrix is therefore close to zero. Moreover, changes in gene expression between cells tend to be low. It is possible for cells from different cell types to have only a few percent of the UMIs to support their difference, and differences within a cell type may be even weaker. All methods that analyze single-cell data should address this sparsity.

We will briefly illustrate how this sparsity is considered in method development for scRNA-seq. The gene by cell matrix that is the experiment's output contains many zero values. A common debate among developers of single-cell analysis algorithms revolves around modeling these zero values. Some methods use a zero-inflation model for the distribution of RNA-seq UMI count,

while others consider the high number of zeros a direct result of the low number of sampled molecules, together with the possibly low number of actual molecules of a gene in a cell [54]. Although the variance of the UMI count seems too high to stem only from sampling variance, the zero-inflations opponents reply that bulk RNA-seq data is characterized by over-dispersion, and that the high zero count in single-cell data may just reflect this high variance, and is unrelated to zero inflation. Another challenge of scRNA-seq is in estimating the expression of a single gene in a cell. Because of the data's sparsity, the measured value will often be either 0 or 1, and will be very noisy. Other methodological considerations that single-cell methods face include how to calculate similarities between cells [55], [56], and more recently how to scale to millions of cells while keeping runtime and space requirements feasible [57], [58].

Our work on single-cell data focuses on Hi-C (to be described later), but also uses scRNA-seq. The framework we use to analyze scRNA-seq data is called metacell [57], [59]. At its core is the idea that instead of looking at single cells, we can group together dozens of highly similar cells. Each group of such similar cells is called a *metacell*. If we sample enough cells, cells in a metacell would be highly similar, and all the metacells together will cover the space (often termed "manifold") of possible cellular states. The main advantage of metacells is that they are no longer sparse, since UMIs are accumulated across the dozens of cells they contain, thus solving the challenge of data sparsity once a metacell model was created.

1.2. Multi-omics clustering for cancer subtyping

1.2.1. Cancer and cancer subtypes

Cancer is a group of diseases characterized by increased cell proliferation. It is currently the second leading cause of death worldwide, after cardiovascular diseases [60]. Moreover, while the death rate from cardiovascular diseases has been decreasing significantly in the last few decades thanks to research into its diagnosis and treatment, death rates from cancers remain disappointingly steady [61].

The classical partition of cancer into different diseases is based on the tissue of origin. Cancers attack almost every tissue and cell type in the body, though with different rates. Most cancers occur at epithelial tissues, which line most organs in the body. Less common are cancers of blood cells, and rarer still are sarcomas, which are cancers of connective tissues. The tissues in which cancers are most common are breast, prostate, lung, and colorectal, but all major organs have cancers [62].

In spite of the many tissues affected by cancer, the “cancer” umbrella term is used to describe the disease in all of them. The reason for this is the common features of all cancer types. The most prominent feature is increased cellular proliferation, which is at the root of the tumor’s growth [63]. But there are many other common characteristics for different cancers, which have been summarized as the “hallmarks of cancer” in a seminal work [64]. This work has been updated twice since its initial publication, to reflect our increasing understanding of cancer biology [65], [66]. The initial hallmarks were: sustained growth signals, evading growth suppressors, resisting cell death, inducing angiogenesis (growth of blood vessels), enabling replicative immortality and activating invasion and metastasis. These hallmarks generally all reflect the notion of sustaining growth, and resisting the cellular mechanisms that inhibit it. The first update to the hallmarks added deregulation of cellular energetics and avoiding immune destruction, as well as two “enabling characteristics”, which are genome instability and tumor-promoting inflammation. The second update also added polymorphic microbiome, senescent cells, unlocking phenotypic plasticity and epigenetic reprogramming (Figure 5).

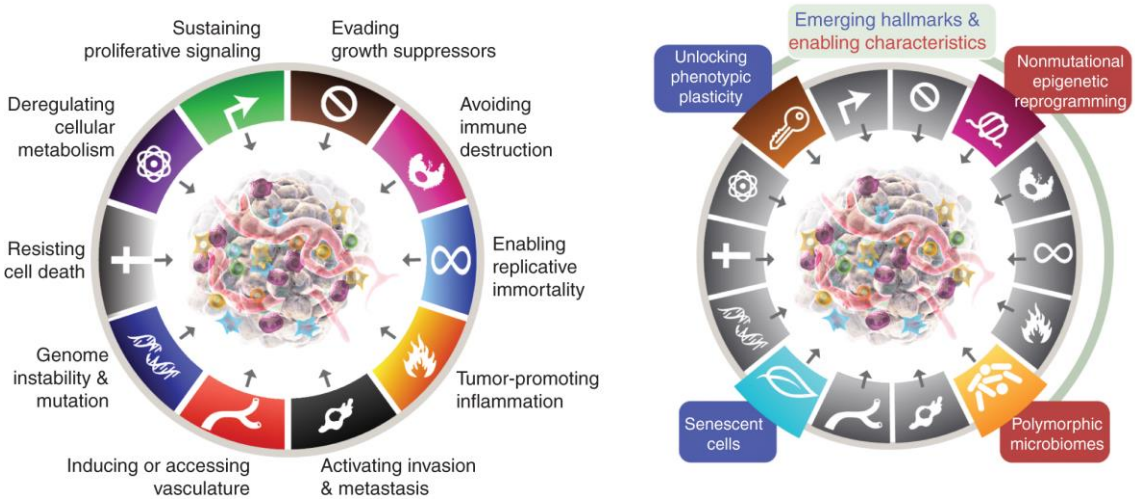


Figure 5: The hallmarks of cancer and their enabling characteristics. Left: the hallmarks and characteristics described in the first two “Hallmarks of Cancer” publications. Right: the new hallmarks and characteristics introduced in the latest “Hallmarks of Cancer” paper. Source: [66].

The above hallmarks describe common biological characteristics of different cancers. But the differences in tumor biology between tissues are sufficient to treat the diseases as different. For example, different genes are mutated in different tissues, transcriptional aberrations differ [67], and tumors from different source tissues metastasize to different locations. Importantly, these differences between tissues are not reflected only in the biology of the tumors, but also in the response to treatment. Drugs that are effective in one tissue are not necessarily effective in another, and the treatment regime largely depends on the tissue of origin [68].

However, the partition into tumor types does not stop at the tissue level. Even within a tissue and cell type there is high heterogeneity between tumors, both in their biology and in their response to therapy [69]. This variance led to the definition of *cancer subtypes* – further partitions of classical cancer types. One example for subtypes is seen in breast cancer [70]. There, a subset of the tumors expresses receptors for the estrogen hormone, which sustain the tumor's growth. For these tumors, Tamoxifen is a potent drug. It works by preventing the binding of estrogen to its receptor in the tumor, thus depriving the tumor of its growth sustaining signals. The same drug would not be effective in other breast tumors, which do not have high expression of the estrogen receptor. Much like estrogen, two other types of breast cancers are characterized by a receptor they express and that sustains their growth – one is the progesterone receptor and the other is the HER2 gene which is a growth factor receptor. A fourth breast cancer type is called "triple-negative", and is characterized by the lack of any of the three growth sustaining receptors we mentioned.

Breast cancer is not the only cancer with subtypes. Other examples include Acute Promyelocytic Leukemia, a subtype of Acute Myeloid Leukemia (AML) that responds to ATRA therapy [71], and Philadelphia syndrome, a subtype of Chronic Myeloid Leukemia that responds to Imatinib [72]. In these examples the subtype is associated with a treatment, but this is not necessarily the case.

Many cancer subtypes were defined before omic technologies were invented. These definitions usually depend on the tumor's histopathology, or on the presence of a single protein or chromosomal aberration. In contrast, omic data measures many features, which can all be used to characterize tumors. An early example of using omic data in cancer research is the work of Van't Veer et al. [73]. In this work, the authors developed a classifier that could predict with high accuracy whether a patient with breast cancer will develop metastases based on gene expression data. A patient with low risk of metastases may be spared of additional therapy after the initial treatment, which exacerbates side effects with the goal of decreasing metastases.

Van't Veer's work is an example for how to use omic data for patient classification, but omics data can also be used to detect cancer subtypes de-novo using clustering methods. Such an analysis, again for breast cancer, was performed by Perou et al. [74]. This analysis identified subtypes of breast cancer based on gene expression, and these subtypes were associated with the presence of the classical breast cancer subtypes proteins (e.g. estrogen).

Cluster analyses have since been performed on many cancer types, using many omics. The TCGA project published a paper for each cancer type, and in these works clustering tumors based on omic data was a standard part of the analysis (e.g., [75]–[78]). Not all cancer types revealed

clusters that are strongly matching known tumor biology, and neither do all cancer types contain a clear discrete partition into subtypes, but clustering omics data has become a widely used method to better understand cancer's biology.

Many methods were used to cluster cancer data [79]. These include classical clustering algorithms, such as k-means, hierarchical clustering, spectral clustering, self-organizing maps, and consensus clustering. They also include methods that were developed specifically for biological data, such as CLICK [80].

1.2.2. Multi-omic clustering algorithms

Early works for omic cancer subtyping used single-omic data such as gene expression. The TCGA initiative, which collected multi-omic datasets, also used at first only single-omics clustering to define cancer subtypes, and then analyzed the distribution of features from other omics on the obtained clustering. But later works from TCGA used integrative multi-omic clustering. For example, lung adenocarcinoma data was clustered using copy number, DNA methylation and mRNA expression and was found to contain six clusters [81]. Cluster membership was significantly associated with mutations in specific genes, even though this omic was not used for the clustering.

Another prominent example for multi-omic clustering in TCGA is the work of Hoadley et al. (Figure 6) [82]. In this work, data from 12 cancer types was clustered using five omics – DNA copy number, DNA methylation, gene expression, miRNA expression and protein expression. The goal of this work was not to define cancer subtypes in a classical cancer type, but to look for clusters across cancer types. Interestingly, a cluster that included tumors from several cancers was found. This cluster included tumors of squamous cells: lung squamous cell carcinoma, head and neck cancer, and a subset of the bladder cancers.

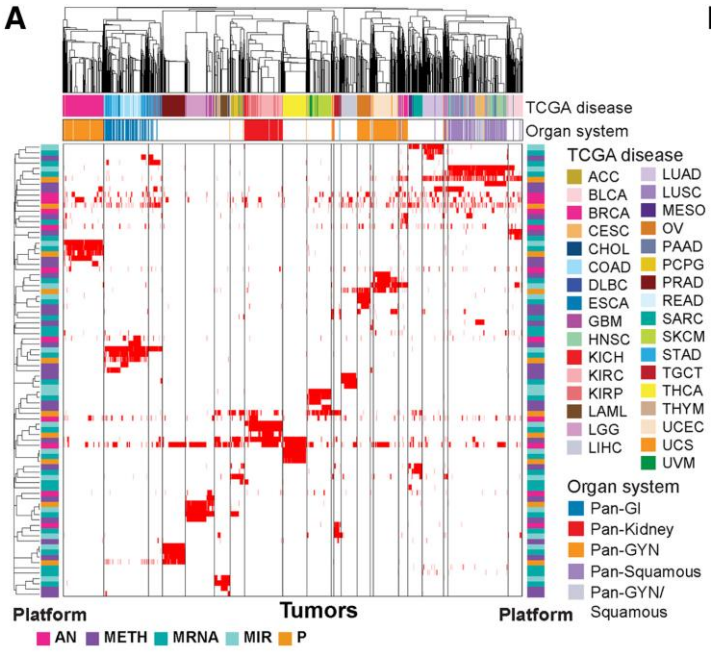


Figure 6: Multi-omics clustering of pan-cancer data from TCGA. Columns are tumors, and rows represent features from five different omics. Source: [82].

The algorithmic approaches used for multi-omic clustering are diverse. These approaches are explained in more detail in Chapter 2, where citations on each approach are provided, but we will briefly review them here. First, integration methods can be classified based on when the different omics are integrated: early integration, intermediate integration or late integration. Early integration methods first concatenate features from all the different omics, and then use single-omic clustering methods. Their advantage is that there are many off-the-shelf, well-tested clustering methods whose input is a single matrix and can be used. Their disadvantage is that they further increase the dimensionality, and do not handle the different distributions of the different omics. Late integration methods first cluster each omic separately, and then integrate their clustering results. Again, an advantage is the ability to use demonstrated single-omic methods. Moreover, here different single-omic clustering methods can be used for different omics, and can be even methods that support only this one omic. The main disadvantage is the inability to integrate heterogeneous signals from across different omics. Finally, intermediate integration methods try to build a model that incorporates all omics, without explicitly concatenating them.

Intermediate algorithms can also be classified into major categories. Dimension reduction methods assume that the data come from a low dimension, and that the different omics represent different mappings of that low dimensional data into higher dimension. Each omic has a different mapping function, and thus the different distributions of the omics are handled. The mapping functions are often assumed to be linear, so that the model could be fit, thus limiting

the expressibility of the models. A second, not entirely disjoint category, is made of statistics-based methods. These methods often suggest a generative model for the different omics, while modeling the inter-omic correlations. A statistical method worth mentioning here is PARADIGM [83]. This method most explicitly presents a model of the cell, and considers known relations between cellular pathways as part of its model. A Third category of models is similarity-based methods. These methods calculate similarities in each omic, and these similarities are then used for clustering in an integrative manner. The similarity calculation offers several advantages. First, it abrogates the need for a parametric model of the data, such as the linear mappings in dimension reduction-based methods. Second, similarity values are more easily compared between different omics. Third, different similarity measures can be used for different omics, thus making the similarity calculation omic-specific. On the other hand, these methods do not offer insights at the feature level, such as which features had strong effect on the obtained clustering. They also do not show connections between feature, either within an omic or between omics.

A similar problem to multi-omics clustering, called multi-view clustering, was investigated in the machine-learning community [84]. In this problem, samples need to be clustered based on different types of measured data, and each type is called a view. For example, if we consider data where samples are movies, one view can be the movie's images, and another view can be the movie's script. Multi-view methods are usually not specific to one data type, but can be applied to diverse data types, including omic data. Therefore, multi-view clustering methods can be used as multi-omic clustering methods.

1.3. Single-cell Hi-C in embryonic development

1.3.1. Genome organization and Hi-C

DNA is a molecule that appears in all living creatures. In Eukaryotes, the DNA is at the cell's nucleus, and it is often divided into several chromosomes. The DNA's physical structure and location in the nucleus were speculated for a long time to be involved in cellular regulation, but tools to investigate its structure were lacking. In the 1980s fluorescence in situ hybridization (FISH) methods produced the first evidence for the existence of chromosomal territories [85], [86]. That is, that specific chromosomes tend to be located in specific parts of the nucleus. Moreover, these methods showed that parts of the DNA that contain highly expressed genes tend to be located in the center of the nucleus, while parts of the chromosome without highly expressed genes tend to be at the nucleus' periphery [87].

In 2002 Dekker et al. developed Capturing Chromosome Conformation (3C), a technique to measure the physical distance between pairs of segments of the genome [88]. The experiment works basically as follows: First, DNA is cross-linked using Formaldehyde. The DNA is then cut and ligated again. Since the cut DNA molecules can move only a little between the cutting and the ligation, the probability that two pieces of DNA will be ligated to one another is proportional to their physical distance before the cutting took place. Using PCR with primers for the two bits of DNA of interest, the physical distance between the two DNA parts can be assessed.

3C was extended into circular 3C (4C) [89]. While 3C measures the physical contact between a specific pair of genomic loci (one vs. one), 4C measures the contact of one locus of interest with all other DNA in the cell (one vs. many). Following 4C, Hi-C, which measures contacts between all pairs of loci in the genome (many vs many), was introduced by Lieberman-Aiden et al. [90]. Hi-C works similarly to 3C, but it sequences all ligated DNA molecules, and the proportion with which two DNA loci are seen in the same sequenced molecules is used to estimate their physical distance. Following Hi-C, other methods in the field of genome organization were developed. ChIA-PET [91] and HiChIP [92] are similar to Hi-C, but they measure genomic physical interactions mediated by a specific protein of interest. Micro-C uses MNase to digest cross-linked DNA, instead of the cutting that is done in Hi-C using restriction enzymes, and allows for genome structure investigation at higher resolution [93].

Investigations into the genome structure revealed several main concepts. First is the division of the genome into compartments [94]. As mentioned previously when discussing FISH, highly transcribed regions of the genome are found in the center of the nucleus, and lowly transcribed regions are in the periphery. These two parts of the genome are known as the A and B compartments, respectively, and can be found in Hi-C data by clustering the genome into two parts that tend to have many contacts within themselves and a low number of contacts between them (see Figure 7). Further research found that the compartments correspond to euchromatin and heterochromatin, that they are overrepresented with specific histone modifications (e.g. H3K9me3 in heterochromatin), and that the A compartments replicates before the B compartment during S-phase [95], [96]. Some works attempted to describe chromosomal compartments in higher resolution using more than two compartments [97], but A/B remains the most widely used compartmentalization.

Another concept related to chromosomal organization and observed in Hi-C is that of Topologically Associating Domain (TAD) [98]. These are contiguous stretches of DNA that tend to self-interact, and that manifest as blocks in the Hi-C contact matrix (Figure 7). The mechanism

for the formation of such TADs is of high interest, and was found to involve Cohesin rings that advance along the chromosome until they are blocked by CTCF proteins bound to specific binding sites on the DNA [99]–[101]. While the formation of TADs is now becoming clearer, their functional role is still investigated. Among its suggested roles is limiting possible enhancer-promoter interactions to pairs that are mainly in the same TAD. TADs were shown to form a coherent unit of DNA replication – DNA in the same TAD will tend to replicate together in the cell cycle [96]. A concept related to TADs is that of DNA loops. These are cases where two genomic loci are found to be physically proximal, while their adjacent DNA is not, distinguishing it from a TAD.

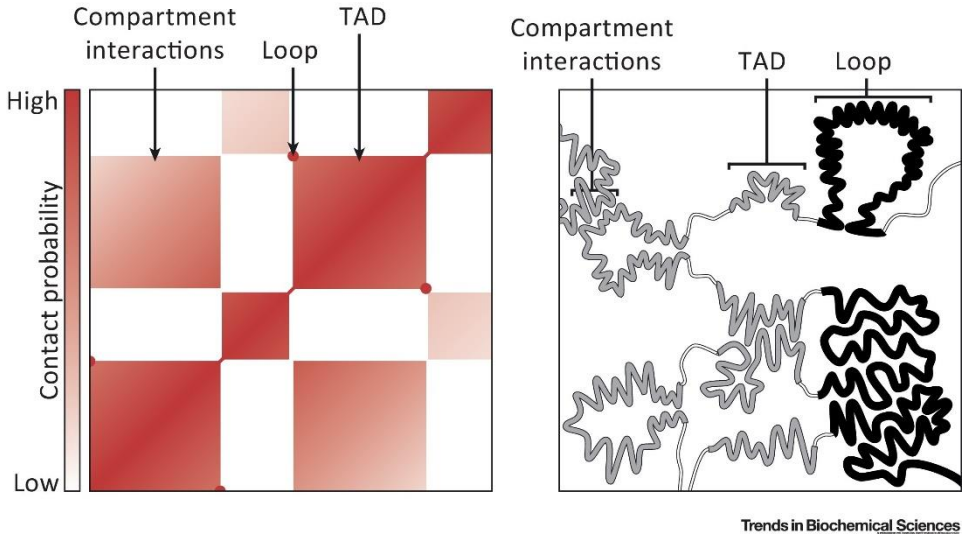


Figure 7: Left: schematic figure for the Hi-C contact matrix. The rows and columns of the matrix represent genomic bins, and the value (color intensity) is the number of contacts observed between the two genomic bins. The figure illustrates how compartments, TADs, and loops look in the matrix. Right: a schematic figure of the physical structure of the DNA that may give rise to compartments, TADs, and loops as they are observed in the contact matrix. Source: [102].

The roles of compartments, TADs and loops in genomic regulation is under active research. Notably, TADs were shown to be relatively consistent across different cell types [97], while compartments show higher variance – genomic bins whose genes are more highly expressed in a cell type will tend to be more strongly associated with the A compartment, consistent with the connection we mentioned previously between gene expression and compartments. Supporting the role of genome organization in gene regulation is its disruption seen in diseases. For example, TAD disruption is commonly observed in cancers [103], either through local deletions or as part of larger genome rearrangements, and TAD disruptions were also found in developmental diseases [104].

Hi-C data gives rise to several computational challenges. First, calling compartments and TADs from the Hi-C data is non-trivial. Regions of the genome with close genomic coordinates (that is,

that are close on the 2D structure of the genome) will also tend to be close in 3D. But to describe TADs and compartments, we need to find physical contacts that are more frequent than would be expected based on their 2D proximity. In that sense, Hi-C differs from other omic data in that its measured signal is inherently probabilistic (and not only due to molecule sampling). A read that is sequenced in RNA-seq is evidence for a transcript of some gene. A read sequenced in Hi-C is not indicative of physical contact between these two regions, but only of an increased probability that these regions are close.

To this end, various ways to normalize the contact frequency for the 2D coordinate distance were employed. For example, the most common method to detect compartments involves normalizing contacts for distances, converting the Hi-C matrix into a correlation matrix, and computing the first or second principal component of that matrix. The sign assigned to each genomic region indicates whether it belongs to the A or B compartment [90].

Computational tasks in Hi-C analysis are not limited to the detection of organizational entities such as compartments and TADs, but also include their comparison. For example, several methods were developed to call differential interactions in Hi-C data [105]. Such differential analyses are paramount to understand how genome organization varies across cell types, and in disease.

1.3.2. Mouse embryonic development

Embryonic development investigates the process by which a zygote becomes a mature organism. In addition to the applications for human fertility and developmental disorders that research into embryology enables, it is also an important model to investigate cellular differentiation – the process by which cells specialize to perform a specific function.

Early research in embryology, conducted before the era of molecular biology, established the concept of differentiation potential. A prominent example is the Spermann and Mangold experiment, in which they took cells from the dorsal side of an amphibian embryo and transplanted them in a different place on the embryo [106]. The transplanted cells led to the development of second set of body structures, demonstrating that the cellular fate of the transplanted cells was already determined. Such experiments and others demonstrate the concept of differentiation potential - a cell has the potential to create a specific set of cell types, and this set decreases during development and differentiation. While early embryonic cells can create a whole embryo, later ones can create only more specialized cell types.

With the advent of molecular biology, the cellular processes by which differentiation occurs could be investigated. Some of the factors that were found to direct these processes include transcription factors (TFs), secreted morphogens and growth factors, and cell-cell interactions [107]–[109]. Our understanding of the role of TFs can be demonstrated by the work of Takahashi et al. [110]. Research into TFs active in embryonic stem cells (ESC) allowed for a screen of TFs that will cause mature cells to revert to ESCs (Figure 8). This research has transformed research in embryology and in human biology in general.

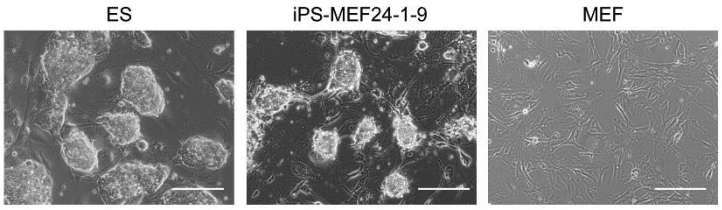


Figure 8: Morphology of mouse ES cells (left), mature fibroblasts reprogrammed to ES cells, and mouse embryonic fibroblasts (right). Source: [110].

Morphogens are secreted molecules whose gradual change affect cell fate decisions. Some of the most prominent pathways involving morphogens are the Wnt, SHH and BMP pathways. For example, gradients of SHH proteins were shown to be involved in limb development [111]. Cell-cell interactions are important for development both because of the molecules exchanged between cells, and because of the mechanical forces that cells exert on one another [107]. SHH, which we mentioned in the context of limb development, is also secreted by the notochord and affects motor neuron development [112]. The Hippo signaling pathways, which relies on cell density and the mechanical stress it causes, is important for the first fate decision to take place in the embryo – that of differentiating to either the inner cell mass (which will proceed to form the embryo), or to the trophectoderm (which will form the extraembryonic tissues) [113].

Epigenetic factors are also of great importance in embryogenesis. For example, DNA methylation was observed to undergo two major reprogramming events – one in the creation of the primordial germ cells (PGCs), and one in the early steps of development post fertilization [114]. The causative role of DNA methylation in development was demonstrated using knock-outs of major methylation modifying enzymes, such as the methyltransferases DNMT1, DNMT3A and DNMT3B, and the demethylases TET1-3 [115]. Other epigenetic factors, such as histone modification, and their relation to chromatin modifying enzymes such as the Polycomb group complex, were also shown to have an integral role in embryogenesis [116], [117].

The large cell type diversity and the small number of cells make embryonic development especially suited to benefit from single-cell analyses. Indeed, single-cell studies have been used

extensively to study embryonic development, including mammalian development, which is more difficult to study with traditional embryology research methods (Figure 9) [118], [119]. These studies characterized in high detail the cell types and transcriptional programs that shape development.

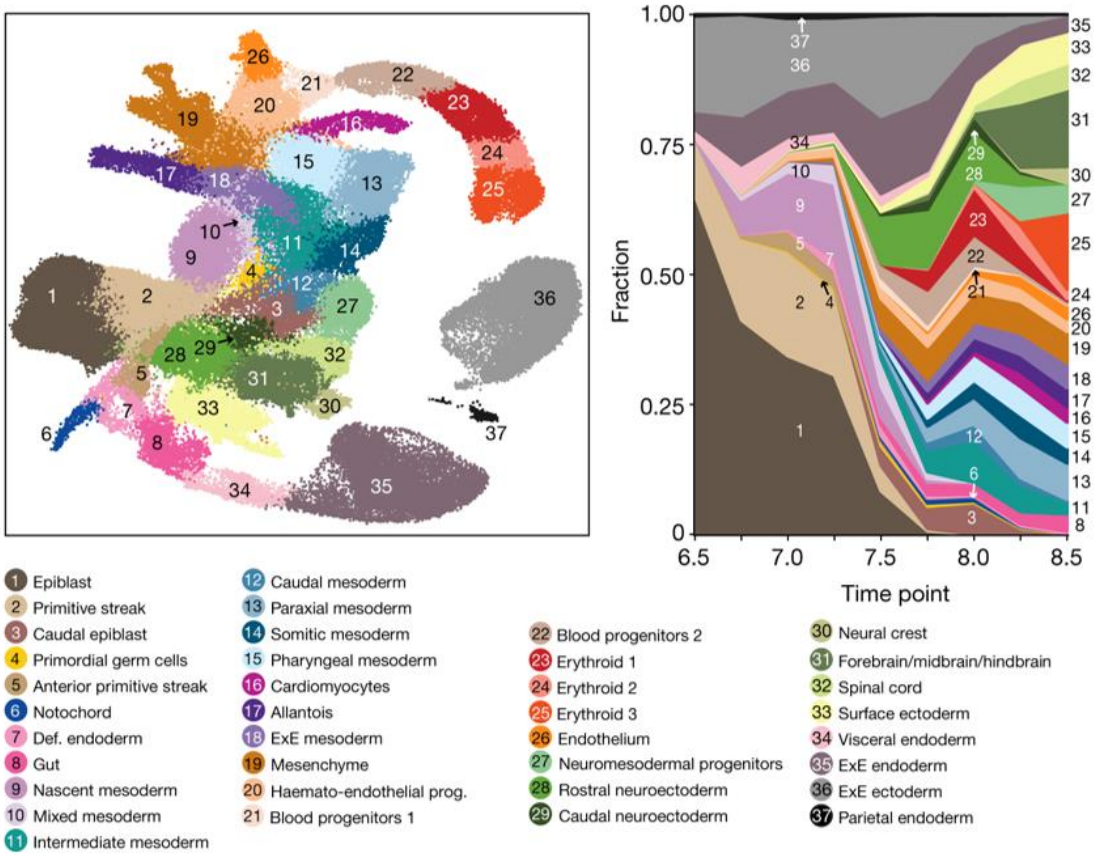


Figure 9: left: UMAP plot of scRNA-seq data from about 120K mouse embryo cells, taken from mice at ages E6.5-E8.5. Cells are colored by their annotated cell type. Right: cell type proportions during mouse embryonic development, calculated from scRNA-seq data. Source: [119].

1.3.3. Single-cell Hi-C

While the role of several epigenetic factors in embryonic development is under investigation, the role of the genome organization has not been sufficiently studied. Hi-C is traditionally performed in bulk, and the average signal from many embryonic cell types is not illuminating. Sorting embryos for specific cell types is also problematic, because of the low number of cells in an embryo. The chromosomal conformation in embryos is therefore poorly understood. As mentioned previously, Hi-C experiments showed large differences between mature tissues. How these changes are established during embryonic development, and what is their role in differentiation, remains to be discovered. To overcome the shortcomings of bulk analysis, single-cell Hi-C is needed.

The first work to perform single-cell Hi-C (scHi-C) was done by Nagano et al. [120]. In this work, the authors developed the first scHi-C protocol, and applied it to 74 mouse splenic CD4+ T-cells. Their analysis revealed low variation between cells in terms of intradomain contacts, but showed variance in terms of interdomain contacts. The conserved intradomain structure is consistent with bulk studies showing similar TAD structure across cell types. The varying interdomain contacts suggest that some compartmental differences exist between cells, and not only between cell types. The work also bridged between the inter-chromosomal structure observed in microscopy and in bulk Hi-C data. Chromosomes in the microscope appear to be well separated, but bulk Hi-C data finds many contacts between chromosomes. Nagano et al. suggested that at the single-cell level, contacts between chromosomes are mostly seen between specific pairs of chromosomes that are presumably physically close. The abundance of contacts between chromosomes seen in bulk Hi-C is due to different chromosome pairs being adjacent in different cells.

A follow-up work from Nagano et al. analyzed the effect of the cell cycle on the genome organization [121]. The scHi-C protocol in this work was improved into a version very similar to bulk Hi-C. In this protocol, DNA crosslinking, cutting and ligation are performed in bulk. Cells are then sorted into 96-well plates, where a tagmentation step takes place, such that each cell is marked with a specific barcode, similar to other single-cell experimental methods.

Nagano et al. performed scHi-C on 2000 mouse embryonic stem cells in this study. The authors developed a method to order cells along the cell cycle, based on two main statistics (Figure 10). The first is the distribution of the chromosomal distances between the two ends of all sequenced reads from a cell. The second is the ratio between reads sequenced from early replicating regions compared to late replicating regions. The authors' approach was confirmed using cells that were sorted using FACS and known cell cycle measures. The study then continued to delineate changes in chromosomal structure during the cell cycle, given the inferred computational ordering. The investigators identified that insulation (which measures the strength of the TAD structure) increases rapidly following exit from M-phase, further increases during G1, decreases during replication and decreases more rapidly when entering M-phase. The compartment structure presents different dynamics, with an increase in compartmentalization from the exit from M, all the way until late G2, and then a decrease again when re-entering M.

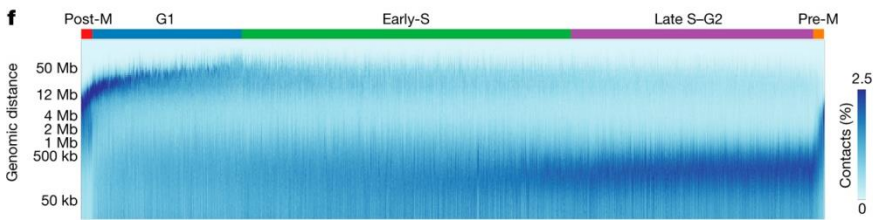


Figure 10: Contact distance distribution in scHi-C data. Each column represents a cell, and each row a specific range of genomic distances. The values (colors) represent the fraction of contacts in a cell that come from each range of genomic distances. Cells are ordered by their inferred point in the cell cycle. Source: [121].

Other works developed techniques to describe the genome organization of single-cells. Stevens et al. used an approach based on imaging followed by Hi-C to investigate mouse embryonic stem cells[122]. In contrast to previous reports, they found variance in TAD structure between cells, but high conservation of the A/B structure. Ramani et al. used an approach based on combinatorial indexing, where each cell has two distinct barcodes[123]. Two cells can receive the same barcode, but the probability that two cells will receive the same two barcodes is low. They create six scHi-C libraries from synthetic mixtures of different cell lines. Lee et al. developed a method that measures both DNA methylation and Hi-C for single cells and applied it to the human prefrontal cortex[124]. To the best of our knowledge, this is the first work that profiled a tissue with diverse cell types, and showed that they can be distinguished based on their Hi-C profiles. However, the prefrontal cortex is made of terminally differentiated cells, and no work used scHi-C to investigate a differentiating system, including in embryogenesis.

Algorithms for the analysis of scHi-C data are scarce. The noisy signal, sparsity, and paucity of scHi-C data are all challenges facing such algorithms. Nevertheless, several methods were developed. scHiCluster performs imputation on scHi-C data using convolution (smoothing across near neighbors) followed by random walks-based imputation[125]. scHiCluster also uses dimension reduction on the imputed data to study it. Another algorithm does not perform imputation, but is interested only in dimension reduction, and uses topic modeling for that purpose[126]. More recently, a method called Higashi was developed[127]. Higashi is based on hypergraph representation learning and a neural network architecture, and like scHiCluster it imputes the Hi-C matrix. However, there is still much need for algorithms that can analyze scHi-C data.

1.4. Summary of articles included in this thesis

1. **Multi-omic and multi-view clustering algorithms: review and cancer benchmark**
   Nimrod Rappoport and Ron Shamir

Nucleic Acids Research, Volume 46, Issue 20, 16 November 2018, Pages 10546–10562 [1].

Recent high throughput experimental methods have been used to collect large biomedical omics datasets. Clustering of single omic datasets has proven invaluable for biological and medical research. The decreasing cost and development of additional high throughput methods now enable measurement of multi-omic data. Clustering multi-omic data has the potential to reveal further systems-level insights, but raises computational and biological challenges. Here, we review algorithms for multi-omics clustering, and discuss key issues in applying these algorithms. Our review covers methods developed specifically for omic data as well as generic multi-view methods developed in the machine learning community for joint clustering of multiple data types. In addition, using cancer data from TCGA, we perform an extensive benchmark spanning ten different cancer types, providing the first systematic comparison of leading multi-omics and multi-view clustering algorithms. The results highlight key issues regarding the use of single- versus multi-omics, the choice of clustering strategy, the power of generic multi-view methods and the use of approximated p-values for gauging solution quality. Due to the growing use of multi-omics data, we expect these issues to be important for future progress in the field.

2. NEMO: cancer subtyping by integration of partial multi-omic data

Nimrod Rappoport and Ron Shamir

Bioinformatics, Volume 35, Issue 18, September 2019, Pages 3348–3356 [2].

Motivation - Cancer subtypes were usually defined based on molecular characterization of single omic data. Increasingly, measurements of multiple omic profiles for the same cohort are available. Defining cancer subtypes using multi-omic data may improve our understanding of cancer, and suggest more precise treatment for patients. **Results** - We present NEMO (NEighborhood based Multi-Omics clustering), a novel algorithm for multi-omics clustering. Importantly, NEMO can be applied to partial datasets in which some patients have data for only a subset of the omics, without performing data imputation. In extensive testing on ten cancer datasets spanning 3168 patients, NEMO achieved results comparable to the best of nine state-of-the-art multi-omics clustering algorithms on full data and showed an improvement on partial data. On

some of the partial data tests, PVC, a multi-view algorithm, performed better, but it is limited to two omics and to positive partial data. Finally, we demonstrate the advantage of NEMO in detailed analysis of partial data of AML patients. NEMO is fast and much simpler than existing multi-omics clustering algorithms, and avoids iterative optimization.

Availability and implementation - Code for NEMO and for reproducing all NEMO results in this paper is in github: https://github.com/Shamir-Lab/NEMO.

3. **Inaccuracy of the log-rank approximation in cancer data analysis**

   Nimrod Rappoport and Ron Shamir

   Mol Syst Biol. (2019) 15: e8754 [3].

   Comparing the survival between different groups of patients is widely used in cancer research, and as a means to benchmark cancer clustering algorithms. The most commonly used statistical test for such comparisons is the log-rank test. In this work we show that most software tools use an asymptotic version of the test, which is highly inaccurate in datasets with the number of patients observed in cancer datasets. We show that the reported p-values overstate the significance of the results, highlight previous false discoveries made using this test, and provide an implementation for an exact test for multiple groups.

4. **MONET: Multi-omic module discovery by omic selection**

   Nimrod Rappoport, Roy Safra, Ron Shamir

   PLOS Computational Biology 16(9): e1008182 [4].

   Recent advances in experimental biology allow creation of datasets where several genome-wide data types (called omics) are measured per sample. Integrative analysis of multi-omic datasets in general, and clustering of samples in such datasets specifically, can improve our understanding of biological processes and discover different disease subtypes. In this work we present MONET (Multi Omic clustering by Non-Exhaustive Types), which presents a unique approach to multi-omic clustering. MONET discovers modules of similar samples, such that each module is allowed to have a clustering structure for only a subset of the omics. This approach differs from most existent multi-omic clustering algorithms, which assume a common structure across all omics, and from

several recent algorithms that model distinct cluster structures. We tested MONET extensively on simulated data, on an image dataset, and on ten multi-omic cancer datasets from TCGA. Our analysis shows that MONET compares favorably with other multi-omic clustering methods. We demonstrate MONET's biological and clinical relevance by analyzing its results for Ovarian Serous Cystadenocarcinoma. We also show that MONET is robust to missing data, can cluster genes in multi-omic dataset, and reveal modules of cell types in single-cell multi-omic data. Our work shows that MONET is a valuable tool that can provide complementary results to those provided by existent algorithms for multi-omic analysis.

5. **Single cell Hi-C identifies plastic chromosome conformations underlying the gastrulation enhancer landscape**

Nimrod Rappoport*, Elad Chomsky*, Takashi Nagano, Charlie Seibert, Yaniv Lubling, Yael Baran, Aviezer Lifshitz, Wing Leung, Zohar Mukamel, Ron Shamir, Peter Fraser, Amos Tanay. * denotes equal contribution.

Accepted to Nature Communications.

Embryonic development involves massive proliferation and differentiation of cell lineages. This must be supported by chromosome replication and epigenetic reprogramming, but how proliferation and cell fate acquisition are balanced in this process is not well understood. Here we use single cell Hi-C to map chromosomal conformations in post-gastrulation mouse embryo cells and study their distributions and correlations with matching embryonic transcriptional atlases. We find that embryonic chromosomes show a remarkably strong cell cycle signature. Despite that, replication timing, chromosome compartment structure, topological associated domains (TADs) and promoter-enhancer contacts are shown to be variable between distinct epigenetic states. About 10% of the nuclei are identified as primitive erythrocytes, showing exceptionally compact and organized compartment structure. The remaining cells are broadly associated with ectoderm and mesoderm identities, showing only mild differentiation of TADs and compartment structures, but more specific localized contacts in hundreds of ectoderm and mesoderm promoter-enhancer pairs. The data suggest that while fully committed embryonic lineages can rapidly acquire specific chromosomal conformations, most embryonic cells are showing plastic signatures driven by complex and intermixed enhancer landscapes.

Chapter 2

Multi-omic and multi-view clustering algorithms: review and cancer benchmark

10546–10562 *Nucleic Acids Research, 2018, Vol. 46, No. 20* *Published online 8 October 2018*
doi: 10.1093/nar/gky889

SURVEY AND SUMMARY

Multi-omic and multi-view clustering algorithms: review and cancer benchmark

Nimrod Rappoport and Ron Shamir*

Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Received July 18, 2018; Revised September 17, 2018; Editorial Decision September 19, 2018; Accepted September 20, 2018

ABSTRACT

Recent high throughput experimental methods have been used to collect large biomedical omics datasets. Clustering of single omic datasets has proven invaluable for biological and medical research. The decreasing cost and development of additional high throughput methods now enable measurement of multi-omic data. Clustering multi-omic data has the potential to reveal further systems-level insights, but raises computational and biological challenges. Here, we review algorithms for multi-omics clustering, and discuss key issues in applying these algorithms. Our review covers methods developed specifically for omic data as well as generic multi-view methods developed in the machine learning community for joint clustering of multiple data types. In addition, using cancer data from TCGA, we perform an extensive benchmark spanning ten different cancer types, providing the first systematic comparison of leading multi-omics and multi-view clustering algorithms. The results highlight key issues regarding the use of single- versus multi-omics, the choice of clustering strategy, the power of generic multi-view methods and the use of approximated p-values for gauging solution quality. Due to the growing use of multi-omics data, we expect these issues to be important for future progress in the field.

INTRODUCTION

Deep sequencing and other high throughput methods measure a large number of molecular parameters in a single experiment. The measured parameters include DNA genome sequence (1), RNA expression (2,3), DNA methylation (4) etc. Each such kind of data is termed 'omic' (genomics, transcriptomics, methylomics, respectively). As costs decrease and technologies mature, larger and more diverse omic datasets are available.

Computational methods are imperative for analyzing such data. One fundamental analysis is clustering - finding coherent groups of samples in the data, such that samples within a group are similar, and samples in different groups are dissimilar (5). This analysis is often the first step done in data exploration. Clustering has many applications for biomedical research, such as discovering modules of co-regulated genes and finding subtypes of diseases in the context of precision medicine (6). Clustering is a highly researched computational problem, investigated by multiple scientific communities, and a myriad algorithms exist for this task.

While clustering each omic separately reveals patterns in the data, integrative clustering using several omics for the same set of samples has the potential to expose more fine-tuned structures that are not revealed by examining only a single data type. For example, cancer subtypes can be defined based on both gene expression and DNA methylation together. There are several reasons why a clustering based on multiple omics is desirable. First, Multi-omics clustering can reduce the effect of experimental and biological noise in the data. Second, different omics can reveal different cellular aspects, such as effects manifest at the genomic and epigenomic levels. Third, even within the same molecular aspect, each omic can contain data that are not present in other omics (e.g. mutation and copy number). Fourth, omics can represent data from different organismal levels, such as gene expression together with microbiome composition.

A problem akin to multi-omics clustering was investigated independently by the machine learning community, and is termed 'multi-view clustering' (see (7) and 'A Survey on Multi-View Clustering'). Multi-view clustering algorithms can be used to perform clustering of multi-omic data. In the past, methods developed within the machine learning community have proven useful in the analysis of biomedical

*To whom correspondence should be addressed. Tel: +972 3 640 5383; Fax: +972 3 640 5384; Email: rshamir@tau.ac.il

© The Author(s) 2018. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

datasets. However, by and large, multi-view clustering have not penetrated bioinformatics yet.

In this paper, we review methods for multi-omics clustering, and benchmark them on real cancer data. The data source is TCGA (The Cancer Genome Atlas) (8)—a large multi-omic repository of data on thousands of cancer patients. We survey both multi-omics and multi-view methods, with the goal of exposing computational biologists to these algorithms. Throughout this review, we use the terms *view* and *multi-view* instead of omic and multi-omics in the context of Machine Learning algorithms.

Several recent reviews discussed multi-omics integration. (9–11) review methods for multi-omics integration, and (12) review multi-omics clustering for cancer application. These reviews do not include a benchmark, and do not focus on multi-view clustering. (13) reviews only dimension reduction multi-omics methods. To the best of our knowledge, (14) is the only benchmark performed for multi-omics clustering, but it does not include machine learning methods. Furthermore, we believe the methods tested in the benchmark do not represent the current state of the art for multi omics clustering. Finally, (7) is a thorough review of multi-view methods, directed to the Machine Learning community. It does not discuss algorithms developed by the bioinformatics community, and does not cover biological applications.

REVIEW OF MULTI-OMICS CLUSTERING METHODS

We divide the methods into several categories based on their algorithmic approach. *Early integration* is the most simple approach. It concatenates omic matrices to form a single matrix with features from multiple omics, and applies single-omic clustering algorithms on that matrix. In *late integration*, each omic is clustered separately and the clustering solutions are integrated to obtain a single clustering solution. Other approaches try to build a model that incorporates all omics, and are collectively termed *intermediate integration*. Those include: (i) methods that integrate sample similarities, (ii) methods that use joint dimension reduction for the different omics datasets and (iii) methods that use statistical modeling of the data.

The categories we present here are not clear-cut, and some of the algorithms presented fit into more than one category. For example, iCluster (15) is an early integration approach that also uses probabilistic modeling to project the data to a lower dimension. The algorithms are described in the categories where we consider them to fit most.

Multi-omics clustering algorithms can also be distinguishable by the set of omics that they support. *General* algorithms support any kind of omics data, and are therefore easily extendible to novel future omics. *Omic specific* algorithms are tailored to a specific combination of data types, and can therefore utilize known biological relationships (e.g. the correlation between copy number and expression). A mixture of these two approaches is to perform feature learning in an omic specific way, but then cluster those features using general algorithms. For example, one can replace a gene expression omic with an omic that scores expression in cellular pathways, and thus take advantage of existing biological knowledge.

Throughout this review, we use the following notation: a multi-omic dataset contains M omics. n is the number of samples (or patients for medical datasets), p_m is the number of features in the m'th omics, and X^m is the n x p_m matrix with measurements from the m'th omic. X^m_{ij} is therefore the value of the j'th feature for the i'th patient in the m'th omic. $p = \Sigma_{m=1}^{M} p_m$ is the total number of features, and X is the $n \times p$ matrix formed by the concatenation of all X^m matrices. Throughout the paper, for a matrix A, we use A^t to designate its transpose, and consistently with the X^m notation, we use A^m for matrix indexing (and not for matrix powering). Additional notation is chosen to follow the original publications and common conventions.

Figure 1 summarizes pictorially the different approaches to multi-omics clustering. A summary table of the methods reviewed here is given in Table 1.

Early integration

Early integration is an approach that first concatenates all omic matrices, and then applies single-omic clustering algorithms on that concatenated matrix. It therefore enables the use of existing clustering algorithms. However, this approach has several drawbacks. First, without proper normalization, it may give more weight to omics with more features. Second, it does not consider the different distribution of data in the different omics. Finally, it increases the data dimension (the number of features), which is a challenge even in some single-omic datasets. When applying early integration algorithms designed specifically for multi-omics data, or when running single-omic methods on a concatenated matrix, these drawbacks must be addressed. Normalization of the features in different omics can assist in handling the different distributions, and feature selection can be used to decrease the dimension and to give different omics an equal prior opportunity to affect the results.

An additional way to handle the high dimension of the data is by using regularization, i.e. adding additional constraints to a problem to avoid overfitting (76). Specifically, LASSO (Least Absolute Shrinkage and Selection Operator) regularization creates models where the number of features with non-zero effect on the model is low (77), and regularization of the nuclear norm is often used to induce data sparsity. Indeed, LASSO regularization is used by iCluster (15) (reviewed in a later section), and LRACluster uses nuclear norm regularization (reviewed in this section). While any clustering algorithm can be applied using early integration, we highlight here algorithms that were specifically developed for this task.

LRACluster (16) uses a probabilistic model, where numeric, count and binary features have distributions determined by a latent representation of the samples Θ. For example, X^m_{ij} is distributed $\propto exp(-\frac{1}{2}(X^m_{ij} - \Theta^m_{ij})^2)$, where Θ^m is of the same dimensions as X^m. The latent representation matrix is encouraged to be of low rank, by adding a regularization on its nuclear norm. The objective function for the algorithm is $-\log(\text{model's likelihood}) + \mu \cdot |\Theta|_*$ where Θ is the concatenation of all Θ^m matrices, and $|\cdot|_*$ is the nuclear norm. This objective is convex and provides a global optimal solution, which is found using a fast gradient-ascent algorithm. Θ is subsequently clustered using k-means. This

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

Table 1. Multi-omic clustering methods

Method	Description	Refs.	Implementation
Early integration			
LRAcluster•	Data originate from low rank matrix, omic data distributions modeled based on it	(16)	R
Structured sparsity	Linear transformation projects data into a cluster membership orthogonal matrix	(17)	Matlab
Alternate optimization			
MV k-means, MV EM	Alternating *k*-means and EM. Each iteration is done w.r.t. a different view	(18)	NA
Late integration			
COCA	Per omic clustering solutions integrated with hierarchical clustering	(19)	NA
Late fusion using latent models	Per omic clustering solutions integrated with PLSA	(20)	NA
PINS•	Integration of co-clustering patterns in different omics. The clusterings are based on perturbations to the data	(21)	R
Similarity-based methods			
Spectral clustering generalizations	Generalizations of similarity based spectral clustering to multi-omics data	(22–25)	Matlab
Spectral clustering with random walks	Generalizations of spectral clustering by random walks across similarity graphs	(26,27)	NA
SNF•	Integration of similarity networks by message passing	(28,29)	R, Matlab
rMKL-LPP•	DR using multiple kernel learning; similarities maintained in lower dimension	(30)	**
Dimension reduction			
General DR framework	General framework for integration with DR	(31)	NA
JIVE	Variation in data partitioned into joint and omic-specific	(32)	Matlab,R (33)
CCA•	DR to axes of max correlation between datasets. Generalizations: Bayesian, kernels, >2 omics, sparse solutions, deep learning, count data	(34–43), CCA for count data	R, two omics (44), R, multiple omics
PLS	DR to axes of max covariance between datasets. Generalizations: kernels, >2 omics, sparse solutions, partition into omic-specific and joint variation	(45–52)	R, two omics, Matlab, multiple omics
MCIA	DR to axes of max covariance between multi-omic datasets	(53)	R
NMF generalizations•	DR using generalizations of NMF to multi-omic data	(54–57), EquiNMF, (58,59)	MultiNMF (Matlab)
Matrix tri- factorization	DR. Each omic describes the relationship between two entities	(60)	NA
Convex methods	DR with convex objective functions, allowing unique optimum and efficient computation	(16,61,62)	Matlab
Low-rank tensor MV clustering	Factorization based on low-rank tensors	(63)	Matlab
Statistical methods			
iCluster/Plus/Bayes•	Data originate from low dimensional representation, which determines the distribution of the observed data	(15,64,65)	R
PARADIGM	Probabilistic model of cellular pathways using factor graphs	(66)	REST API
Disagreement between clusters	Methods based mainly on hierarchical Dirichlet processes; clustering in different omics need not agree	(67–71)	BCC (R)
Survival-based	Probabilistic model; patient survival data used in the clustering process	(72,73)	SBC (R)
Deep learning			
Deep learning methods	Neural networks used for integration. A variant of CCA, early integration and middle integration approaches	(37,74,75)	DeepCCA (Python)

DR: dimension reduction; EM: expectation maximization; MV: multi-view; PLSA: Probabilistic Latent Semantic Analysis; CCA: Canonical Correlation Analysis; PLS: Partial Least Squares; NMF: non-negative matrix factorization. •Methods included in the benchmark. Single-omic *k*-means and spectral clustering were also included in the benchmark. ** Available from the authors upon request.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

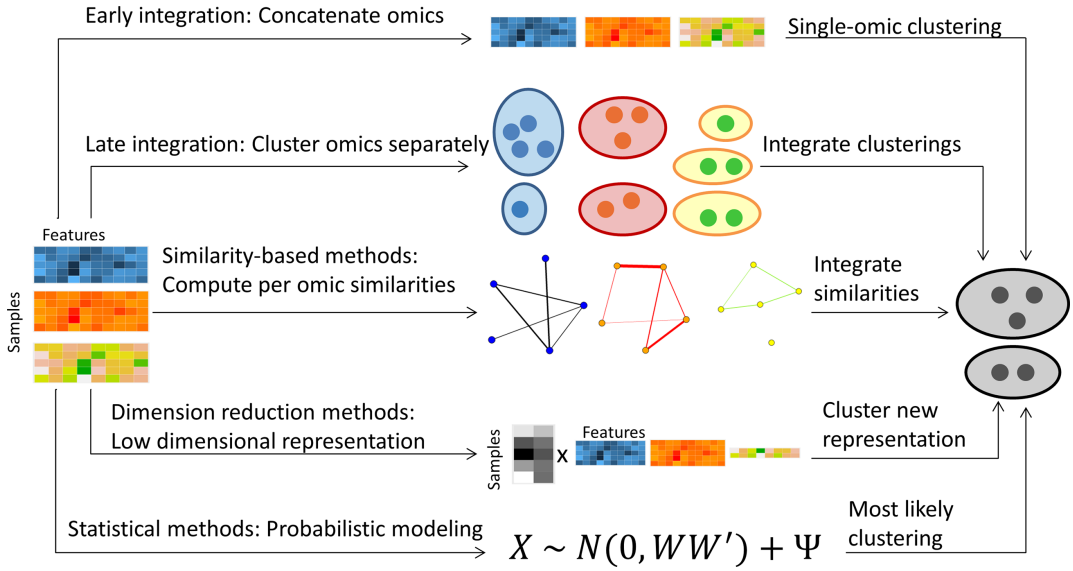


Figure 1. Overview of multi-omics clustering approaches.

method was used to analyze pan-cancer TCGA data from eleven cancer types using four different omics, and to further find subtypes within these cancer types.

In (17), all omics are concatenated to a matrix X and the algorithm minimizes the following objective: $||XW + 1_n b^t - F||_2^2 + \gamma||W||_{G_1}$. W is a p x k projection matrix, F is an n x k cluster indicator matrix such that $F^tF = I_k$, 1_n is a column vector of length n of 1's, b is an intercept column vector of dimension k and γ is a scalar. The algorithm therefore seeks a linear transformation such the projected data are as close to a cluster indicator matrix as possible. That indicator matrix is subsequently used for clustering. The regularization term uses the G_1 norm, which is the l_2 norm for W entries associated with a specific cluster and view, summed over all views and clusters. Therefore, features that do not contribute to the structure of a cluster will be assigned with low coefficients in W.

Alternate optimization

Early research for integration of two views was performed in (78). This work improved classification accuracy for semi-supervised data with two views using an approach termed co-training, and inspired others to analyze multi-view data. One of the first attempts to perform multi-view clustering was (18). In this work, EM (expectation maximization) and k-means, which are widely used single-omic clustering algorithms, were adapted for multi-view clustering. Both EM and k-means are iterative algorithms, where each iteration improves the objective function value. The suggested multi-view versions perform optimization in each iteration with respect to a different omic in an alternating manner. This approach loses theoretical guarantees for convergence, but was found to outperform algorithms that use each view separately, and also naive early integration methods that cluster the concatenated matrix of the two views. Interestingly, (18) report improved results using the multi-view clustering algorithms on single-view datasets that were randomly split to simulate multi-view data. This was the first evidence for improved clustering using multiple views, and for the utility of a multi-view algorithm in clustering single-view data. While this work was very influential, other preliminary multi-view clustering methods (e.g. (22,31)) were since shown to achieve better results on datasets where the gold standard is known.

Late integration

Late integration is an approach that allows to use existing single-omic clustering algorithms on single-omic data. First, each omic is clustered separately using a single-omic algorithm. Different algorithms can be used for each omic. Then, the different clusterings are integrated. The strength of late integration lies in that any clustering algorithm can be used for each omic. Algorithms that are known to work well on a particular omic can therefore be used, without having to create a model that unifies all of these algorithms. However, by utilizing only clustering solutions in the integration phase we can lose signals that are weak in each omic separately.

COCA (19) was applied to pan-cancer TCGA data, to investigate how tumors from different tissues cluster, and whether the obtained clusters match the tissue of origin. The algorithm first clusters each omic separately, such that the m'th omic has c_m clusters. The clustering of sample i for omic m is encoded in a binary vector v_{im} of length c_m, where $v_{im}(j) = 1$ if i belongs to cluster j and 0 otherwise. The concatenation of the v_{im} vectors across all omics results in a binary cluster indicator vector for sample i. The $n \times c$ binary matrix B of these indicator vectors, where $c = \Sigma_{i=1}^{M} c_m$, is used as input to consensus clustering (79) to obtain the final clustering of the samples. Alternatively, in (20) a model based on Probabilistic Latent Semantic Analysis (80) was proposed for clustering B. These two methods allow any clustering algorithm to be used on each single omic, and therefore have an advantage when a method is known to

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

perform well for a particular omic. Additionally, they can be used given the clustering solution only when the raw omic data are unavailable.

PINS (21) integrates clusters by examining their connectivity matrices for the different omics. Each such matrix S^m is a binary n x n matrix, where $S^m_{ij} = 1$ if patients i and j are clustered together in omic m, and 0 otherwise. These S^m matrices are averaged to obtain a single connectivity matrix, which is then clustered using different methods based on whether the different S^m matrices highly agree with each other or not. The obtained clusters are tested if they can be further split into smaller clusters. To determine the number of clusters for each omic and for the integrated clustering, perturbations are performed on the data by adding Gaussian noise to it, and the number of clusters is chosen such that the resulting clustering is robust to the perturbations. Unlike the previously presented late integration methods, PINS requires the original data and not only the clustering of each omic, since it performs perturbations to the data.

Several methods for ensemble clustering were developed over the years, and are reviewed in (81). While these were not originally developed for this purpose, they can be used for late multi-omics clustering as well.

Similarity-based methods

Similarity-based methods use similarities or distances between samples in order to cluster data. These methods compute the similarities between samples in each omic separately, and vary in the way these similarities are integrated. The integration step uses only similarity values. Since in current multi-omic datasets, the number of samples is much smaller than the number of features, these algorithms are usually faster than methods that consider all features while performing integration. However, in such methods it may be more difficult to interpret the output in terms of the original features. An additional advantage of similarity-based methods is that they can easily support diverse omic types, including categorical and ordinal data. Each omic only requires a definition of a similarity measure.

Spectral clustering generalizations. Spectral clustering (82) is a widely used similarity-based method for clustering single-view data. The objective function for single-view spectral clustering is $max_U trace(U^t L U)$ s.t. $U^t U = I$, where L is the Laplacian (83) of the similarity matrix of dimension $n \times n$, and U is of dimension $n \times k$, where k is the number of clusters in the data. Intuitively, it means that samples that are similar to one another have similar row vectors in U. This problem is solved by taking the k first eigenvectors of L (details vary between versions that use the normalized and the unnormalized graph Laplacian), and clustering them with a simple algorithm such as k-means. The spectral clustering objective was shown to be a relaxation of the discrete normalized cut in a graph, providing an intuitive explanation for the clustering. Several multi-view clustering algorithms are generalizations of spectral clustering.

An early extension to two views performs clustering by computing a new similarity matrix, using the two views' similarities (22). Denote by W_1 and W_2 the similarity matrices for the two views. Then the integrated similarity, W, is defined as $W_1 W_2$. Spectral clustering is performed on the block matrix

$$\begin{bmatrix} 0 & W \\ W^t & 0 \end{bmatrix}$$

Note that each eigenvector for this matrix is of length $2n$. Either half of the vector or an average of the two halves are used instead of the whole eigenvectors for clustering using k-means. Note that this method is limited in that it only supports two views.

(23) generalizes spectral clustering for more than two views. Instead of finding a global U matrix, a matrix U^m is defined for each omic. The optimization problem is:

$$max_{U^1,\ldots,U^M} \Sigma_m trace(U^{mt} L^m U^m)$$
$$+\lambda \cdot \text{Reg} \quad \text{s.t. } \forall m \; U^{mt} U^m = I.$$

L^m is the graph Laplacian for omic m and Reg is a regularization term equal to either $\Sigma_{m_1 \neq m_2} U^{m_1} U^{m_1 t} U^{m_2} U^{m_2 t}$ or $\Sigma_m U^m U^{mt} U^* U^{*t}$ with the additional constraint that U^* is an n x k matrix such that $U^{*t} U^* = I$.

Chikhi (24) uses a different formulation, which does not require a different U^m for each omic, but instead uses the same U for all matrices. The following objective function is used:

$$max_U \Sigma_m trace(U^t L^m U) \text{ s.t. } U^t U = I$$

This is equivalent to performing spectral clustering on the Laplacian $\Sigma_m L^m$. The obtained clusters are then further improved in a greedy manner, by changing the assignment of samples to clusters, while looking directly at the discrete normalized cut objective, rather than the continuous spectral clustering objective.

Li (25) suggests a runtime improvement over (23). Instead of looking at the similarity matrix for all the samples, a small set of 'representative' vectors, termed salient points, are calculated by running k-means on the concatenation of all omics and selecting the cluster centers. A similarity matrix is then computed between all these samples in the data and their s nearest salient points. Denote this similarity matrix for the m'th omic by W^m, and let Z^m be its normalization such that rows sum to 1. These matrices are of dimension $n \times$ the number of salient points. Next, the matrices

$$\begin{bmatrix} 0 & Z^m \\ Z^{mt} & 0 \end{bmatrix}$$

are given as input to an algorithm with the same objective as (24). This way, similarities are not computed between all pairs of samples.

The methods above differ in several ways. (23) allows each omic to have a different low dimensional representation, and has a parameter that controls the trade-off between how similar these representations are, and how similarities in the original data are maintained in U^m. Therefore, it allows to express cases where the omics are not assumed to have the same similarity structure (e.g., two samples can be similar in one omic but different in another). On the other hand, Chikhi (24) assumes the same similarity structure, and its greedy optimization step can result in an improved solution in such cases. (25) can be used when the number of samples is exceptionally large.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

Zhou and Burges (26) views similarity matrices as networks, and examines random walks on these networks. Random walks define a stationary distribution on each network, which captures its similarity patterns (84). Since that stationary distribution is less noisy than the original similarity measures, Zhou and Burges (26) uses them instead to integrate the networks. Xia (27) also examines random walks on the networks, but argues that the stationary distribution in each network can still be noisy. Instead, the authors compute a consensus transition matrix, that has minimum total distance to the per-omic transition matrices and is of minimal rank. Random walks are highly related to spectral clustering; using a normalized variant of the graph's Laplacian in spectral clustering results in a solution in which random walks seldom cross between clusters (82). These random walk-based methods are currently competitive with other spectral clustering methods.

Similarity Network Fusion. SNF (Similarity Network Fusion) first constructs a similarity network for every omic separately (28,29). In each such network, the nodes are samples, and the edge weights measure the sample similarity. The networks are then fused together using an iterative procedure based on message passing (85). The similarity between samples is propagated between each node and its k nearest neighbors.

More formally, denote by $W^{(m)}$ the similarity matrix for the m'th omic. Initially a transition probability matrix between all samples is defined by:

$$P_1^{(m)}(i,j) = \begin{cases} \frac{W^{(m)}(i,j)}{2\Sigma_{k\neq i} W^{(m)}(i,k)}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases}$$

and a transition porbability matrix between nearest neighbors is defined by:

$$S^{(m)}(i,j) = \begin{cases} \frac{W^{(m)}(i,j)}{\Sigma_{k\neq i} W^{(m)}(i,k)}, & j \in N_i \\ 0, & \text{otherwise} \end{cases}$$

where N_i are i's k nearest neighbors in the input X^m matrices. The P matrices are updated iteratively using message passing between the nearest neighbors: $P_{q+1}^{(m)} = S^{(m)} \frac{\Sigma_{k\neq m} P_q^{(k)}}{M-1} S^{(m)q}$ where $P_q^{(m)}$ is the matrix for omic m at iteration q. This process converges to a single similarity network, summarizing the similarity between samples across all omics. This network is partitioned using spectral clustering.

In (29), SNF is used on gene expression, methylation and miRNA expression data for several cancer subtypes from TCGA. In addition to partitioning the graph to obtain cancer sutbypes, the authors show that the fused network can be used for other computational tasks. For example, they show how to fit Cox proportional hazards (86), a model that predicts prognosis of patients, with a constraint such that similar patients in the integrated network will have similar predicted prognosis.

rMKL-LPP. Kernel functions implicitly map samples to a high (possibly infinite) dimension, and can efficiently measure similarity between the samples in that dimension. Multiple kernel learning uses several kernels (similarity measures), usually by linearly combining them, and is often used in supervised analysis. (30) developed rMKL-LPP (regularized Multiple Kernel Learning with Locality Preserving Projections), which uses multiple kernel learning in unsupervised settings. The algorithm performs dimension reduction on the input omics such that similarities (defined using multiple kernels) between each sample and its nearest neighbors are maintained in low dimension. This representation is subsequently clustered with k-means. rMKL-LPP allows the use of diverse kernel functions, and even multiple kernels per omic. A regularization term is added to the optimization problem to avoid overfitting. The authors run the algorithm on five cancer types from TCGA, and show that using multiple kernels per omic improves the prognostic value of the clustering, and that regularization improves robustness.

Dimension reduction-based methods

Dimension reduction-based methods assume the data have an intrinsic low dimensional representation, with that low dimension often corresponding to the number of clusters. The views that we observe are all transformations of that low dimensional data to a higher dimension, and the parameters for the transformation differ between views. This general formulation was proposed by (31), which suggest to minimize $\Sigma_{m=1}^{M} w_m l(X^m, f_m(B))$, where B is a matrix of dimension $n \times p$, f_m are the parametrized transformations, and w_m are weights for the different views, and l is a loss function. The work further provides an optimization algorithm when the f_m transformations are given by matrix multiplication. That is, $f_m(B) = BP^m$, and l is the squared Frobenius norm applied to $X^m - BP^m$. Once B is calculated, single-omic clustering algorithm can be applied to it. This general framework is widely used. Since the transformation is often assumed to be linear, many of the dimension reduction methods are based on matrix factorization. Dimension reduction methods work with real-valued data. Applying these methods to discrete binary or count data is technically possible but often inappropriate.

An advantage of linear dimension reduction methods is that they provide some interpretation for the dominant features in each cluster. For example, in the general framework just presented, each entry in the P^m matrix can be considered as the weight of a feature in a cluster. Such interpretation is missing from similarity-based methods, which ignore the original features once the similarities between samples were calculated. Therefore, dimension reduction methods may be useful when an association between clusters and features is needed.

JIVE. (32) assumes that the variation in each omic can be partitioned to a variation that is joint between all omics, and an omic-specific variation: $X^{m^t} = J^m + A^m + E^m$ where E^m are error terms. Let J and A be the concatenated J^m and A^m matrices, respectively. The model assumes that $JA^t = 0$, that is, the joint and omic specific variations are uncorrelated, and that $rank(J) = r$ and $rank(A_i) = r_i$ for each omic, so that the structure of each omic and the total joint variation are of low rank. In order for the weight of the different omics to be equal, the input omic matrices are normalized to have equal Frobenius norm. A penalty term is

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

added to encourage variable sparsity. This method was applied to gene expression and miRNA data of Glioblastoma Multiforme brain tumors, and identified the joint variation between these omics.

Correlation and covariance-based. Two of the most widely used dimension reduction methods are Canonical Correlation Analysis (CCA) (34) and Partial Least Squares (PLS) (45). Given two omics X^1 and X^2, in CCA the goal is to find two projection vectors u^1 and u^2 of dimensions p_1 and p_2, such that the projected data has maximum correlation:

$$argmax_{u^1,u^2} corr(X^1u^1,\ X^2u^2)$$

These projections are called the first canonical variates, and are the axis with maximal correlation between the omics. The k'th pair of canonical variates, u_k^1 and u_k^2 are found such that correlation between $X^1u_k^1$ and $X^2u_k^2$ is maximal, given that the new pair is uncorrelated (that is, orthogonal) to the previous canonical variates. Chaudhuri *et al.* (87) proved and showed empirically that if the data originate from normal or log concave distributions, the canonical variates can be used to cluster the data. CCA was formulated in a probabilistic framework such that the optimization solutions are maximum likelihood estimates (88), and further extended to a Bayesian framework (35). An additional expansion to perform CCA in high dimension is Kernel CCA (36). A deep-learning based CCA method, DeepCCA, was recently developed (37). Rather than maximize the correlation between linear projections of the data, the projections are taken to be functions of the data calculated using neural networks, and the optimization process optimizes the parameters for these networks.

Solving CCA requires inversion of the covariance matrix for the two omics. Omics data usually have a higher number of features than samples, and these matrices are therefore not invertible. To apply CCA to omics data, and to increase the interpretability of CCA's results, sparsity regularization was added (38,39).

CCA supports only two views. Several works extend it to more than two views, including MCCA (39) which maximizes the sum of pairwise correlations between projections and CCA-RLS (40). Luo *et al.* (41) generalize CCA to tensors in order to support more than two views.

Another line of work on CCA, with high relevance for omics data, investigated relationships between the features while performing the dimension reduction. ssCCA (structure constrained sparse CCA) allows to incorporate into the model known relationships between features in one of the input omics, and force entries in the u^i vector for that view to be close for similar features. This model has been developed by (42) and utilized microbiome's phylogenies as the feature structure. Another model that considers relationship between features was developed in (43). In this work, rather than defining similarities between features, they are partitioned into groups. Regularization is performed such that both irrelevant groups and irrelevant features within relevant groups are removed from the model. Finally, Podosinnikova et. al, in 'Beyond CCA: Moment matching for multi-view models', extended CCA to support count data, which are common in biological datasets.

PLS also follows a linear dimension reduction model, but maximizes the covariance between the projections, rather than the correlation. More formally, given two omics X^1 and X^2, PLS computes a sequence of vectors u_k^1 and u_k^2 for $k = 1, 2, \ldots$ such that $cov(X^1u_k^1,\ X^2u_k^2)$ is maximal, given that ${u_k^1}^t u_k^1 = 1$, ${u_k^2}^t u_k^2 = 1$, and $cor(X^1u_k^1,\ X^1u_l^1) = 0$ for $l < k$. That is, new projections are not correlated with previous ones. PLS can be applied to data with more features than samples even without sparsity constraints. A sparse solution is nonetheless desirable, and one was developed (46,47). O2-PLS increases the interpretability of PLS by partitioning the variation in the datasets into joint variation between them, and variations that are specific for each dataset and that are not correlated with one another (48). While PLS and O2-PLS were originally developed for chemometrics, they were recently used for omics data as well (89,90). PLS was also extended to use the kernel framework (49), and a combined version of kernel PLS and O2 PLS was developed (50).

Like CCA, PLS was developed for two omics. MBPLS (Multi Block PLS) extends the model to more than two omics (91), and sMBPLS adds sparsity constraints. sMBPLS was developed specifically for omics data (51). It looks for a linear combination of projections of non-gene-expression omics that has maximal correlation with a projection of gene expression omic. An extension of O2PLS also exists for multi-view datasets (52).

Both CCA and PLS can be used in cases where high interpretability is wanted. The different u_k^1 and u_k^2 vector pairs are those along which the correlation (or covariance) between patients is maximal. They can therefore be used to associate between features from the different views.

An additional method that is based on maximizing covariance in low dimension is MCIA (53), an extension of co-inertia analysis to more than two omics (92). It aims to find projections for all the omics such that the sum of squared covariances with a global variation axis is maximal: $max_{u^m,v} \Sigma_{m=1}^{M} cov^2(X^m u^m, v)$. The projections of different omics can be used to evaluate the agreement between the different omics (the distance between projections reflects the level of disagreement between omics). Each of the projections can be used as a representation for clustering.

Non-negative Matrix Factorization. Non-negative Matrix Factorization (NMF) assumes that the data have an intrinsic low dimensional non-negative representation, and that a nonnegative matrix projects it to the observed omic (93). It is therefore only suitable for non-negative data. For a single omic, denote by k the low dimension. The formulation is $X \approx WH$, where X is the $n \times p$ observed omic matrix, W is $n \times k$ and H is $k \times p$. The objective function is $||X - WH||_2^2$, and it is minimized by updating W and H in an alternating manner, using multiplicative update rules, such that solutions remain non negative after each update (94). The low dimension representation W can be clustered using a simple single-omic algorithm. Like other dimension reduction methods, the W and H matrices can be used to better understand the weight of each feature in each cluster. The non-negativity constraint makes this weight more interpretable.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

Several methods generalize this model to multi-omic data. MultiNMF (54) uses the following generalization: Each omic X^m is factorized into $W^m H^m$. This model is equivalent to performing NMF on each omic separately. Integration between the omics is done by adding a constraint that the W^m matrices are close to a 'consensus' matrix W^*. The objective function is therefore: $\Sigma_{m=1}^{M} ||X^m - W^m H^m||_2^2 + \lambda \Sigma_{m=1}^{M} ||W^m - W^*||_2^2$. Kalayeh *et al.* (55) generalizes this method to support weights for features' and samples' similarity. (56) extend MultiNMF by further requiring that the low dimensional representation W^* maintains similarities between samples (samples that are close in the original dimension must be close in W^*). This approach combines factorization and similarity-based methods.

Joint NMF (57) uses a different formulation, where a sample has the same low dimensional representation for all omics: $X^m \approx WH^m$. Note that by writing $X = WH$ where X and H are obtained by matrix concatenation, this model is equivalent to early integration. Joint NMF is not directly used for clustering. Rather, the data are reduced to a large dimension ($k = 200$) and high values in W and H^m are used to associate samples and features with modules that are termed 'md-modules'. The authors applied Joint NMF on miRNA, gene expression and methylation data from ovarian cancer patients, and showed that functional enrichment among features that are associated with md-modules that is more significant than the enrichment obtained in single-omic modules. In addition, patients in certain modules have significantly different prognosis compared to the rest of the patients. Much like (56) extends multiNMF, EquiNMF extends Joint NMF such that similarities in the original omics are maintained in lower dimension. (58) extends NMF to the case where different views can contain different samples, but constrains certain samples from different views to belong to the same cluster based on prior knowledge. Finally, PVC (59) performs partial multi-view clustering. In this setting, not all samples necessarily have measurements for all views.

The difference between MultiNMf and Joint NMF resembles the difference described previously between similarity-based methods. MultiNMF allows for different omics to have different representations, where the similarity between them is controlled by a parameter. It can therefore be used in cases where the different omics are not expected to have the same low dimensional representation.

Matrix tri-factorization. An alternative factorization approach presented in (60) is tri-matrix factorization. In this framework, each input omic is viewed as describing a relationship between two entities, which are its rows and columns. For example, in a dataset with two omics, gene expression and DNA methylation of patients, there are three entities which are the patients, the genes and the CpG loci. The gene expression matrix describes a relationship between patients and genes, while the methylation matrix describes a relationship between patients and CpG loci.

Each omic matrix R_{ij} of dimension $n_i \times n_j$ that describes the relationship between entities i and j is factorized as $R_{ij} = G_i S_{ij} G_j^t$, where G_i and G_j provide a low dimensional representation for entities i and j respectively and are of dimensions $n_i \times k_i$ and $n_j \times k_j$, and S_{ij} is an omic-specific matrix of dimension $k_i \times k_j$. As in NMF, the G_i matrices are non-negative. The same G_i matrix is used in all omics with entity i, and in this way data integration is achieved. In the above example, both the gene expression and DNA methylation omics will use the same G matrix to represent patients, but different matrices to represent genes and CpG loci. In this model, an additional matrix describing the relationship between genes and CpGs could optionally be used. This is a major advantage of matrix tri-factorization, as it allows to incorporate prior known relations between different entities, without changing the input omic matrices. (60) adds constraints to the formulation that can encourage entities to have similar representations. This framework was applied to diverse problems in bioinformatics, including in supervised settings: It was used to perform gene function prediction (60), and for patient survival regression (95).

Convex formulations. A drawback of most factorization-based methods is that their objective functions are not convex, and therefore optimization procedures do not necessarily reach a global optimum, and highly depend on initialization. One solution to this issue is by formulating dimension reduction as a convex problem. White *et al.* (61) relaxes CCA's conditions and defines a convex variant of it. Performance was assessed on reducing noise in images, but the method can also be used for clustering. However, like CCA, the method only supports two views. Guo (62) presents a different convex formulation for dimension reduction, for the general factorization framework presented earlier, which minimizes $\Sigma_{m=1}^{M} ||X^m - BP^m||_F^2 + \gamma ||B||_{2,1}$. $\| \cdot \|_{2,1}$ is the $l_{2,1}$ norm, namely the sum of the Euclidean norms of the matrix rows. This relaxation therefore supports multiple views. LRAcluster (16) also uses matrix factorization and has a convex objective function.

Tensor-based methods. A natural extension of factorization methods for multi-omic data is to use tensors, which are higher order matrices. One such method is developed in (63). This method writes each omic matrix as $X^m = Z^m X^m + E^m$, $diag(Z^m) = 0$, where Z^m is an n x n matrix and E^m are error matrices. The idea is that each sample in each omic can be represented as a linear combination of other samples (hence the $diag(Z^m) = 0$ constraint), and that its representation in that base (Z^m) can then be used for clustering. To integrate the different views, the different Z^m matrices are merged to a third-order tensor, Z. The objective function encourages Z to be sparse, and the E^m error matrices to have a small norm.

Statistical methods

Statistical methods model the probabilistic distribution of the data. Some of these methods view samples as originating from different clusters, where each cluster defines a distribution for the data, while other methods do not explicitly use the cluster structure in the model. An advantage of the statistical approach is that it allows to include biological knowledge as part of the model when determining the distribution functions. This can be done either using Bayesian priors or by choosing probabilistic functions, e.g.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

using normal distribution for gene expression data. An additional advantage of statistical frameworks is their ability to make 'soft', probabilistic decisions. For example, a statistical method can not only assign a sample to a cluster, but can also determine the probability that the sample belongs to the cluster. For most formulations, parameter estimation is computationally hard, and different heuristics are used. Several models under the Bayesian framework allow for samples to belong to different clusters in different omics.

iCluster and iCluster+. iCluster (15) assumes that the data originate from a low dimension representation, which determines the cluster membership for each sample: $X^{m'} = W^m Z + \epsilon^m$, where Z is a k x n matrix, W^m is an omic specific p_m x k matrix, k is the number of clusters and ϵ^m is a normally distributed noise matrix. This model resembles other dimension reduction models, but here the distribution of noise is made explicit. Under this model iCluster maximizes the likelihood of the observed data with an additional regularization for sparse W^m matrices. Optimization is performed using an EM-like algorithm, and subsequently k-means is run on the lower dimension representation of the data Z to get the final clustering assignments. iCluster was applied to breast and lung cancer, using gene expression and copy number variations. iCluster was also recently used to cluster more than ten thousand tumors from 33 cancers in a pan-cancer analysis (96). Note that by concatenating all W^m matrices to a single W matrix, and rewriting the model as $X^t = WZ + \epsilon$, iCluster can be viewed as an early integration approach.

iCluster's runtime grows fast with the number of features, and therefore feature selection is essential before using it, as was shown in (29). Shen *et al.* (15) only use genes located on one or two chromosomes in their analysis.

Since iCluster's model uses matrix multiplication, it requires real-values features. An extension called iCluster+ (64) includes different models for numeric, categorical and count data, but maintains the idea that data originate from a low dimension matrix Z. For categorical data, iCluster+ assumes the following model:

$$Pr(X_{ij}^m = c|z_i) = \frac{exp(\alpha_{jcm} + \beta_{jcm} \cdot z_i)}{\Sigma_l exp(\alpha_{jlm} + \beta_{jlm} \cdot z_i)}$$

while for numeric data the model remains linear with normal error:

$$x_{ijm} = \gamma_{jm} + \delta_{jm} \cdot z_i + \epsilon_{ijm}, \epsilon_{ijm} \sim N(0, \sigma_{jm}^2)$$

A regularization term encouraging sparse solution is added to the likelihood, and a Monte-Carlo Newton–Raphson algorithm is used to estimate parameters. The Z matrix is used as in iCluster for the clustering. The latest extension of iCluster, which builds on iCluster+, is iClusterBayes (65). This method replaces the regularization in iCluster+ with full Bayesian regularization. This replacement results in faster execution, since the algorithm no longer needs to fine tune parameters for iCluster+'s regularization.

PARADIGM. PARADIGM (66) is the most explicit approach to modeling cellular processes and the relations among different omics. For each sample and each cellular pathway, a factor graph that represents the state of different entities within that pathway is created. As a degenerate example, a pathway may include nodes representing the mRNA levels of each gene in that pathway, and nodes representing those genes' copy number. Each node in the factor graph can be either activated, nominal or deactivated, and the factor graph structure defines a distribution over these activation levels. For example, if a gene has high copy number it is more likely that it will be highly expressed. However, if a repressor for that gene is highly expressed, that gene is more likely to be deactivated. PARADIGM infers the activity of non-measured cellular entities to maximize the likelihood of the factor graph, and outputs an activity score for each entity per patient. These scores are used to cluster cancer patients from several tissues.

PARADIGM's model can be used for more than clustering. For example, PARADIGM-shift (97) predicts loss-of-function and gain-of-function mutations, by finding genes whose expression value as predicted based on upstream entities in the factor graph is different from their predicted expression value using downstream entities. However, PARADIGM relies heavily on known interactions, and requires specific modeling for each omic. It is also quite limited to the cellular level; For example, it is not clear how to incorporate into the model an omic describing the microbiome composition of each patient.

Combining omic-specific and global clustering. All the methods discussed so far assume that there exists a consistent clustering structure across the different omics, and that analyzing the clusters in an integrative way will reveal this structure more accurately than analyzing each omic separately. However, this is not necessarily the case for biomedical datasets. For example, it is not clear that the methylation and expression profiles of cancer tumors really represent the same underlying cluster structure. Rather, it is possible that each omic represents a somewhat different cluster structure. Several methods take this view point using Bayesian statistics.

Savage *et al.* (67) define a hierarchical Dirichlet process model, which supports clustering on two omics. Each sample can be either *fused* or *unfused*. Fused samples belong to the same cluster in both omics, while unfused samples can belong to different clusters in different omics. Patterns of fused and unfused samples reveal the concordance between the two datasets. This model is extended in PSDF (68) to include feature selection. Savage *et al.* (67) apply the model to cluster genes using gene expression and ChIP-chip data, while (68) clusters cancer patients using expression and copy number data.

In MDI (69) each sample can have different cluster assignments in different omics. However, a prior is given such that the stronger an association between two omics is, the more likely a sample will belong to the same cluster in these two omics. This association strength adjusts the prior clustering agreement between two omics. In addition to these priors, MDI's model uses Dirichlet mixture model, and explicitly represents the distribution of the data within each cluster and omic. Since samples can belong to different clusters in different omics, no global clustering solution is re-

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

turned by the algorithm. Instead, the algorithm outputs sets of samples that tend to belong to the same cluster.

A different Bayesian formulation is given by BCC (70). Like MDI, BCC assumes a Dirichlet mixture model, where the data originate from a mixture of distributions. However, BCC does assume a global clustering solution, where each sample maps to a single cluster. Given that a sample belongs to a global cluster, its probability to belong to that cluster in each omic is high, but it can also belong to a different cluster in that omic. Parameters are estimated using Gibbs sampling (98). BCC was used on gene expression, DNA methylation, miRNA expression and RPPA data for breast cancer from TCGA.

Like MDI and BCC, Clusternomics (71) uses a Dirichlet mixture model. Clusternomics suggests two different formulations. In the first, each omic has a different clustering solution, and the global clusters are represented as the Cartesian product of clusters from each omic. This approach does not perform integration of the multi-omic datasets. In the second formulation, global clusters are explicitly mapped to omic-specific clusters. That way, not all possible combinations of clusters from different omics are considered as global clusters.

Survival-based clustering. One of the areas multi-omics clustering is widely used for is discovering disease subtypes. In this context, we may expect different disease subtypes to have a different prognosis, and this criterion is often used to assess clustering solutions. Ahmad and Fröhlich (72) develop a Bayesian model for multi-omics clustering that considers patient prognosis while clustering the data. Patients within a cluster have both similar feature distribution and similar prognosis. This approach is not entirely unsupervised, as it considers patient survival data, which are also used to assess the solutions. Coretto *et al.* (73) also develop a probabilistic clustering method that considers survival, and that supports a large number of features compared to (72), which only uses a few dozen features. As the survival data are used as input to the model, it is not surprising that this approach gives clusters with more significantly different survival than other approaches. This was demonstrated on Glioblastoma Multiforme data by (72) and for data from several cancer types by (73), both from TCGA.

Deep multi-view methods

A recent development in machine learning is the advent of deep learning algorithms (99). These algorithms use multi-layered neural networks to perform diverse computational tasks, and were found to improve performance in several fields such as image recognition (100) and text translation (101). Neural networks and deep learning have also proven useful for multi-view applications (102), including unsupervised feature learning (37), (103). Learned features can be used for clustering, as described earlier for DeepCCA. Deep learning is already used extensively for biomedical data analysis (104).

Recent deep learning uses for multi-omics data include (74) and (75). Chaudhary *et al.* (74) use an autoencoder, which is a deep learning method for dimension reduction. The authors ran it on RNA-seq, methylation and miRNA-seq data in order to cluster Hapatocellular Carcinoma patients. The architecture implements an early integration approach, concatenating the features from the different omics. The autoencoder outputs a representation for each patient. Features from this representation are tested for association with survival, and significantly associated features are used to cluster the patients. The clusters obtained have significantly different survival. This result is compared to a similar analysis using the original features, and features learned with PCA (Principal Component Analysis) rather than autoencoders. However, the analysis in this work is not unsupervised, since the feature selection is based on patient survival.

Liang *et al.* (75) use a different approach. They analyze expression, methylation and miRNA ovarian cancer data using Deep Belief Networks (105) which explicitly consider the multi-omic structure of the data. The architecture contains separate hidden layers, each having inputs from one omic, followed by layers that receive input from all the single-omic hidden layers, thus integrating the different omics. A 3D representation over $\{0, 1\}$ is learned for each patient, partitioning the patients into $2^3 = 8$ clusters. The clustering results are compared to k-means clustering on the concatenation of all omics, but not to other multi-omics clustering methods.

Deep learning algorithms usually require many samples and few features. They use a large number of parameters, which makes them prone to overfitting. Current multi-omic datasets have the opposite characteristics—they have many features and at least one order of magnitude less samples. The works presented here use only a few layers in their architectures to overcome this limitation, in comparison to the dozens of layers used by state-of-the-art architectures for imaging datasets. As the number of biomedical samples increases, deep multi-view learning algorithms might prove more beneficial for biomedical datasets.

BENCHMARK

In order to test the performance of multi-omics clustering methods, we compared nine algorithms on ten cancer types available from TCGA. We also compared the performance of the algorithms on each one of the single-omic datasets that make up the multi-omic datasets, for algorithms that are applicable to single-omic data. The nine algorithms were chosen to represent diverse approaches to multi-omics clustering. Within each approach, we chose methods with available software and clear usage guidelines (e.g. we chose PINS over COCA as a late integration method since COCA does not explicitly state how each single omic should be clustered), and that are widely used, so that a comparison of these methods will be most informative to the community. Three algorithms are early integration methods: LRAcluster, and k-means and spectral clustering on the omics concatenated into a single matrix. For similarity-based algorithms we used SNF and rMKL-LPP. For dimension reduction we used MCCA (39) and MultiNMF. We chose iClusterBayes as a statistical method, and PINS as a late integration approach.

The ten datasets contain cancer tumor multi-omics data, where each dataset is a different cancer type. All datasets

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

contain three omics: gene expression, DNA methylation and miRNA expression. The number of patients range from 170 for AML to 621 for BIC. Full details on the datasets and cancer type acronyms appear in Supplementary File 2.

To assess the performance of a clustering solution, we used three metrics. First, we measured differential survival between the obtained clusters using the logrank test (106). Using this test as a metric assumes that if clusters of patients have significantly different survival, they are different in a biologically meaningful way. Second, we tested for the enrichment of clinical labels in the clusters. We chose six clinical labels for which we tested enrichment: gender, age at diagnosis, pathologic T, pathologic M, pathologic N and pathologic stage. The four latter parameters are discrete pathological parameters, measuring the progression of the tumor (T), metastases (M) and cancer in lymph nodes (N), and the total progression (pathologic stage). Enrichment for discrete parameters was calculated using the χ^2 test for independence, and for numeric parameters using Kruskal-Wallis test. Not all clinical parameters were available for all cancer types, so a total of 41 clinical parameters were available for testing. Finally, we recorded the runtime of each method. We did not consider in the assessment computational measures for clustering quality, such as heterogeneity, homogeneity or the silhouette score (107), since the different methods perform different normalization on the features (and some even perform feature selection). Full details about the survival and phenotype data appear in Supplementary File 2.

To derive a p-value for the logrank test, the χ^2 test for independence, and the Kruskal-Wallis test, the statistic for these three tests is assumed to have χ^2 distribution. However, for the logrank test and χ^2 test this approximation is not accurate for small sample sizes and unbalanced cluster sizes, especially for large values of the test statistic (this was shown for example in (108) for the logrank test in the case of two clusters). The p-values we report here are therefore estimated using permutation tests (i.e., we permuted the cluster labels between samples and used the test statistic to obtain an empirical p-value). We indeed observed large differences between the p-values based on permutation testing and based on the approximation, for both the logrank test and enrichment of clinical parameters. More details on the permutation tests appear in Supplementary File 1. After permutation testing, the p-values for the clinical labels were corrected for multiple hypotheses (since several labels were tested) using Bonferroni correction for each cancer type and method at significance level 0.05. Results for the statistical analyses are in Supplementary File 3.

We applied all nine methods to the ten multi-omics datasets, and to the thirty single-omic matrices comprising them. The only exceptions were MCCA, which we could not apply to single-omic data, and PINS, which crashed consistently on all BIC datasets*. All methods were run on a Windows machine, except for iCluster which was run on a Linux cluster utilizing up to 15 nodes in parallel. In general, we chose parameters for the methods as suggested by the authors. In case the authors suggested a parameter search, such search was performed, and the best solution was chosen as suggested by the authors, without considering the survival and clinical parameters that are used for assessment. The runtime we report for the methods includes the parameter search. The rationale is that the benchmark aims to record how a user would run the methods in terms of both results quality and total runtime. Details on hardware, data preprocessing and application of the methods appear in Supplementary File 1. Full clustering results appear in Supplementary File 4. All the processed raw data are available at http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html, and all software scripts used are available at https://github.com/Shamir-Lab/Multi-Omics-Cancer-Benchmark/.

Figure 2 depicts the performance of the benchmarked methods on the different cancer datasets, and Figures 3 and 4 summarize the performance for multi-omics data and for each single-omic separately across all cancer types. No algorithm consistently outperformed all others in either differential survival or enriched clinical parameters. With respect to survival, MCCA had the total best prognostic value (sum of -log10 p-values = 17.53), while MultiNMF was second (16.07) and LRACluster third (15.72). The sum of p-values can be biased due to outliers, so we also counted the number of datasets for which a method's solution obtains significantly different survival. These results are reported in Table 2. Here, with the exception of iClusterBayes, all methods that were developed for multi-omics or multi-view data had at least four cancer types with significantly different survival. MCCA and LRACluster had five. These cancer types are not identical for all the algorithms.

rMKL-LPP achieved the highest total number of significant clinical parameters, with 16 parameters. Spectral clustering came second with 14 and LRAcluster had 13. MCCA and MultiNMF, which had good results with respect to survival, had only 12 and 10 enriched parameters, respectively. rMKL-LPP did not outperform all other methods for all cancer types. For example, it had one enriched parameter for SKCM, while several other methods had two or three. We also considered the number of cancer types for which an algorithm had at least one enriched clinical label (Table 2). rMKL-LPP, spectral clustering, LRACluster and MCCA had enrichment in 8 cancer types, despite MCCA having a total of only 12 enriched parameters. Overall, rMKL-LPP outperformed all methods except MCCA, LRACluster and multiNMF with respect to both survival and clinical enrichment. MCCA, LRACluster and multiNMF had better prognostic value, but found less enriched clinical labels.

Each method determines the number of clusters for each dataset. These numbers are presented in Table 3. The numbers vary drastically among methods, from 2 or 3 (iCluster and MultiNMF) to more than 10 on average (MCCA). MCCA, LRACluster and rMKL-LPP partitioned the data into a relatively high number of clusters (average of 10.6, 9.4 and 6.7 respectively), and had good performance, which may indicate that clustering cancer patients into more clusters improves prognostic value and clinical significance. The

* Correction after publication: We performed all the benchmarks on a 64-bit computer, using the 32-bit version of R. In later tests we observed that PINS did not crash on 64-bit R, and it only crashed on 32-bit R due to insufficient memory. The clustering that PINS obtained on the breast cancer dataset had 4 enriched clinical parameters, and the p-value for the logrank test on that clustering was 0.05.).

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

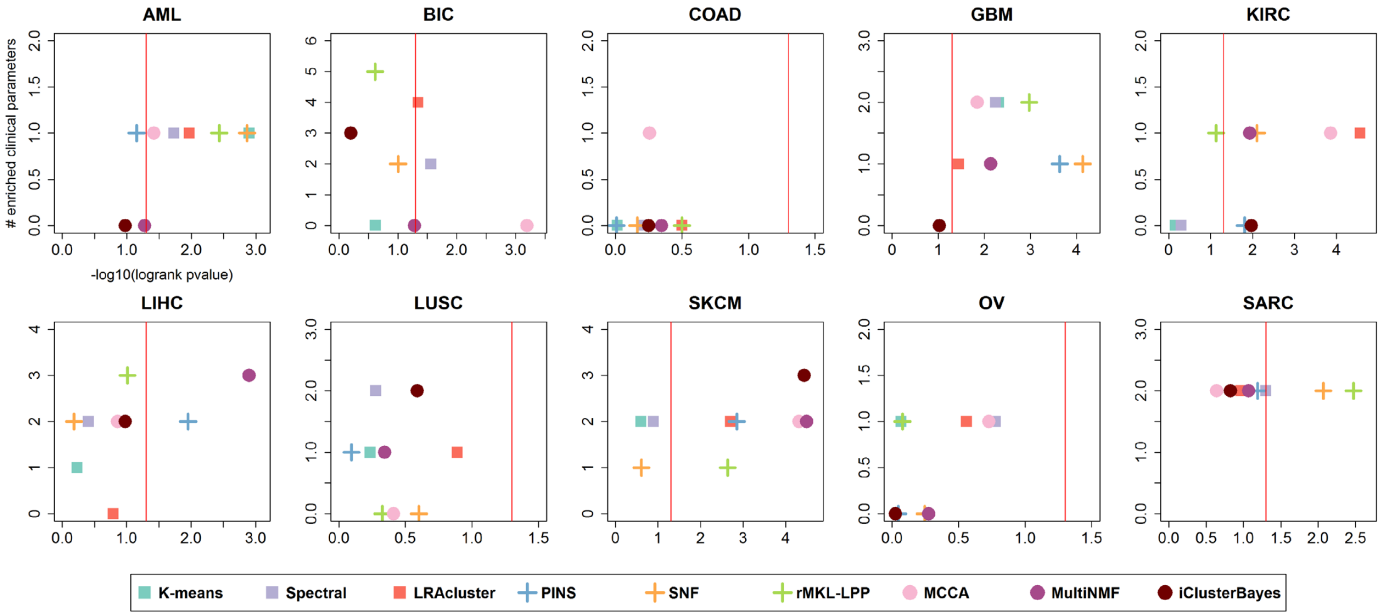


Figure 2. Performance of the algorithms on ten multi-omics cancer datasets. For each plot, the x-axis measures the differential survival between clusters ($-\log_{10}$ of logrank's test P-value), and the y-axis is the number of clinical parameters enriched in the clusters. Red vertical lines indicate the threshold for significantly different survival (P-value ≤ 0.05)

Table 2. Cancer types with significant results per algorithm

	k-means	Spectral	LRAcluster	PINS	SNF	rMKL-LPP	MCCA	MultiNMF	iClusterBayes
Significantly different survival	2	3	5	4	4	4	5	4	2
Significant clinical enrichment	7	8	8	6	7	8	8	6	5

For each benchmarked algorithm, the number of cancer subtypes for which its clustering had significantly different prognosis (first row) and had at least one enriched clinical label (second row) are shown.

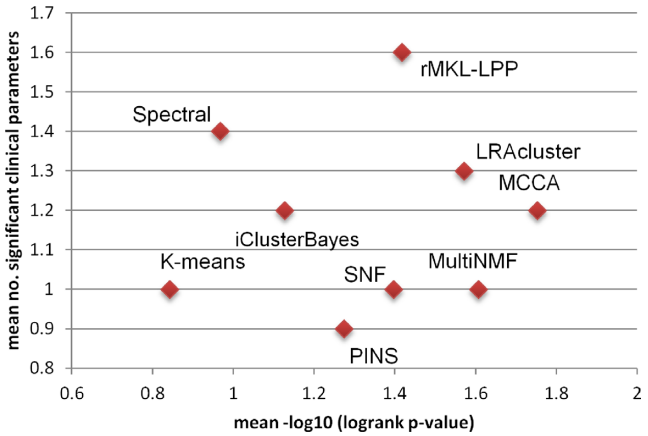


Figure 3. Mean performance of the algorithms on ten multi-omics cancer datasets. The x-axis measures the differential survival between clusters (mean $-\log_{10}$ of logrank's test P-value), and the y-axis is the mean number of clinical parameters enriched in the clusters.

higher number of clusters is controlled in the logrank and clinical enrichment tests by having more degrees of freedom for its χ^2 statistic.

The runtime of the different methods is reported in Table 4. Note that as mentioned earlier, iClusterBayes was run on a cluster, while the other methods were run on a desktop computer. All methods except for LRAcluster and iCluster took less than ten minutes per dataset on average. LRAcluster and iClusterBayes took about 56 and 72 minutes per dataset, respectively.

Figure 4 also shows the performance of the benchmarked methods for single-omic data. While several methods had worse performance on single-omic datasets, some achieved better performance. For example, the highest number of enriched clinical parameters for both single and multi-omic datasets (18) was achieved by rMKL-LPP on gene expression. The gene expression solution also had better prognostic value than the multi-omic solution.

To further test how analysis of single-omic datasets compares to multi-omic datasets, we chose for each dataset and method the single omic that gave the best results for survival and clinical enrichment. In this analysis, rMKL-LPP had both the highest total number of enriched clinical parameters (21), and the highest total survival significance (21.86). The runtime, number of clusters, and survival and clinical enrichment analysis for single-omic datasets appear in Supplementary Files 1 and 3. These results suggest that analysis of multi-omics data does not consistently provide better prognostic value and clinical significance compared to analysis of single-omic data alone, especially when different single-omics are used for each cancer types.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

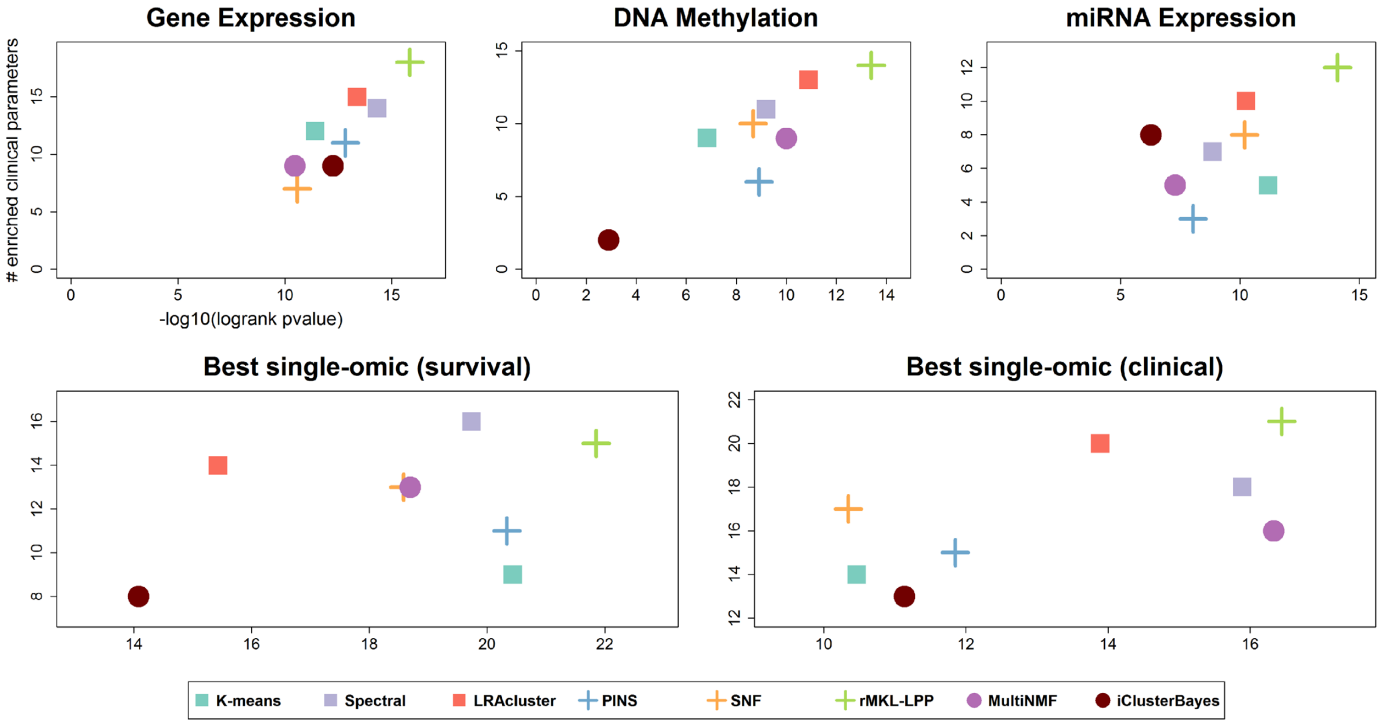


Figure 4. Summarized performance of the algorithms across ten cancer datasets. For each plot, the x-axis measures the total differential prognosis between clusters (sum across all datasets of $-\log_{10}$ of logrank's test *P*-value), and the y-axis is the total number of clinical parameters enriched in the clusters across all cancer types. (**A**–**C**) Results for single-omic datasets. (**D**) Results when each method uses the single omic that achieves the highest significance in survival. (**E**) Same with respect to enrichment of clinical labels.

Table 3. Number of clusters chosen by the benchmarked algorithms on ten multi-omics cancer datasets

	AML	BIC	COAD	GBM	KIRC	LIHC	LUSC	SKCM	OV	SARC	Means
K-means	5	2	2	5	2	2	2	2	2	2	2.6
Spectral	9	3	2	5	2	2	2	2	4	2	3.3
LRAcluster	7	7	5	11	3	12	12	15	9	13	9.4
PINS	4	NA	4	2	2	5	4	15	2	3	4.6
SNF	4	2	3	2	4	2	2	3	3	3	2.8
rMKL-LPP	6	7	6	6	11	6	6	7	6	6	6.7
MCCA	11	14	2	11	15	15	12	2	9	15	10.6
MultiNMF	2	2	2	3	2	3	2	2	2	2	2.2
iClusterBayes	2	3	2	2	2	3	2	2	2	2	2.2

The right column is the average number of clusters across all cancer types.

Table 4. Runtime in seconds of the algorithms on ten multi-omics cancer datasets

	AML	BIC	COAD	GBM	KIRC	LIHC	LUSC	SKCM	OV	SARC	Means
K-means	96	1306	153	212	102	407	444	723	303	191	394
Spectral	3	8	3	3	3	5	5	6	4	4	4
LRAcluster	957	11655	1405	1370	991	3959	3353	5892	2299	2004	3388
PINS	41	NA	112	115	59	125	228	317	214	113	147
SNF	5	42	7	7	6	14	13	21	9	8	13
rMKL-LPP	222	192	205	221	191	255	213	333	263	238	233
MCCA	12	43	12	13	13	26	25	25	19	16	20
MultiNMF	19	51	25	19	17	35	27	45	21	23	28
iClusterBayes*	2628	7832	3213	2569	2756	5195	4682	6077	4057	3969	4298

The right column is the average runtime across all cancer types. *For iClusterBayes numbers are elapsed time on a multi-core platform.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

DISCUSSION

We have reviewed methods for multi-omics and multi-view clustering. In our tests on 10 cancer datasets, overall, rMKL-LPP performed best in terms of clinical enrichment, and outperformed all methods except MCCA and MultiNMF with respect to survival. The high performance of MCCA and MultiNMF is remarkable, as these are multi-view methods that were not specifically developed for omics data (though MCCA was applied to it).

Throughout this review we provided guidelines about the advantages and disadvantages of different approaches and algorithms. In the benchmark, no single method consistently outperformed all others on any of the assessment criteria. While some methods were shown to do well, we cannot conclude from this that they should be always preferred. We also could not identify one 'best' integration approach, but it is interesting to note that the top two performers with respect to survival were dimension reduction methods.

Careful consideration should be given when applying multi-view clustering methods to multi-omic data, since these data have characteristics that multi-view methods do not necessarily consider. The most prominent of these characteristics is the large number of features relative to the number of samples. For example, CCA inverts the covariance matrix of each omic. This matrix is not invertible when there are more features than samples, and sparsity regularization is necessary. Another feature of multi-omic data is the dependencies between features in different omics, but several multi-view algorithms assume conditional independence of the omics given the clustering structure. This dependency is rarely considered, since it greatly increases the complexity of models. An additional characteristic of current omic data types is that due to cellular regulation, they have an intrinsic lower dimensional representation. The characteristic is utilized by many methods.

In our benchmark, single-omic data alone sometimes gave better results than multi-omics data. This was intensified when for each algorithm the 'best' single-omic for each cancer type was chosen. These results question the current assumptions underlying multi-omics analysis in general and multi-omics clustering in particular.

Several approaches may lead to improved results for multi-omics analysis. First, methods that suggest different clusterings in different omics were developed and reviewed here, but were not included in the benchmark, since it is not clear how to compare algorithms that do not output a global clustering solution to those that do. These methods may be more sensitive to strong signals appearing in only some of the omics. Second, future algorithms can perform omic selection in the same manner that algorithms today perform feature selection. In the benchmark, we let each method choose a single-omic for each cancer type given the results of the analysis, which are usually not available for real data. Methods that filter omics with contradicting signals might obtain a clearer clustering. Finally, while some methods for multi-omics clustering incorporate prior biological knowledge, few of them incorporate knowledge regarding the relationship between omics, or between features in different omics. Several statistical methods include some form of biological modeling by describing the distribution of the omics, and MDI tunes the similarity of clustering solutions in different omics based on the omics similarity. However, these methods do not model the biological relationships between omics. A notable exception is PARADIGM, which formulates the relationships between different omics. However, it also requires accurate prior knowledge about biochemical consequences of interactions, which is often unavailable. Methods that model relations between omics might benefit from additional biological knowledge, even without modeling whole pathways. For example, one can incorporate in a model the fact that promoter methylation is anti-correlated with gene expression. As far as we know, such methods were only developed for copy-number variation and gene expression data (e.g. (109)), and not in the context of clustering.

We detected large differences between the p-values derived from the χ^2 approximation compared to the *P*-values derived from the permutation tests in the statistical tests we used. The differences were especially large due to the small sample size, small cluster sizes (in solutions with a high number of clusters) and due to a low number of events (high survival) for the logrank test. These p-values are used by single and multi-omic methods to assess their performance, and the logrank p-value is often the main argument for an algorithm's merit. The large differences between the *P*-values question the validity of analyses that are based on the χ^2 approximation, at least for TCGA data. Future work must use exact or permutation-based calculations of the *P*-value in datasets with similar characteristics to those used here for the benchmark.

The benchmark we performed is not without limitations. Gauging performance using patient survival is somewhat biased to known cancer subtypes, which may have been used in treatment decisions. Additionally, cancer subtypes that are biologically different may have similar survival. This is also true for enrichment of clinical parameters, although we attempted to choose parameters that would not lead to this bias. However, these measures are widely used for clustering assessment, including in the papers describing some of the benchmarked methods. Another limitation of the benchmark is that it only examines clustering, while some of the methods have additional goals and output. For example, in dimension reduction algorithms, the low dimensional data can be used to analyze features, and not only patients, e.g. by calculating axes of variation common to several omics. With respect to feature analysis, multi-omic algorithms can have an advantage over single-omic algorithms that we did not test. Finally, though we selected the parameters of each benchmarked method according to the guidelines given by the authors, judicious fine-tuning of the parameters may improve results.

DATA AVAILABILITY

All the processed raw data are available at http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

ACKNOWLEDGEMENTS

The results published here are based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. We thank Nora K. Speicher for providing the rMKL-LPP tool and Ron Zeira for helpful comments.

FUNDING

United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel and the United States National Science Foundation (NSF); Bella Walter Memorial Fund of the Israel Cancer Association (in part); Edmond J. Safra Center for Bioinformatics at Tel-Aviv University (to N.R.) (in part). Funding for open access charge: BSF–NSF and ICA grant listed under Funders (in part).

Conflict of interest statement. None declared.

REFERENCES

1. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
2. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
3. Allison,D.B., Cui,X., Page,G.P. and Sabripour,M. (2006) Microarray data analysis: From disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.
4. Yong,W.-S., Hsu,F.-M. and Chen,P.-Y. (2016) Profiling genome-wide DNA methylation. *Epigenet. Chromatin*, **9**, 26.
5. Jain,A.K., Murty,M.N. and Flynn,P.J. (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.
6. Prasad,V., Fojo,T. and Brada,M. (2016) Precision oncology: origins, optimism, and potential. *Lancet Oncol.*, **17**, e81–e86.
7. Zhao,J., Xie,X., Xu,X. and Sun,S. (2017) Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, **38**, 43–54.
8. Network,T.C.G.A. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
9. Huang,S., Chaudhary,K. and Garmire,L.X. (2017) More is better: recent progress in multi-omics data integration methods. *Front. Genet.*, **8**, 84.
10. Bersanelli,M., Mosca,E., Remondini,D., Giampieri,E., Sala,C., Castellani,G. and Milanesi,L. (2016) Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*, **17**, S15.
11. Li,Y., Wu,F.-X. and Ngom,A. (2016) A review on machine learning principles for multi-view biological data integration. *Brief. Bioinformatics*, 325–340.
12. Wang,D. and Gu,J. (2016) Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quant. Biol.*, **4**, 58–67.
13. Meng,C., Zeleznik,O.A., Thallinger,G.G., Kuster,B., Gholami,A.M. and Culhane,A.C. (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinformatics*, **17**, 628–641.
14. Tini,G., Marchetti,L., Priami,C. and Scott-Boyer,M.-P. (2017) Multi-omics integration-a comparison of unsupervised clustering methodologies. *Brief. Bioinformatics*, doi:10.1093/bib/bbx167.
15. Shen,R., Olshen,A.B. and Ladanyi,M. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.
16. Wu,D., Wang,D., Zhang,M.Q. and Gu,J. (2015) Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: Application to cancer molecular classification. *BMC Genomics*, **16**, 1022.
17. Wang,H., Nie,F. and Huang,H. (2013) Multi-view clustering and feature learning via structured sparsity. *Proc. ICML '13*, **28**, 352–360.
18. Bickel,S. and Scheffer,T. (2004) Multi-view clustering. *Proc. ICDM 2004*, 19–26.
19. Hoadley,K.A., Yau,C., Wolf,D.M., Cherniack,A.D., Tamborero,D., Ng,S., Leiserson,M.D., Niu,B., McLellan,M.D., Uzunangelov,V. *et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, **158**, 929–944.
20. Bruno,E. and Marchand-Maillet,S. (2009) Multiview clustering: A late fusion approach using latent models categories and subject descriptors. In: *Proc. ACM SIGIR '09*. ACM Press, NY, pp. 736–737.
21. Nguyen,T., Tagett,R., Diaz,D. and Draghici,S. (2017) A novel approach for data integration and disease subtyping. *Genome Res.*, **27**, 2025–2039.
22. de Sa,V.R. (2005) Spectral Clustering with Two Views. In: *Proceedings of the Workshop on Learning with Multiple Views, 22nd ICML*. pp. 20–27.
23. Kumar,A., Rai,P. and Daumé,H. III (2011) Co-regularized multi-view spectral clustering. In: *Proc. NIPS '11*. USA, pp. 1413–1421.
24. Chikhi,N.F. (2016) Multi-view clustering via spectral partitioning and local refinement. *Inform. Process. Manage.*, **52**, 618–627.
25. Li,Y., Nie,F., Huang,H. and Huang,J. (2015) Large-scale multi-view spectral clustering with bipartite graph. In: *Proc. AAAI 15*. pp. 2750–2756.
26. Zhou,D. and Burges,C.J.C. (2007) Spectral clustering and transductive learning with multiple views. In: *Proc. ICML '07*. pp. 1159–1166.
27. Xia,R., Pan,Y., Du,L. and Yin,J. (2014) Robust multi-view spectral clustering via low-rank and sparse decomposition. *AAAI Conf. Artif. Intell.*, 2149–2155.
28. Bo,Wang, Jiayan,Jiang, Wei,Wang, Zhi-Hua,Zhou and Zhuowen,Tu (2012) Unsupervised metric fusion by cross diffusion. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2997–3004.
29. Wang,B., Mezlini,A.M., Demir,F., Fiume,M., Tu,Z., Brudno,M., Haibe-Kains,B. and Goldenberg,A. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.
30. Speicher,N.K. and Pfeifer,N. (2015) Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, **31**, i268–i275.
31. Long,B., Yu,P.S. and Zhang,Z.M. (2008) A General Model for Multiple View Unsupervised Learning. In: *Proceedings of the 2008 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 822–833.
32. Lock,E.F., Hoadley,K.A., Marron,J.S. and Nobel,A.B. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.*, **7**, 523–542.
33. O'Connell,M.J. and Lock,E.F. (2016) R. JIVE for exploration of multi-source molecular data. *Bioinformatics*, **32**, 2877–2879.
34. Hotelling,H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321.
35. Klami,A., Virtanen,S. and Kaski,S. (2013) Bayesian canonical correlation analysis. *J. Mach. Learn.*, **13**, 723–773.
36. Lai,P.L. and Fyfe,C. (2000) Kernel and Nonlinear Canonical Correlation Analysis. *Int. J. Neural Syst.*, **10**, 365–377.
37. Andrew,G., Arora,R., Bilmes,J. and Livescu,K. (2013) Deep canonical correlation analysis. In: *Proc. ICML '13*. Vol. **28**, pp. 1247–1255.
38. Parkhomenko,E., Tritchler,D. and Beyene,J. (2009) Sparse canonical correlation analysis with application to genomic data integration. *Stat. Applic. Genet. Mol. Biol.*, **8**, 1–34.
39. Witten,D.M. and Tibshirani,R.J. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Applic. Genet. Mol. Biol.*, **8**, Article28.
40. Vía,J., Santamaría,I. and Pérez,J. (2007) A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Netw.*, **20**, 139–152.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

41. Luo,Y., Tao,D., Ramamohanarao,K., Xu,C. and Wen,Y. (2016) Tensor canonical correlation analysis for multi-view dimension reduction. In: *Proc. ICDE 2016*. pp. 1460–1461.
42. Chen,J., Bushman,F.D., Lewis,J.D., Wu,G.D. and Li,H. (2013) Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, **14**, 244–58.
43. Lin,D., Zhang,J., Li,J., Calhoun,V.D., Deng,H.W. and Wang,Y.P. (2013) Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, **14**, 245.
44. Rohart,F., Gautier,B., Singh,A. and Lê Cao,K.-A. (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computat. Biol.*, **13**, e1005752.
45. Wold,S., Sjöström,M. and Eriksson,L. (2001) PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.*, **58**, 109–130.
46. Lê Cao,K.-A., Rossouw,D., Robert-Granié,C. and Besse,P. (2008) A sparse PLS for variable selection when integrating omics data. *Stat.Applic. Genet.Mol. Biol.*, **7**, doi:10.2202/1544-6115.1390.
47. Lê Cao,K.-A., Martin,P.G., Robert-Granié,C. and Besse,P. (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, **10**, 34.
48. Trygg,J. (2002) O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J. Chemometrics*, **16**, 283–293.
49. Rosipal,R., Trejo,L.J., Cristianini,N., Shawe-Taylor,J. and Williamson,B. (2001) Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Mach. Learn. Res.*, **2**, 97–123.
50. Rantalainen,M., Bylesjö,M., Cloarec,O., Nicholson,J.K., Holmes,E. and Trygg,J. (2007) Kernel-based orthogonal projections to latent structures (K-OPLS). *J. Chemometrics*, **21**, 376–385.
51. Li,W., Zhang,S., Liu,C.-C. and Zhou,X.J. (2012) Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**, 2458–2466.
52. Löfstedt,T. and Trygg,J. (2011) OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation. *J. Chemometrics*, **25**, 441–455.
53. Meng,C., Kuster,B., Culhane,A.C. and Gholami,A. (2014) A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, **15**, 162.
54. Liu,J., Wang,C., Gao,J. and Han,J. (2013) Multi-View Clustering via Joint Nonnegative Matrix Factorization. In: *Proc. ICDM '13*. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 252–260.
55. Kalayeh,M.M., Idrees,H. and Shah,M. (2014) NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 184–191.
56. Huang,J., Nie,F., Huang,H. and Ding,C. (2014) Robust Manifold Nonnegative Matrix Factorization. *ACM Trans. Knowledge Discov. Data*, **8**, 1–21.
57. Zhang,S., Liu,C.-C., Li,W., Shen,H., Laird,P.W. and Zhou,X.J. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379–9391.
58. Zhang,X., Zong,L., Liu,X. and Yu,H. (2015) Constrained NMF-based multi-view clustering on unmapped data. In: *Proc. AAAI '15*. Vol. **4**, pp. 3174–3180.
59. Li,S.-Y., Jiang,Y. and Zhou,Z.-H. (2014) Partial multi-view clustering. In: *Proc. AAAI '14*. AAAI Press, pp. 1968–1974.
60. Žitnik,M. and Zupan,B. (2015) Data fusion by matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**, 41–53.
61. White,M., Yu,Y., Zhang,X. and Schuurmans,D. (2012) Convex multi-view subspace learning. In: *Proc. NIPS '12*. USA, pp. 1673–1681.
62. Guo,Y. (2013) Convex subspace representation learning from multi-view data. *AAAI 2013*, 387–393.
63. Zhang,C., Fu,H., Liu,S., Liu,G. and Cao,X. (2015) Low-rank tensor constrained multiview subspace clustering. In: *Proc. ICCV '15*. IEEE, pp. 1582–1590.
64. Mo,Q., Wang,S., Seshan,V.E., Olshen,A.B., Schultz,N., Sander,C., Powers,R.S., Ladanyi,M. and Shen,R. (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 4245–4250.
65. Mo,Q., Shen,R., Guo,C., Vannucci,M., Chan,K.S. and Hilsenbeck,S.G. (2018) A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, **19**, 71–86.
66. Vaske,C.J., Benz,S.C., Sanborn,J.Z., Earl,D., Szeto,C., Zhu,J., Haussler,D. and Stuart,J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.
67. Savage,R.S., Ghahramani,Z., Griffin,J.E., de la Cruz,B.J. and Wild,D.L. (2010) Discovering transcriptional modules by Bayesian data integration. *Bioinformatics*, **26**, i158–i167.
68. Yuan,Y., Savage,R.S. and Markowetz,F. (2011) Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.*, **7**, e1002227.
69. Kirk,P., Griffin,J.E., Savage,R.S., Ghahramani,Z. and Wild,D.L. (2012) Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**, 3290–3297.
70. Lock,E.F. and Dunson,D.B. (2013) Bayesian consensus clustering. *Bioinformatics*, **29**, 2610–2616.
71. Gabasova,E., Reid,J. and Wernisch,L. (2017) Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLOS Comput. Biol.*, **13**, e1005781.
72. Ahmad,A. and Fröhlich,H. (2017) Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering. *Bioinformatics*, **33**, 3558–3566.
73. Coretto,P., Serra,A. and Tagliaferri,R. (2018) Robust clustering of noisy high-dimensional gene expression data for patients subtyping. *Bioinformatics*, doi:10.1093/bioinformatics/bty502.
74. Chaudhary,K., Poirion,O.B., Lu,L. and Garmire,L.X. (2018) Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.*, **24**, 1248–1259.
75. Liang,M., Li,Z., Chen,T. and Zeng,J. (2015) Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **12**, 928–937.
76. Bickel,P.J., Li,B., Tsybakov,A.B., van de Geer,S.A., Yu,B., Valdés,T., Rivero,C., Fan,J. and van der Vaart,A. (2006) Regularization in statistics. *Test*, **15**, 271–344.
77. Tibshirani,R. (1996) Regression Selection and Shrinkage via the Lasso. *J. R. Stat. Soc. B*, **58**, 267–288.
78. Blum,A. and Mitchell,T. (1998) Combining labeled and unlabeled data with co-training. In *Proc. COLT '98*. ACM Press, NY, 92–100.
79. Monti,S., Tamayo,P., Mesirov,J. and Golub,T. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.
80. Hofmann,T. (1999) Probabilistic latent semantic analysis. In: *Proc. UAI '99*. Morgan Kaufmann Publishers Inc., San Francisco, pp. 289–296.
81. Vega-Pons,S. and Ruiz-Shulcloper,J. (2011) A Survey of clustering ensemble algorithms. *Int. J. Pattern Recognit. Artif. Intell.*, **25**, 337–372.
82. von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.
83. Mohar,B. (1991) The Laplacian spectrum of graphs. *Graph Theory Combinatorics Applic.*, **2**, 871–898.
84. Lo Asz,L. (1993) Random walks on graphs: a survey. *Combinatorics*, 1–46.
85. Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann Publishers.
86. Cox,D.R. and Oakes,D. (1984) *Analysis of Survival Data*. Chapman and Hall
87. Chaudhuri,K., Kakade,S.M., Livescu,K. and Sridharan,K. (2009) Multi-view clustering via canonical correlation analysis. In: *Proc. ICML '09*. pp. 1–8.
88. Bach,F.R. and Jordan,M.I. (2006) A probabilistic interpretation of canonical correlation analysis. *Dept. Statist. Univ. California Berkeley CA Tech. Rep.*, **688**, 1–11.
89. Bylesjö,M., Eriksson,D., Kusano,M., Moritz,T. and Trygg,J. (2007) Data integration in plant biology: The O2PLS method for combined modeling of transcript and metabolite data. *Plant J.*, **52**, 1181–1191.
90. el Bouhaddani,S., Houwing-Duistermaat,J., Salo,P., Perola,M., Jongbloed,G. and Uh,H.-W. (2016) Evaluation of O2PLS in omics data integration. *BMC Bioinformatics*, **17**, S11.
91. Hwang,D., Stephanopoulos,G. and Chan,C. (2004) Inverse modeling using multi-block PLS to determine the environmental conditions that provide optimal cellular function. *Bioinformatics*, **20**, 487–499.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

92. Dray,S., Chessel,D. and Thioulouse,J. (2003) Co-inertia analysis and the linking of ecological data tables. *Ecology*, **84**, 3078–3089.
93. Seung,H.S. and Lee,D.D. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.
94. Lee,D.D. and Seung,H.S. (2001) Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Proc. Syst.*, 535–541.
95. Žitnik,M. and Zupan,B. (2015) Survival regression by data fusion. *Syst. Biomed.*, **2**, 47–53.
96. Hoadley,A. *et al.* (2018) Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, **173**, 291–304.
97. Ng,S., Collisson,E.A., Sokolov,A., Goldstein,T., Gonzalez-Perez,A., Lopez-Bigas,N., Benz,C., Haussler,D. and Stuart,J.M. (2012) PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*, **28**, i640–i646.
98. Geman,S. and Geman,D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6*, 721–741.
99. LeCun,Y., Bengio,Y. and Hinton,G. (2015) Deep learning. *Nature*, **521**, 436–444.
100. Krizhevsky,A., Sutskever,I. and Geoffrey,E. H. (2012) ImageNet classification with deep Convolutional neural Networks. In: *Proc. NIPS '12*. Vol. **1**, pp. 1097–1105.
101. Sutskever,I., Vinyals,O. and Le,Q.V. (2014) Sequence to sequence learning with neural networks. In: *Proc. NIPS'14*. MIT Press, Cambridge, pp. 3104–3112.
102. Ngiam,J., Khosla,A., Kim,M., Nam,J., Lee,H. and Ng,A.Y. (2011) Multimodal deep learning. *Proc. ICML '11*, 689–696.
103. Wang,W., Arora,R., Livescu,K. and Bilmes,J. (2016) On deep multi-view representation learning: objectives and optimization. *Proc. ICML '16*, 1083–1092.
104. Ching,T., Himmelstein,D.S., Beaulieu-Jones,B.K., Kalinin,A.A., Do,B.T., Way,G.P., Ferrero,E., Agapow,P.-M., Zietz,M., Hoffman,M.M. *et al.* (2018) Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface*, **15**, 20170387.
105. Hinton,G.E., Osindero,S. and Teh,Y.-W. (2006) A fast learning algorithm for deep belief nets. *Neural Comput.*, **18**, 1527–1554.
106. Hosmer,D.W., Lemeshow,S. and May,S. (2008) *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Wiley-Interscience.
107. Rousseeuw,P.J. and Peter (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
108. Vandin,F., Papoutsaki,A., Raphael,B.J. and Upfal,E. (2015) Accurate Computation of Survival Statistics in Genome-Wide Studies. *PLOS Comput. Biol.*, **11**, 1–18.
109. Aure,M.R., Steinfeld,I., Baumbusch,L.O., Liestøl,K., Lipson,D., Nyberg,S., Naume,B., Sahlberg,K.K., Kristensen,V.N., Børresen-Dale,A.-L. *et al.* (2013) Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. *PLoS ONE*, **8**, 1–15.

Downloaded from https://academic.oup.com/nar/article/46/20/10546/5123392 by Weizmann Institute of Science user on 19 June 2023

Chapter 3

NEMO: cancer subtyping by integration of partial multi-omic data

Bioinformatics, 35(18), 2019, 3348–3356
doi: 10.1093/bioinformatics/btz058
Advance Access Publication Date: 30 January 2019
Original Paper

OXFORD

Gene expression

NEMO: cancer subtyping by integration of partial multi-omic data

Nimrod Rappoport and Ron Shamir*

Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, 69978, Israel

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on September 30, 2018; revised on December 23, 2018; editorial decision on December 29, 2018; accepted on January 25, 2019

Abstract

Motivation: Cancer subtypes were usually defined based on molecular characterization of single omic data. Increasingly, measurements of multiple omic profiles for the same cohort are available. Defining cancer subtypes using multi-omic data may improve our understanding of cancer, and suggest more precise treatment for patients.

Results: We present NEMO (NEighborhood based Multi-Omics clustering), a novel algorithm for multi-omics clustering. Importantly, NEMO can be applied to partial datasets in which some patients have data for only a subset of the omics, without performing data imputation. In extensive testing on ten cancer datasets spanning 3168 patients, NEMO achieved results comparable to the best of nine state-of-the-art multi-omics clustering algorithms on full data and showed an improvement on partial data. On some of the partial data tests, PVC, a multi-view algorithm, performed better, but it is limited to two omics and to positive partial data. Finally, we demonstrate the advantage of NEMO in detailed analysis of partial data of AML patients. NEMO is fast and much simpler than existing multi-omics clustering algorithms, and avoids iterative optimization.

Availability and implementation: Code for NEMO and for reproducing all NEMO results in this paper is in github: https://github.com/Shamir-Lab/NEMO.

Contact: rshamir@tau.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent technological advances have facilitated the production of multiple genome-wide high throughput biological data types, collectively termed 'omics'. These include genomics, transcriptomics, proteomics and many more. Analysis of omics datasets was proven invaluable for basic biological research and for medicine. Until recently, research in computational biology has focused on analyzing a single omic type. While such inquiry provides insights on its own, methods for integrative analysis of multiple omic types may reveal more holistic, systems-level insights.

Omic profiles of large cohorts collected in recent years can help to better characterize human disease, facilitating more personalized treatment of patients. In oncology, analysis of large datasets has led to the discovery of novel cancer subtypes. The classification of tumors into these subtypes is now used in treatment decisions (Parker *et al.*, 2009; Prasad *et al.*, 2016). However, these subtypes are usually defined based on a single omic (e.g. gene expression), rather than through an integrative analysis of multiple data types. The large international projects like TCGA (McLendon *et al.*, 2008) and ICGC (Zhang *et al.*, 2011) now provide multi-omic cohort data, but better methods for their integrated analysis are needed. Novel, improved methods that employ multiple data types for cancer subtyping can allow us to better understand cancer biology, and to suggest more effective and precise therapy (Kumar-Sinha and Chinnaiyan, 2018; Senft *et al.*, 2017).

1.1 Multi-Omics clustering approaches

There are several approaches to multi-omics clustering (see the reviews by Huang *et al.*, 2017; Rappoport and Shamir, 2018; Wang and Gu, 2016). The simplest approach, termed *early integration*,

© The Author(s) 2019. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

concatenates all omic matrices and applies single-omic clustering on the resulting matrix. LRAcluster (Wu *et al.*, 2015) is an example of such a method, which probabilistically models the distribution of numeric, count and discrete features. Early integration increases the dimensionality of the data, and ignores the different distributions of values in different omics.

Late integration methods cluster each omic separately, and then integrate the clustering results, for example using consensus clustering (Monti *et al.*, 2003). PINS (Nguyen *et al.*, 2017) is a late integration method that defines connectivity matrices as describing the co-clustering of different samples within an omic, and integrates these matrices. Late integration ignores interactions that are weak but consistent across omics.

Middle integration approaches build a single model that accounts for all omics. These models include joint dimension reduction of omic matrices and similarity (kernel) based analyses. Dimension reduction approaches include jNMF, MultiNMF (Liu *et al.*, 2013; Zhang *et al.*, 2012), iCluster (Shen *et al.*, 2009), and its extensions iClusterPlus and iClusterBayes (Mo *et al.*, 2013, 2018). CCA is a classic dimension reduction algorithm (Hotelling, 1936), which linearly projects two omics to a lower dimension such that the correlation between the projections is maximal. MCCA (Witten and Tibshirani, 2009) generalizes CCA to more than two omics. Because of the high number of features and the complexity of dimension reduction algorithms, feature selection is required. Similarity based methods handle these shortcomings by working with inter-patient-similarities. These methods have improved runtime, and are less reliant on feature selection. Examples are SNF (Wang *et al.*, 2014) and rMKL-LPP (Speicher and Pfeifer, 2015). SNF builds a similarity network of patients per omic, and iteratively updates these networks to increase their similarity until they converge to a single network, which is then partitioned using spectral clustering. rMKL-LPP uses dimension reduction, such that similarities between neighboring samples is maintained in lower dimension. For that purpose, it employs multiple kernel learning, using several different kernels per omic, and providing flexibility in the choice of the kernels. All the middle integration methods above use iterative optimization algorithms, and in some cases guarantee only convergence to local optimum.

To the best of our knowledge, to date, all middle integration methods for multi-omics clustering developed within the bioinformatics community assume full datasets, i.e. data from all omics were measured for each patient. However, in real experimental settings, often for some patients only a subset of the omics were measured. We call these *partial datasets* in the rest of the paper. This phenomenon is already prevalent in existing multi-omic datasets, such as TCGA (McLendon *et al.*, 2008), and will increase as cohorts grow. Being able to analyze partial data is of paramount importance, due to the high cost of experiments, and the unequal cost for acquiring data for different omics. Naive solutions like using only those patients with all omics measured or imputation have obvious disadvantages.

A close problem to multi-omics clustering was researched in the machine learning community. In the area of 'multi-view learning' (reviewed in Zhao *et al.*, 2017), methods for multi-view clustering actually solve the multi-omic clustering problem. PVC (Li *et al.*, 2014) is such a method for clustering in the presence of partial data, which is based on joint nonnegative matrix factorization, such that the objective function only considers observed values. This method has not been previously applied on multi-omic data.

1.2 Our contribution

We present NEMO (NEighborhood based Multi-Omics clustering), a simple algorithm for multi-omics clustering. NEMO does not require iterative optimization and is faster than prior art. NEMO is inspired and bulids on prior similarity-based multi-omics clustering methods such as SNF and rMKL-LPP. NEMO's novelty lies in its simplicity, and in its support of partial data. Its implementation, as well as code to reproduce the results in this paper, are available in github: https://github.com/Shamir-Lab/NEMO.

We evaluated the performance of NEMO by comparing it to a wide range of multi-omics clustering methods on several cancer data types. On full datasets, despite its simplicity, NEMO performed comparably to leading multi-omics clustering algorithms. In order to evaluate performance on partial multi-omic data, we compared NEMO to PVC and to data imputation followed by clustering using several methods. In most tests on synthetic data and on real cancer data, NEMO had clear advantage. Finally, we analyzed NEMO's clustering solution for Acute Myeloid Leukemia, and showed the merit of using multiple omics with partial data.

2 Materials and methods

NEMO works in three phases. First, an inter-patient similarity matrix is built for each omic. Next, the matrices of different omics are integrated into one matrix. Finally, that network is clustered.

2.1 NEMO – full omics datasets

The input to NEMO is a set of data matrices of n subjects (samples or patients). Given L omics, let X_l denote the data matrix for omic l. X_l has dimensions $p_l \times n$, where p_l is the number of features for omic l. $P = \Sigma_l p_l$ is the total number of features.

Denote by x_{li} the profile of sample i in omic l (column i in X_l). Let η_{li} denote its k nearest neighbors within omic l, where Euclidean distance is used to measure profile closeness. For omic l, an $n \times n$ similarity matrix S_l is defined as follows:

$$S_l(i,j) = \frac{1}{\sqrt{2\pi}\sigma_{ijl}} \exp\left(-\frac{||x_{li} - x_{lj}||^2}{2 \cdot \sigma_{ijl}^2}\right) \quad (1)$$

where σ_{ijl}^2 is defined by:

$$\sigma_{ijl}^2 = \frac{1}{3} \cdot \left(\frac{1}{k}\sum_{r \in \eta_{li}} ||x_{li} - x_{lr}||^2 + \frac{1}{k}\sum_{r \in \eta_{lj}} ||x_{lj} - x_{lr}||^2 + ||x_{li} - x_{lj}||^2\right) \quad (2)$$

This similarity measure is based on the radial basis function kernel (Buhmann, 2003). σ_{ijl}^2 is a normalizing factor, which controls for the density of samples by averaging the squared distance of the ith and jth samples to their nearest neighbors and the squared distance between these two samples (Wang *et al.*, 2012, 2014; Yang *et al.*, 2008).

Next, we define the relative similarity matrix, RS_l, for each omic:

$$RS_l(i,j) = \frac{S_l(i,j)}{\sum_{r \in \eta_{li}} S_l(i,r)} \cdot I(j \in \eta_{li}) + \frac{S_l(i,j)}{\sum_{r \in \eta_{lj}} S_l(r,j)} \cdot I(i \in \eta_{lj}) \quad (3)$$

where I is the indicator function. $RS_l(i,j)$ measures the similarity between i and j relative to i's k nearest neighbors and to j's k nearest neighbors. Since different omics have different data distributions, the relative similarity is more comparable between omics than the original similarity matrix S.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

In the next step, NEMO calculates the $n \times n$ average relative similarity matrix ARS as:

$$ARS = \frac{1}{L}\sum_{l} RS_l \quad (4)$$

RS_l can be viewed as defining a transition probability between samples, such that the probability to move between samples is proportional to their similarity. Such transition distributions are widely used to describe random walks on graphs (Lo Asz, 1993). ARS is therefore a mixture of these distributions (Zhou and Burges, 2007).

Given ARS, the clusters are calculated by performing spectral clustering on ARS (von Luxburg, 2007). We use the spectral clustering variant that is based on the eigenvectors of the normalized Laplacian, developed by Ng *et al.* (2001).

To determine the number of clusters, we use a modified eigengap method (von Luxburg, 2007). The number of clusters is set to $argmax_i(\lambda_{i+1} - \lambda_i) \cdot i$, where λ are ARS eigenvalues. Intuitively, this objective maintains the idea of the eigengap while encouraging the solution to have a higher number of clusters. This is desired since we observed that often some increase in the number of clusters compared to that prescribed by the eigengap method improved the prognostic value for cancer data. The number of clusters determined by this method is at least as high as the number determined using the eigengap method.

As suggested by Wang *et al.* (2014), we set the number of neighbors in each omic to be $k = \frac{\#samples}{\#clusters}$ in case the number of clusters is known. When the number of clusters is not known, we use $k = \frac{\#samples}{6}$, using 6 as a crude estimate for the number of clusters observed in cancer datasets. We show NEMO's robustness to that parameter.

2.2 NEMO – partial datasets

NEMO can handle samples that were measured on only a subset of omics. Specifically, we require that each pair of samples has at least one omic on which they were both measured. Note that this holds in particular if there is an omic for which all samples have measurements, which is often the case for gene expression data. Under these conditions, RS_l is computed as in the full-data scenario, but ARS is now only averaged on the observed values. Denote by $JM(i, j)$ the omic types available for both samples. Then:

$$ARS(i,j) = \frac{1}{|JM(i,j)|}\sum_{l \in JM(i,j)} RS_l(i,j) \quad (5)$$

Note that we require that all samples that have measurements for some omic, have measurements for the same set of features in that omic, such that even in the partial data settings each X_l is a full matrix, albeit with fewer rows. For example, the expression of the same set of genes is measured for all patients with RNA-seq data. When patients have different sets of measured features in the same omic, either intersection of the features or imputation of missing values is required.

On partial datasets, each omic l may have a different number of samples $\#samples(l)$. The number k of nearest neighbors is chosen per omic. Generalizing the full data setting, for omic l we set $k = \frac{\#samples(l)}{6}$.

2.3 Time complexity

Computing the distance between a pair of patients in omic l takes $O(p_l)$, so calculating the distance between all patients in all omics takes $O(n^2 \cdot P)$. The k nearest neighbors of each patient and its average distance to them in a specific omic can be computed in time $O(n)$ per patient (Blum *et al.*, 1973), for a total of $O(n^2 \cdot L)$. Given the distances, the nearest neighbors, and the average distance to them, each σ^2_{ijl} can be computed in $O(k)$ time. Each entry in RS_l is also calculated in $O(k)$. ARS calculation therefore requires $O(n^2 \cdot P)$, and spectral clustering takes $O(n^3)$, so the total time is $O(n^2 \cdot P + n^3)$.

Other similarity-based methods such as SNF and rMKL-LPP need the same $O(n^2 \cdot P)$ time to calculate the distances. However, the iterative procedure in both SNF and rMKL-LPP requires $O(n^3)$ per iteration.

2.4 Clustering assessment

In datasets where the true clustering is known, to gauge the agreement between a clustering solution and the correct cluster structure, we used the adjusted Rand index (ARI) (Hubert and Arabie, 1985).

To assess clustering solutions for real cancer samples, we used survival data and clinical parameters reported in TCGA. We used the logrank test for survival (Hosmer *et al.*, 2008) and enrichment tests for clinical parameters. We used the χ^2 test for independence to calculate enrichment of discrete clinical parameters, and Kruskal-Wallis test for numerical parameters. It was previously observed that the χ^2 approximation for the statistic of these tests produces biased P-values that overestimate the significance (Rappoport and Shamir, 2018; Vandin *et al.*, 2015). In order to better approximate the real P-values, we performed permutation tests on the clustering solution, and reported the fraction of permutations for which the test statistic was greater or equal to that of the original clustering solution as the empirical P-value. Full details on the permutation testing appear in Rappoport and Shamir (2018).

3 Results

We applied NEMO in several settings. First, we compared it to nine multi-omics clustering algorithms on ten *full* cancer datasets. We next compared NEMO to several methods on simulated partial data, on multi-view image data and on real cancer datasets with parts of the data artificially removed. Finally, we used NEMO on a real partial cancer dataset.

3.1 Full datasets

We applied NEMO to ten TCGA datasets spanning 3168 patients. The datasets are for the following cancer types: Acute Myeloid Leukemia (AML), Breast Invasive Carcinoma (BIC), Colon Adenocarcinoma (COAD), Glioblastoma Multiforme (GBM), Kidney Renal Clear Cell Carcinoma (KIRC), Liver Hepatocellular Carcinoma (LIHC), Lung Squamous Cell Carcinoma (LUSC), Skim Cutaneous Melanoma (SKCM), Ovarian serous cystadenocarcinoma (OV) and Sarcoma (SARC). For each dataset, we analyzed three omics: gene expression, methylation and miRNA expression. When some of the patients lacked measurements for some of the omics, we included only those patients that had data from all omics. We have previously used these datasets to benchmark multi-omics clustering methods (Rappoport and Shamir, 2018). Datasets sizes varied between 170 and 621 samples. See Rappoport and Shamir (2018) for full details and preprocessing. Results for the execution of all methods on all datasets appear in Supplementary, Tables 1–4. Clustering results for NEMO on all datasets are in Supplementary File S2.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

Table 1. Aggregate statistics of the tested multi-omics clustering methods across ten cancer datasets

	K-means	Spectral	LRAcluster	PINS	SNF	rMKL-LPP	MCCA	MultiNMF	iClusterBayes	NEMO
Significantly different survival	2	3	5	5	4	4	5	4	2	**6**
Significant clinical enrichment	7	**8**	**8**	7	7	**8**	**8**	6	5	**8**
Number of clusters	2.6 (1.3)	3.3 (2.3)	9.4 (3.8)	4.6 (3.8)	2.8 (0.8)	6.7 (1.6)	10.6 (5.0)	2.2 (0.4)	2.2 (0.4)	4.5 (2.8)
Runtime (s)	394 (374)	**4** (2)	3388 (3295)	449 (958)	13 (11)	233 (43)	20 (10)	28 (12)	4298 (1703)	10 (4)

Note: First row: number of datasets with significantly different survival. Second row: number of datasets with at least one enriched clinical label. Third row: mean number of clusters. Fourth row: mean runtime. Best performers in each category are marked in bold. The numbers in parentheses are one standard deviation.

Table 2. Results of applying the ten algorithms on cancer datasets

Alg/Cancer	AML	BIC	COAD	GBM	KIRC	LIHC	LUSC	SKCM	OV	SARC	Means	#sig
kmeans	1/**2.9**	0/0.6	0/0	2/**2.3**	0/0.2	1/0.2	1/0.2	2/0.6	1/0.1	2/1.3	1/0.8	7/2
spectral	1/**1.7**	2/**1.6**	0/0.2	2/**2.2**	0/0.3	2/0.4	2/0.3	2/0.9	1/0.8	2/1.3	1.4/1	8/3
lracluster	1/**2**	4/**1.3**	0/0.5	1/**1.4**	1/**4.6**	0/0.8	1/0.9	2/**2.7**	1/0.6	2/1	1.3/**1.6**	8/5
pins	1/1.2	4/1.3	0/0	1/**3.6**	0/**1.8**	2/**2**	1/0.1	2/**2.8**	0/0	2/1.2	0.9/1.3	7/5
snf	1/**2.9**	2/1	0/0.2	1/**4.1**	1/**2.1**	2/0.2	0/0.6	1/0.6	0/0.2	2/**2.1**	1/**1.4**	7/4
mkl	1/**2.4**	5/0.6	0/0.5	2/**3**	1/1.1	3/1	0/0.3	1/**2.6**	1/0.1	2/**2.5**	1.6/**1.4**	8/4
mcca	1/**1.4**	0/**3.2**	1/0.3	2/**1.8**	1/**3.9**	2/0.9	0/0.4	2/**4.3**	1/0.7	2/0.6	1.2/**1.8**	8/5
nmf	0/1.3	0/1.3	0/0.3	1/**2.1**	1/**1.9**	3/**2.9**	1/0.3	2/**4.5**	0/0.3	2/1.1	1/**1.6**	6/4
iCluster	0/1	3/0.2	0/0.2	0/1	0/**2**	2/1	2/0.6	3/**4.4**	0/0	2/0.8	1.2/1.1	5/2
nemo	1/**2.1**	3/**1.4**	0/0.2	1/**2**	1/1.2	3/**3.3**	0/0.4	3/**3.9**	1/0.1	2/**1.8**	1.5/**1.6**	8/6

Note: The first number in each cell is the number of significant clinical parameters detected, and the second number is the −log10 *P*-value for survival, with bold numbers indicating significant results. Means are algorithm averages. #sig is the number of datasets with significant clinical/survival results. We use 0.05 as the threshold for significance.

We compared NEMO on each dataset to nine different multi-omics clustering methods. As early integration methods we used LRAcluster, and k-means and spectral clustering on the concatenation of all omic matrices. For late integration we used PINS. We used MCCA, MultiNMF and iClusterBayes as joint dimension reduction methods. Finally, SNF and rMKL-LPP represented similarity-based integration. We set k, the number of neighbors in NEMO to $k = \frac{\#samples}{6}$. For all methods, we chose the number of clusters in the range 2-15 using the methods recommended by the authors. The results of the nine methods were taken from our benchmark study (Rappoport and Shamir, 2018), where full details on the execution of all methods are available. (For MCCA, LRAcluster and k-means the results are slightly different, since here we increased the number of k-means repeats they perform in order to increase their stability.)

To assess the clustering solutions we compared the survival curves of different clusters, and performed enrichment analysis on clinical labels (see Section 2). To avoid biases, we chose the same set of clinical parameters for all cancers: age at initial diagnosis, gender and four discrete clinical pathological parameters. These parameters quantify the progression of the tumor (pathologic T), cancer in lymph nodes (pathologic N), metastases (pathologic M) and total progression (pathologic stage). In each cancer type we tested the enrichment of each parameter that was available for it.

Table 1, Figure 1 and Table 2 summarize the performance of the ten algorithms on the ten datasets. NEMO found a clustering with significant difference in survival for six out of ten cancer types, while all other methods found at most five. None of the methods found a clustering with significantly different survival for the COAD, LUSC and OV datasets. The *P*-value for KIRC, the only other dataset for which NEMO did not reach significance, was 0.063. NEMO had an average logrank *P*-value of 1.64, second after MCCA (1.75). NEMO found at least one enriched clinical parameter in eight of the ten datasets, the highest number found and tied with spectral clustering, LRACluster, rMKL-LPP, PINS and MCCA. The average number of enriched clinical parameters for NEMO was 1.5, second only to rMKL-LPP with 1.6. Standard deviations across the different datasets for Figure 1 appear in Supplementary Figure S1.

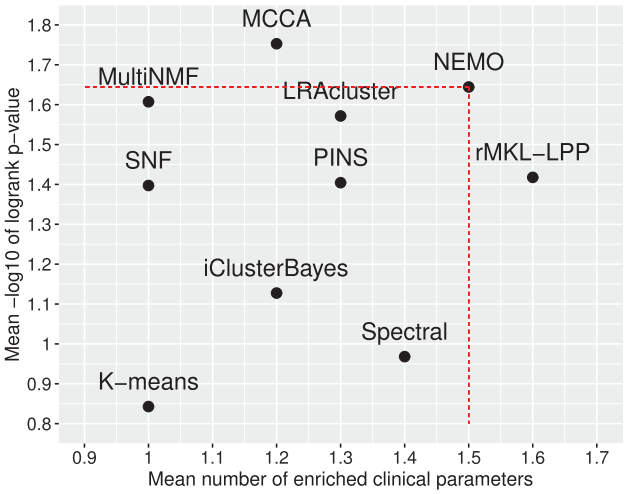


Fig. 1. Mean performance of the ten algorithms on ten cancer datasets. Y axis: average significance of the difference in survival among clusters (−log10 logrank test's *P*-values). X axis: average number of enriched clinical parameters in the clusters. The dotted lines highlight NEMO's performance

Compared to the other methods, NEMO tended to choose an intermediate number of clusters per dataset (average 4.5, see Table 1). This number of clusters is small enough so that the clusters will be highly interpretable, but still capture the heterogeneity among cancer subtypes.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

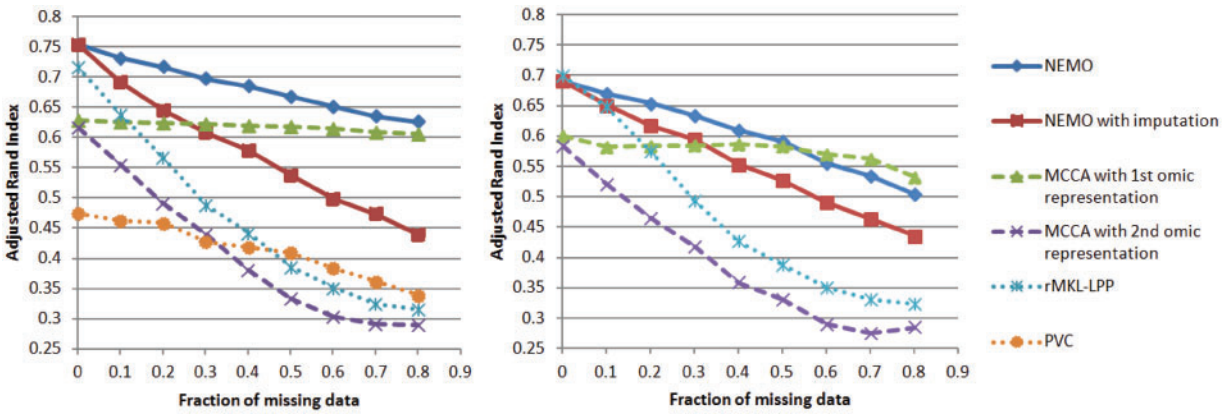


Fig. 2. Performance on simulated partial data. We executed the algorithms with an increasing fraction of samples missing data in one of the omics, and compared the resulting clustering to the ground truth using ARI. The left plot uses two omics, and the right plot uses three omics where the third one contains only noise

NEMO had the the second fastest average runtime after spectral clustering of the concatenated omics matrix. (The same was true for the geometric mean runtime, see Supplementary Table S5). All methods except for LRAcluster and iClusterBayes took only a few minutes to run on datasets with hundreds of samples and tens of thousands of features. However, due to NEMO's simple integration step, it was the fastest of all non-trivial integration methods, including other similarity-based methods (SNF and rMKL-LPP). The runtime improvement over SNF was minor for most datasets in the experiment, and was mainly seen in the largest dataset (BRCA), where SNF took 43 s and NEMO 19. For rMKL-LPP, the time reported does not include the similarity computation, as this code was not provided by the authors, but was implemented by us, so its total runtime is higher. Details regarding the hardware used appear in Supplementary File S1. We note that since NEMO's integrated network is sparse, its spectral clustering step can be further improved using methods for spectral clustering of sparse graphs (e.g. Lanczos, 1950). This advantage in runtime, and NEMO's improved asymptotic runtime compared to other similarity-based methods (see Section 2) will become more important as the number of patients in medical datasets increases.

3.2 Simulated partial datasets

We next evaluated NEMO's performance on simulated partial datasets. We tested two scenarios. In the first we created two clusters using multivariate normal noise around the clusters' centers, and then created two omics by adding to these data different normal noise for each omic (see Supplementary File S1). The simulation is therefore designed such that both omics share the same underlying clustering structure. In the second scenario, we added a third omic that does not distinguish between the clusters. To simulate partial data, we removed the second omic data in an increasing fraction, which we denote θ, of randomly chosen samples, for θ ranging between 0 and 0.8. We generated 10 different full datasets, and for each dataset and for each value of θ we performed ten repeats. Here we report the average ARI between the computed and the correct clustering for each θ.

We compared NEMO's performance to PVC, and also to MCCA and rMKL-LPP, the top performers on the full real data. To run PVC on the dataset, we subtracted the minimal observed value from each omic, making all values non-negative, and set PVC's λ parameter to 0.01. Since PVC's implementation supports only two omics, we ran it only in the first scenario. To run rMKL-LPP and MCCA on partial data, we completed the missing values using KNN imputation on the concatenated omics matrix. We used KNN imputation since it was shown to perform well in omic data (Troyanskaya *et al.*, 2001). We ran the procedure on the concatenated matrix because it allows imputation of values for samples that lack one of the omics, using the similarity of a sample to other samples in other omics, and assuming that the different omics are correlated. As the number of clusters in the simulated data was known to be 2, we set NEMO's parameter k to half the number of samples as described in the Section 2. Full details about PVC's execution appear in Supplementary File S1. MCCA was applied twice, using two low dimensional representations (See Supplementary File S1 for details).

Figure 2 shows that NEMO outperformed other methods in both simulations. Furthermore, NEMO performed better on data that were not imputed than on data that were imputed. This shows the advantage of using NEMO directly on partial datasets, rather than performing imputation. In both scenarios, the performance of all methods deteriorated as the fraction of missing data increased. A notable exception was MCCA when using the low dimensional representation of the first omic. We believe this representation was barely affected by the second omic. Interestingly, adding a third omic that contributes no information to the clustering solution decreased the performance, but this decrease was minor for all methods.

PVC performed poorly compared to NEMO in the two-omics simulation. In fact, PVC with all data for both omics performed worse than NEMO with 80% missing data in the second omic. We suspect that since PVC is based on linear dimension reduction, it does not capture the spheric structure of the clusters. Since PVC's implementation supports only two omics, we could not test it in the second scenario.

3.3 Image dataset

To further test NEMO's performance in a complicated dataset where the true clustering is known, we ran the different methods on Handwritten, an image dataset which contains 2000 images of the digits 0–9, with 200 images of each digit. This dataset is widely used by the machine learning community to benchmark multi-view methods (Zhao *et al.*, 2017). We used two 'omics' for this dataset. The first contains 240 pixel averages for windows of size 2 × 3. The second contains 76 Fourier coefficients of the images. For performance reasons, we used only 500 randomly sampled images. We simulated partial data by randomly removing data for half the

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

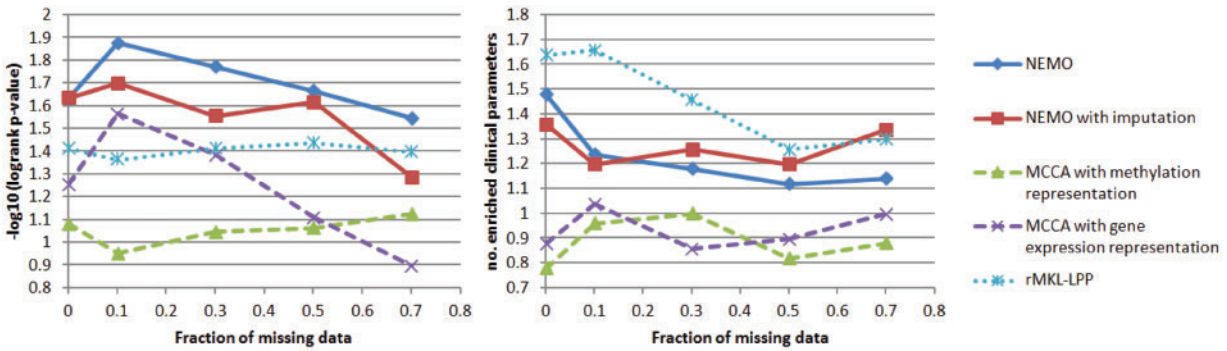


Fig. 3. Performance on partial cancer datasets as a function of the fraction of samples missing data in one of the omics. Left: survival analysis. Right: clinical parameters. Results are averages across ten three-omics cancer datasets

samples from the second omic. Like in the simulated data, NEMO was compared to the methods with best performance on the full cancer datasets. Data for methods other than NEMO and PVC was imputed using KNN, and the data were clustered assuming ten clusters. We repeated this clustering process ten times, each time selecting at random the samples that were removed in the second omic. Supplementary Table S6 contains the mean ARI between the obtained clusters and the true clustering. On this dataset, with full data, rMKL-LPP and NEMO were comparable, and they both greatly ourperformed the other methods. On partial data, NEMO was best.

3.4 Partial cancer datasets

We next compared NEMO to other methods on partial cancer datasets, by simulating data loss on the ten full TCGA datasets analyzed above. We tested two scenarios, (i) using three omics for all subtypes, and (ii) using only two omics, to allow comparison with PVC. We randomly sampled a fraction θ of the patients and removed their second omic data. The other omic(s) data were kept full. The θ values tested were between 0 and 0.7. In all datasets, the first omic was DNA methylation, and the second (from which samples were removed) was gene expression. In the three-omic scenario, the last omic was miRNA expression. We repeated each test five times, and the *P*-values reported here are the geometric averages of the observed *P*-values.

Full details on how each method was executed are in Supplementary File S1. We set the number of clusters in PVC to be the same as determined by NEMO, since no method to determine the number of clusters was suggested for PVC. We used survival analysis and enrichment of clinical labels to assess the quality of the clustering solutions. Full results for this analysis are in Supplementary File S3.

Figure 3 shows the mean results on three omics across all ten cancer types. NEMO performed best with respect to survival, followed by NEMO with imputation. rMKL-LPP performed best with respect to clinical parameters, followed by NEMO with and without imputation.

Note that in contrast to simulated data, here the performance of the methods did not consistently deteriorate as more data were removed. This is somewhat surprising, as gene expression (the omic that was partially removed) is believed to be the most informative omic. While on average performance across the cancer types was not consistent, we did see a decrease in performance on some of the datasets. The difference between the performance of MCCA for full data here ($\theta = 0$) and its previous results (Fig. 1) is due to the fact that MCCA optimizes its objective with respect to one omic at a time, which makes the solution sensitive to the order of the omics. We also ran MCCA using the original omic order (see Supplementary Fig. S8). Still, NEMO outperformed MCCA with respect to survival in all runs except on full data with all three omics (the setting for the original full data experiments).

In the second scenario, out of the datasets that had statistically significant survival results, NEMO was best performer for AML, GBM and SARC, while PVC was best for BIC and SKCM (Supplementary Fig. S4). PVC (using the number of clusters determined by NEMO) had best mean survival and clinical enrichment across all datasets (Supplementary Fig. S6). This shows the merit of PVC (and of NEMO's method to determine the number of clusters) in datasets with two omics. Interestingly, for both NEMO and PVC, the mean performance across all ten full two-omics datasets was better than the performance of all methods on full three-omics datasets in terms of survival (Fig. 3 and Supplementary Fig. S6; see also Supplementary Fig. S7 for MCCA with the reverse omic order).

Performing imputation increases the runtime of the algorithms. For example, the average time (across 5 runs) to perform imputation for the BIC dataset with methylation and mRNA expression omics, when $\theta = 0.5$, was 560 s. This is a necessary preprocessing step for methods that do not directly support missing data. In contrast, not only do NEMO and PVC not require imputation, but they also run faster as the fraction of missing data increases. The runtime of NEMO on the same two-omic BIC dataset decreased from 42 s with full data to 21 s with $\theta = 0.7$. PVC was slower than NEMO, and its runtime decreased from 92 to 27 s.

3.5 Robustness analysis

We sought to assess NEMO's robustness with respect to the parameter k, the number of neighbors, and with respect to the number of clusters. We first tested robustness on the simulated data. We executed NEMO on the three-omic simulated data described previously. We used $k = 10, 20, \ldots, 200$ and compared the obtained clustering to the known clustering using the Adjusted Rand Index. Supplementary Figure S9 shows that in that setting, NEMO was highly robust to the choice of k, except for low values.

We next performed clustering on the ten cancer datasets using $k = 25, 35, \ldots, 105$, a range that includes all k values we used in the full and partial data analyses. Supplementary Figure S10 shows the *P*-values for logrank test for each value of k. Generally, the performance of NEMO was robust with respect to k. In a few cases, such as the GBM, SKCM and SARC datasets, the results varied more depending on k. This is partially explained by the different number of clusters NEMO chose for different values of k (see Supplementary Fig. S12). We conclude that usually changing k has little effect on

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

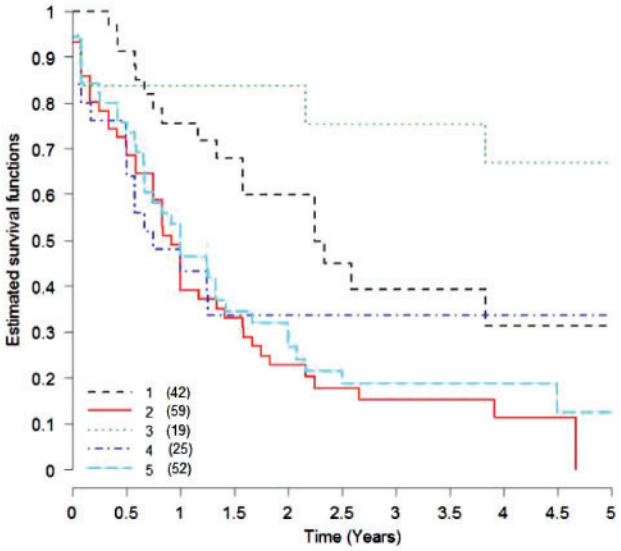


Fig. 4. Kaplan-Meier plot for the five clusters obtained by NEMO on the AML partial dataset (logrank *P*-value = 3.5e−4). The number of patients in each cluster is shown in parentheses

the number of clusters and on the significance. In those cases where significance changed with k, it was usually a result of change in the number of clusters chosen (compare Supplementary Figs S10, S11 and S12). We next clustered the ten datasets using a number of clusters ranging from 2 to 15. Supplementary Figures S13 and S14 show the effect of the different number of clusters on the survival analysis and number of enriched clinical parameters. We note that NEMO is less robust to the number of clusters chosen than to k.

3.6 Acute myeloid leukemia analysis

We applied NEMO to an AML cohort of 197 patients from TCGA. This is a partial dataset, containing 173 patients with gene expression profiles, 194 with methylation and 188 with miRNA profiles. As it is partial, the dataset cannot be directly clustered using other algorithms for multi-omics clustering. To apply these methods, one must limit analysis to a sub-cohort of 170 patients that have full data, or perform imputation. NEMO suggested five clusters for this dataset; their sizes appear in Figure 4. When plotting survival curves of the clusters (Fig. 4), we found them to be significantly different (*P*-value = 3.5e−4). The significance was higher than obtained by all other nine clustering methods on the full data subcohort (lowest *P*-value 1.3e−3 using k-means). This shows the higher significance gained from analyzing more samples, including partial data.

We compared the prognostic value of NEMO's clusters to that of the FAB (French-American-British) classification. FAB is a well-accepted clinical classification for AML tumors (Bennett *et al.*, 1976), which is based on quantification of blood cells. We performed logrank test using the FAB label as clustering solution, and obtained a *P*-value 5.4e−2, which shows NEMO's favorable prognostic value. Executing NEMO using only a single omic, results for gene expression, methylation and miRNA expression data had logrank *P*-values 3.4e−2, 3.4e−3 and 3.7e−3 respectively. These results demonstrate the improved clustering obtained by NEMO using multi-omic data.

We performed enrichment analysis for each NEMO cluster using the PROMO tool, which allows systematic interrogation of all clinical labels (Netanely *et al.*, 2016). In addition to the significantly differential survival, the clusters were found to be enriched in other clinical labels. Cluster 1 had particularly young patients, and showed favorable prognosis. Cluster 2 had poor prognosis, older patients, and was enriched with FAB label 'M0 undifferentiated'. 17 out of 19 patients with label 'M0 undifferentiated' appeared in this cluster. This label corresponds to the undifferentiated acute myeloblastic AML subtype, which is known to have poor prognosis (Bene *et al.*, 2001). Cluster 3 showed favorable prognosis, and was enriched with the M3 FAB label, which corresponds to the acute promyelocytic leukemia (APL) subtype. All 19 patients in this cluster were labeled with M3, and only one patient outside cluster 3 had this label. APL is caused by a translocation between the genes RARA on chromosome 17 and PML on chromosome 15, and is known to have favorable prognosis (Wang and Chen, 2008). Cluster 4 was enriched with the M5 label, which corresponds to acute monocytic leukemia. Indeed, it was also enriched with a high monocyte count. Finally, cluster 5 was enriched with patients with no known genetic aberrations. All the clustering results and enriched clinical labels are included in Supplementary Files S4 and S5.

4 Discussion

We presented the NEMO algorithm for multi-omics clustering, and tested it extensively on cancer datasets and in simulation. NEMO is much simpler than existing multi-omics clustering algorithms, has comparable performance on full datasets, improved performance on partial datasets without requiring missing data imputation, and faster execution.

The main insight NEMO uses is that the local neighborhood of each sample best captures its similarity patterns in each omic. We believe that NEMO's performance stems largely from this insight. Previous methods used local similarities, and NEMO suggests that the performance of these methods was largely due to that use, rather than to other steps performed by these algorithms.

NEMO's simplicity makes it more flexible and more easily adapted to different circumstances. It requires only the definition of a distance between two samples within an omic, and can therefore support additional omics, numerical, discrete and ordinal features, as well as more complicated feature types, such as imaging, EMR data and microbiome. In addition to enabling clustering, the network produced by NEMO represents the similarity between samples across all omics, and can thus be used for additional computational tasks. Future work will test the usability of NEMO on discrete data types, and of its output network for tasks other than clustering.

We showed that NEMO can be used to analyze partial multi-omic datasets, i.e. ones in which some samples lack measurements for all omics. Partial datasets are ubiquitous in biology and medicine, and methods that analyze such datasets hold great potential. This challenge is exacerbated by the high cost of high-throughput experiments. While the price of some experiments is decreasing, it is still high for other omics. Methods that analyze partial datasets may affect experimental design and reduce costs, and, as we demonstrated, they can outperform full-data methods applied only to the subset of samples that have all omics. The demand for algorithms that analyze partial datasets is likely to further increase, as more high throughput methods become prevalent, and the number of omics in biomedical datasets increases.

NEMO has several limitations. First, in partial data, each pair of samples must have at least one omic in common. This assumption holds if one omic was measured for all patients, which is often the case for gene expression. Second, the choice of k, the number of nearest neighbors, requires further study. NEMO currently chooses

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

the same k for all samples, implicitly assuming that all cluster sizes are equal. Choosing different k for different samples based on the estimated size of their cluster may further improve NEMO's results. Third, unlike some dimension reduction methods, NEMO does not readily provide insight on feature importance. Given a clustering solution, importance of features to clusters can be computed using differential analysis.

We compared NEMO to PVC in the context of missing data. PVC was developed within the machine learning community for the task of partial multi-view clustering, and has not been applied to omic data previously. Remarkably, on average, in terms of survival analysis, PVC (using the number of clusters of NEMO) outperformed all other methods on the partial cancer datasets with two omics, while NEMO was better on the simulated partial datasets. As PVC is limited to two omics, extension of that NMF-based algorithm to more omics and to include a mechanism for determining the number of clusters is desirable.

In some of the cancer datasets the results obtained using only mRNA expression and DNA methylation were superior to those achieved when also considering miRNA expression. In addition, in some of the datasets we did not observe a significant decrease in performance when removing a fraction of the gene expression data for cancer patients. This phenomenon suggests that multi-omics clustering does not necessarily improve with more omics (see also Rappoport and Shamir, 2018). A possible explanation is that the different omics are highly correlated, such that additional omics do not add signal. At least for some of the cancer types, this was not the case. Alternatively, it is possible that omics contain contradicting or independent signals, such that removal of data from one omics strengthens the overall structure of the data. While NEMO performed well with an additional omic that contains no signal, future work is needed to deal with omics that contain independent or contradicting signals.

5 Conclusion

Clustering cancer patients into subgroups has the potential to define new disease subtypes that can be used for personalized diagnosis and therapy. The increasing diversity of omics data as well as their reduced cost creates an opportunity to use multi-omic data to discover such subgroups. NEMO's simplicity, efficiency and efficacy on both full and partial datasets make it a valuable method for this challenge.

Acknowledgements

The results published here are based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. We thank Ron Zeira for helpful comments. The contribution of N.R. is part of Ph.D. thesis research conducted at Tel Aviv University.

Funding

This work was supported in part by grant 2016694 of the United States - Israel Binational Science Foundation (BSF), Jerusalem, Israel and the United States National Science Foundation (NSF), by the Naomi Prawer Kadar Foundation and by the Bella Walter Memorial Fund of the Israel Cancer Association. N.R. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

Conflict of Interest: none declared.

References

Bene,M.-C. *et al.* (2001) Acute myeloid leukaemia M0: haematological, immunophenotypic and cytogenetic characteristics and their prognostic significance: an analysis in 241 patients. *Br. J. Haematol.*, **113**, 737–745.

Bennett,J.M. *et al.* (1976) Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br. J. Haematol.*, **33**, 451–458.

Blum,M. *et al.* (1973) Time bounds for selection. *J. Comput. Syst. Sci.*, **7**, 448–461.

Buhmann,M.D. (2003) *Radial Basis Functions: Theory and Implementations. Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, UK.

Hosmer,D.W. *et al.* (2008) *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Wiley-Interscience, New York, NY, USA.

Hotelling,H. (1936) Relations between two sets of variates. *Biometrika*, **28**, 321.

Huang,S. *et al.* (2017) More is better: recent progress in multi-omics data integration methods. *Front. Genet.*, **8**, 84.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Kumar-Sinha,C. and Chinnaiyan,A.M. (2018) Precision oncology in the age of integrative genomics. *Nat. Biotechnol.*, **36**, 46–60.

Lanczos,C. (1950) An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl. Bureau Standards 1950*, **45**, 255–281.

Li,S.-Y. *et al.* (2014) Partial multi-view clustering. In: *Proc. Proc. Assoc. Adv. Artif. Intell., 2014*. AAAI Press, pp. 1968–1974.

Liu,J. *et al.* (2013) Multi-view clustering via joint nonnegative matrix factorization. In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, Philadelphia, PA, pp. 252–260.

Lo Asz,L. (1993) Random walks on graphs: a survey. *Combinatorics*, **2**, 1–46.

McLendon,R. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Mo,Q. *et al.* (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci. USA*, **110**, 4245–4250.

Mo,Q. *et al.* (2018) A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, **19**, 71–86.

Monti,S. *et al.* (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.

Netanely,D. *et al.* (2016) Expression and methylation patterns partition luminal-a breast tumors into distinct prognostic subgroups. *Breast Cancer Res.*, **18**, 74.

Ng,A.Y. *et al.* (2001) On spectral clustering: analysis and an algorithm. In: *Proc. Conf. Neural Information Processing Systems*. MIT Press, Cambridge, Massachusetts, pp. 849–856.

Nguyen,T. *et al.* (2017) A novel approach for data integration and disease subtyping. *Genome Res.*, **27**, 2025–2039.

Parker,J.S. *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

Prasad,V. *et al.* (2016) Precision oncology: origins, optimism, and potential. *Lancet Oncol.*, **17**, e81–e86.

Rappoport,N. and Shamir,R. (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.*, **46**, 10546–10562.

Senft,D. *et al.* (2017) Precision oncology: the road ahead. *Trends Mol. Med.*, **23**, 874–898.

Shen,R. *et al.* (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

Speicher,N.K. and Pfeifer,N. (2015) Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, **31**, i268–i275.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for dna microarrays. *Bioinformatics*, **17**, 520–525.

Vandin,F. *et al.* (2015) Accurate computation of survival statistics in genome-wide studies. *PLOS Comput. Biol.*, **11**, 1–18.

von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

Wang,B. *et al*. (2012) Unsupervised metric fusion by cross diffusion. In: *Proceeding IEEE Conference on Computer Vision and Pattern Recognition*. IEEE 2012, pp. 2997–3004.

Wang,B. *et al*. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Wang,D. and Gu,J. (2016) Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quant. Biol*., **4**, 58–67.

Wang,Z.-Y. and Chen,Z. (2008) Acute promyelocytic leukemia: from highly fatal to highly curable. *Blood*, **111**, 2505–2515.

Witten,D.M. and Tibshirani,R.J. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol*., **8**, 28.

Wu,D. *et al*. (2015) Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics*, **16**, 1022.

Yang,X. *et al*. (2008) Improving shape retrieval by learning graph transduction. In: *Proc. 10th Eur. Conf. Comput. Vis. (ECCV), 2008*, Forsyth,D. (eds), pp. 788–801.

Zhang,J. *et al*. (2011) International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database J. Biol. Databases Cur*., **2011**, bar026.

Zhang,S. *et al*. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*., **40**, 9379–9391.

Zhao,J. *et al*. (2017) Multi-view learning overview: recent progress and new challenges. *Inf. Fusion*, **38**, 43–54.

Zhou,D. and Burges,C.J.C. (2007) Spectral clustering and transductive learning with multiple views. In: *ICML '07: Proceedings of the 24th international conference on Machine learning, 2007*,, pp. 1159–1166.

Downloaded from https://academic.oup.com/bioinformatics/article/35/18/3348/5304361 by Weizmann Institute of Science user on 19 June 2023

Chapter 4

Inaccuracy of the log-rank approximation in cancer data analysis

Check for updates

Correspondence

OPEN ACCESS

molecular systems biology

Inaccuracy of the log-rank approximation in cancer data analysis

Nimrod Rappoport & Ron Shamir

Comparing survival patterns between groups of individuals is ubiquitous in biomedical research. A significant difference in survival can show the efficacy of a drug or the biological relevance of a biomarker. In cancer research, clustering of patient profiles is used to discover disease subtypes (Prasad *et al*, 2016), and a significant difference in survival between clusters is usually considered a strong indication for a clustering algorithm's merit (Gabasova *et al*, 2017; Argelaguet *et al*, 2018). In these settings, the standard means to compare survival between groups of patients is the log-rank test (Hosmer *et al*, 2008). We refer here to the conditional version of the test (see Appendix).

The log-rank test is very broadly used. A Google Scholar search for "logrank test statistic" identifies > 22,000 citations, and a PubMed search in titles or abstracts for "logrank" or "log-rank" identifies > 30,000 papers, and 3,357 published in 2018 alone. The real number of studies that use this test is likely even higher. The *P*-value of the log-rank test statistic is commonly approximated by the chi-square distribution. We show here that in important contexts that approximation is poor and can be misleading.

The chi-square approximation provides a good fit when there are a large number of events in each patient group and the group sizes are balanced. Heinze *et al* (2003) and Wang *et al* (2010) developed exact permutation tests that condition on the observed follow-up in each group. While they showed that the asymptotic log-rank test is inaccurate, the extent of this inaccuracy in practice, for real modern datasets that contain hundreds of patients and more than two clusters, is unclear.

We have recently benchmarked nine methods for clustering multi-omic data across ten cancer cohorts from TCGA (The Cancer Genome Atlas Network, 2008; Rappoport & Shamir, 2018). Since survival information was available for the patients, we used the log-rank test chi-square approximation to evaluate each solution. In addition, we implemented the exact test developed by Heinze *et al* (2003) for more than two groups. We validated on simulated data that the implementation preserves the false-positive rate better than the asymptotic version (see Appendix), and used the implementation to compute the exact test's *P*-value (EP) of the log-rank score for each solution on each cancer cohort. The results (Fig 1A) show large gaps between the EP and asymptotic *P*-value (AP). In fact, the APs for 48 out of the 90 clustering solutions were not within their 95% confidence intervals constructed using the permutation test. This inaccuracy was exacerbated for small *P*-values: 30 out of the 37 significant APs (≤ 0.05) did not fall within their 95% confidence intervals. In all these cases, the EPs were higher (less significant). In 17 out of the 37, the difference between the EP and the AP was at least 2-fold. Three of the 37 cases reported as significant according to the asymptotic approximation (8%) were actually not significant according to the permutation tests.

Some asymptotic results were rather extreme. The MCCA method (Witten & Tibshirani, 2009) on the KIRC cancer dataset gave a clustering solution that obtained AP < 2e-16, but EP = 6.8e-5. The distribution of the APs over one million permutations of the KIRC cluster labels is shown in Fig 1B. By definition, that distribution should be uniform under the null hypothesis. However, 10.9% of the permutations received an AP ≤ 0.05.

We performed an additional test using the breast cancer dataset, which contains 621 patients. For each number k of clusters from 2 to 20, we partitioned the samples at random into $k-1$ clusters of 10 patients and one large cluster with all other patients, and computed the APs. We repeated the process for many random permutations of the patient labels and calculated the fraction of permutations with AP ≤ 0.05 (see Appendix). The results are shown in Fig 1C. In spite of the large size of the breast cancer dataset, the probability to report a clustering as significant was markedly higher than 0.05 and increased as the number of clusters k increased. For $k = 4$, a common number of clusters for breast cancer datasets, the probability for AP ≤ 0.05 was already > 0.08.

How common is the use of the asymptotics in software tools? The R "survival" package, the Python "lifelines" package, SPSS, SAS and Stata all use the asymptotic test and report the same *P*-values. While several packages do implement non-asymptotic tests (see Appendix), they are less widely used. We conclude that the vast majority of the studies that perform the log-rank test use the asymptotic *P*-value.

A systematic search for cases where the use of the asymptotic test led to wrong conclusions is challenging: most studies do not publish survival data, and these data have no standard format. However, we were able to find recent cases where the test led to wrong or overstated reports. Joachim *et al* (2018) reported that use of the chemotherapeutic agent Topotecan resulted in a significant survival benefit in a murine model of endotoxemia. While the log-rank AP was 0.042 for the presented data, the EP was actually 0.059. Gabasova *et al* (2017) developed a novel method for multi-omic

Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. E-mail: rshamir@tau.ac.il
DOI 10.15252/msb.20188754 | Mol Syst Biol. (2019) 15: e8754

© 2019 The Authors. Published under the terms of the CC BY 4.0 license

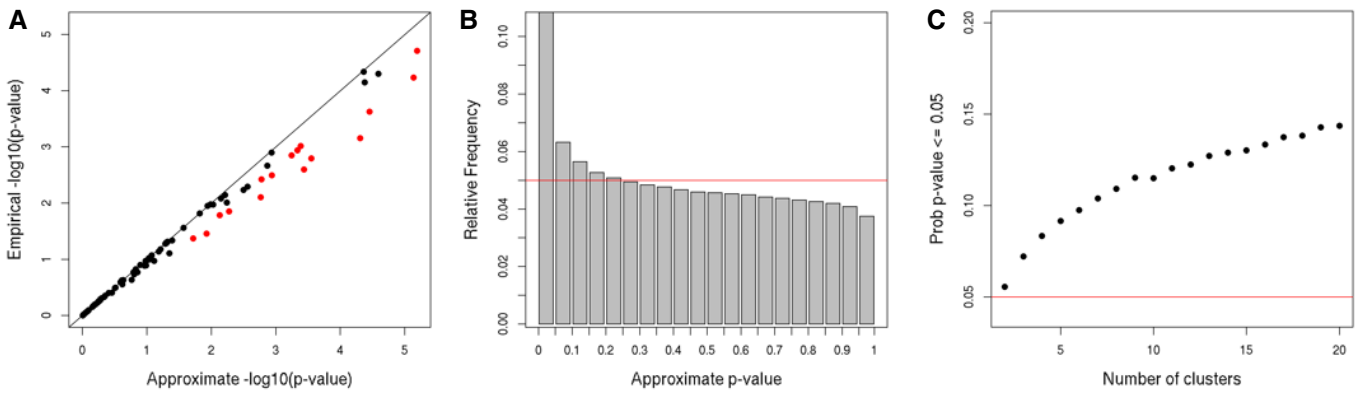


Figure 1. Asymptotic *P*-values (APs) compared to *P*-values based on permutation tests (EPs).
(A) APs and EPs for clustering solutions of nine algorithms over ten cancer datasets. Red dots: 2AP < EP. MCCA's solution on KIRC is omitted. Confidence intervals for the EPs are small such that they are contained in the dots. (B) Distribution of APs across permutations of MCCA's solution on KIRC dataset. The red line represents the expected theoretical distribution. (C) The probability to observe AP ≤ 0.05 in random clustering solutions with different number of clusters on the breast TCGA dataset (see text).

clustering and used it to cluster breast cancer data. The authors reported a *P*-value of 0.038. As the version of log-rank used was not specified, and the clustering solution was not provided, we could not calculate the EP. Instead, we permuted the group labels a large number of times, and for each permutation computed the AP of the conditional log-rank, which is the more appropriate version to use in this scenario (see Appendix). For 13.5% of the permutations, the computed AP was ≤ 0.038, which shows that reporting the AP is not sufficient in this case to show a clustering solution's merit. Hence, erroneous significance conclusions due to the use of AP occur both in biomedical research and in algorithm development. Overstatement of significance is likely even more common.

The difference between asymptotic and exact tests is not unique to the log-rank test. Rather, it is important for all statistical tests that rely on asymptotics, when sample size is small. In the log-rank test, inaccuracy is not affected only by the sample size, but also by the number of events within each group, and by imbalance in the group sizes. In some other statistical tests, there is higher community awareness of inaccuracies. For example, the R implementation of the chi-square test for independence issues warnings when used with small sample sizes. Such awareness should be raised for all asymptotic statistical tests.

Aside from the inaccuracy caused by using the asymptotic test, there are additional factors that one should consider when using the log-rank test. The null hypothesis for the test with multiple groups is that the survival function is the same for all groups. The test will therefore reject the null hypothesis even in cases where only a single group differs from the others. Another factor to consider is that the test has low power when the different survival functions cross one another. Analysis of differential survival for a clustering solution should therefore be accompanied by visualizing the Kaplan–Meier curve, and not by solely reporting the log-rank *P*-value, whether it is asymptotic or exact.

The log-rank test is widely used to compare survival of different patient groups and to assess disease subtyping. It is perhaps the leading evaluation criterion that guides development of new computational methods for clustering patients. For large datasets with many events in each group, the asymptotic log-rank test computes an accurate *P*-value. However, our results show that *P*-values based on the chi-square approximation are highly inaccurate in evaluating clustering solutions of popular methods on real cancer datasets. It is therefore essential that future analyses compute and report *P*-values using exact tests.

Data and software availability

TCGA data after preprocessing for all cancer types are available here: http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. Code to reproduce the analyses presented in this paper, and our implementation of the permutation-based log-rank test for more than two groups, are available in GitHub: https://github.com/Shamir-Lab/Logrank-Inaccuracies/tree/master.

Expanded View for this article is available online.

Acknowledgements
We thank Malka Gorfine for helpful discussion. The results published here are based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. The study was supported in part by the United States—Israel Binational Science Foundation (BSF) and the United States National Science Foundation (NSF), by the Israel Science Foundation (Grant 1339/18) and by the Israel Cancer Association—donation of Avraham Rotstein. N.R. was supported in part by a PhD fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University.

Author contributions
NR and RS conceived the project and wrote the manuscript. NR performed the analysis. RS supervised the project.

References

Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W (2018) Stegle O (2018) Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 14: 6

Gabasova E, Reid J, Wernisch L (2017) Clusternomics: integrative context-dependent clustering for heterogeneous datasets. *PLoS Comput Biol* 13: e1005781

Heinze G, Gnant M, Schemper M (2003) Exact log-rank tests for unequal follow-up. *Biometrics* 59: 1151–1157

© 2019 The Authors

17444292, 2019, 8, Downloaded from https://www.embopress.org/doi/10.15252/msb.20188754 by Weizmann Institute Of Science, Wiley Online Library on [19/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hosmer DW, Lemeshow S, May S (2008) *Applied survival analysis: regression modeling of time-to-event data*. New York, NY: Wiley-Interscience

Joachim RB, Altschuler GM, Hutchinson JN, Wong HR, Hide WA, Kobzik L (2018) The relative resistance of children to sepsis mortality: from pathways to drug candidates. *Mol Syst Biol* 14: 5

Prasad V, Fojo T, Brada M (2016) Precision oncology: origins, optimism, and potential. *Lancet Oncol* 17: e81 – e86

Rappoport N, Shamir R (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* 47: 1044

The Cancer Genome Atlas Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455: 1061 – 1068

Wang R, Lagakos SW, Gray RJ (2010) Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. *Biostatistics* 11: 676 – 692

Witten DM, Tibshirani RJ (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol* 8: Article 28

License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

17444292, 2019, 8, Downloaded from https://www.embopress.org/doi/10.15252/msb.20188754 by Weizmann Institute Of Science, Wiley Online Library on [19/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Chapter 5

MONET: Multi-omic module discovery by omic selection

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

MONET: Multi-omic module discovery by omic selection

Nimrod Rappoport, Roy Safra, Ron Shamir*

The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

* rshamir@tau.ac.il

Abstract

Recent advances in experimental biology allow creation of datasets where several genome-wide data types (called omics) are measured per sample. Integrative analysis of multi-omic datasets in general, and clustering of samples in such datasets specifically, can improve our understanding of biological processes and discover different disease subtypes. In this work we present MONET (Multi Omic clustering by Non-Exhaustive Types), which presents a unique approach to multi-omic clustering. MONET discovers modules of similar samples, such that each module is allowed to have a clustering structure for only a subset of the omics. This approach differs from most existent multi-omic clustering algorithms, which assume a common structure across all omics, and from several recent algorithms that model distinct cluster structures. We tested MONET extensively on simulated data, on an image dataset, and on ten multi-omic cancer datasets from TCGA. Our analysis shows that MONET compares favorably with other multi-omic clustering methods. We demonstrate MONET's biological and clinical relevance by analyzing its results for Ovarian Serous Cystadenocarcinoma. We also show that MONET is robust to missing data, can cluster genes in multi-omic dataset, and reveal modules of cell types in single-cell multi-omic data. Our work shows that MONET is a valuable tool that can provide complementary results to those provided by existent algorithms for multi-omic analysis.

Check for updates

OPEN ACCESS

Citation: Rappoport N, Safra R, Shamir R (2020) MONET: Multi-omic module discovery by omic selection. PLoS Comput Biol 16(9): e1008182. https://doi.org/10.1371/journal.pcbi.1008182

Editor: Teresa M. Przytycka, National Center for Biotechnology Information (NCBI), UNITED STATES

Received: April 28, 2020

Accepted: July 22, 2020

Published: September 15, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008182

Copyright: © 2020 Rappoport et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Digit dataset is available at: https://archive.ics.uci.edu/ml/machine-learning-databases/mfeat/. scNMT data are available at: https://github.com/BIRSBiointegration/Hackathon/tree/master/scNMT-seq. TCGA Breast

This is a *PLOS Computational Biology* Methods paper.

Introduction

Modern experimental methods can measure a myriad of genome-wide molecular parameters for a biological sample. Each type of such parameters is called "omic" and is measured by a different method. Analysis of omic data improved our understanding of biological processes and human disease, and is now used in therapeutic decisions [1]. While each experiment usually measures only one omic, several experiments can be performed on the same biological sample, resulting in multi-omic datasets. Large consortia such as TCGA and ICGC collected multi-omic data from tens of thousands of tumors [2,3]. Analysis of these data can further improve our understanding of cancer biology and suggest novel treatments.

cancer microarray data are available at: http://firebrowse.org/?cohort=BRCA&download_dialog=true. All other TCGA data are available at: http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html.

Funding: Study was supported in part by the Israel Science Foundation (grant 1339/18 and grant 3165/19 within the Israel Precision Medicine Partnership program), German-Israeli Project DFG RE 4193/1-1. NR was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics, Tel Aviv University, and by the Planning and Budgeting Committee (PBC) fellowship for excellent PhD students in Data Sciences. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Many algorithms have been developed in recent years to analyze multi-omic data, and most prominently, to detect subtypes of cancer, a task termed multi-omic clustering [4,5]. The vast majority of multi-omic clustering algorithms assume that a *common underlying structure* exists across all omics, and use all omic datasets to reveal this structure. Among the algorithms developed under this assumption are SNF and NEMO [6,7], as well as matrix factorization based methods such as MOFA+ [8], iClusterBayes [9] and MultiNMF [10]. However, this assumption does not always hold. For example, expression and mutation data do not seem to share the same structure. Even more closely related omics, such as expression and methylation, differ. This is demonstrated by the low agreement in clustering solutions that are produced based on different omics [11,12], and was also shown in a number of recent papers [13,14]. Moreover, in a recent benchmark we performed, we observed that solutions based on single omics can sometimes be more clinically relevant than solutions based on multiple omics [5]. Algorithms that can cluster patients while *accounting for the disagreement between omics* are therefore required.

Several recent methods addressed the distinct structure in different omics by using Bayesian statistics and modeling the different omics and their correlations. Savage et al. performed clustering on two omics, while allowing samples to be *fused* or *unfused* [15]. A fused sample belongs to a cluster spanning both omics, while unfused samples can belong to different clusters in the two omics. PSDF extended this framework to support feature selection [16].

MDI supports more than two omics [17]. Each omic has its own clustering, but clusters in different omics match each other. The probability that a sample will belong to matching clusters in two different omics has a prior that is higher the more these two omics are similar. In TWL [14], as in MDI, each omic also has its own clustering, and clusters in different omics match each other. A prior is placed such that samples are more likely to belong to the same cluster in different omics. BCC assumes a model with a global clustering and a clustering for each omic separately, and the global clustering serves as a Bayesian prior for each omic-specific clustering [18]. Clusternomics represents the global clustering as a Cartesian product of the omic-specific clusters, and can also map several such clusters into the same global cluster [13]. These methods have several limitations. MDI and TWL include only omic specific clusters, without providing a global clustering solution, and leave it to the user to choose between multiple clustering solutions. MDI, TWL and BCC further require that clusters in different omics match each other. Clusternomics' approach of representing global clusters as a Cartesian product of omic-specific clusters is less suited to find signals that are weak but consistent across many omics, and results in a high number of clusters. All methods except PSDF require a sample to belong to a coherent cluster in each of the omics, and PSDF is limited to only two omics. Furthermore, all available methods are based on Bayesian statistics, which requires explicit modeling of each omic, and is slow to optimize.

Here we present MONET (Multi Omic clustering by Non-Exhaustive Types), an algorithm for detection of patient modules for multi-omic cancer data. MONET uses ideas from MATISSE [19], an algorithm to detect gene modules, and generalizes its algorithmic approach to multi-omic data. In MONET's unique approach to multi-omic clustering, the goal is to form patient *modules*, such that each module can use only a *subset* of the omics. Thus, MONET can find patient modules with a common structure across some omics, and disregard other omics in that module, allowing different omic subsets for different modules. Note that this differs from ignoring an omic altogether, because an omic that is not used for one patient module can be used for other modules. MONET's solution allows outlier patients, who do not belong to any module.

We show that MONET finds biologically and clinically relevant patient modules in several datasets, giving results that compare favorably to those obtained from existent multi-omic

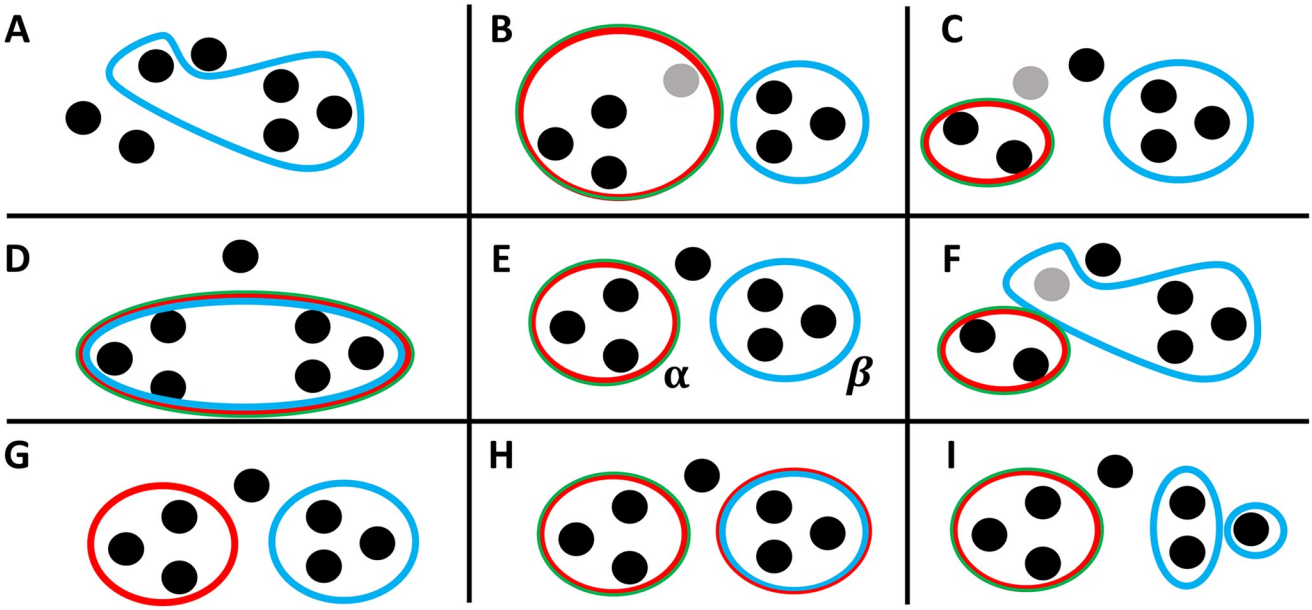


Fig 1. Actions performed by MONET when detecting heavy modules. Dots represent samples, and enclosing circles represent modules. The colors of the enclosing circle represent the omics covered by the module. Panel E shows the current state–two modules, where the left module (α) is covered by two omics and the right module (β) by one. An additional sample is lonely, i.e., does not belong to any module. Each other panel shows one action. B: the grey sample is added to module α. C: the grey sample is removed from module α. F: the grey sample moves into module β. I: module β is split. H: an omic is added to module β. G: an omic is removed from module α. D: modules α and β are merged. A: module α is discarded. In the shown case one of its samples is added to module β, and the other two become lonely. Actions for splitting module with omic or by adding omic are not shown.

https://doi.org/10.1371/journal.pcbi.1008182.g001

clustering methods. Furthermore, we show that MONET is useful for other biomedical tasks, as it successfully finds modules of genes, and of cells in single-cell data.

Methods

Overview

The input to MONET is a set of L omic matrices. Matrix l has n samples and p_l features. The output is a set of modules, where each module is a subset of the samples. Modules are disjoint, and not all samples necessarily belong to a module. Samples not belonging to a module are called *lonely*. Each module M is characterized by its samples, denoted *samples*(M), and by a set of omics that it covers, denoted *omics*(M). Intuitively, *samples*(M) are similar to one another in *omics*(M).

MONET works in two phases. It first constructs an edge-weighted graph per omic, such that nodes are samples and weights correspond to the similarity between samples in that omic. In the second phase, it detects modules by looking for heavy subgraphs common to multiple omic graphs.

Omic graphs

MONET constructs a graph G_l for each omic l separately. G_l is a full graph on n nodes. Denote by $sim_l(u, v)$ some similarity measure between samples u and v in omic l. The weight assigned to edge (u, v) in omic l, denoted by $w_l(u, v)$, is given by a function of the similarity between these two samples which we term "weighting scheme". This function is denoted f:

$$w_l(u, v) = \mathrm{f}(sim_l(u, v))$$

The weight of a module is defined as:

$$weight(M) = \Sigma_{l \in omics(M)} \Sigma_{u,v \in samples(M)} w_l(u, v)$$

The optimization problem

MONET's objective function is to find a disjoint set of modules M_1,M_2. . . maximizing $\Sigma_i weight(M_i)$.

Importantly, we require that the weighting scheme returns values that are both positive and negative. High positive values indicate that the two samples are similar and should belong to the same module for omic l, while low negative values indicate the converse. A module with a positive weight therefore contains samples that are on average highly similar in the omics covered by the module. If all edge weights are positive, modules will always improve their scores by adding more samples and omics. Note that we present MONET here as a combinatorial optimization problem, but for some weighting schemes, the weight of each edge has a probabilistic interpretation. In such cases, the weight of a module is interpreted as the score for a log-likelihood ratio test for whether $samples(M)$ form a module on $omics(M)$, under the simplifying assumption that modules and sample pairs are independent. More details on this probabilistic formulation are in the appendix.

To construct the omics graphs, any weighting scheme can be used. The scheme we used here is as follows. We first apply NEMO [7], a multi-omic clustering algorithm we recently developed, to each omic separately R times, each time on randomly selected 80% of the samples. We set $c_l^r(u, v)$ to 1 if samples u and v clustered together in the r'th run on omic l, and to 0 otherwise. Denote by $avg(c_l^r)$ the average value of the c_l^r matrix, and by $R(u, v)$ the set of NEMO executions in which both u and v were sampled. We set $w_l(u, v) = mean_{r \in R(u,v)}(c_l^r(u, v) - avg(c_l^r)) - C$. The constant C controls the balance between modules that cover one omic (higher C value) and modules that cover multiple omics (lower C value). Here we used $C = 0.2$ and $R = 100$. For the classification experiments we used a different weighting scheme, which is based on a Gaussian mixture model. Its full details are in the appendix.

Heavy module detection

Given all the omic graphs, MONET now detects modules with high weight by maximizing the objective function $\Sigma_M weight(M)$. There is no constraint on the number of modules, or an upper bound on module sizes, so the weighting scheme must create both positive and negative edges, otherwise the trivial optimal solution is a single module containing all patients and covering all omics. The problem of detecting heavy subgraphs in this setting is NP-hard even for the case of a single graph [19]. We therefore developed an iterative greedy heuristic for detecting heavy modules. The algorithm is initialized with a set of modules termed seeds. After seed finding, at every iteration MONET considers several possible actions, described below, that can increase the objective function. It then performs an action that provides the greatest improvement.

■ Seed finding: Seeds are found iteratively. The first seed is determined by constructing a graph where edge weights are the sum of the edge weights in all individual omics, randomly selecting a first sample, and constructing a module containing all omics, which contains the first sample and its k neighbors with highest positive edge weights. All samples that were assigned to a module are removed from the graph, and the next seed module is sought. The procedure ends once S seeds were found. In this work we used $S = 15$ seeds for all datasets, and $k = floor\left(\frac{n}{15}\right)$.

■ Optimization actions: Once a set of seeds is found, MONET improves the modules iteratively in a greedy manner. In each iteration, a module M' is selected at random, and MONET calculates the gain in the objective function from a set of possible actions concerning the module. It then chooses the action with maximal gain. It stops when no action provides a gain in any module. The actions considered are (see Fig 1):

- Add a sample to M'. All lonely samples are considered. Since we observed that this action is commonly chosen in initial iterations when S and k are both small, we allowed up to 10 (or $\frac{n}{50}$ if $n>1000$) samples to be added in a single action, to reduce the number of iterations.

- Remove a sample from M'.

- Move sample from module M' to another module, or move a sample from another module to M'. All possible samples and modules are considered. Similarly to adding samples, we allow up to 10 (or $\frac{n}{50}$ if $n>1000$) sample switches in a single action.

- Add an additional omic to a module. All omics are considered.

- Remove an omic from a module. All the covered omics of the module are considered.

- Merge modules M' and M''. The set of samples for the new module is $samples(M')\cup samples(M'')$. The omics for the new module are one of the following: 1. $omics(M')\cup omics(M'')$ 2. $omics(M')\cap omics(M'')$ 3. $omics(M')$ 4. $omics(M'')$. All four options are considered.

- Split M' into two modules. For this action, a graph is constructed with nodes $samples(M')$, and where the weight of the edge between u and v is $\Sigma_{l\in omics(M')}w_l(u, v)$. In this graph we find a heavy subgraph M'', and create two modules, M'' and $M\backslash M''$. The omics of both modules are $omics(M')$.

- Discard M'. Each sample u in M' is moved to the module M'' with the highest sum of weights from u to M'' using $omics(M'')$. If all these sums are negative, u is made lonely.

- Create a new module using all lonely samples. MONET finds a heavy subgraph in each omic separately, and a module is created from the heaviest subgraph found.

- Split M' by adding an omic. For every omic $l\notin omics(M')$, MONET looks at the subgraph induced by $samples(M')$ on G_l, denoted $G_l[samples(M')]$, and detects in it a heavy subgraph. Denote the nodes of the heavy subgraph by U. We then split M' into two modules. In one module the nodes are U, and the omics are $omics(M')\cup\{l\}$. In the second module the nodes are $samples(M')\backslash U$ and the omics are $omics(M')$.

- Split M' with an omic. As in the previous action, a heavy subgraph with nodes U is found in $G_l[samples(M')]$, but here for every $l\in omics(M')$. Two modules are constructed. In one the nodes are $samples(M')\backslash U$ and omics are $omics(M')$. In the other samples are U and the only omic is l that produced the heavy subgraph.

MONET uses a parameter η for the minimum module size. Actions that reduce the number of samples below η are not executed, and module splits are considered under this restriction. Here we used $\eta = \max\left(round\left(\frac{n}{30}\right), 10\right)$.

To find a heavy subgraph in a graph, we use a heuristic based on Charikar's 2-approximation to the problem of maximum density subgraph [20]. We iteratively find the node with lowest (weighted) degree and remove it from the graph, until no node is left. We then choose the heaviest of the sequence of subgraphs obtained during this process. The complexity of the heuristic on an n-node weighted full graph is $O(n^2)$.

The MONET algorithm is guaranteed to converge to a local maximum, because the sum of weights within all modules is increasing in each iteration. The algorithm stops when no action on any module improves the objective.

In each iteration, all actions that do not involve finding heavy subgraphs consider each edge in each of the omic graphs a constant number of times. The complexity of all these actions is therefore $O(\Sigma_l(n+|E_l|))$, where E_l is the number of edges in G_l. The complexity of splitting a

module and of creating a new module involves finding a heavy subgraph and is thus $O(\Sigma_l(n+|E_l|)+n^2)$. For the last two actions, for the same reason, the same complexity is needed for each omic considered for the split, and the overall complexity is $O(L(\Sigma_l(n+|E_l|)+n^2))$, which is therefore the overall complexity of each iteration. For full graphs, this gives a worst case complexity of $O(L^2n^2)$. The space complexity is $O(Ln^2)$.

In a post-processing step we perform empirical significance testing to filter modules. Given a module, we sample 500 modules of the same size and omics, and only keep the module if its weight is in the highest 1%. In practice we only performed the testing for modules of minimal size ($\eta = 10$ here), as we never found larger non-significant modules. Samples that do not belong to any module after filtering are marked as lonely.

Since the algorithm for finding heavy modules is only guaranteed to converge to a local maximum, the algorithm is repeated multiple times, and the best solution is returned. Unless otherwise specified, we used 15 repeats for the analyses performed in this work.

Additional MONET features

■ Partial datasets: MONET can handle datasets where only a subset of the omics were measured for some samples. Such samples are added to all omic graphs, but in omics where these samples were not measured their nodes have no edges. This way, omics in which no data were measured for a sample do not affect the decision of assigning the sample to a module.

■ Sample classification after clustering: Once modules were calculated from the data, MONET can naturally classify new samples into modules. For each module M, MONET calculates the gain in $weight(M)$ from adding the new sample u to M: $\Sigma_{v\in samples(M)}w_l(u, v)$, and classifies the sample to the module with maximal gain. If the gain is always negative, the sample is not classified to any module. This computation takes $O(nL)$ given that the edge weights were already calculated.

Testing methodology

We applied MONET and several other algorithms to simulated, image and cancer datasets that are described later. Here we outline the way we evaluated the results.

Clustering assessment. To assess a clustering solution where the true clustering of the data is known, we used the Adjusted Rand Index (ARI) [21]. Note that the ARI can compare solutions with different number of clusters and different cluster sizes. On cancer datasets from TCGA we performed survival analysis to assess the distinction in survival between the different groups of samples, and tested enrichment of known clinical parameters. For the survival analysis we used a permutation-based approach to perform the log-rank test, since the widely used asymptotical version of this test tends to overstate significance, and specifically for TCGA data [22–24]. We also used permutation testing to assess the enrichment of clinical parameters [5]. The clinical parameters we considered were gender, age at diagnosis, pathological stage and pathologic M, N and T. In addition we considered known subtype definitions—PAM50 for breast cancer [25] and the French-American-British classification (FAB) for AML [26].

Partial datasets experiments. For cancer datasets, we sampled 40% of the patients, partitioned them into three equal groups, and removed every group from one of the omics. For the image dataset we removed 20% of the samples in each omic independently. We then applied MONET to the data and calculated ARI with MONET's solution on all data. We repeated this experiment 10 times.

Classification experiments. to perform experiments on a dataset we first applied MONET to it. Denote MONET's solution by Sol_{all}. We then partitioned the samples in the dataset into 10 equal folds. For every fold i, we applied MONET to all samples except those in

the fold, and denote the solution by Sol_i. We define the *stability* of the fold to be $ARI(Sol_{all},Sol_i)$ where the ARI is computed using only samples that appear in both Sol_{all} and Sol_i. We then classified the held out samples to the modules from Sol_i, and denote the solution after classification by $\hat{Sol_i}$. We define the *Rand Index following classification (RFC)* of the fold to be $ARI(Sol_{all}, \hat{Sol_i})$, where the ARI is now measured across all samples. For datasets where the ground truth is known we also measured $ARI(ground\_truth,Sol_i)$, and $ARI(ground\_truth, \hat{Sol_i})$, and term them the *pre-classification accuracy* (preCA) and *post-classification accuracy* (postCA) respectively.

Simulations. The simulations are described in the appendix.

Ovarian cancer analysis. To check the clustering solution for enrichment of clinical parameters we used chi-squared test for discrete features (e.g. tumor stage) and Kruskal-Wallis for numeric ones (e.g. age). We also used chi-square to test for enrichment of mutations and used Benjamini-Hochberg to correct for multiple hypotheses. To find genes and miRNAs that are highly expressed in a module, we performed a one sided t-test (with $\alpha = 0.05$) comparing the expression level in the module and the rest of the samples (after log normalization) and corrected for multiple hypotheses with Benjamini-Hochberg. Survival analysis was performed as described for the other TCGA datasets. To determine differential survival while controlling for the age and stage, we fitted a Cox multivariate proportional hazard model.

Results

Simulated datasets

We first performed two simulations to test MONET's approach to multi-omics clustering. In the first, we simulated 300 samples from five equal-size modules in two omics. Module 1 covers only the first omic, module 2 only the second omic, and modules 3–5 cover both omics (**Fig A in S1 Appendix**). We added five outlier samples that do not belong to any module. MONET correctly identified the modules (ARI = 0.92) and their corresponding omics (**Fig B in S1 Appendix**). In another experiment, we simulated 150 samples from five modules in three omics (**Fig C in S1 Appendix**). Module 1 covers all omics. Modules 2–5 cover all omics, are indistinguishable in omic 1, but belong to different clusters in omics 2 and 3. Only a small number of features separate the modules in omic 2, so the signal in omic 2 is weak. When presented with only omics 1 and 2, MONET identified module 1 but chose to treat modules 2–5 as one module that only covers the first omic (**Fig D in S1 Appendix**). When faced with omic 3 as well, the ARI equaled 1, and MONET identified these samples as coming from different modules that cover all omics (except for one module whose samples were very different in omic 2, which does not cover that omic) (**Fig E in S1 Appendix**). These simulations highlight MONET's approach to multi-omic integration, where sample modules can cover only a subset of the omics, based on the strength of the clustering structure in these omics. Full details on the simulations are in the appendix.

Digits dataset

We next tested MONET in a dataset where the ground truth is known. The dataset [27] contains six types of features ("omics") of 2000 images of the handwritten digits 0–9. For most tests, we used 400 images. See additional details in the appendix.

We applied MONET and seven other methods to the data. We chose BCC, MDI, Clusternomics and TWL, which model disagreement between omics. We also chose SNF and NEMO to represent general multi-omic clustering methods. SNF is widely used, and we recently showed NEMO's high performance [7]. We also included MOFA+ [8], a widely used multi-

omic dimension reduction method. While MOFA was not developed specifically to cluster samples, its low dimensional representation can be used to cluster samples. Each method clustered the data into 10 groups. Note that MONET cannot get as input the number of modules, so we instead shifted the edge weights of the omic graphs to encourage about 10 modules (see details in the appendix). Fig 2A shows that MONET outperformed the other methods that model omic disagreement, and was comparable to SNF and NEMO. When ignoring lonely samples, MONET was slightly better than SNF and NEMO. Several modules found by MONET covered only a subset of the omics, suggesting a different structure in different omics (**Fig E in S1 Appendix**). Methods modeling omic disagreement were much slower than SNF, NEMO and MONET, which required a few seconds or minutes (Fig 2B).

In order to test MONET's scalability to thousands of samples, we also executed MONET on all 2000 images in the dataset. MONET took almost six hours to run, compared to less than ten minutes on 400 images. This was mainly due to increased runtime per iteration, but also because more iterations were required for convergence (**Fig G in S1 Appendix**). The performance was largely unchanged, with the ARI decreasing from 0.79 to 0.78.

Cancer datasets

We next executed the same eight methods on real cancer datasets from TCGA, each containing three omics: mRNA expression, DNA methylation and miRNA expression. We used ten cancer types: Acute Myeloid Leukemia (AML), Breast Invasive Carcinoma (BIC), Colon Adenocarcinoma, Glioblastoma Multiforme (GBM), Kidney Renal Clear Cell Carcinoma (KRCCC), Liver Hepatocellular Carcinoma, Lung Squamous Cell Carcinoma (LUSC), Skin Cutaneous Melanoma, Ovarian serous cystadenocarcinoma and Sarcoma. Dataset sizes ranged from 170 to 621 patients. Full details on the datasets are available in our recent benchmark [5]. We used differential survival between clusters as an assessment criterion for the quality of a clustering solution (see Methods). MONET's modules for all cancer datasets are available in S1 Supporting Data.

As we can see in Fig 2C, MONET and NEMO had the highest number of cancer types with significantly different survival (at significance level 0.05), with 6 such types. MDI came next with 5, and the other methods had 3–4. Remarkably, in our recent benchmark, eight other multi-omic clustering methods, including the factorization-based methods iClusterBayes and MultiNMF, achieved significance for at most five cancer types. NEMO and MONET were also the best performers in terms of the number of subtypes with enriched clinical parameters (Fig 2D). The cancer types for which MONET and NEMO obtained a significant difference in survival were not identical. While both had different survival in AML, GBM, liver hepatocellular carcinoma and Sarcoma, NEMO found differential survival in BIC and melanoma, and MONET in KRCCC and ovarian cancer. Such a difference was also evident in the clinical parameters: NEMO found an enrichment in melanoma, while MONET found in LUSC. These results suggest that NEMO and MONET can be used complementarily. In terms of runtime, SNF and NEMO required seconds per dataset, MONET and MOFA+ a few minutes, and the remaining methods were an order of magnitude slower (Fig 2E).

The number of clusters chosen varied considerably among algorithms (**Fig H in S1 Appendix**). SNF had a mean of 2.8, TWL 3.4, NEMO, MONET and BCC 4–5, MOFA+ 5.7, MDI 8.9 and Clusternomics 26.5. The high numbers of MDI and Clusternomics are possibly due to attempting to model clustering in each individual omic. The log-rank p-value, number of enriched clinical labels, running time and number of clusters for each method and dataset are presented in **Tables A-D in S1 Appendix**.

MONET discovered modules that use different combinations of omics (Fig 2F). Most of the modules were based on only a single omic, and for several cancer types all modules covered

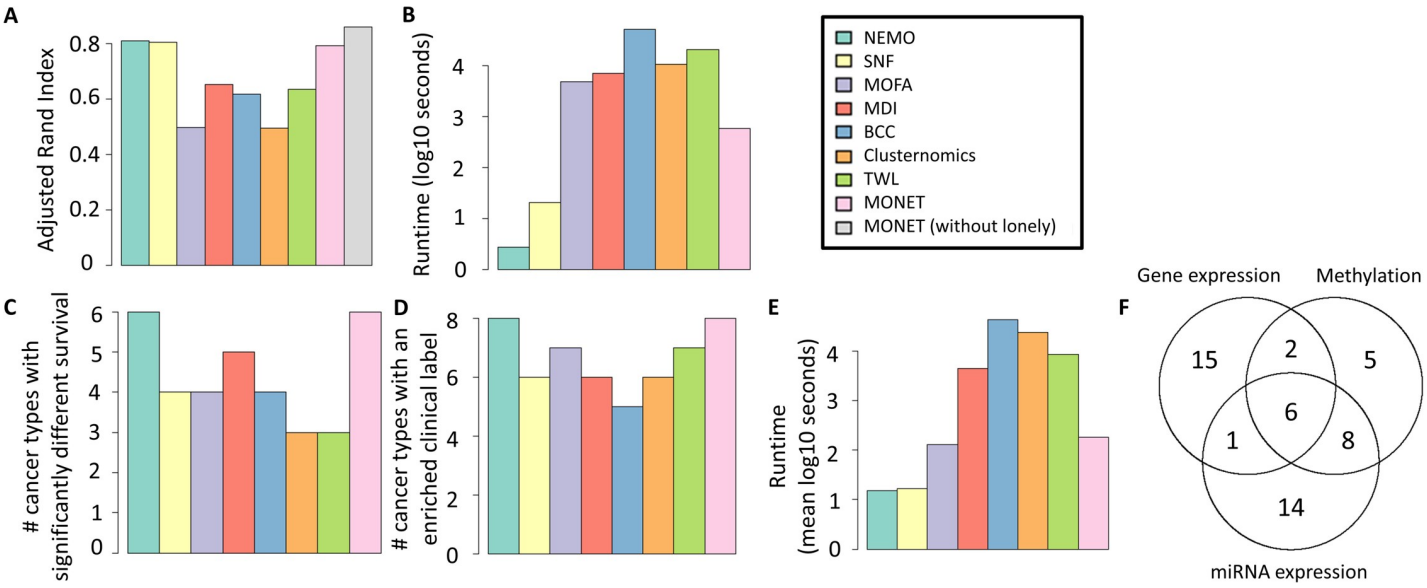


Fig 2. Performance results. A-B: Digits dataset. A: ARI of methods for multi-omic clustering. B: Run time. C-F: Results on ten TCGA cancer datasets. C: Number of cancer subtypes for which each method found a clustering with statistically different survival. D: Number of cancer subtypes for which each method found a clustering with an enrichment of a known clinical label. E: Run time. F: Number of MONET modules that cover each subset of omics.

https://doi.org/10.1371/journal.pcbi.1008182.g002

only one omic. For some cancer types, this omic was the same for all modules, signifying a strong clustering structure in that omic. In none of the cancer types the solution contained only modules that covered all omics. These results suggest that different omics may have different structures, and that MONET reveals such differences. MONET also reported several (between 0 and 14) lonely samples per cancer (**Fig I in S1 Appendix**).

Since MONET is only guaranteed to converge to a local optimum, we experimented with using different numbers of restarts. In addition to the above results, which used 15 restarts, we also executed MONET with 1 and 50 restarts. For both 1 and 50 restarts, 6 cancer datasets were significantly associated with survival. The number of datasets with enriched clinical labels was 8 for one restart, and increased from 8 to 9 for 50 restarts, suggesting that MONET may benefit from more iterations. However, the clustering results of different restarts were generally similar to one another (**Fig J in S1 Appendix**).

Additional analysis of the cancer results

We examined in more detail the clustering solution of MONET on the 287-patient ovarian cancer dataset. MONET found four modules in this dataset, with sizes 22, 63, 77 and 115, named M1-M4, and identified 10 samples as outliers (see **Fig K in S1 Appendix** for the feature heatmaps). While SNF and MDI seek to integrate structure across all omics (Fig 3A), MONET chooses the omics covered by each module. In its solution all modules cover the gene expression omic, and M1 also covers miRNA expression (Fig 3B). To assess the clinical relevance of MONET's modules, we examined the distribution of different clinical parameters across the modules. The modules showed significant differential survival (p = 0.036, Fig 3C), with M2 showing significantly better survival than the others (p = 4e-3). The modules showed differential survival even after correcting for age at diagnosis and clinical stage (p = 2e-4 using a Cox proportional hazards model). None of the other clustering algorithms found a solution with a significant difference in survival (Fig 3D). The modules were not significantly dependent of the clinical stage (0.056, chi-square test, 0.08 for Kruskal-Wallis), and they were enriched for

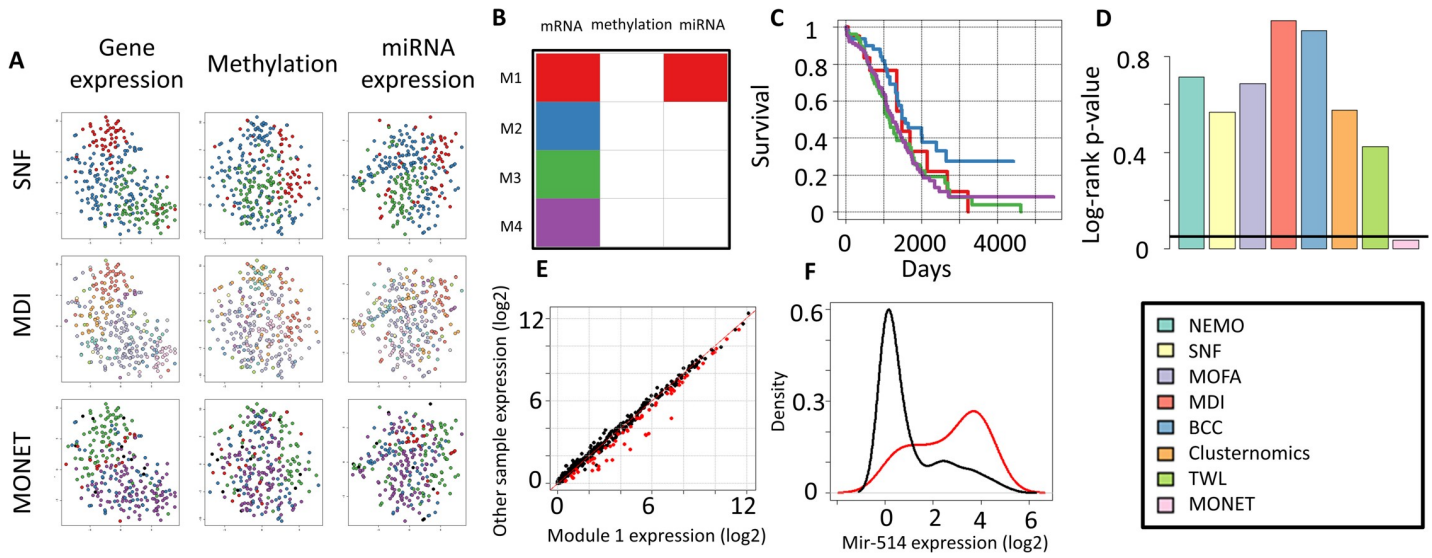


Fig 3. Analysis of ovarian cancer. A. t-sne [32] visualization of the solutions obtained by SNF, MDI and MONET on the data. Samples are colored by their assigned module. In MONET's panels, lonely samples are black. B. Omics covered by each MONET module. Columns are omics and rows are modules. C. Kaplan-Meier plot for the different MONET modules. D. p-value of the log-rank test for the clustering solutions of different methods. E. Comparison of miRNA expression for samples in MONET's Module 1 (x axis) and other samples (y axis). Genes that are significantly highly expressed in Module 1 are colored in red. F. Distribution of mir-514 expression in samples in Module 1 (red) and in other samples (black).

https://doi.org/10.1371/journal.pcbi.1008182.g003

venous invasion status (8e-4, chi-square test, **Table E in S1 Appendix**) and for age at initial diagnosis (p = 7e-3 by Kruskal-Wallis, **Fig L in S1 Appendix**). No module was enriched for any mutation from a list of known driver mutations reported in TCGA's analysis of ovarian cancer [11] (see **Table F in S1 Appendix**).

We next characterized each module in more detail using clinical parameters and GO enrichment analysis (performed with Gorilla [28]) of differentially expressed genes with high module expression (see Methods). M1 had younger patients (p = 0.02, Wilcoxon test). It was the only module that included the miRNA omic. We found 21 miRNAs that were highly differentially expressed in M1's patients (**Fig 3E, Table G in S1 Appendix**), including mir-514, which was far higher on samples in M1 compared to all other samples (**Fig 3F**). It was recently reported to regulate proliferation and cisplatin chemoresistance in ovarian cancer [29]. M2 had significantly better survival, and its highly expressed genes were enriched for immune response. M3 was characterized by older samples (p = 4e-3, Wilcoxon test) without venous invasion (p = 2e-4, chi-square), and upregulation of genes involved in microtubule-based process (e.g. TUBB2B, TUBB4A). Finally, samples in M4 were enriched for venous invasion (p = 0.02, chi-square) and high expression of immune response and extracellular matrix organization related genes (e.g. MMP9 and multiple collagen subunits).

To understand the differences between M2 and M4, we found genes differentially expressed between them. M4 had higher expression of genes related to cell adhesion (e.g. collagen subunits), extracellular matrix (ECM) organization, and regulation of developmental process (e.g. WNT7A, WNT7B). Both the extracellular matrix and WNT signaling were previously reported to regulate ovarian cancer progression [30,31], and may explain the difference in venous invasion and survival between the modules. The high expression of ECM proteins may link M4 with the previously reported Mesenchymal subtype [11].

We also executed NEMO and MONET on each individual omic in the ovarian cancer data. MONET found a significant separation in survival for each omic individually (p-value 0.04–

0.05 in all omics), while NEMO did not find such separation for any. This shows MONET's effectiveness as a single-omic clustering approach (in this setting it is very similar to Matisse).

We observed that in several cases MONET used omics that were especially relevant for a specific dataset. For example, MONET's solution on GBM used only methylation in all modules. We executed spectral clustering and NEMO on each GBM omic separately and both algorithms found a solution with significant difference in survival only for the methylation dataset (p-value < 0.001 in both cases). Note however that MONET's solution often uses multiple omics (see Fig 2F for all cancer datasets and **Figs M-N in S1 Appendix** for the solutions on BIC and Sarcoma).

One of the main advantages of Bayesian methods is that they associate a posterior probability for each sample to belong to each cluster. MONET also provides a quantitative measure for the association between a sample and a module: the sum of weights between the sample and all the module's samples across all omics covered by the module. A similar association score can also be calculated for each omic separately (see **Fig O in S1 Appendix** for the scores for the ovarian cancer dataset). These scores can assist in better understanding of the data, on top of the binary module memberships. For example, **Fig O in S1 Appendix** shows that M1 has a weak structure in both the omics it covers, while the three other modules differ greatly in gene expression. The score also suggests that M3 samples have some similarity in methylation, as is also suggested by **Fig K in S1 Appendix**, though this level of similarity is not sufficient for M3 to cover the methylation omic. These observations appear consistent with the t-SNE plot for the data (Fig 3A).

Partial datasets

Often in multi-omic datasets, some samples have measurements for only a subset of the omics. Such datasets are called *partial*. MONET can address such datasets by assigning edge weight 0 to samples in the omics that were not measured. We tested this ability using the Sarcoma dataset, which had modules covering all omics, and using the digits dataset. In each dataset we randomly removed samples from some omics (see Methods), applied MONET, and compared its solution to the solution using all samples, and to the ground truth in case of the digits dataset. The results are presented in Fig 4A and Fig 4B.

MONET's output on the digits dataset was quite robust, with only a slight deterioration in performance. The Sarcoma results were stable as well, but the stability highly varied between the omics from which samples were removed. Samples removed from the gene expression omic had lower ARI compared to samples removed from other omics, possibly indicating that MONET's solution is highly affected by that omic for that dataset. The ARI slightly differed for samples in the digits dataset as well depending on the omic from which they were removed (**Fig P in S1 Appendix**). These results suggest that MONET can be robustly applied to partial datasets.

Classification

Given a clustering solution, MONET supports classification of new samples into modules (see Methods). We tested MONET's robustness and classification on the Sarcoma and digits datasets. For each dataset we performed an unsupervised version of 10-fold cross validation. We define the *stability* of a fold as the ARI between MONET's solution on all samples and MONET's solution for the current fold (which excludes 10% of the samples). We define the *Rand Index following classification* (RFC) of a fold as the ARI between MONET's solution on all samples and its solution on the fold following the classification of the 10% held out samples (see Methods). For the digits dataset, we also compared the result of every fold to the ground truth,

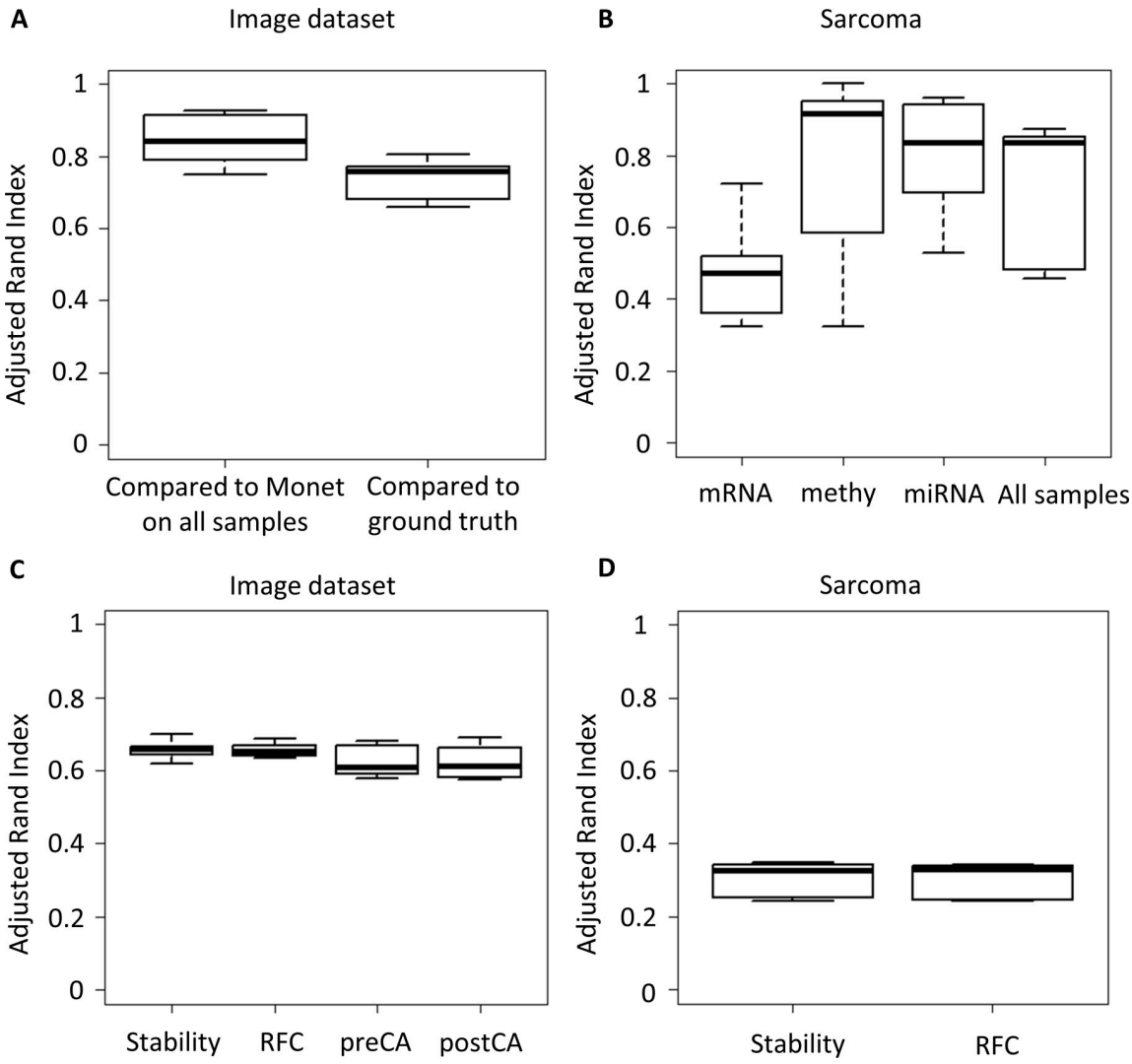


Fig 4. Performance of MONET on partial datasets and in classification. A. ARI on a partial version of the digits dataset compared to its solution on the full dataset and to the ground truth. B. ARI on a partial version of the Sarcoma dataset compared to its solution on all samples. Shown is the ARI for samples that were dropped from each one of the omics (three left boxplots), and for all the samples in the dataset (rightmost boxplot). C. Performance in classification experiments on the digits dataset. See Methods for the assessment criteria. D. Performance in classification experiments on the Sarcoma dataset. All boxplots are distributions over 10 random runs.

https://doi.org/10.1371/journal.pcbi.1008182.g004

with and without the 10% of held out samples, and term them the *pre-classification accuracy* (preCA) and *post-classification accuracy* (postCA). Note that we used here the Gaussian mixture weighting scheme (which is described in the appendix), as in order to perform classification MONET calculates the edge weights for the new samples.

The results are presented in Fig 4C and Fig 4D. In the runs on the digits dataset, both the stability and RFC are high. 45 (11%) of the images were not classified to a module, as no module with positive classification score was found for them. In the runs on the Sarcoma dataset the results are only moderately stable, but the RFC is as high as the stability. This suggests that the classification is accurate, and that decrease in performance stems largely from the different clustering structure that is obtained from sampling the datasets. All samples were classified in this dataset. Overall, these results show that MONET's framework can be used to perform classification given new samples.

Other biological tasks: Gene and single cell clustering

We next tested MONET on additional biological tasks. We used MONET to cluster 1532 genes measured by both RNA-seq and microarrays of the BIC TCGA dataset that exhibited high variance in both these omics. We used BIC because of its large sample size, and to demonstrate MONET's utility for in-depth analysis on an additional cancer type. MONET reported five main gene modules (**Fig 5A**, **Fig Q in S1 Appendix**). We used Gorilla [28] to perform enrichment analysis for these gene modules. Reassuringly, we found enrichment of biological processes that vary across breast cancer patients in several modules, including "mitotic cell cycle process", "immune system process", and "extracellular matrix organization". As expected, all gene modules covered both omics.

Finally, we applied MONET to single-cell data. Argelaguet et al. recently developed scNMT, a method that measured gene expression, DNA methylation and DNA accessibility at single cell resolution, and applied it to mouse embryos at embryonic days 4.5–7.5 [33]. We applied MONET to the gene expression and promoter methylation data of 619 single cells (**Fig 5B** and **5C**). The modules obtained were highly enriched for specific cell types and embryonic days of development (**Tables H-J in S1 Appendix**). Several modules, across different cell types and stages of development, covered both omics, reflecting the widespread changes in expression and methylation during the onset of gastrulation [34,35]. Other modules used only gene expression, suggesting an overall stronger distinction between cell types at the expression level. One module covered only DNA methylation. This module comprised cells from different cell

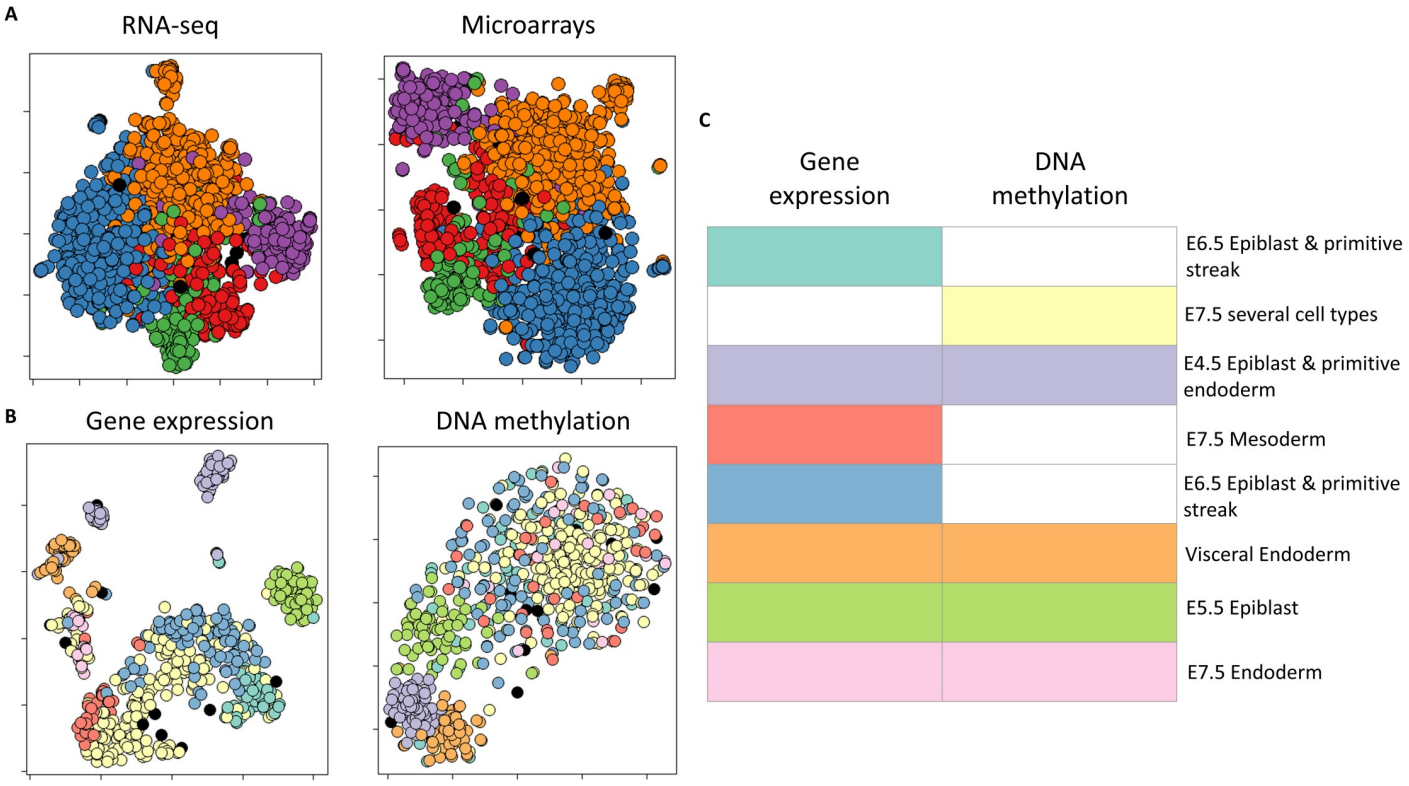


Fig 5. Using MONET to cluster genes and single cells. A. Gene clustering. t-sne visualization of MONET's gene modules on the BIC dataset. Genes are colored by MONET's output. Lonely samples are colored in black. B-C. Single cell clustering based on gene expression and DNA methylation of promoters, using the scNMT mouse embryonic development dataset. B. Like A, for MONET's solution on the dataset. C. Module omics identified by MONET. Rows represent modules and columns correspond to omics. Colored panels indicate that the module covers the omic.

https://doi.org/10.1371/journal.pcbi.1008182.g005

types at E7.5, including cells from all germ layers, again highlighting that while the transcriptional signatures of different cell types differ at that stage, the promoter methylation profile of the different germ layers is still quite similar [33]. Overall, these results demonstrate that MONET can be applied and lead to insights in diverse biological scenarios.

Discussion

We presented MONET, a novel multi-omic clustering algorithm. MONET can identify modules with structures present in some of the omics, without imposing these structures on other omics. MONET can also identify samples that do not fit any detected module. State-of-the-art methods that seek clusters across all omics often perform quite well, and structures that span all omics have been observed in many studies. We view these approaches as complementary to MONET, and suggest using both for multi-omic analysis. That is, data analysis can benefit from using both MONET as well as other algorithms that seek a common structure, and each of these approaches will reveal different aspects of the data.

It is challenging to interpret omics data and its clustering in the face of disagreement between omics. From a data analysis point of view, as we noted before, one can use different tools for the analysis. Methods that assume agreement between omics can be used, together with different formulations for omic disagreement: omic-specific clusters, omic-specific deviations from a global clustering solution, or clusters that apply in only a subset of the omics. From a biological point of view, a different structure between omics can reveal insight on biological regulation and disease. For example, for biological regulation, it is interesting to discover gene modules that are co-expressed but are not highly correlated on the protein level. As another example, in disease, the GBM G-CIMP subtype is associated with IDH mutations and a characteristic methylation phenotype, while its expression profile does not define the subtype as distinctly [36].

The edge weighting in MONET's omic graphs can be done by schemes tailored to the omic and data, allowing flexibility in the analysis. The weighting schemes used here to cluster patients, genes, and single-cells show MONET's ability in different biomedical domains. The weighting scheme can also shift the balance between modules with single or multiple omics, or place more emphasis on one particular omic.

Most multi-omic analysis methods assume that samples are present in all omics. This is rarely the case in datasets available today, such as TCGA. It is also likely that partial datasets will be prevalent in single-cell analysis, where measuring multiple omics from a cell is just beginning and is experimentally challenging. MONET's ability to analyze partial datasets will make it valuable in this setting.

MONET has several limitations. Using different weighting schemes allows flexibility, but it can be challenging to choose one that balances finding omic-specific signals and signals reinforced by different omics. The optimization problem MONET solves is NP-hard, so the algorithm is heuristic. Adding new actions to MONET's heavy subgraph algorithm can improve its output. While MONET is faster than methods modeling disagreement between omics and can easily be run on today's datasets, which contain hundreds of samples, it is currently not scalable to more than a few thousand samples. Future work can improve MONET's runtime, for example by removing edges in the omic graphs, or by discretizing the edge weights, which allows a more efficient implementation of Charikar's algorithm. The potential of MONET for classification warrants further validation in the cancer context. Finally, as MONET does not model the features in the dataset, understanding the molecular differences between modules requires additional analysis.

Code availability

Code for MONET and for reproducing all results in this paper is in Github: https://github.com/Shamir-Lab/MONET.

Supporting information

S1 Appendix. Additional implementation details, and supporting figures and tables.
(DOCX)

S1 Supporting data MONET's clustering results on the TCGA and scNMT datasets.
(ZIP)

Acknowledgments

The results published here are based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. The contribution of N.R. is part of Ph.D. thesis research conducted at Tel Aviv University.

Author Contributions

Conceptualization: Nimrod Rappoport, Ron Shamir.

Formal analysis: Nimrod Rappoport, Roy Safra.

Funding acquisition: Ron Shamir.

Methodology: Nimrod Rappoport, Roy Safra.

Project administration: Ron Shamir.

Software: Roy Safra.

Supervision: Ron Shamir.

Visualization: Nimrod Rappoport, Roy Safra.

Writing – original draft: Nimrod Rappoport, Ron Shamir.

Writing – review & editing: Nimrod Rappoport, Ron Shamir.

References

1. Prasad V, Fojo T, Brada M. Precision oncology: origins, optimism, and potential. Lancet Oncol. 2016; 17: e81–e86. https://doi.org/10.1016/S1470-2045(15)00620-8 PMID: 26868357
2. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, M. Mastrogianakis G, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455: 1061–1068. https://doi.org/10.1038/nature07385 PMID: 18772890
3. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database (Oxford). 2011; 2011: bar026. https://doi.org/10.1093/database/bar026 PMID: 21930502
4. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet. 2017; 8: 84. https://doi.org/10.3389/fgene.2017.00084 PMID: 28670325
5. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018; 46: 10546–10562. https://doi.org/10.1093/nar/gky889 PMID: 30295871
6. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014; 11: 333–337. https://doi.org/10.1038/nmeth.2810 PMID: 24464287

7. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. Schwartz R, editor. Bioinformatics. 2019; 35: 3348–3356. https://doi.org/10.1093/bioinformatics/btz058 PMID: 30698637
8. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020; 21: 111. https://doi.org/10.1186/s13059-020-02015-1 PMID: 32393329
9. Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics. 2017; 19: 71–86. https://doi.org/10.1093/biostatistics/kxx017 PMID: 28541380
10. Liu J, Wang C, Gao J, Han J. Multi-View Clustering via Joint Nonnegative Matrix Factorization. Proceedings of the 2013 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2013. pp. 252–260. https://doi.org/10.1137/1.9781611972832.28
11. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474: 609–615. https://doi.org/10.1038/nature10166 PMID: 21720365
12. Netanely D, Avraham A, Ben-Baruch A, Evron E, Shamir R. Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. Breast Cancer Res. 2016; 18: 74. https://doi.org/10.1186/s13058-016-0724-2 PMID: 27386846
13. Gabasova E, Reid J, Wernisch L. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. Morris Q, editor. PLOS Comput Biol. 2017; 13: e1005781. https://doi.org/10.1371/journal.pcbi.1005781 PMID: 29036190
14. Swanson DM, Lien T, Bergholtz H, Sørlie T, Frigessi A. A Bayesian two-way latent structure model for genomic data integration reveals few pan-genomic cluster subtypes in a breast cancer cohort. Bioinformatics. 2019; 35: 4886–4897. https://doi.org/10.1093/bioinformatics/btz381 PMID: 31077301
15. Savage RS, Ghahramani Z, Griffin JE, de la Cruz BJ, Wild DL. Discovering transcriptional modules by Bayesian data integration. Bioinformatics. 2010; 26: i158–i167. https://doi.org/10.1093/bioinformatics/btq210 PMID: 20529901
16. Yuan Y, Savage RS, Markowetz F. Patient-Specific Data Fusion Defines Prognostic Cancer Subtypes. Markel S, editor. PLoS Comput Biol. 2011; 7: e1002227. https://doi.org/10.1371/journal.pcbi.1002227 PMID: 22028636
17. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012; 28: 3290–3297. https://doi.org/10.1093/bioinformatics/bts595 PMID: 23047558
18. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013; 29: 2610–2616. https://doi.org/10.1093/bioinformatics/btt425 PMID: 23990412
19. Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol. 2007; 1: 8. https://doi.org/10.1186/1752-0509-1-8 PMID: 17408515
20. Charikar M. Greedy Approximation Algorithms for Finding Dense Components in a Graph. Lecture Notes in Computer Science. 2000. pp. 84–95. https://doi.org/10.1007/3-540-44436-X_10
21. Hubert L, Arabie P. Comparing partitions. J Classif. 1985; 2: 193–218. https://doi.org/10.1007/BF01908075
22. Heinze G, Gnant M, Schemper M. Exact log-rank tests for unequal follow-up. Biometrics. 2003; 59: 1151–7. https://doi.org/10.1111/j.0006-341x.2003.00132.x PMID: 14969496
23. Vandin F, Papoutsaki A, Raphael BJ, Upfal E. Accurate Computation of Survival Statistics in Genome-Wide Studies. Boutros PC, editor. PLOS Comput Biol. 2015; 11: e1004071. https://doi.org/10.1371/journal.pcbi.1004071 PMID: 25950620
24. Rappoport N, Shamir R. Inaccuracy of the log-rank approximation in cancer data analysis. Mol Syst Biol. 2019; 15. https://doi.org/10.15252/msb.20188754 PMID: 31464374
25. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol. 2009; 27: 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370 PMID: 19204204
26. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. Br J Haematol. 1976; 33: 451–8. Available: http://www.ncbi.nlm.nih.gov/pubmed/188440 https://doi.org/10.1111/j.1365-2141.1976.tb03563.x PMID: 188440
27. Van Breukelen M, Duin RPW, Tax DMJ, Den Hartog JE. Handwritten digit recognition by combined classifiers. Kybernetika. 1998; 34: 381–386. Available: https://dml.cz/dmlcz/135219
28. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009; 10: 48. https://doi.org/10.1186/1471-2105-10-48 PMID: 19192299

29. Xiao S, Zhang M, Liu C, Wang D. MiR-514 attenuates proliferation and increases chemoresistance by targeting ATP binding cassette subfamily in ovarian cancer. Mol Genet Genomics. 2018; 293: 1159–1167. https://doi.org/10.1007/s00438-018-1447-0 PMID: 29752546

30. Yoshioka S, King ML, Ran S, Okuda H, MacLean JA, McAsey ME, et al. WNT7A regulates tumor growth and progression in ovarian cancer through the WNT/β-catenin pathway. Mol Cancer Res. 2012; 10: 469–82. https://doi.org/10.1158/1541-7786.MCR-11-0177 PMID: 22232518

31. Cho A, Howell VM, Colvin EK. The Extracellular Matrix in Epithelial Ovarian Cancer—A Piece of a Puzzle. Front Oncol. 2015; 5: 245. https://doi.org/10.3389/fonc.2015.00245 PMID: 26579497

32. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008; 9: 2579–2605. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

33. Argelaguet R, Clark SJ, Mohammed H, Stapel LC, Krueger C, Kapourani C-A, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature. 2019; 576: 487–491. https://doi.org/10.1038/s41586-019-1825-8 PMID: 31827285

34. Mohammed H, Hernando-Herraez I, Savino A, Scialdone A, Macaulay I, Mulas C, et al. Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. Cell Rep. 2017; 20: 1215–1228. https://doi.org/10.1016/j.celrep.2017.07.009 PMID: 28768204

35. Smith ZD, Meissner A. DNA methylation: Roles in mammalian development. Nature Reviews Genetics. 2013. pp. 204–220. https://doi.org/10.1038/nrg3354 PMID: 23400093

36. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010; 17: 510–22. https://doi.org/10.1016/j.ccr.2010.03.017 PMID: 20399149

Chapter 6

Single cell Hi-C identifies plastic chromosome conformations underlying the gastrulation enhancer landscape

Single cell Hi-C identifies plastic chromosome conformations underlying the gastrulation enhancer landscape

Nimrod Rappoport*[1,2], Elad Chomsky*[1], Takashi Nagano[3, 4], Charlie Seibert[5], Yaniv Lubling[1], Yael Baran[1], Aviezer Lifshitz[1], Wing Leung[3, 4], Zohar Mukamel[1], Ron Shamir[2], Peter Fraser[4, 5], Amos Tanay[1]

1 Department of Computer Science and Department of Biological Regulation, Weizmann Institute of Science, Rehovot, Israel.

2 The Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel.

3 Laboratory for Nuclear Dynamics, Institute for Protein Research, Osaka University, Osaka, Japan.

4 Nuclear Dynamics Programme, The Babraham Institute, Cambridge, UK.

5 Department of Biological Science, Florida State University, Tallahassee, FL, USA.

* Equal contribution

Corresponding authors:

Amos Tanay amos.tanay@weizmann.ac.il,

Peter Fraser pfraser@bio.fsu.edu

ABSTRACT

Embryonic development involves massive proliferation and differentiation of cell lineages. This must be supported by chromosome replication and epigenetic reprogramming, but how proliferation and cell fate acquisition are balanced in this process is not well understood. Here we use single cell Hi-C to map chromosomal conformations in post-gastrulation mouse embryo cells and study their distributions and correlations with matching embryonic transcriptional atlases. We find that embryonic chromosomes show a remarkably strong cell cycle signature. Despite that, replication timing, chromosome compartment structure, topological associated domains (TADs) and promoter-enhancer contacts are shown to be variable between distinct epigenetic states. About 10% of the nuclei are identified as primitive erythrocytes, showing exceptionally compact and organized compartment structure. The remaining cells are broadly associated with ectoderm and mesoderm identities, showing only mild differentiation of TADs and compartment structures, but more specific localized contacts in hundreds of ectoderm and mesoderm promoter-enhancer pairs. The data suggest that while fully committed embryonic lineages can rapidly acquire specific chromosomal conformations, most embryonic cells are showing plastic signatures driven by complex and intermixed enhancer landscapes.

INTRODUCTION

The organization of mammalian chromosomes[1] must accommodate physical nuclear packaging constraints alongside three major sources of dynamics – transcription[2], replication[3] and differentiation[4–6]. Recent advances in microscopy[7], and different conformation capture technologies[8] have provided improved understanding of the way chromosomes fold in general, leading to models for organization at multiple scales; from chromosomal territories and interchromosomal spaces[9], through active and inactive (also known as A and B) intra-chromosomal compartments, and cohesin/CTCF mediated loop structures[10]. These models explain observations on the distribution of chromosomal contacts and domain insulation that give rise to topological associated domains (TADs)[11–14]. Moreover, parallel advances in mapping the dynamics of genome replication show a high degree of linkage between chromosomal compartments, TADs, and genome replication time control[15,16], highlighting genome replication as a key driver of the linkage between chromosomal structures and cellular proliferation. Quantification of the mitosis and replication cycle in chromosomes using synchronized cells[17,18] and single cell Hi-C[19,20] was used to combine the effects of mitotic compaction and genome replication into one model describing the effect of cellular proliferation on chromosomal structure. Overall, current data indicate that chromosomes are continuously being remodeled in all phases of the cell cycle - during exit from the mitotic state (M-G1 phase), while replicating (S-phase), and when re-entering the mitotic state (G2-M phase).

A cycling dynamics of chromosome structure is therefore unavoidable for proliferating cell populations. This dynamics can be challenging if cells should combine proliferation with the acquisition of stable transcriptional and epigenetic identities. A classical model for a process that must balance remarkable proliferation with rapid differentiation is embryonic development. Recent advances in single cell RNA-seq have provided unbiased and detailed maps of the earliest stages of transcriptional sorting during embryo gastrulation[21]. These data confirmed and refined the classical observations on the emergence of an epiblast cell population and its rapid diversification into the three germ-layers by embryonic day 7.5. It also showed that diversification within the germ-layer is rapid and almost immediate, including early expansion of embryonic blood and several distinct mesodermal lineages, the differentiation of basic ectodermal neuronal progenitors, and the emergence of endodermal precursors from primitive precursors and convergent extra-embryonic endoderm lineages[22]. Since these dramatic transcriptional events are occurring while cells are dividing at maximal rates (at least every 8 hours on average), the chromosomal structure underlying them must simultaneously support replication and cell-fate acquisition. But it is currently not understood if and when

chromosome conformation/structure in embryonic lineages differentiates and stabilizes. It is unclear if cell-type specific chromosomal structures that were observed in-vitro[23,24] or in mature tissues[25,26] emerge before cells establish transcriptional identities, during (and in direct correlation with) transcriptional sorting, or only several cell cycles after cells commit to their fate transcriptionally.

Here we use single cell Hi-C to explore the chromosomal organization of post-gastrulation embryonic cells. We developed algorithms that combine analysis of replication time traces with contact distributions to enable de-novo clustering of single cells in the embryo while minimizing bias by cell cycle signatures. This leads to two main observations on the timing and structure of the initial cell type specific chromosomal structures in the embryo. First, we discover that a highly distinct chromosomal conformation is characterizing primitive erythrocytes, showing that in principle, conformation can be specified and stabilized rapidly in differentiated cell types in the embryo. In contrast to this effect, most of the embryo nuclei show much milder conformational heterogeneity that is associated primarily with broad clustering into mesoderm and ectoderm architectures. We show that the overall conformations of single cell Hi-C clusters representing the mesoderm and ectoderm layers are remarkably similar at the level of compartments and TADs. Nevertheless, we show that promoter-enhancer contacts that link ecto- or mesoderm specific promoter activity with specific enhancer markup are enriched for differential long-range contacts. Further analysis suggests that enhancers that are specific to diverse gastrulation lineages are interleaved within one group of TADs, while enhancers that are more accessible in the pluripotent epiblast state are demarcated from these genomic domains in a second group of TADs. Together the data suggest that while committed embryonic lineages may acquire specific chromosomal conformations rapidly, the majority of the embryonic lineages in gastrulation share a common and possibly more plastic chromosomal structure.

RESULTS

Cell cycle signatures dominate embryonic chromosome conformations

We applied single-cell Hi-C to assay chromosomal conformation in three E9.5 C57BL/6J mouse embryos. We processed 3456 embryonic cells, out of which 87.15% passed quality control (QC) **(Supplementary Fig 1A-J**). We sequenced at a depth that allowed recovery of a median of 91K contacts per nucleus, with an overall low rate of trans-chromosomal contacts (median 7.85%), demonstrating high library quality (**Fig 1A**). Across all cells, we captured 310M contacts, with 8% trans-chromosomal contacts. We initially phased nuclei along the cell cycle using our previously reported strategy[19], observing high degree of similarity between the

parameters of the cell cycle model originally inferred for mouse embryonic stem cells (mESCs) and the embryonic cells. For example, we observed that 6.2% of the nuclei are enriched (20% or more) for contacts in genomic distances ranging between 2-12 Mb (**Fig 1B**), defining a canonical mitotic cycle as previously observed for mESCs. Ordering embryo nuclei based on their distribution of contact distances as in Nagano et al.[19] (**Supplementary Fig 1K**) recapitulated the cell cycle dynamics involving transition between a G1 conformation landscape defined by long range (>12Mb) contacts and the S-phase regime involving gradual increase in short range (<2Mb) contacts. To allow robust comparison of the replication time trends between ESC and Embryos we identified genomic regions that are constitutively replicating early or late in S-phase according to both datasets (defined as "strict early" and "strict late", Methods, **Supplementary Fig 2A-B**). Analysis of the ratio between Hi-C coverage in these genomic regions in embryo cells showed partial consistency with the trend observed in ESC, where we observed increase in the ratio through mid-S phase and decrease toward G2 (**Fig 1C**). Interestingly, in the embryo this trend was perturbed by a population of nuclei with high early/late coverage ratios and atypically low fraction of short-range contacts in cells that were initially annotated as G1. These data reinforced our earlier observations on the dominance of cell cycle signatures in scHi-C, but also suggested the canonical signature may be shadowing additional conformational heterogeneity within the embryonic nuclei pool.

Clustering scHi-C profiles using S-phase cluster seeding and RNA atlas projection.

To enable de-novo clustering of scHi-C profiles with reduced cell-cycle bias, we developed a two-stage approach (denoted *S-phase cluster seeding*). We seeded scHi-C clusters using analysis of replication time trends in mid-S phase cells and expanded these seeds to clusters using A-compartment association scores (A-scores, Methods). We applied this approach to a combined data set of ESC and embryo cells (**Supplementary Fig 2C-H**), deriving a model defined by three main clusters, one involving a distinct group of embryo nuclei with non-canonical cell cycle phasing (C3, **Supplementary Fig 2I**), and the other two representing clustering of the remaining embryo (C2) and ES (C1) nuclei. As expected, M-phase nuclei were poorly separated into clusters, but otherwise G1-S cell cycle variation was captured as intra-cluster structure (**Fig 1D**).

To annotate nuclei clusters and explore their underlying gene regulatory programs, we acquired and sequenced single cell RNA from two E9.0 embryos and from ESCs using MARS-seq and created a map of transcriptional states using Metacell[27] (**Supplementary Fig 3A-B**). We identified differentially expressed genes and genomic bins encompassing them for each expression metacell and projected scHi-C clusters on the transcriptional maps by calculating relative A-scores on these genomic bins. Remarkably, this strategy associated unambiguously

C3 conformations with primitive erythrocyte (pEry) expression (**Supplementary Fig 3C**), but showed that the remaining transcriptional landscape in the embryo could not be matched by strong conformation clusters within C2. This was further confirmed by re-analysis of a reference gastrulation scRNA-seq atlas (E6.5-8.25, **Supplementary Fig 3D-E**). Overall, despite the rich transcriptional embryonic space, C2 nuclei were reflecting variation that was approximately similar in extent to the transcriptionally homogeneous ESC states represented in the C1 cluster and only primitive erythrocytes stood out as a distinct conformation cluster.

Differential contacts in pluripotent and embryonic nuclei

Many genomic bins showed average transcriptional change in non-pEry embryo cells compared to ESCs (806 and 1289 bins with over 4-fold decrease and increase respectively, **Fig 1E**). Global comparison of *A-scores* in the ESC (C1) and embryo (C2) clusters (**Fig 1F**) showed however conservation of the A/B compartment structure, with 85% of the genomic bins showing less than 0.1 change in A-score, and only 0.2% showing over 0.3 change. Analysis of A-score in loci stratified according to expression levels (**Fig 1G**, excluding bins with differential expression) suggested a clear distinction between A-linked expressed and B-linked non-expressed loci. Further analysis also indicated that bins containing genes over expressed in the ESC or embryo will have a higher A-score in that sample (**Fig 1H**, KS D=0.4, $p<<0.01$). This association was observed based on relatively small changes in A-score and despite the lack of loci showing major A-B compartment switches. We next compared embryo and ESC estimated replication time per genomic bin (defined as the early-score, methods). This suggested a similar trend of expression linkage (**Fig 1I-K,** KS=0.36, $p<<0.01$). Together these data show that despite the mild magnitude of compartment and replication-time remodeling in embryos compared to ESCs, it still reflects transcriptional regulation in these cells.

We next searched for localized differential chromatin contacts in ESCs and embryo cells by pooling contacts from single cell clusters and performing Shaman[28,29] normalization and enrichment analysis. Using a threshold of differential enrichment score of 50, we identified 3267 pairs of loci losing contacts and 1914 pairs of loci gaining contacts in embryos compared to ESCs, suggesting many cases of local conformation remodeling. We observed that genomic bins with higher A score in ESCs are involved in significantly more differential contacts than loci with constitutively high A-score or loci gaining A-score in the embryo (**Supplementary Fig 4A**, two-sided Kolmogorov–Smirnov test). A screen for differential contacts at ESC-regulated TSSs highlighted cases of conformation changes with potential regulatory impact (**Supplementary Data 1**). For example, we observed specific contacts and insulation structure isolating the pluripotency genes *Rex1/Zfp42*, *Tet2* and *Dppa2/4* from

surrounding B-compartment associated regions in ESC nuclei (**Fig 1L**, see **Supplementary Fig 4B** for conformation changes in loci conserving their A-association**).** These data suggested that specific contacts, with possible linkage to gene regulation and in particular to the repression of the pluripotency program, are observed in embryonic nuclei. This is occurring even when global structural features such as compartment, replication and insulation (**Supplementary Fig 4C**) are changing only mildly.

Primitive Erythrocyte chromosomes show compact and highly organized folding.

In contrast to the weak separation of scHi-C clusters C1 and C2, the pEry cluster C3 was defined by a well separated group of 264 single cells (reclassified using total A-scores per cell, **Supplementary Fig 4D**). This separation was supported by a large number of genomic bins with modified A-score in pEry compared to other embryo cells (**Fig 2A**, 4.8% with A-score delta > 0.3). We estimated mean expression in pEry and non-pEry E9 embryo metacells (**Fig 2B**) and noted that in pEry, genomic bins bearing expressed genes at any level show remarkable alignment to the A-compartment (**Fig 2C**). Estimation of single cell early/late coverage ratios (**Fig 2D**), showed that cells classified as pEry are enriched in S-phase, but are also represented in other phases. Genome replication landscapes (quantified by early-scores) were more conserved than A-scores (**Fig 2E**), but genomic bins containing expressed genes showed earlier replication, in concordance with their increased A-score (**Fig 2F**). Beyond its unique compartment structure, the pEry single cell cluster was also characterized by uniformly high fractions (30-60%) of contacts over > 2Mb (**Fig 2G)**. The data also showed high variance for pEry long range contact distances, with no distance bin representing over 6% of the contacts (**Fig 2H**). This property distinguishes the long-range contacts in pEry maps from those observed in embryonic or ESC G1 cells during exit from mitosis. Despite the higher rate of long range intra-chromosomal contacts, pEry nuclei show low rates of trans-chromosomal contacts (**Fig 2I**). Together these observations indicate pEry chromosomes form compact and highly organized territory structures, with A/B compartments that are strongly demarcated and reflective of transcriptional activity patterns.

pEry funnel-like A-compartment structures are anchored at TSSs and cryptic loci

To understand further the sharp increase in pEry A-compartment association specificity, we identified 357 loci with the highest increase in pEry specific A-scores. Clustering of A-scores profiles over 400kb around such pEry A-specific loci showed that about 50% of the sites (**Fig 2J**, clusters A1-A3) involved sharp A-linked pEry hotspots that reside in the B compartment in non-pEry embryo cells. The remaining sites typically represented increase in A-score for a larger domain bounded by the identified pEry A-linked peak (clusters A4-A8). Projection of differential TSS expression on the clustered genomic interval confirmed that the majority

(92%) of pEry A-peaks were associated with an expressed TSS (**Fig 2K**). Interestingly it also suggested many hotspots of A-association could not be explained by any known localized transcriptional driver. We then computed the mean pEry and non-pEry contact enrichment patterns for the loci clusters (**Fig 2L**). In pErys, this revealed an unexpected trend involving a funnel-like structure representing aligned contacts around the focal A-compartment contact hotspot. Contact enrichment maps around the same loci in non-pEry nuclei showed these sites are located within embryo insulators and between two loop structures (**Fig 2L - right**). We visualized individual loci showing major funnel-like conformational remodeling around key genes (**Fig 2M**) and multi-peak loci (**Supplementary Fig 4E**), but also in hotspots that represented uncharacterized regulatory effects (**Fig 2M, bottom).** This suggested that strong A-compartment alignment in pEry is not driven solely by transcription, and must therefore also involve some other trans-acting factors (e.g., we observed enrichment for erythrocyte TF binding, **Supplementary Fig 4F-G**). For control, we clustered profiles of 272 loci with top non-pEry A-score increase, indicating lack of similar funnel-like effects in the embryo conformation cluster (**Supplementary Fig 4H**). To validate that the highly specific conformations in C3 nuclei are indeed representing primitive erythrocytes in a non-biased fashion, we sorted directly 118 primitive erythrocytes cells from E10.5 embryos and generated new single cell Hi-C profiles from them (**Supplementary Fig 5A-D**). The data confirmed that sorted pEry cells represent the same sharp A/B compartment structure as the one characterized in non-sorted cells and reconfirmed the presence of remarkable funnel-like structures in these cells (**Supplementary Fig 5E-F**). Of note, Guo et al. recently reported a similar funnel structure in thymocytes and B cells, which they termed “chromatin jets”, suggesting its prevalence in hematopoietic cells[30].

Refined embryo clustering by model-based analysis of replication dynamics

Since embryo transcriptional states are highly heterogeneous at E9, we made several attempts to enhance resolution within cluster C2, searching for conformation variation that can be linked with differentiating cell types on the background of massive proliferation signatures. Direct clustering of single cell coverage profiles in S-phase cells and UMAP visualization of these cells (**Fig 3A**) suggested cell-to-cell variation may be present in the data, but showed that it is superimposed over strong cell-cycle gradients, even when restricting analysis to replicating cells alone. We therefore developed a sensitive algorithm that considers both the replication cycle and the potential cell-type structure explicitly and quantitatively (**Supplementary Fig 6A-B**). The algorithm infers a probabilistic mixture model in which each cell is associated with a cluster and a latent replication timing variable defined as the *s-score*. Each cluster specifies the replication timing of each genomic bin, such that once a cell’s s-score is inferred, the algorithm can compare its observed bins read coverage to the values

predicted by a linear replication process that is timed in a bin-specific way (Methods). The algorithm tries to fit the observed data by clustering cells de-novo while simultaneously inferring their s-scores and the cluster-specific replication timing parameters. We used cross-validation to tune model parameters and verify the algorithm robustness (**Supplementary Fig 6C-I**). This resulted in good matching of observed and modeled replication regimes (**Supplementary Fig 6J**) for a model including 3 clusters denoted C2.1, C2.2 and C2.3. Importantly, the model's inferred s-scores facilitate normalization of the coverage statistics for each cell. UMAP projection of such normalized profiles show a clear, cell-cycle independent cluster structure (**Fig 3B**). Analysis of the observed cluster structure suggested C2.1 and C2.3 are distributed homogeneously along the replication cycle (**Fig 3C**). Cluster C2.2 showed skewed distribution enriched for late-S profiles and additional analysis indicated cells within the cluster are of lower coverage and potentially lower quality (**Supplementary Fig 7A**). We note that we could not derive robust results using alternative methods for clustering scHi-C data, which are based on differential compartment structure and are lacking explicit cell cycle modelling[31,32] (**Supplementary Fig 7B-C)**.

Ectoderm and mesoderm/endoderm scHi-C clusters

We estimated replication time per genomic bin (early-score) in the C2.1-3 clusters to facilitate their further annotation. These estimations were consistent for C2.1 and C2.3 (**Fig 3D**), but showed C2.2 cells are skewed to high and low coverage values (as expected by their bias to mid to late-S phase, **Fig 3E**). We identified groups of genomic bins with C2.1 or C2.3-specific early replication, showing that pooling coverage on these groups provided replication-time dependent separation of the clusters (**Fig 3F**). Moreover, mean A-score over the same genomic bin groups showed matching separation of single cells (i.e. early replicating loci in C2.1 were also more A-associated in C2.1 and conversely for C2.3, **Fig 3G**). This allowed expansion of our clustering to additional S phase cells (**Fig 3H**). A similar approach was not applicable to G1 cells (**Supplementary Fig 7D**). Overall this strategy yielded a total of 431 C2.1 cells and 504 C2.3 cells for further analysis. We searched for cluster-specific replication time in groups of loci representing correlated gene modules inferred from scRNA-seq data (**Fig 3I-K, Supplementary Fig 7E-G**). This unambiguously associated cells in cluster C2.1 with ectoderm gene expression programs and cells in C2.3 with mesoderm or endoderm programs. Cells in C2.2 were not associated with any gene expression program. The annotation of clusters C2.1 and C2.3 was supported by comparing their compartments to data from Neural progenitor cells and hematopoietic cells from E14.5 embryos (**Supplementary Fig 7H-J**)[24,33].

Lineage-specific scHi-C conformation differences are weak

To test the potential for detecting additional cell type structure given the limited breadth and depth of our scHI-C sample, we performed simulations with downsampled data. These experiments show that the data and algorithms are sufficiently sensitive to allow detection of clusters similar to C2.1 and C2.3 even when these involved as little as 50-75 cells (~5-10% of the modeled cells, **Supplementary Fig 8A-B**). This suggests that other possible chromosomal differences between cell types are weaker, or are present in scarcer cell populations. Further analysis suggested that even for the mesoderm and ectoderm clusters, contact landscapes could be remarkably similar, even around loci that support dramatic transcriptional regulation (e.g., *Igf2* or *Crabp2*, **Fig 3L**). Quantitatively, only 1% of the genome (divided into 40kb bins) show A-score different of 0.2 or more between the clusters, compared to 3.6% in a comparison of the E14.5 NPC and HSC maps (**Supplementary Fig 8C**).

Consistent with their overall similarity in conformation landscapes, further dissection of the mesoderm or ectoderm into cell types using our mixture model approach (**Supplementary Fig 8D-E**) was deriving only cell-cycle dependent refinements of the C2.1 and C2.3 clusters. To improve on this, we used inferred replication time parameters to normalize coverage profiles per cell in each cluster (**Supplementary Fig 9A-B**). Hierarchical clustering of the resulted data did identify an intra-mesoderm lineage structure, including a small cluster strongly matching the endothelial transcriptional state (**Supplementary Fig 9C-D**). It can therefore be hypothesized that replication time and compartment structure of refined embryonic lineages may be detected using sensitive algorithms and deeper single cell sampling. But the data strongly suggest that the magnitude of conformational changes between such refined lineages will remain small, in particular compared to the highly distinct pEry state we described above.

Three-way identification of regulated long-range interactions

Pooling contacts in ectoderm and mesoderm scHi-C clusters provided us with a strategy for identifying germ-layer specific chromatin interactions. We first identified 256 and 236 loci with higher ectoderm or mesoderm/endoderm A-score respectively. Annotation of these sites (**Supplementary Data 2**) revealed several important regulatory genes, for example the mesoderm TF *Twist1*, and the epiblast/ectoderm TF *Sox2* (**Fig 4A-B**). Comparative analysis of contact maps in these loci showed again a very high degree of consistency between the global conformation of the two clusters. Nevertheless, we could identify refined alteration in contact distributions of the promoter of *Twist1* with a putative regulatory element (shown by virtual 4C, **Fig 4B**). To generalize this observation, we used published histone modification maps from ectoderm (hind-, mid- and forebrain) and mesoderm (heart, limb) tissues, and identified cell type specific putative enhancers (Methods, **Supplementary Fig 10A-B**). We

also screened for identified genes with germ-layer specific expression, and combined them with the epigenomic maps by mapping each enhancer to its closest promoter. Proximal pairs of enhancers and promoters with matching ecto- or mesoendo- specific activity could then be defined (**Fig 4C**). To complete a three-way integrative screen on putatively interacting pairs, we next computed the contact enrichment in the C2.1 and C2.3 contact maps for each of the matching promoter-enhancer pairs. We observed significant contact enrichment in the ectoderm scHi-C cluster for ectoderm specific promoter-enhancer pairs, and mesoderm contact enrichment in the mesoderm pairs (**Fig 4D**). We also implemented a direct statistical test for contact frequency around putative ectoderm and mesoderm hotspots (**Supplementary Fig 10C-D**), which supported a similar observation. We note that the two methods differ in their normalization strategy and power, and their identified hits are only partly overlapping. When using Shaman comparison, we detected 173 and 338 promoter-enhancer pairing with 3-way support for ecto- and mesoderm regulatory activity respectively (**Supplementary Data 3**), including many examples linked with key regulators of cell type specific transcriptional programs (See examples in **Fig 4E**). These putative interactions should be interpreted carefully. First, while we believe comparisons using Shaman scores are more sensitive, these cannot be fully controlled statistically. Second, we note that only 1.5% of the highest intensity (Shaman score difference > 40) differential ectoderm-mesoderm contacts were annotated within one of our enhancer-promoter pairs, illustrating that the complex conformational landscape in these clusters involves many uncharacterized contacts despite showing only weak compartment and TAD differences.

Polycomb markup and ectoderm specific long-range contacts in the *Tbx*3-5 locus

Our 3-way analysis of regulated promoter-enhancer pairs suggested contact enrichment is positively linked with lineage-specific gene activation in most cases. It is however possible that contact enrichment will be associated with gene repression, as postulated previously for polycomb domains[34–36]. We therefore screened for ectoderm/mesoderm differential H3K27me3 loci (using hind-, mid- and forebrain / heart and limb) with proximal anti-correlated promoter expression pattern (**Supplementary Fig 10E-G**). This screen yielded several candidate locus pairs showing high H3K27me3 occupancy in correlation with proximal gene repression and low contact intensity (**Supplementary Data 4**), where most of these cases were of lower specificity than the positive interactions observed for activated genes. A reciprocal effect was detected in the *Tbx3-Tbx5* locus, where polycomb marks and gene repression were associated with increased rather than decreased contact intensity. This locus codes for two transcription factors with sophisticated transcriptional control, where *Tbx3* is expressed in the epiblast and most mesodermal tissues, and *Tbx5* is specific to pharyngeal mesoderm and cardiomyocytes (**Supplementary Fig 10H**). In the mesoderm cluster,

consistently with previous reports[37], we observed two TAD structures (contacts over L1 and L2, **Fig 4F, Supplementary Fig 10I**) physically separating the two TFs. In the ectoderm, however, the near-complete repression of both genes is correlated with the emergence of a new *Tbx3*-*Tbx5* contact (L3), and severe attenuation of the L1 contact. The internal structure at *Tbx5* (L2.1) is unperturbed. While we have not observed other repressive chromatin structures of similar intensity, this example suggests that de-novo establishment of chromatin interactions may be facilitated in the context of either the polycomb or some other uncharacterized repressive machinery.

Gastrulation cell-type specific accessibility hotspots are intertwined within TADs

We reasoned that the linkage between extensive transcriptional diversification in gastrulation and the rather rudimentary observed chromosomal conformation diversity must involve the chromosomal and genomic distribution of active regulatory elements and promoters. Using single-cell ATAC/RNA-seq multiomics data[38], we derived clusters of cell type specific chromosome accessibility peaks with specific distributions over the key gastrulating cell types (**Fig 5A**, **Supplementary Fig 10J**). We then tested the A-score distribution of the loci in each cluster of peaks. Comparing ESC and embryo A-scores (**Fig 5B**) we discovered stronger A-linkage in ESC for cluster 27, 37 and 38, which are enriched for accessibility in extraembryonic tissues and early gastrulation state (e.g. Epiblast). Comparing embryo and pEry A-scores (**Fig 5C**) showed strong pEry A-linkage in clusters 8, 9 and 5, which represent erythrocyte or combined hematoendothelial peak specificity. Importantly, the extent of A-association differential for pEry clusters was significantly higher than that observed between ESCs and embryo cells. Comparison of mesoderm and ectoderm A-scores (**Fig 5D**) showed several clusters with compatible A-score and accessibility preferences including clusters 69, 75 and 76 for the ectoderm, and clusters 95, 98, 99 and 117 for the mesoderm. This analysis also highlighted more complex combinatorics such as the one observed for cluster 73 (accessible in both ectoderm and endoderm).

The compartment association analysis of the ATAC peak clusters confirmed that we can observe strikingly cell-type specific accessibility hotspots in loci with very mild compartment association differences. Since chromosomal organization is observed at scales of at least 10s of kilobases and TADs are typically organizing hundreds of kilobases into looped units, we reasoned that this effect could be explained if accessible hotspots with differential cell type activity were intertwined within large chromosomal units rather than demarcated into cell type specific domains. To test this idea, we computed log enrichment ratios for genomic proximity between clustered ATAC peaks. These values are positive if ATAC peaks from one cluster are more likely than expected by chance to be localized within 200kb of peaks from another

cluster in the same TAD. Negative values represent under-representation of pairs from the same cluster at <200kb distance and within the same TAD. As shown in **Fig 5E**, this analysis showed that peaks clusters with activity in embryonic cell types but not extra-embryonic types (P2), or peaks clusters with strong embryonic cell type specific accessibility (P3) are overall demarcated from constitutively accessible sites (P1) or loci that are active specifically in the extra embryonic or early embryonic states (P4). While there are additional proximity relationships within the embryonic peak clusters, the primary organizational principle seems to package the thousands of regulatory elements driving gastrulation in relative proximity, while isolating them from pluripotency or constitutive regulatory elements.

DISCUSSION

In order to characterize how chromosome conformations are reorganized immediately following gastrulation, we generated single cell Hi-C maps from more than 3000 mouse E9.5 embryo cells. We modeled the derived maps along two major axes: first, we aimed to account for the conformation changes occurring during the replication and mitotic cycle; second, we searched for clusters of conformations that can be associated with the rich transcriptional landscape in the embryo at this stage. Separating these two simultaneous dynamics in the embryo (or other tissues) remains a major analytical challenge. Identification of cell types in cells approaching mitosis or exiting it is not realistic at this stage. But new algorithms we introduce here can use robust changes in genome replication time to cluster mid S-phase cells and then derive contact matrix-based (in particular differential A-compartment association) signatures from S-phase clusters. Based on these signatures, cells from nearly all parts of the cell cycle can be classified into balanced models of cell types. Once a cluster structure is inferred, we can pool contacts from single cells into conformation maps and explore cluster-specific differential compartments, long-range contacts and putative promoter-enhancer interactions at high resolution.

Our analysis of Hi-C maps in mouse post-gastrulation highlights several aspects of the relationship between chromosome conformation and embryonic differentiation. First, while the genome organization of ES cells compared to the embryo reflects changes in regulation of key pluripotency genes, the organization within the embryo is largely homogeneous. This suggests that differences in chromosomal conformation between ES cells and E9.5 cells are greater than those between different cell types immediately after gastrulation and at the onset of organogenesis. The exceptions to this homogeneity within the embryo are the distinctively folded primitive erythrocytes. Erythrocytes are unexpected positive controls for the ability to precisely detect a cell type specific conformation when it exists. The unique, compact and highly organized structure of pEry chromosomes cannot be explained by gene expression

alone. In contrast to other differentiating embryonic tissues that continuously respond to signals from neighboring cells and tissue contexts, erythrocytes are fully committed to their functional fate, which may explain their highly distinct (and potentially less plastic) conformation. It is unclear if the erythrocyte chromosome condensation and enucleation program[39] is related to the conformation we observe at E9.5, since definitive erythrocytes only appear several days after. Similar effect could be expected in other terminally differentiated cell types, such as cardiomyocytes or endothelium. But our analysis could not detect a cardiomyocyte conformation cluster, and the small cluster that we linked with endothelial programs could not be associated with a highly distinctive conformation, but was clustered as part of the mesoderm state. It is possible that the reason for this is that these cell states are differentiating much later than pEry cells.

Within the embryo-proper we detected two clusters that match broad ectoderm and mesoderm/endoderm genome regulatory programs. The considerable transcriptional diversity within the mesoderm (and to a lesser extent the ectoderm and endoderm at E9) at this stage was correlated very weakly with conformation sub-clusters within these two clusters. Our current scHi-C data is limited in its depth and number of cells, in particular compared to scRNA-seq or scATAC modern datasets and our analysis suggests that sampling more embryonic cells may lead to characterization of additional statistically significant conformation clusters. But subsampling and in-depth analysis show that such potential additional conformation clusters are unlikely to represent high intensity differential conformation features (as those we detected for pEry cells). The differences between the clusters in terms of replication time regime and compartment structure were small and we had to use sensitive algorithms to deconvolve them from the more apparent cell cycle signature. Interestingly, on the background of such homogeneous conformation landscape we detected hundreds of lineage specific promoter-enhancer contacts that showed matching expression and epigenetic markup in the respective tissues. This argues for an important role for localized embryonic contacts within an initially homogeneous TAD and compartment structure in the embryo. However, the epigenetic stability of such local contacts and the existence of factors regulating them (in addition to the known TFs binding the relevant enhancers) are still unclear. Furthermore, only a small fraction of differential contacts could be explained by enhancer-promoter interactions. It also remains to be seen how specific localized contacts and their higher order structures[29,40,41] contribute to later emergence of broader contact structures, as previously observed in the brain and other tissues. Conversely, since we showed in the case of erythrocytes that chromosomes can in principle be reprogrammed quickly, it will be interesting to understand how conformation in the embryo remains relatively homogeneous despite the activity of specific gene regulatory program, which epigenetic factors may facilitate

the maintenance of such flexible conformation and whether this is linked with the retained developmental plasticity of most embryonic cells at this stage.

METHODS

1. Experimental methods

Cell extraction, fixation and permeabilization

Pregnant C57BL/6 mice were sacrificed at day 9.5 post-coitum and three embryos were dissected under a microscope, in accordance with the Babraham Institute Animal Welfare and Ethical Review Body. The yolk sack was mechanically removed from each embryo, leaving the embryo proper only, and the embryos' morphology was validated to match that of a wildtype E9.5 embryo. To create single-cells suspension, each embryo was moved to a 1.5ml tube containing 200µl of trypsin-EDTA (0.05% trypsin, 0.02% EDTA) and incubated at 37ºC for 5 minutes. 800µl of cold MEF medium was then added to each tube to inactivate the trypsin.

To fix the cells, the cell suspensions of all three embryos were combined and MEF medium at room temperature was added to a final volume of 21ml. 3ml of 16% formaldehyde were added (2% formaldehyde final concentration) and the mixture was incubated for 10 minutes at room temperature, followed by quenching with 127mM glycine for 5 minutes on ice and washing with cold PBS + 0.001% BSA. Cells were then permeabilized in 10 mM Tris-Cl pH 8, 10 mM NaCl, 0.2% IGEPAL CA-630 and cOmplete EDTA-free protease inhibitor cocktail (Roche) for 30 min on ice with intermittent agitation, and spun to collect a nuclei pellet.

Single-cell Hi-C library preparation

scHi-C libraries were prepared in a fashion similar to the one previously described[19]. Briefly, the nuclei were washed with 1.24x NEBuffer 3 (New England Biolabs) and suspended in 400µl of that buffer. 6µl of 20% SDS and then 40µl of 20% Triton X-100 were added to the suspension, with an incubation of 60 minutes at 37ºC with constant agitation following the addition of each of these detergents. Next 50µl of 25U/µl MboI (New England Biolabs) was added and the suspension incubated at 37ºC overnight with constant agitation.

To label the digested DNA ends, dCTP, dGTP, dTTP and biotin-14-dATP (Thermo fisher) were added to the suspension (final concentration of 28.36µM per nucleoside triphosphate) along with DNA polymerase I, large (Klenow) fragment (New England Biolabs, final concentration 0.095U/µl) and the sample incubated at 37ºC for 60 minutes with occasional mixing. The sample was then spun and the supernatant

partially removed, leaving a volume of 50µl, followed by the addition of 100µl 10x T4 DNA ligase reaction buffer (New England Biolabs), 10µl 100x BSA (New England Biolabs), 10µl of 1U/µl T4 DNA ligase (Thermo Fisher) and water to a final volume of 1 ml, and incubated at 16ºC overnight. Finally, the nuclei were filtered through a 30µm cell strainer and single nuclei were sorted into individual empty wells in 384 well plates using an BD Influx cell sorter. The plates were sealed and stored at -80ºC until further processing.

To prepare single-cell Hi-C libraries from single nuclei in plate wells, 2.5µl of PBS was added to each well and the plate was sealed and incubated at 65ºC overnight. DNA was then tagmented using the Nextera XT kit (Illumina) by adding 5µl of TD and 2.5µl of ATM per well and incubating at 55ºC for 5 minutes, followed by cooling to 10ºC and adding of 2.5µl of NT per well. Hi-C ligation junctions were then captured by Dynabeads M-280 streptavidin beads (Thermo Fisher; 10µl of original suspension per well). Beads were prepared by washing with 1x BW buffer (5mM Tris-Cl pH 7.5, 0.5 mM EDTA, 1M NaCl), resuspended in 4x BW buffer (20mM Tris-Cl pH 7.5, 2mM EDTA, 4M NaCl; 4µl per sample), and then mixed with the 12.5µl per-well sample and incubated at room temperature overnight with gentle agitation. The beads were then washed four times with 40µl of 1x BW buffer, washed twice with 40µl of 10mM Tris-Cl pH 7.5 and resuspended in 12.5µl of 10mM Tris-Cl pH 7.5. Single-cell Hi-C libraries were amplified from the beads by adding 7.5µl of Nextera PCR Master Mix, 2.5µl of Index 1 primer and 2.5µl of Index 2 primer (a different combination of index 1 and index 2 per well) followed by 12 PCR cycles. The beads were then magnetically removed and the supernatant from all 384 wells combined. The combined supernatant was purified using AMPure XP beads (Beckman Coulter; 0.6 times volume of the supernatant) according to the manufacturer's instructions and resuspended in 100µl of 10mM Tris-Cl pH 7.5. Finally, the sample was purified again using AMPure XP beads (1.0 times volume of supernatant) and resuspended in 11µl of 10mM Tris-Cl pH 7.5.

Embryo dissection and collection of primitive erythrocytes

Pregnant females were anesthetized with isoflurane using the open-drop system, followed by decapitation in accordance with a protocol approved by the Florida State University Animal Care and Use Committee (ACUC). Uterine horns were removed, rinsed in room temperature PBS and embryos were isolated and transferred to a

droplet of DMEM-high glucose, 10% FBS, 2mM L-glutamine, 1X MEM-Eagle non-essential amino acids and 12ug/mL heparin (Sigma #H31493). Placenta and extraembryonic tissues were removed, embryos were decapitated and circulating peripheral blood was allowed to flow into the droplet of the room temperature media from the severed vitelline and umbilical veins. Media was collected, pooled, and brought to a volume of 21.875mL with room temperature media. Cells were fixed by adding a final concentration of 2% paraformaldehyde for ten minutes. Fixation was quenched by bringing the solution to a final concentration of 0.127M glycine, then incubating on ice for 5 minutes. Cells were pelleted, washed with PBS and pelleted. Cells were flash-frozen and kept at -80C. Cells were thawed and stained for CD71 and TER119. Cells were first blocked with 1mL of PBS-FT (5% FBS, 0.1% Tween-20) for 1 hour, then stained with 1:200 anti-CD71-PE (Invitrogen, 12-0711-82) and 1:200 anti-TER119-APC (Invitrogen, 17-5921-82) for 2 hours at room temperature. Cells were washed and resuspended in 500uL PBS-F (2% FBS) and Hoechst (15ug/mL) and subjected to FACS by Aria (BD Biosciences). Primitive erythrocytes (CD71+, TER119+) were collected and pooled into a 50mL falcon for scHi-C processing following the established protocol (Nagano et al., 2017).

MARS-seq

MARS-seq on E9.0 embryos was performed as previously described[42] sorting 15 plates from 2 129S4/SvJae embryos and sequencing a total of 5760 cells, out of which we retained for analysis 4781 cells with at least 1000 unique molecular identifiers (UMIs) each (median coverage 4574 UMIs). The experiment was performed in accordance with the institutional animal care and use guidelines of the Weizmann Institute of Science.

2. Sequencing and basic computational analysis

scHi-C sequence processing, quality control and cell cycle phasing

We processed the scHi-C data as described previously[19]. Briefly, paired-end reads were demultiplexed to single cells using cell specific barcodes. Reads were broken to segments using matches to MboI recognition site (GATC), and segments were mapped to the genome using Bowtie2. Duplicate contacts were discarded.

We next performed quality control (QC) on each single cell. Cells were filtered based on their coverage (total number of reads), fraction of non-digested contacts, maximal

chromosomal coverage aberration, and the contact distance bin with highest number of contacts.

To partition cells into different phases of the cell cycle, and order the cells within the phases, we calculated for each cell the fraction of "near" reads (with distance < 2Mb), the fraction of mitotic reads (with distance 2-12Mb), mean contact distance for distances at least 4.5 Mb, and the fraction of contacts from a predefined set of early replicating regions. These statistics were used to phase cells into post-mitotic, G1, early to mid-S, mid-S to G2 and pre-mitotic phases, and to order cells within each phase. We note that this approach to phasing was only used as a preliminary stage for the algorithms described below.

Metacell analysis

We applied the Metacell algorithm[27] to organize E9 single cell profiles in 77 metacells (excluding 69 outlier cells), that we summarized into quantitative expression profiles and visualized as previously described[27]. We also downloaded published single cell profiles from the mouse gastrulation atlas[21] and generated 1306 atlas metacells on 110,291 QC-positive cells. Atlas metacells were annotated by majority voting on the published annotations of their cells, defining for each metacell m, the function $\mathrm{atlas.type}(m)$. Each atlas metacell *i* defined a gene expression distribution e_{gi}^{atlas} over the set of the 2237 feature genes *g* used while constructing the metacell graph. For annotation of the E9 map, we identified for each E9 single cell profile the atlas metacell with maximal correlations $\mathrm{ann}_j = \mathrm{atlas.type}(\mathrm{argmax}_i[\mathrm{cor}(\log(u_g + 1), \log(\epsilon + e_{gi}^{\mathrm{atlas}})])$, where u_g is the UMI vector for the E9 cell and $\epsilon = 10^{-5}$ is a regularization factor. We then annotated each E9 metacell with the atlas annotation $\mathrm{atlas.type}$ that was linked with most of its cells.

Definitions and derivation of the strict early and strict late genomic subsets

We partitioned the genome into bins of size 200kb (or 40kb, depending on application) and counted scHi-C coverage per bin and cell in a matrix. We performed down-sampling of scHi-C data such that each cell has 75k contacts and defined:

$\mathrm{DSN} = \mathrm{dsn}_j^i$

as the number of contacts that map to genomic bin j in cell i after downsampling.

We next identified strict-early and strict-late genomics bins. This was done by clustering the genomic bins j using the vectors dsn_j^i into 4 groups using hierarchical clustering. The two clusters showing the highest and lowest coverage were shown to represent the previously observed[19] A and B compartment structures respectively. These clusters behaved consistently (e.g. show enrichment (for A) and anti-enrichment (for B) in S-phase cells) between the pool of embryo and ESC cells. We will denote that derived genomic bins subsets $\mathrm{early}^{\mathrm{strict}}$ and $\mathrm{late}^{\mathrm{strict}}$.

We defined the early/late ratio of a cell as:

$$\mathrm{el}^i = \log 2\left(\frac{\sum_{j\in \mathrm{early}^{\mathrm{strict}}} \mathrm{dsn}_j^i}{\sum_{j\in \mathrm{late}^{\mathrm{strict}}} \mathrm{dsn}_j^i}\right)$$

and classified mid S-phase cells as:

$K^S = \{i\ s.t.\ el^i > 1.8\}$.

K^{non-S} was defined as all other cells.

A-score and early-score for genomic bins

For each genomic bin j and each cell i, we count the number of long range intra-chromosomal contacts (>1Mb) observed between fragment ends in the bins and fragment ends in the $\mathrm{early}^{\mathrm{strict}}$ and $\mathrm{late}^{\mathrm{strict}}$ genome compartments, defining count vectors cA_j^i and cB_j^i.

The A-score of a genomic bin is determined given a set C of scHi-C profiles (possibly all) as:

$$\mathrm{scoreA}_j^C = \sum_{\{i\in C\}} cA_j^i \,/\left[\sum_{\{i\in C\}} cA_j^i + \sum_{\{i\in C\}} cB_j^i\right].$$

The early-replication score (shortened early-score) of a bin is computed given a group C of cells (typically all or part of cells classified as S-phase) by comparing the relative coverage of the bin in C to its relative coverage in G1 cells:

$$\mathrm{scoreE}_j^C = \log\left((|G_1|/|C|)\sum_{\{i\in C\}} \mathrm{dsn}_j^i / \sum_{\{i\in G_1\}} \mathrm{dsn}_j^i\right)$$

Mapping gene expression to genomic bins and scHi-C clusters

We used UCSC gene annotation to determine for each gene (as defined by the MARS-seq or 10X pipeline) a transcription start site (TSS) coordinate. Gene expression profiles were generated as the fraction of UMI per gene observed in scRNA-seq metacells or group of metacells[27]. Given an expression profile e_g we defined a profile

over genomic bins e_j by taking the maximal expression of all genes mapping to TSSs on the bin *j*.

To match expression profiles and scHi-C clusters, we used clusters of TSSs showing coordinated or enriched expression to compute $\mathrm{mean}(\mathrm{scoreA}^{C}_{\{j\in \mathrm{TSSbins}\}})$ or $\mathrm{mean}\left(\mathrm{scoreE}^{C}_{\{j\in \mathrm{TSSbins}\}}\right)$ for each scHi-C cluster C. $\mathrm{TSSbins}$ sets were generated in two ways. First, given a metacell model, we normalized expression (log transforming and subtracting the mean over all metacell profiles), and selected the top 50 enriched TSSs that had enrichment value larger than 0.5. These TSS sets were used to compute the matching between A and early scores and the erythrocyte scHi-C cluster in **Fig 1**. Second, we clustered genes based on their metacell log2 UMI enrichment profiles (**Fig 3I**), generating clusters that were curated manually and derived $\mathrm{TSSbins}$ sets from them for analysis of A-score and early score differences (**Fig 3J,K**).

3. Hi-C contact matrices analysis

Shaman analysis

To calculate enrichment of genome contacts in a Hi-C contact matrix, and to visualize chromosomal conformations, we used the Shaman algorithm[24,28]. We pooled all cells in each cell cluster, and down-sampled the contacts to the same number in each cell cluster pool. We then applied Shaman to the down-sampled contact pools. Briefly, Shaman shuffles contacts while maintaining the marginal coverage distribution and the contact distance distribution, creating a random shuffled contact matrix. The enrichment of a contact is then scored using a KS statistic on the k-nearest neighbors of that contact in the original down-sampled contact matrix and shuffled contact matrix. The Shaman results we report here were derived using an improved MCMC sampler that provide better convergence (in particular on matrices with a smaller number of contacts). In short, the algorithm uses efficient data structures to compute precisely the MCMC update rule. This approach is replacing the previously used strategy of adaptive calibration of a correction term for the function assigning probability for contacting at any genomic distance.

Insulation

We calculated insulation as described previously[24,29]. For a genomic locus, we counted the number of contacts where one contact is up to 200kb upstream of the

coordinate and the other up to 200kb downstream. We next counted the number of contacts where both contacts are in distance up to 200kb from the coordinate. The log ratio between these two numbers is the insulation score. We performed this calculation genome wide in 40kb jumps.

Virtual 4C

To calculate the 4C trace at a specific genome coordinate *x*, we looked at all contacts which satisfy either of the next conditions:

a. One of the fragment ends is at distance < 3e3 from *x*, and the distance between the fragment ends is < 1e5.

b. One of the fragment ends is at distance < 1e4 from *x*, and the distance between the fragment ends is between 1e5 and 5e5.

c. One of the fragment ends is at distance < 3e4 from *x*, and the distance between the fragment ends is between 5e5 and 1e6.

To screen for conformation differences for two scHi-C clusters in a set of target loci, we calculated the difference between their virtual 4Cs. We partitioned the 4C trace to bins based on contact distance (2.5e4, 5e4, 1e5, 2e5, 3e5, 5e5, 7.5e5, 1e6, on both 3' and 5'). We averaged the Shaman scores within every bin, and defined the distance of the conformations for two clusters as the maximum difference (in absolute value) over all bins.

4. Parameters and specific figure panel analysis

Clustering ESC, Embryos and erythrocytes (Figure 1)

We processed the scHi-C data, and performed QC and cell cycle phasing as described above. For generating clusters in **Fig 1**, we used S-phase seeding (**Supplementary Methods**), with the following inputs: K^S, K^{non-S}, DSN, and the matrices cA_j^i and cB_j^i.

To identify primitive erythrocytes, we identified a bin cluster (of the 11 A-score-based bin clusters, see **Supplementary Methods**) that had high C3-specific A-score, and similarly a bin cluster with low C3-specific A-score. We calculated the pooled A-score of each of these two bin clusters in each single cell (denoted $\mathrm{cell\_A}_m^i$ in the **Supplementary Methods**), and used a linear separator to classify cells based on these two scores as either C3 or non-C3 (**Supplementary Fig 4D**). Embryo cells that were not classified as C3 were assigned to C2, and ESC cells to C1.

We generated the genomic bin expression value, A-score and early-score in ESC and non-pEry embryo as described above, in 40kb resolution. We defined genomic bins with at least 4-fold change in expression as ESC- and embryo-induced. Similarly, we defined embryo A-specific bins and ESC A-specific bins as having at least 0.2 difference in A-score.

We screened for genes with different Shaman score in ESC and embryo using comparisons of virtual 4Cs as described above. We similarly calculated differences in Shaman scores for embryo and ESC A-specific bins (**Supplementary Fig 4A**), but looking at the 4C profile of each bin only up to 500kb upstream and downstream.

Erythrocyte analysis (Figure 2)

As before, we generated the genomic bin expression value, A-score and early-score in pEry and non-pEry in 40kb resolution. We defined genomic bins with at least 4-fold change in expression as pEry- and non-pEry-induced. We also identified bins that were not expressed in either of the clusters.

To create **Fig 2J** we identified 40kb genomic bins with A-score that is at least 0.35 higher in Erys compared to embryo. We merged adjacent bins meeting this criterion, and for every set of merged bins found the bin with highest difference in A-score between Erys and embryo. For every such bin, we looked at the average A-score of its 3' and 5' bins up to 400kb. We reversed A-score traces (mirroring 3'/5') to create a matrix in which for all rows, the upstream 5' A-score is higher. We concatenated the A-score trace in pErys and non-pErys, and clustered the concatenated traces using kmeans into 8 clusters. We performed a similar analysis when taking genomic bins with A-score that is at least 0.35 higher in embryo compared to Erys (**Supplementary Fig 4H**).

To create **Fig 2L**, we partitioned the contact matrix into 20kb x 20kb bins, and created for each of the loci in **Fig 2J**, a matrix of average Shaman scores in the 1Mb around it (on both sides). We then averaged the scores in such matrices for the loci in each of the **Fig 2J** clusters.

Gata1 and Tal1 ChIP-seq. We scored and normalized 20bp bins for their *Gata1* and *Tal1* ChIP-seq score using data from ENCODE. We used ChIP-seq scores as previously described, computing ChIP coverage percentiles p for each bin, and defined the score as $-\log_2(1-p)$. We defined *Gata1* and *Tal1* binding sites as those

having score > 8. For 40kb genomic bins we computed a binding score as the maximum ChIP-seq score of all binding sites contained in it.

Clustering the embryo proper (Figure 3)

We applied the replication trend mixture model (**Supplementary Methods**) on embryo scHi-C profiles classified as non-pEry, non-G1 and non-M as described above. We further selected cells with sufficient coverage (at least 8 contacts per 200kb bin on average), and mid-S phase classification as in Nagano et al 2017[19].

We set n_{ij} as the number of contacts in cell i that map to genomic bin j and excluded the X chromosome, or any bin with mean coverage < 8. To set p_j, we calculated the fraction of contacts that mapped to each genomic bin across all G1 cells that are not erythrocytes.

To initialize the model we clustered cells hierarchically using distances based on correlations between rows in a normalized n_{ij} matrix. Normalization provided initial heuristic correction to the cell cycle effect by ordering cells according to their scHi-C fraction of short range contacts and subtracting for each locus the running mean (using a window of 20 cells).

Given this clustering solution, we initialized $E[z_{ik}]$ such that each cell belongs to its cluster with probability 0.5, and to all other clusters with equal probability. In case $k = 2$, each cell belongs to its cluster with probability $\frac{2}{3}$. To initialize s_i we ordered the cells by the fraction of short-range contacts they make, and assigned them values between 1.2 and 1.8 according to their order, assuming that all parts of the replication program are equally represented in the data.

We performed cross validation on the hyperparameters, as described in the **Supplementary Methods**, and selected $L = 12$, $R = 11$, $\lambda = 40$.

To generate Umap projections of mid-S phase cells, we normalized n_{ij} coverage by G1 mean coverage, selected bins with high variance to mean ratio, and calculated a cell-cell correlation matrix using these values. We then used the R package umap with default parameters (and random seed = 42). We repeated this analysis using data normalized based on inferred s-score (see **Supplementary Methods**).

Plotting replication trends for early replicating bins. To plot **Fig 3F**, we identified bins that are in replication regime 2 (out of 12) in C2.1 (left plot; C2.3 for the right plot), and

are in replication regime ≥ 4 in all other clusters, and for every cell calculated the total fraction of contacts from these bins.

Executing other scHi-C clustering algorithms. We executed schicluster and scHi-C topic modeling. We ran schicluster and topic modeling with resolutions 1Mbp and 0.5Mbp respectively, as performed in the publications of these methods.

Cluster annotation. To annotate the C2.1, C2.2 and C2.3 clusters, we used 15 TSS bin set $G_1,..G_{15}$ derived from the E9 metacell model data as described above. To account for possible differences in the s-score distribution in each cell cluster, we ordered genomic bins $j \in \bigcup_m G_m$ by their mean early score across clusters, computed for each bin $\mathrm{scoreE}_j^{C_k}$ and subtracted from it the running mean using a window of 200 bins, defining $\mathrm{scoreE}_j'^{C_k}$. We then computed $\mathrm{mean}(\mathrm{scoreE}'^{C_k}_{\{j \in G_m\}})$, and normalized rows to create the matrix shown in **Fig 3J**. Similar normalization strategy was used with A-scores to derive the matrix in 3K. We repeated the same analysis for the gastrulation atlas metacell model, using 20 gene modules. We note that in order to test possible functional association of the C2.2 cluster, in this analysis we only used 42% of its cells showing a stronger correlation structure. Similar results were obtained using the entire C2.2 cluster.

To compare C2.1 and C2.3 to E14.5 data[24,33], we computed A-scores for genomic bins of length 40Kbp in four samples: C2.1 cells, C2.3 cells, E14.5 HSCs and E14.5 NPCs. To compute these scores, we used the strict-early and strict-late genomic bins that we used to calculate A-scores previously. Because of the large difference in depth between our data and the E14.5 data, we downsampled the contacts of each genomic bin such that the total number of strict-early and strict-late contacts a genomic bin makes is the same in the four different samples. The downsampled contacts were used to calculate the A-scores.

To estimate our assay's sensitivity, we sampled 100, 75, 50 and 25 cells from cluster C2.1, and applied the replication mixture model to a dataset including this subset with all cells from C2.2 and C2.3. We performed a similar analysis for C2.3.

To search for additional sub-structure in cluster C2.1 and C2.3 (**Supplementary Fig 9**) we applied hierarchical clustering to cell-cell correlations derived using s-score normalized copy number profiles. We partitioned C2.1 into 3 subclusters, and C2.3 into 9 subclusters. To annotate the C2.1 subclusters, we correlated their A-score and coverage fold changes with differential gene expression of ectodermal cell types. To

calculate the gene expression profile of a cell type, we calculated the average log2 expression of each gene across all the cell type's metacells in the E9 scRNA-cell data. We then subtracted from each gene its mean expression across ectodermal cell types. This gave each gene its differential expression across all ectodermal cell types. To calculate a genomic bin's A-score fold change in a subcluster, we calculated the A-score by pooling all contacts from the subcluster's cells, and the A-score by pooling all contacts from the other subclusters' cells. The bin's A-score fold change is then the log2 of the ratio between these A-scores. To calculate the coverage fold change, we calculated the relative coverage in the pool of the subcluster's cells as described above, and the relative coverage in the pool of other subclusters' cells, and took the log2 of their ratio. We then only selected genes with at least 2-fold change in gene expression in some cell type, and correlated their relative expression with the A-score and coverage fold changes of the bins containing these genes. Both the A-score and coverage were calculated for genomic bins of size 200kb. To annotate the C2.3 subclusters we did a similar analysis, but used only genes with at least 4-fold change in gene expression in some cell type.

Screening for differential ecto/meso contacts (Figure 4)

We identified enhancers using Chip-seq ENCODE data from ectoderm (forebrain, midbrain, hindbrain) and mesoderm (heart and limb) tissues[43]. We calculated the ChIP-seq scores (log2(1-percentile)) in 20bp resolution for each of the 5 tissues. We called enhancers as contiguous genomic intervals (or peaks) showing H3k4me1 scores > 7 (that is, the top 1/128 bins). We scored each peak H3k4me1 occupancy in mesoderm (maximum between the values of the two tissues) and ectoderm (maximum among the values of the three tissues). To define mesoderm and ectoderm specific peaks we required a score of at least 9 in one set of tissues and a difference of at least 3 between the scores of the two tissue sets. Overall this approached generated 24059 and 9506 meso and ecto- specific enhancers respectively.

To identify meso- and ecto- specifically expressed genes (and TSSs), we identified three metacells representing the mesoderm transcriptional state and three others representing the ectoderm state. We computed the maximum expression per gene in each set. 826 genes were showing at least two-fold difference between the two profiles, and their TSSs were considered for enhancer associations below.

To create a set of potential promoter-enhancer interactions, we linked each enhancer peak with its closest TSSs. We searched for such TSSs 50k-500k upstream and downstream of the enhancer locus (so up to two genes were linked with each peak). We did not use pairings spanning less than 50kb since scHi-C resolution at such distances is limited. We also note that our pairing heuristic is by no means exhaustive, and is meant only to generate a shortlist of putative pairs.

Finally, we identified mesoderm and ectoderm specific enhancer-promoter candidate pairs as those involving a differential enhancer and a specifically expressed gene according to the definitions above. We defined the Shaman score of a putative enhancer-promoter interaction as the score of the contact between coordinates closest to the enhancer and promoter (using Euclidean distances)), and computed it for both meso and ecto. We selected pairs with Shaman score at least 15 higher in the expected cell cluster, or with an absolute shaman score at least 15 if the other score is negative, as having three way support.

To derive a p-value for the number of contacts between an enhancer and a promoter, we counted the number of contacts in the pooled ectoderm and mesoderm cells in a 50kb window (25kb to each direction) around the enhancer-promoter in the contact matrix. Denote these numbers for an enhancer-promoter pair *ep* by ecto_{ep} and meso_{ep}. We similarly counted the number of contacts for all other enhancers and their associated TSSs. To test whether an enhancer-promoter pair had a high number of contacts in ectoderm, we calculated $\mathrm{ecto}_{ep} - \mathrm{meso}_{ep}$, and compared it to the empirical distribution of $\mathrm{ecto}_{e'p'} - \mathrm{meso}_{e'p'}$ for all background enhancer promoter pairs ***e'p'*** that had the same $\mathrm{ecto}_{e'p'} + \mathrm{meso}_{e'p'}$ value as *ep*. To increase power, for **Supplementary Fig 10C-D** we only looked at enhancer-promoter pairs for which $\mathrm{ecto}_{ep} + \mathrm{meso}_{ep} \geq 80$. For the background distribution to be accurate, we only considered $\mathrm{ecto}_{ep} + \mathrm{meso}_{ep}$ values with more than 100 other enhancer-promoter pairs with similar $\mathrm{ecto}_{e'p'} + \mathrm{meso}_{e'p'}$ value. We performed a similar analysis for mesoderm.

We performed a similar analysis to identify H3k27me3-gene pairs. Genes were selected similarly. H3k27me3 regions were selected as those with ChIP-seq score > 7. We designated ecto- or mesoderm specific regions as those having ChIP-seq score > 3 higher than the other tissue. The selected pairs are those where the gene is lowly expressed and the H3k27me3 signal is higher.

To find all hotspots with support for different chromosomal conformation between ectoderm and mesoderm, we looked only at contacts in distance 1e4 to 1e6. We compared the Shaman score of every meso contact to the Shaman score of its nearest ecto contact. We detected regions with high Shaman difference iteratively. In each iteration we identified the contact with the maximal Shaman difference between meso and ecto, and removed all contacts where both their ends are in distance < 5e4 from the maximal-difference contact. We continued with this process until no contact with Shaman difference > 40 remained. This resulted in 5200 hits that we used in order to estimate the fraction of differential contacts explicable by known three-way supported promoter-enhancer pairing in the text.

Analysis of multiome-data and integration with pooled Hi-C clusters.

We used scRNA-seq and scATAC-seq profiles from a recent paper by the Reik group[38], to generate the analysis in **Fig 5**, applying the following steps:

1. Using metacell-2[44] with default parameters and target metacell size of 320K UMIs to organize scRNA-seq profiles into 1404 metacells.

2. Using the RNA-based grouping of cells to collect single cell ATAC reads and create a genomic track for each metacell.

3. Identifying all genomic intervals with ATAC-coverage (total over all metacells) larger than 300 and identified the maximal coverage 300bp within each such interval as a *peak*. Overall this provided us with 94600 peaks.

4. Grouping RNA metacells into 300 clusters using hierarchical clustering of the RNA signatures. RNA clusters were associated with cell type by comparison to gastrulation manifolds and TF expression profiles.

5. We then pooled ATAC reads over the clusters and extracted the reads within identified peaks into an accessibility count matrix. We removed cell clusters supported by fewer than 82K ATAC reads, retaining for analysis 285 clusters.

6. Normalizing peak ATAC coverage in each cluster by normalizing (dividing by total reads for the cluster) and transforming the frequencies p to log2(1e-5+p).

7. Running kmeans++ with a large number of clusters (K=120) over the normalized accessibility profiles. Deriving mean peak cluster profile by averaging the log normalized ATAC values.

8. Filtering peak clusters with less than 100 peaks (only 1 case). Annotating peak clusters as variable whenever at least four metacell clusters showed mean ATAC

value smaller than -16 and the difference between minimum and maximum value over the cluster was larger than 0.7. All other clusters were considered constitutive.

9. Computing the A-score of each peak in ESC, Embryo, pEry, C2.1 and C2.3. Analysis of the A-score distributions in each cluster is used to generate **Fig 5B-D**.

10. Identifying all pairs of peaks within less than 200kb genomic distance. Summarizing the number of such pairs between elements of each pair of clusters into a matrix of observed "proximities". Multiplying each element in the matrix by the total matrix counts divided by the product of its row and column total counts. Log transforming the resulted *enrichment ratio*, followed by hierarchical clustering of the submatrices defined by the constitutive and variable peak clusters (separately) in order to generate the heat map of **Fig 5E**.

DATA AVAILABILITY

The scHi-C and scRNA-seq data generated in this study have been deposited in the GEO database under accession code GSE148793. The ESC scHi-C data used in this study are available in the GEO database under accession code GSE94489. The previously published embryo gastrulation scRNA-seq data used in this study are available in the ArrayExpress database under accession code E-MTAB-6967. The scRNA / scATAC multiome data used in this study are available in the GEO database under accession code GSE205117. The neural progenitor cells' Hi-C data used in this study are available in the GEO database under accession code GSE96107. The hematopoietic Hi-C data used in this study are available in the GEO database under accession code GSE119201.

CODE AVAILABILITY

All code supporting the analysis of this work is available in Github: https://github.com/tanaylab/scHiC_embryo.

REFERENCES

1. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).
2. Kim, S. & Shendure, J. Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Mol. Cell* **76**, 306–319 (2019).
3. Marchal, C., Sima, J. & Gilbert, D. M. Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **20**, 721–737 (2019).
4. Despang, A. *et al.* Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
5. Symmons, O. *et al.* The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev. Cell* **39**, 529–543 (2016).
6. Rodríguez-Carballo, E. *et al.* The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes Dev.* **31**, 2264–2281 (2017).
7. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, (2018).
8. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* (2019) doi:10.1038/s41576-019-0195-2.
9. Monahan, K., Horta, A. & Lomvardas, S. LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature* **565**, 448–453 (2019).
10. Mirny, L. A., Imakaev, M. & Abdennur, N. Two major mechanisms of chromosome organization. *Curr. Opin. Cell Biol.* **58**, 142–152 (2019).
11. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nat. Genet.* **52**, 8–16 (2020).
12. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).

13. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–80 (2012).
14. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–5 (2012).
15. Pope, B. D. *et al.* Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).
16. Dileep, V. & Gilbert, D. M. Single-cell replication profiling to measure stochastic variation in mammalian replication timing. *Nature Communications* **9**, (2018).
17. Naumova, N. *et al.* Organization of the mitotic chromosome. *Science* **342**, 948–953 (2013).
18. Zhang, H. *et al.* Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature* **576**, 158–162 (2019).
19. Nagano, T. *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
20. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nature Methods* **14**, 263–266 (2017).
21. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
22. Nowotschin, S. *et al.* The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
23. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
24. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e24 (2017).
25. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports* **17**, 2042–2059 (2016).
26. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

27. Baran, Y. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).

28. Cohen, N. M. *et al.* SHAMAN: bin-free randomization, normalization and screening of Hi-C matrices. *bioRxiv* 187203–187203 (2017) doi:10.1101/187203.

29. Olivares-Chauvet, P. *et al.* Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. *Nature* **540**, 296–300 (2016).

30. Guo, Y. *et al.* Chromatin jets define the properties of cohesin-driven in vivo loop extrusion. *Molecular Cell* **82**, 3769-3780.e5 (2022).

31. Zhou, J. *et al.* Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation. *Proceedings of the National Academy of Sciences* **116**, 14011–14018 (2019).

32. Kim, H.-J. *et al.* Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Comput Biol* **16**, e1008173 (2020).

33. Chen, C. *et al.* Spatial Genome Re-organization between Fetal and Adult Hematopoietic Stem Cells. *Cell Rep* **29**, 4200-4211.e7 (2019).

34. Du, Z. *et al.* Polycomb Group Proteins Regulate Chromatin Architecture in Mouse Oocytes and Early Embryos. *Mol. Cell* (2019) doi:10.1016/j.molcel.2019.11.011.

35. Loubiere, V., Martinez, A.-M. & Cavalli, G. Cell Fate and Developmental Regulation Dynamics by Polycomb Proteins and 3D Genome Architecture. *Bioessays* **41**, e1800222 (2019).

36. Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat. Genet.* **47**, 1179–1186 (2015).

37. van Weerd, J. H. *et al.* A large permissive regulatory domain exclusively controls Tbx3 expression in the cardiac conduction system. *Circulation research* **115**, 432–41 (2014).

38. Argelaguet, R. *et al.* Decoding gene regulation in the mouse embryo using single-cell multi-omics. 2022.06.15.496239 Preprint at https://doi.org/10.1101/2022.06.15.496239 (2022).

39. Ji, P. New insights into the mechanisms of mammalian erythroid chromatin condensation and enucleation. *Int Rev Cell Mol Biol* **316**, 159–182 (2015).

40. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).
41. Allahyar, A. *et al.* Enhancer hubs and loop collisions identified from single-allele topologies. *Nat. Genet.* **50**, 1151–1160 (2018).
42. Keren-Shaul, H. *et al.* MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. *Nat Protoc* **14**, 1841–1862 (2019).
43. Gorkin, D. U. *et al. Systematic mapping of chromatin state landscapes during mouse development*. http://biorxiv.org/lookup/doi/10.1101/166652 (2017) doi:10.1101/166652.
44. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. *Genome Biology* **23**, 100 (2022).

ACKNOWLEDGMENTS

This work is dedicated to the memory of Elad Chomsky. We thank members of the Tanay lab for discussion. Work in AT group was supported by the NIH 4DN nucleomic tools program and by the European Research Council grant (scAssembly). Work in PF group was supported by the NIH 4DN Nucleomic tools program. Work in RS group was supported by the German-Israeli Project DFG RE 4193/1-1, ISF (grant No. 1339/18), Len Blavatnik and the Blavatnik Family foundation and ISF (grant No. 3165/19) within the Israel Precision Medicine Partnership program. NR was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics, Tel Aviv University, and by the Planning and Budgeting Committee (PBC) fellowship for excellent PhD students in Data Sciences. The contribution of N.R. is part of Ph.D. thesis research conducted at Tel Aviv University. T.N. was supported by MEXT/JSPS KAKENHI (JP18H02374 and JP19H05744) and a grant from the Takeda Science Foundation.

AUTHOR CONTRIBUTION

A.T, E.C and P.T conceived and designed the project. E.C, T.N, C.S, W.L and Z.M executed all wet lab-based experiments. N.R, E.C, Y.L, Y.B, A.L and A.T conducted all computational analysis. R.S, A.T and P.T supervised the project. N.R and A.T wrote the manuscript with input from all authors.

COMPETING INTERESTS

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1: single cell Hi-C in mouse embryo cells

A: Distribution of the number of unique contacts per cell (left) and fraction of trans-chromosomal contacts per cell (right) in the Embryo scHi-C dataset

B: For each single cell shown are the fraction of contacts in in 2Mb-12Mb ("mitotic") distance band vs. the fraction of contacts between elements less than 2Mb apart ("short-range"). Color coding is based on classification into cell cycle phases as in Nagano et al. 2017.

C: Comparing normalized ratio of scHi-C coverage on early and late replicating loci (X axis) to fraction of short-range contacts.

D: Visualizing clusters of single ES and embryo cells using PCA projection of A-scores from 11 genomic clusters. Cells are color coded according to cluster (left) or the initial annotation of cell cycle phase (right).

E: Plotting gene expression of 40kb bins in ESCs compared to embryo cells (mean across E9 metacells, excluding pEry). Upper and lower dashed lines indicate the threshold for defining transcriptional changes between embryo and ESC.

F: Comparison of A-scores for 40kb genomic bins.

G: 40kb genomic bins were stratified according to embryonic expression level (units are log2 of the expression frequency). The distributions of A-scores in embryos (blue) and ESCs (green) are depicted using boxplots. The (-19, -18] box contains at least n=48K genomic bins, (-11, -10] and (-10, -9] at least n=20, and the rest at least n=200. Box limits are the first and third quartile, center line is the median, whiskers are 1.5 times the interquartile range, and points are outliers.

H: Distributions of differential A-score (ESC minus Embryo) in genomic bins with TSSs showing differential gene expression in embryos compared to ESCs (n=1289 ESC induced bins: green, n=806 Embryo induced bins: blue). Box limits are as in **1G**.

I-K: Similar to F-H but showing data on the early-scores of genomic bins instead of A-scores.

L: Examples of conformation reprogramming at pluripotency loci. For each locus we show Shaman enrichment plots in Embryos (top) and ESC (middle), and the respective A-score trends (bottom; blue – embryo, green – ESCs). Dashed circles represent focal points for differential conformation.

Figure 2: Distinct, compact conformation for primitive erythrocytes.

A: Comparison of 40kb bins A-score in pEry cells vs. non-pEry embryo cells. Upper and lower dashed lines show differences of at least 0.3 in A-score.

B: Comparison of log2 mean expression (fraction of molecules per gene) for 40kb genomic bins.

C: Distribution of genomic bins' A-score as a function of expression levels. A-score and transcription were calculated for 40kb genomic bins. Plots show A-scores stratified by expression, for loci classified with conserved expression (left), Ery induced expression (middle) and Ery-repressed expression (right). In the left panel, the (-19, -18] box contains at least n=48K genomic bins, (-11, -10] and (-10, -9] at least n=20, and the rest at least n=200. In the middle panel, all boxes contain at least n=15 genomic bins, except for the (-11, -10] and (-10, -9] which contain at least n=5. In the right panel, all boxes contain at least n= genomic bins, except for (-12, -11], (-11, 10] and (-10, -9] which contain at least n=35, 10, and 2 respectively. Box limits are the first and third quartile, center line is the median, whiskers are 1.5 times the interquartile range, and points are outliers.

D: Distribution of single cell early/late coverage ratio for pEry (red) and non-pEry (black) cells.

E: Comparing early-scores for 40kb genomic bins in pEry and non-pEry embryo cells.

F: Similar to C, but showing distributions of 40kb genomic bins early-score instead of A-score.

G: Showing the distribution of contacts with distance >2Mb vs mitotic contacts (2-12Mb) in pEry (red) and non-pEry cells (black). Note the general high degree of long range contacts in p-Ery.

H: Showing the fraction of contacts in the most frequent distance bin (defined as "Far tightness" in Nagano et al 2017) compared to the rate of long-range contacts.

I: Distributions of inter-chromosomal contact rates for pEry and non-pEry cell.

J: Shown are color coded A-scores computed for the pEry (left) and non-pEry (right) clusters around loci with pEry specific high A-score (400kb upstream and downstream). Loci are grouped into 8 clusters using K-means clustering.

K: For each of the loci clustered in J we color coded bins with any level of transcription according to the relative expression in pEry and non-pEry cells (blue – higher in non pEry, red - higher in pEry)**.**

L: Loci within each cluster in J were pooled, and their average Shaman score is color coded for pEry and non-pEry cells. The pooled A-score profile is shown at the bottom for every loci cluster in pEry and non-pEry.

M: Examples of loci showing distinct pEry conformation. For every locus, depicted are contact enrichment in non-pEry cells (top), pEry cells (middle) and profile of A-score in the two clusters (bottom). For *Cpox* and *Hbb* we mark contacts with the TSS locus by black diagonal lines.

Figure 3: Ectoderm and mesoderm/endoderm scHi-C clusters in the embryo

A: S-phase cells from the non-pEry cluster were identified and projected on 2D using Umap analysis of their coverage in 1103 loci. Cells are color-coded by their s-score as inferred by our probabilistic model.

B: Umap projection of the same cells as in A, using features normalized given inferred S-score for each cell.

C: Distribution of inferred s-scores for the three non-pEry embryo clusters.

D: Average normalized coverage (early-score) for genomic bins in clusters C2.1 and C2.3.

E: Similar to D, but comparing average C2.1 and C2.3 behavior to C2.2 behavior.

F: Genomic bins that were inferred to be early replicating (methods) in C2.1 (left) or C2.3 (right) were pooled, and for each cell we plotted total coverage as a function of the inferred S-score. Cells are colored by their cluster (C2.1 – green, C2.3 – orange).

G: Distribution of the difference between C2.1 cells and C2.3 cells in early-score (left) and A-score (right) for genomic bins classified as specific to C2.1 (green) or C2.3 (orange). Grey – all bins.

H: Average normalized A-score for the group of genomic bins specific to C2.1 (X) and C2.3 (Y) are depicted for color-coded cells in the three clusters C2.1- 3 (left). A Similar plot is shown for 898 cells that were not included in the set of 699 mid S-phase cells used for clustering (right). Gray lines mark the thresholds used for classification of the expanded C2.1 and C2.3 clusters.

I: Correlation heatmaps for 2353 gene expression profiles over the E9.0 metacell model. Gene module numbers and representative genes are shown on the right. S. ecto: Surface ectoderm. CM: cardiomyocyte. Endot: Endothelium. E Meso: extraembryonic mesoderm.

J: The color-coded matrix represents the difference in average early-score per single cell cluster (columns) for the TSS loci in each gene module from I (rows).

K: Similar to J, but showing difference in average A-score in each cluster.

L: Depicting the contact structure (color-coded Shaman map) in C2.1 (top) and C2.3 (bottom) cells around the Crabp2 and Igf2 TSSs.

Figure 4: Three-way support for specific regulatory contacts

A-B: Comparing A-score (top), contact maps (middle), virtual-4C using Shaman scores, and H3K4me1 ChIP-seq (bottom) around the *Sox2* and *Twist1* loci. The genes, and for *Twist1* also a nearby enhancer, are marked by vertical grey lines.

C: Shown are distributions of genomic distances between a TSS and the nearest putative enhancer classified according to the ectoderm/mesoderm lineage specificity of the two loci as determined by gene expression (for the promoter) and encode ChIP-seq (for the putative enhancer).

D: The distribution of differential C2.1 and C2.3 Shaman score (X axis) on TSS-enhancers pairs with coordinated mesoderm or ectoderm specific activity. Shaman differences is computed only for contacts with positive scores in both C2.1 and C2.3.

E: Examples of virtual 4C plots (top) and H3K4me1 ChIP-seq (bottom, C2.1 followed by C2.3) around 4 ectoderm and 4 mesoderm genes. Gray vertical lines mark the TSS and putative enhancer. Gene-free regions around regulated genes are highlighted by horizontal gray bars.

F: Contact structure around the *Tbx3-Tbx5* locus in the C2.1 and C2.3 clusters. Contacts discussed in the text are marked by dashed circles.

Figure 5: Gastrulation accessibility hotspots are chromosomally intertwined

A: Bottom panel shows the accessibility of peaks (rows) over metacells (columns) (log2 number normalized ATAC-seq reads). Shown are loci from select clusters highlighted in the text. Top panel depicts gene expression of correlated TFs over the same metacells, provided in order to link accessibility clusters with specific cell types.

B: For each cluster of ATAC peaks we computed the fraction of loci with A-compartment score difference larger than 0.1 when comparing ESC and Embryo pooled Hi-C. Clusters with over 0.08 of the loci showing A-score enrichment in ESCs are colored black.

C: Similar to B, but comparing embryo and pEry pooled Hi-C maps.

D: Similar to B, but comparing the embryonic clusters C2.1 (ectoderm) and C2.3 (mesoderm).

E: Left panel is showing mean normalized accessibility for ATAC peak clusters (row) and metacells (column). Right panel is showing for each pair of peak clusters the enrichment of intra-TAD proximity (number of pairs of peaks in the same TAD and within 200kb of each other).

Figure 1

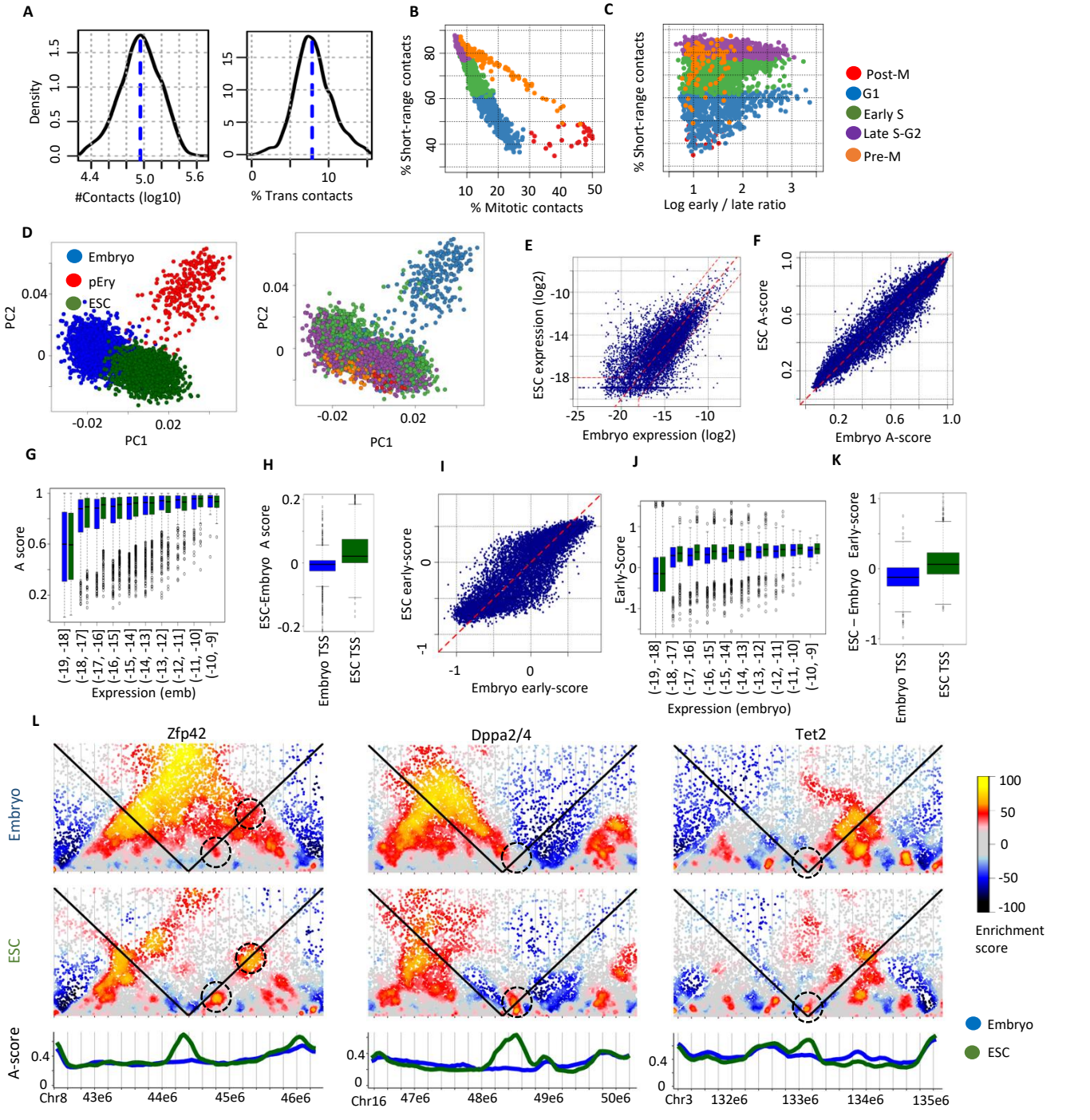

A
Density
#Contacts (log10)
% Trans contacts
B
% Short-range contacts
% Mitotic contacts
C
% Short-range contacts
Log early / late ratio
Post-M
G1
Early S
Late S-G2
Pre-M
D
Embryo
pEry
ESC
PC2
PC1
E
ESC expression (log2)
Embryo expression (log2)
F
ESC A-score
Embryo A-score
G
A score
Expression (emb)
H
ESC-Embryo A score
Embryo TSS
ESC TSS
I
ESC early-score
Embryo early-score
J
Early-Score
Expression (embryo)
K
ESC – Embryo Early-score
Embryo TSS
ESC TSS
L
Zfp42
Dppa2/4
Tet2
Embryo
ESC
A-score
Chr8
Chr16
Chr3
Enrichment score
Embryo
ESC

Figure 2

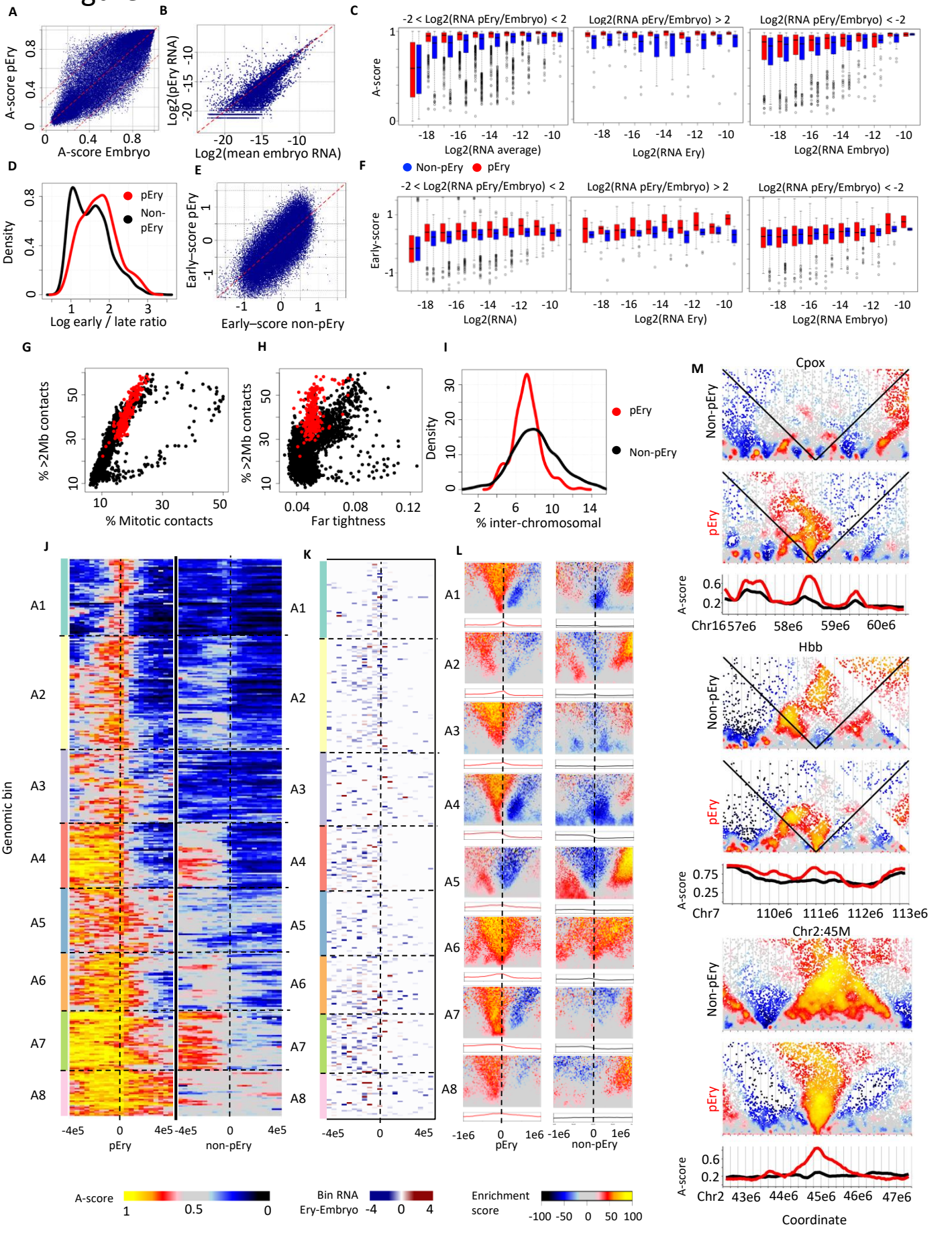

A
A-score pEry
A-score Embryo
B
Log2(pEry RNA)
Log2(mean embryo RNA)
C
-2 < Log2(RNA pEry/Embryo) < 2
Log2(RNA pEry/Embryo) > 2
Log2(RNA pEry/Embryo) < -2
A-score
Log2(RNA average)
Log2(RNA Ery)
Log2(RNA Embryo)
D
Density
pEry
Non-pEry
Log early / late ratio
E
Early–score pEry
Early–score non-pEry
F
Non-pEry
pEry
Early-score
Log2(RNA)
G
% >2Mb contacts
% Mitotic contacts
H
% >2Mb contacts
Far tightness
I
Density
% inter-chromosomal
J
Genomic bin
A1
A2
A3
A4
A5
A6
A7
A8
pEry
non-pEry
K
L
M
Cpox
Non-pEry
pEry
A-score
Chr16 57e6 58e6 59e6 60e6
Hbb
Chr7 110e6 111e6 112e6 113e6
Chr2:45M
Chr2 43e6 44e6 45e6 46e6 47e6
Coordinate
A-score
Bin RNA Ery-Embryo
Enrichment score

Figure 3

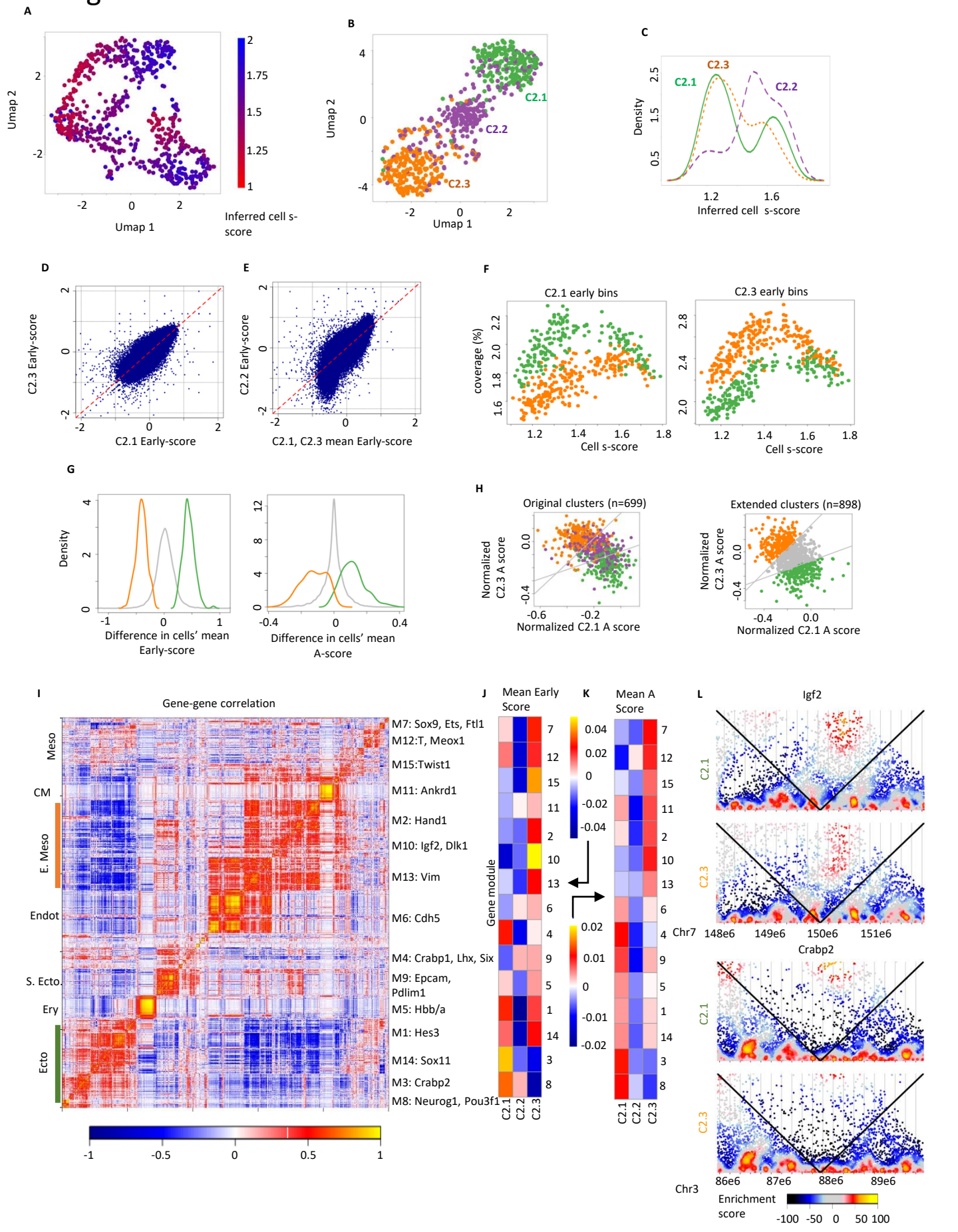

A
Umap 2
Umap 1
Inferred cell s-score
B
C2.1
C2.2
C2.3
C
Density
Inferred cell s-score
D
C2.3 Early-score
C2.1 Early-score
E
C2.2 Early-score
C2.1, C2.3 mean Early-score
F
C2.1 early bins
C2.3 early bins
coverage (%)
Cell s-score
G
Density
Difference in cells' mean Early-score
Difference in cells' mean A-score
H
Original clusters (n=699)
Extended clusters (n=898)
Normalized C2.3 A score
Normalized C2.1 A score
I
Gene-gene correlation
Meso
CM
E. Meso
Endot
S. Ecto.
Ery
Ecto
M7: Sox9, Ets, Ftl1
M12:T, Meox1
M15:Twist1
M11: Ankrd1
M2: Hand1
M10: Igf2, Dlk1
M13: Vim
M6: Cdh5
M4: Crabp1, Lhx, Six
M9: Epcam, Pdlim1
M5: Hbb/a
M1: Hes3
M14: Sox11
M3: Crabp2
M8: Neurog1, Pou3f1
J
Mean Early Score
Gene module
K
Mean A Score
L
Igf2
Chr7
Crabp2
Chr3
Enrichment score

Figure 4

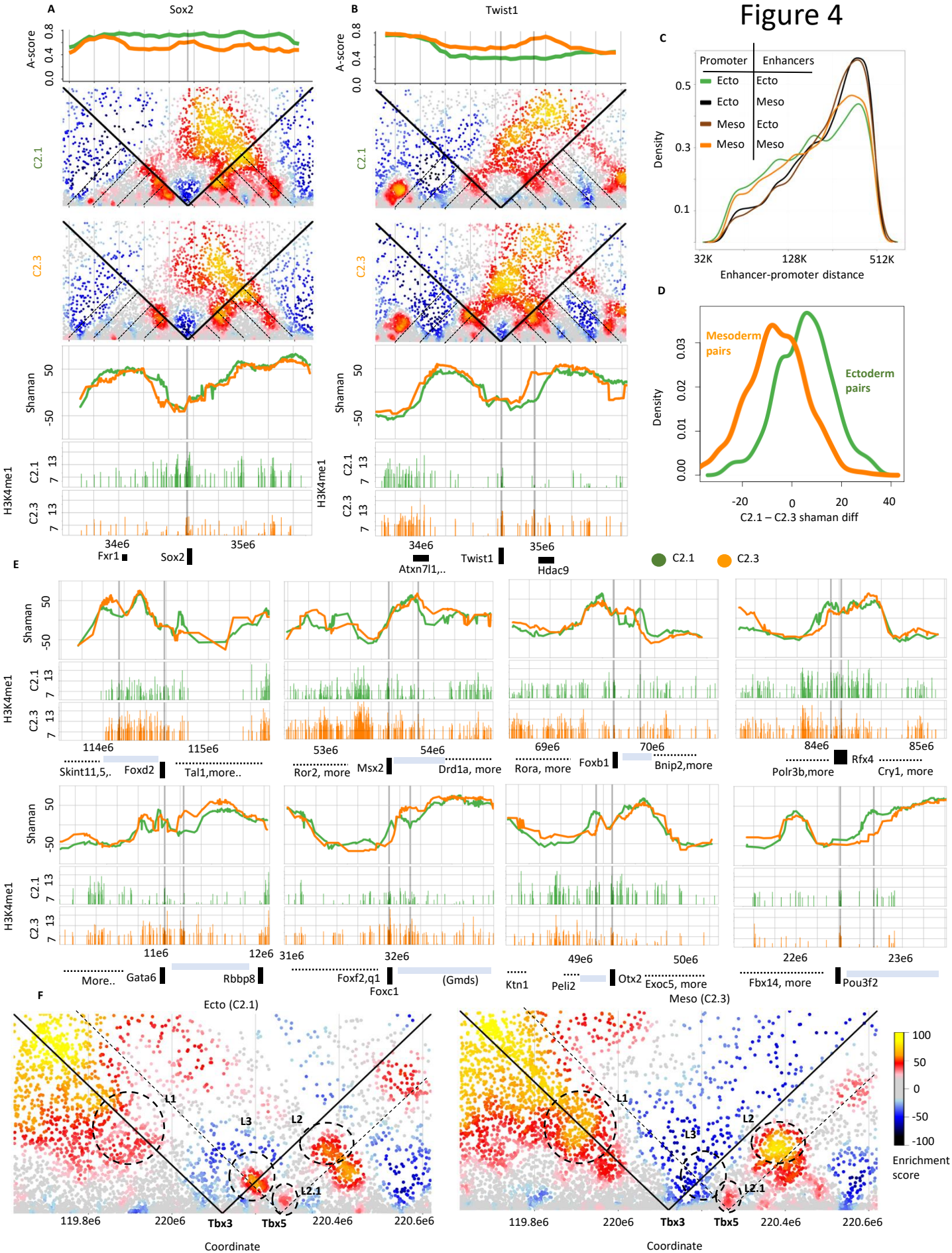
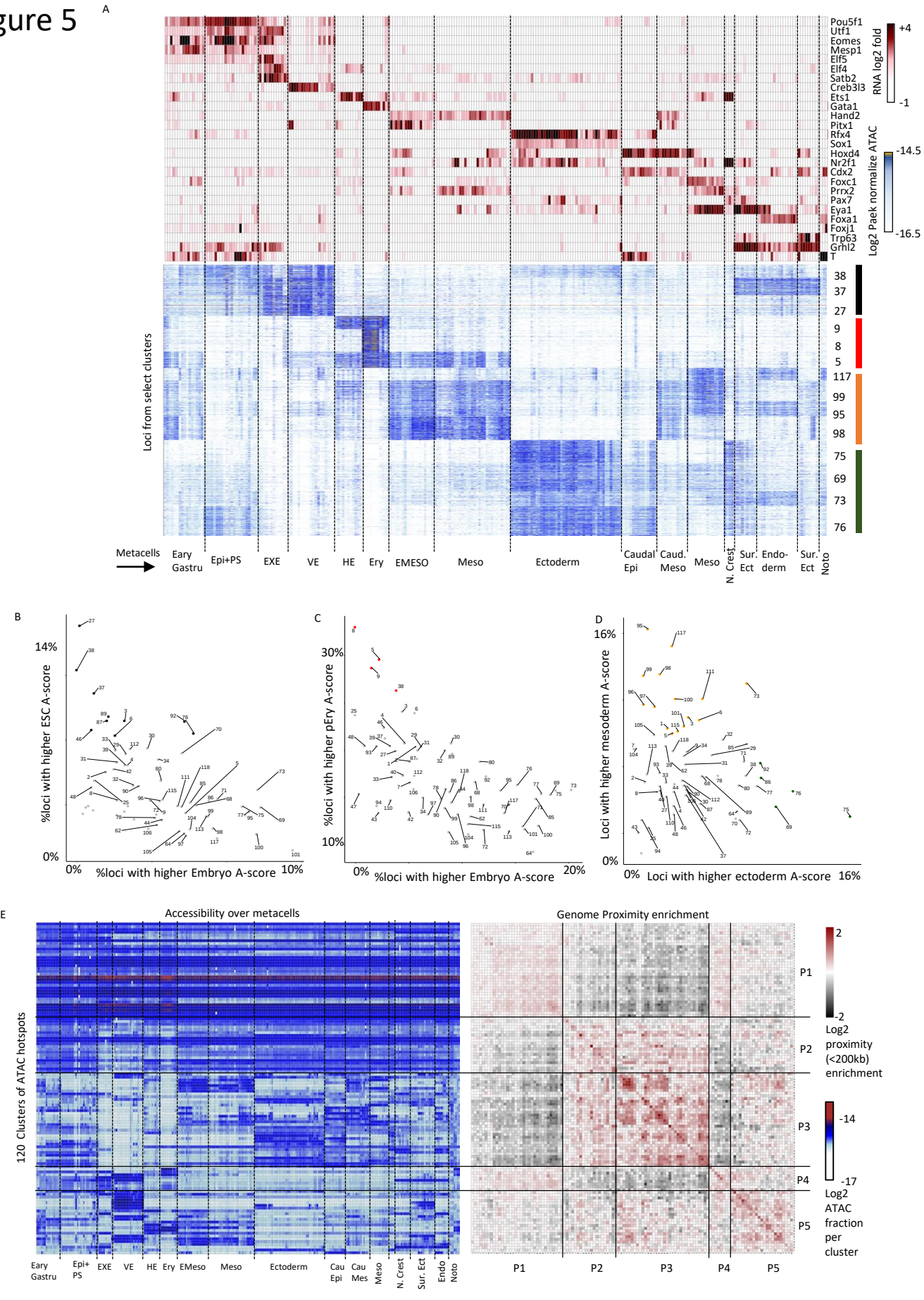

A
Sox2
B
Twist1
A-score
C2.1
C2.3
Shaman
H3K4me1
Fxr1
Sox2
Atxn7l1,..
Twist1
Hdac9
C
Promoter
Enhancers
Ecto
Ecto
Ecto
Meso
Meso
Ecto
Meso
Meso
Density
Enhancer-promoter distance
D
Mesoderm pairs
Ectoderm pairs
C2.1 – C2.3 shaman diff
E
C2.1
C2.3
Skint11,5,..
Foxd2
Tal1,more..
Ror2, more
Msx2
Drd1a, more
Rora, more
Foxb1
Bnip2,more
Polr3b,more
Rfx4
Cry1, more
More..
Gata6
Rbbp8
Foxf2,q1
Foxc1
(Gmds)
Ktn1
Peli2
Otx2
Exoc5, more
Fbx14, more
Pou3f2
F
Ecto (C2.1)
Meso (C2.3)
L1
L2
L3
L2.1
Tbx3
Tbx5
Coordinate
Enrichment score

Figure 5


A
Pou5f1
Utf1
Eomes
Mesp1
Elf5
Elf4
Satb2
Creb3l3
Ets1
Gata1
Hand2
Pitx1
Rfx4
Sox1
Hoxd4
Nr2f1
Cdx2
Foxc1
Prrx2
Pax7
Eya1
Foxa1
Foxj1
Trp63
Grhl2
T
RNA log2 fold
+4
-1
Log2 Paek normalize ATAC
-14.5
-16.5
Loci from select clusters
38
37
27
9
8
5
117
99
95
98
75
69
73
76
Metacells
Eary Gastru
Epi+PS
EXE
VE
HE
Ery
EMESO
Meso
Ectoderm
Caudal Epi
Caud. Meso
Meso
N. Crest
Sur. Ect
Endo-derm
Sur. Ect
Noto
B
%loci with higher ESC A-score
14%
0%
0%
%loci with higher Embryo A-score
10%
C
%loci with higher pEry A-score
30%
10%
0%
%loci with higher Embryo A-score
20%
D
Loci with higher mesoderm A-score
16%
0%
0%
Loci with higher ectoderm A-score
16%
E
Accessibility over metacells
Genome Proximity enrichment
120 Clusters of ATAC hotspots
Eary Gastru
Epi+ PS
EXE
VE
HE
Ery
EMeso
Meso
Ectoderm
Cau Epi
Cau Mes
Meso
N. Crest
Sur. Ect
Endo
Noto
P1
P2
P3
P4
P5
2
-2
Log2 proximity (<200kb) enrichment
-14
-17
Log2 ATAC fraction per cluster

Chapter 7 - Discussion

In this thesis we developed methods for analyzing multi-omic and single-cell datasets. Specifically, we focused on the problems of cancer-subtyping using multi-omic data, and of analyzing scHi-C data during mouse embryonic development. We first performed a benchmark comparing several methods for multi-omic cancer subtyping. We then developed NEMO, an algorithm for multi-omic cancer subtyping that supports partial data, and showed its favorable performance compared to other clustering methods. Next, we highlighted a limitation of the commonly used method for comparing the survival of different groups of patients, and provided an implementation of an exact test that overcomes this limitation. The last work we described on cancer subtyping is MONET, an algorithm that detects patient modules, where patients are allowed to be similar in only a subset of the input omics.

In addition to cancer subtyping, we used MONET to cluster multi-omic single-cell data in embryonic development. This leads to our work on analysis of scHi-C data, which was described in the previous chapter. In that work, we developed methods to detect groups of cells with different genome organization, while accounting for the large variance between cells caused by the cell cycle.

In this chapter, we first summarize the research projects described in this thesis before characterizing possible extensions of them. Then, we outline a few possible directions for future research in the analysis of multi-omic and single-cell data.

7.1. Multi-omics clustering benchmark

Our first work compared performance of different multi-omic clustering methods. We used two criteria to assess the quality of a cancer clustering solution: the number of known enriched clinical parameters, and the differential survival between the clusters. We showed that some multi-view methods, that were not developed specifically for omics data, performed better than some methods developed for such data. We also showed that methods do not necessarily benefit from using all the omics in the input, and that the best omics to use varies between different cancers.

A possible limitation of this work, and also of the clustering algorithms that we developed, is the assessment criteria we used. Differential survival may be a biased measure for several reasons. Cancer patients are being treated for their disease, and the availability of treatment for a cancer

subtype will improve the prognosis of the subtype's patients, while the lack of a treatment will worsen the prognosis. Treatment availability therefore biases this assessment criterion. This criticism is somewhat mitigated by the fact that the existence of a treatment for a subset of the patients indicates that these patients do form an actual type, but other criticisms on differential survival are not as clearly defended. First, it is possible for different cancer subtypes to have similar survival rates. In this case, an algorithm that detects these subtypes will not be rewarded. Second, an algorithm that picks up a signal that is not related to the tumor's biology, but to some unrelated variable that is correlated to survival, will be rewarded. For example, in many cancer types prognosis worsens with age, and an algorithm that clusters patients based on age will find differences in survival. The second assessment criterion, enrichment of known clinical parameters, is similarly biased, and may prefer clustering solutions that better reflect the current understanding of cancer. Still, these assessment criteria are currently the best and most widely accepted way to measure cancer subtyping performance.

7.2. NEMO

In Chapter 3 we introduced NEMO, a method we developed for multi-omic clustering. We compared NEMO's performance to nine other algorithms on ten different cancer types from TCGA, using three omics: gene expression, DNA methylation and miRNA expression. While NEMO's performance was not the best in neither of the two assessment criteria we used, it was second in performance in both, offering overall best results. NEMO is also very fast, and we showed that it supports partial datasets, where there are patients with measurements in only a subset of the omics. Finally, we used NEMO to detect cancer subtypes in Acute Myeloid Leukemia.

Notably, since its publication, NEMO was included in several multi-omic cancer clustering benchmarks conducted by other research groups, with very good results. First, Duan et al. benchmarked ten methods on nine cancer types, using different subsets of four omics. They used clinical significance, accuracy (in cancer types with accepted subtype definitions), robustness and runtime to assess the algorithms' performance, and concluded that NEMO and SNF were best overall [128]. Second, Niessl et al. investigated the tendency of methods to perform better in their introductory paper than in subsequent comparison studies [129]. To study this topic, the authors looked at pairs of methods that are designed for the same task, and assessed each method in the exact same settings that the second method used to measure its own performance in its original publication. Out of four methods, only NEMO's performance did not deteriorate.

NEMO has several limitations. While NEMO supports partial data, it currently requires that all pairs of patients have data in at least one common omic. For this reason, NEMO does not support datasets with two omics, in which some patients have measurements in both omics, while some only in the first omic and some only in the second omic. This is arguably the most common case of partial data, and NEMO could be extended to support such use cases. An additional limitation of NEMO is that it provides a clustering solution, but does not provide insight about the features that support this clustering. Downstream analysis of the clustering can find differential features, but it could have been useful if some biological understanding of the features would come from NEMO directly.

7.3. Inaccuracy of the log-rank test

In Chapter 4 we showed the inaccuracy of the log-rank test when used in modern cancer datasets, which typically have hundreds of patients. The broadly used log-rank implementation is based on an asymptotic approximation and is not an exact test, and therefore it is not expected that its reported p-values will be exactly accurate. However, the extent of the inaccuracy even in datasets with hundreds of patients warrants attention. The test is less accurate when many groups are compared (that is, in our context, in a cancer type with many subtypes), and in cancers with good prognoses, where there is a low number of death events. We also showed examples of previous studies that reported false discoveries (using a significance threshold of 0.05) due to the test's inaccuracy.

The exact version of the log-rank test was described previously [130], [131], and is well known in the statistics community. We provided an implementation of this test in the R programming language, where previously an implementation was available only for the case of two groups. The downside of the exact test is its runtime. Faster tests with higher accuracy in cancer datasets are needed.

7.4. MONET

In Chapter 5 we presented MONET, an additional algorithm for multi-omics clustering. We compared MONET's performance to other algorithms using the same data that we used to assess NEMO's performance. MONET's performance was overall best together with NEMO. However, these two methods performed well on different datasets, so they could potentially be used complementarily. We also showed in more depth how to use MONET on Ovarian Serous Cystadenocarcinoma, a type of ovarian cancer. We finally applied MONET to single-cell multi-omic data, showing that MONET can be used in diverse settings.

MONET is designed to find clustering solutions in which samples are not necessarily similar in all omics, and we find such cases often in cancer data. But such clusters are difficult to interpret biologically. What biological mechanism can cause patients to be similar in gene expression, but not similar in miRNA expression? This is especially baffling given the high interconnectivity among different biological layers. Future work, both biological and computational, will be needed to better understand such cases.

MONET has several computational limitations. It has multiple hyperparameters, and the tuning of these hyperparameters is challenging. MONET requires that its omic similarity graphs will have both positive and negative edges, and tuning the edge weights to obtain this requirement can be performed in many ways. While we suggested a default weighting scheme, we believe that future work can find improved and more robust schemes. An additional limitation is that MONET attempts to heuristically solve an NP-hard problem, and does not guarantee an optimal solution. We added several "actions" to MONET's iterative optimization procedure in order to avoid obvious cases of termination at a local optimum, but additional such actions may further improve MONET's solution, at the expense of a slower runtime.

7.5. Single-cell Hi-C

In Chapter 6 we presented our study on the genome organization at single-cell resolution. In this work we developed a methodology for the analysis of scHi-C from diverse cell types, while controlling for the dominant signal of the cell cycle. We found that primitive erythrocytes have a distinct chromosomal conformation, and then used a subtler approach to find cell groups that we identified as mesodermal and ectodermal. We then connected the genome structure of these groups to other epigenetic regulatory mechanisms – to openness of the genome using single-cell ATAC-seq data, and to histone modifications using ChIP-seq data. We found that in general, tissue specific enhancer-promoter interactions are mediated by physical proximity.

The main conclusion we draw from this study is that we do not observe very high diversity between cell types in terms of the genome organization. This lack of diversity is compared to the high diversity that is already seen at this stage of development in terms of tissue morphology and functionality, and that also manifests itself in scRNA-seq data [119]. This conclusion suggests that genome organization is not a leading epigenetic factor in embryonic differentiation processes, which is driven by other factors. However, a limitation of our study is that it is possible that there are differences in genome organization that our data and analysis were not able to detect. We only sampled a small number of molecules from every cell, a small number of cells (three thousand, compared to hundreds of thousands in scRNA-seq studies), and Hi-C data is

highly noisy in its nature. It is therefore still too early to confidently state that Hi-C is not a driving force in differentiation.

7.6. Future work

Biological research is gradually shifting into a more quantitative, data-driven science. This coincides with decreasing costs and higher availability of high-throughput experimental techniques. These trends will likely lead to more multi-omic and single-cell datasets, and will only increase the need for further methodological innovation. Our work suggests many directions for such future methodological work.

7.6.1. Multi-omic analysis

Our work focused on multi-omic clustering. An advantage of the methods we presented is that they are based on similarity between samples, and can therefore be easily extended to new omics data. But this generality has a downside in that it cannot provide mechanistic understanding for the connection between omics. A different approach than ours can design models for specific omics, incorporating known biological mechanisms. Some previous studies took that approach, e.g. PARADIGM, which we mentioned previously [83], but there is still much room for innovation. Such methods can also help explain the phenomenon of disagreement between omics that we described in our work on MONET.

Besides multi-omic clustering, there are other multi-omic analysis problems that received less attention. Specifically, we are not aware of any multi-omic classification algorithm that performs better than concatenating features from different omics and applying single-omic classification methods. Another important problem is multi-omic visualization. This problem is related to multi-omic dimension reduction, and can be thought of as dimension reduction into two or three dimensions, but the objective in this task is very different from that of dimension reduction. While there are several methods for multi-omic dimension reduction [132], [133], including methods that are also used for multi-omic clustering, work on visualization is scarce [134], [135].

7.6.2. Single-cell analysis

Single-cell methods offer diverse algorithmic problems, and indeed single-cell analysis is a field that enjoys the interest of a growing number of computational researchers. scHi-C datasets are still scarce, and only a handful of experimental groups have the technical expertise and resources to create new data. Nonetheless, several methods have been developed recently that attempt to increase the resolution of single-cell data, and detect structural entities such as TADs and loops [125]–[127]. The approach we presented for single-cell relies heavily on the presence of

many replicating cells, an assumption that is valid in embryonic development but is not valid for almost all mature tissues. Methods that cluster scHi-C data by looking directly on the genome organization, and not on the replication-dependent number of reads, are much needed.

We use the coverage (number of reads) of different genomic bins to analyze scHi-C data. This approach is not specific to Hi-C, and can be readily extended to other single-cell omics data. Indeed, we performed proof-of-concept studies on applying this approach to single-cell methylation data, and think that with some adaptations it can also be applied to single-cell ATAC. The latter data type is somewhat more challenging, because the number of reads is part of the ATAC assay's output, and is affected by the openness of a genomic region. Distinguishing between changes in the number of reads that are due to openness and those that are due to copy number is imperative to extend our approach to ATAC data.

7.6.3. Single-cell multi-omics

Since our work involves multi-omic data and single-cell data, a natural extension of it would be for single-cell multi-omic data, where multiple omics are measured at single-cell resolution. Our work included one such analysis, when we applied MONET to scRNA and single-cell methylation from single cells. Until just a couple of years ago only a handful of labs could produce such data [136], [137]. Only now the relevant experimental methods are becoming more widespread, and still for a small subset of omics. Most notable is the combination of single-cell RNA and ATAC, for which a commercial product is now available, making it the most common multi-omic single-cell data [138]. Multi-omic single-cell data has recently been chosen as the "Method of the Year" by Nature Methods, one of the leading journals for novel methodology [139], and experimental methods for the analysis of such data are now a hot area of research. All the approaches we presented in this work can be extended to this new data type – either by adapting the multi-omic methods for single-cell data, or by adapting the single-cell methods to multi-omic data.

References

[1] N. Rappoport and R. Shamir, “Multi-omic and multi-view clustering algorithms: review and cancer benchmark,” *Nucleic Acids Res.*, vol. 46, no. 20, pp. 10546–10562, Nov. 2018, doi: 10.1093/nar/gky889.

[2] N. Rappoport and R. Shamir, “NEMO: cancer subtyping by integration of partial multi-omic data,” *Bioinformatics*, vol. 35, no. 18, pp. 3348–3356, Sep. 2019, doi: 10.1093/bioinformatics/btz058.

[3] N. Rappoport and R. Shamir, “Inaccuracy of the log-rank approximation in cancer data analysis,” *Mol. Syst. Biol.*, vol. 15, no. 8, p. e8754, Aug. 2019, doi: 10.15252/msb.20188754.

[4] N. Rappoport, R. Safra, and R. Shamir, “MONET: Multi-omic module discovery by omic selection,” *PLOS Comput. Biol.*, vol. 16, no. 9, p. e1008182, 2020, doi: 10.1371/journal.pcbi.1008182.

[5] J. D. Morris and C. K. Payne, “Microscopy and Cell Biology: New Methods and New Questions,” *Annu. Rev. Phys. Chem.*, vol. 70, no. 1, pp. 199–218, 2019, doi: 10.1146/annurev-physchem-042018-052527.

[6] M. S. Smyth and J. H. J. Martin, “x Ray crystallography,” *Mol. Pathol.*, vol. 53, no. 1, pp. 8–14, Feb. 2000.

[7] F. Crick, “Central Dogma of Molecular Biology,” *Nature*, vol. 227, no. 5258, Art. no. 5258, Aug. 1970, doi: 10.1038/227561a0.

[8] B. E. Stranger and E. T. Dermitzakis, “From DNA to RNA to disease and back: The ‘central dogma’ of regulatory disease variation,” *Hum. Genomics*, vol. 2, no. 6, pp. 383–390, Jun. 2006, doi: 10.1186/1479-7364-2-6-383.

[9] K. Danna and D. Nathans, “Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae *,” *Proc. Natl. Acad. Sci.*, vol. 68, no. 12, pp. 2913–2917, Dec. 1971, doi: 10.1073/pnas.68.12.2913.

[10] U. E. Loening, “The determination of the molecular weight of ribonucleic acid by polyacrylamide-gel electrophoresis. The effects of changes in conformation,” *Biochem. J.*, vol. 113, no. 1, pp. 131–138, Jun. 1969.

[11] S. Deepak *et al.*, “Real-Time PCR: Revolutionizing Detection and Expression Analysis of Genes,” *Curr. Genomics*, vol. 8, no. 4, pp. 234–251, Jun. 2007.

[12] O. Smithies, “An improved procedure for starch-gel electrophoresis: further variations in the serum proteins of normal individuals,” *Biochem. J.*, vol. 71, no. 3, pp. 585–587, Mar. 1959, doi: 10.1042/bj0710585.

[13] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977.

[14] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, “Quantitative monitoring of gene expression patterns with a complementary DNA microarray,” *Science*, vol. 270, no. 5235, pp. 467–470, Oct. 1995, doi: 10.1126/science.270.5235.467.

[15] J. Shendure *et al.*, “DNA sequencing at 40: past, present and future,” *Nature*, vol. 550, no. 7676, Art. no. 7676, Oct. 2017, doi: 10.1038/nature24286.

[16] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.

[17] S. Motameny, S. Wolters, P. Nürnberg, and B. Schumacher, “Next Generation Sequencing of miRNAs – Strategies, Resources and Methods,” *Genes*, vol. 1, no. 1, pp. 70–84, Jun. 2010, doi: 10.3390/genes1010070.

[18] A. Chowdhary, V. Satagopam, and R. Schneider, “Long Non-coding RNAs: Mechanisms, Experimental, and Computational Approaches in Identification, Characterization, and Their Biomarker Potential in Cancer,” *Front. Genet.*, vol. 12, 2021, Accessed: Jun. 05, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fgene.2021.649619

[19] A. L. Mattei, N. Bailly, and A. Meissner, “DNA methylation: a historical perspective,” *Trends Genet. TIG*, vol. 38, no. 7, pp. 676–707, Jul. 2022, doi: 10.1016/j.tig.2022.03.010.

[20] P. J. Park, “ChIP–seq: advantages and challenges of a maturing technology,” *Nat. Rev. Genet.*, vol. 10, no. 10, Art. no. 10, Oct. 2009, doi: 10.1038/nrg2641.

[21] M. Tsompana and M. J. Buck, “Chromatin accessibility: a window into the genome,” *Epigenetics Chromatin*, vol. 7, no. 1, p. 33, Nov. 2014, doi: 10.1186/1756-8935-7-33.

[22] B. Aslam, M. Basit, M. A. Nisar, M. Khurshid, and M. H. Rasool, “Proteomics: Technologies and Their Applications,” *J. Chromatogr. Sci.*, vol. 55, no. 2, pp. 182–196, Feb. 2017, doi: 10.1093/chromsci/bmw167.

[23] C. H. Johnson, J. Ivanisevic, and G. Siuzdak, “Metabolomics: beyond biomarkers and towards mechanisms,” *Nat. Rev. Mol. Cell Biol.*, vol. 17, no. 7, Art. no. 7, Jul. 2016, doi: 10.1038/nrm.2016.25.

[24] I. Trbojević-Akmačić *et al.*, “High-Throughput Glycomic Methods,” *Chem. Rev.*, vol. 122, no. 20, pp. 15865–15913, Oct. 2022, doi: 10.1021/acs.chemrev.1c01031.

[25] S. Goodwin, J. D. McPherson, and W. R. McCombie, “Coming of age: ten years of next-generation sequencing technologies,” *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 333–351, May 2016, doi: 10.1038/nrg.2016.49.

[26] R. R. Schaller, “Moore’s law: past, present and future,” *IEEE Spectr.*, vol. 34, no. 6, pp. 52–59, Jun. 1997, doi: 10.1109/6.591665.

[27] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, p. 160, Mar. 2021, doi: 10.1007/s42979-021-00592-x.

[28] V. N. Vapnik, “An overview of statistical learning theory,” *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999, doi: 10.1109/72.788640.

[29] A. Saxena *et al.*, “A review of clustering techniques and developments,” *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017, doi: 10.1016/j.neucom.2017.06.053.

[30] G. Munjal, M. Hanmandlu, and S. Srivastava, “Phylogenetics Algorithms and Applications,” *Ambient Commun. Comput. Syst.*, vol. 904, pp. 187–194, Dec. 2018, doi: 10.1007/978-981-13-5934-7_17.

[31] H. Li and N. Homer, “A survey of sequence alignment algorithms for next-generation sequencing,” *Brief. Bioinform.*, vol. 11, no. 5, pp. 473–483, Sep. 2010, doi: 10.1093/bib/bbq015.

[32] J. Gauthier, A. T. Vincent, S. J. Charette, and N. Derome, “A brief history of bioinformatics,” *Brief. Bioinform.*, vol. 20, no. 6, pp. 1981–1996, Nov. 2019, doi: 10.1093/bib/bby063.

[33] I. M. Johnstone and D. M. Titterington, “Statistical challenges of high-dimensional data,” *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 367, no. 1906, pp. 4237–4253, Nov. 2009, doi: 10.1098/rsta.2009.0159.

[34] K. J. Karczewski and M. P. Snyder, “Integrative omics for health and disease,” *Nat. Rev. Genet.*, vol. 19, no. 5, Art. no. 5, May 2018, doi: 10.1038/nrg.2018.4.

[35] P. T. Spellman *et al.*, “Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization,” *Mol. Biol. Cell*, vol. 9, no. 12, pp. 3273–3297, Dec. 1998, doi: 10.1091/mbc.9.12.3273.

[36] S. Horvath, “DNA methylation age of human tissues and cell types,” *Genome Biol.*, vol. 14, no. 10, p. R115, 2013, doi: 10.1186/gb-2013-14-10-r115.

[37] G. Hannum *et al.*, “Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates,” *Mol. Cell*, vol. 49, no. 2, pp. 359–367, Jan. 2013, doi: 10.1016/j.molcel.2012.10.016.

[38] T. Geiger *et al.*, “Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse,” *Mol. Cell. Proteomics MCP*, vol. 12, no. 6, pp. 1709–1722, Jun. 2013, doi: 10.1074/mcp.M112.024919.

[39] M. Bersanelli *et al.*, “Methods for the integration of multi-omics data: mathematical aspects,” *BMC Bioinformatics*, vol. 17, no. 2, p. S15, Jan. 2016, doi: 10.1186/s12859-015-0857-9.

[40] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge,” *Contemp. Oncol.*, vol. 19, no. 1A, pp. A68–A77, 2015, doi: 10.5114/wo.2014.47136.

[41] Z. Momeni, E. Hassanzadeh, M. Saniee Abadeh, and R. Bellazzi, “A survey on single and multi omics data mining methods in cancer data classification,” *J. Biomed. Inform.*, vol. 107, p. 103466, Jul. 2020, doi: 10.1016/j.jbi.2020.103466.

[42] J. N. Weinstein *et al.*, “The Cancer Genome Atlas Pan-Cancer analysis project,” *Nat. Genet.*, vol. 45, no. 10, Art. no. 10, Oct. 2013, doi: 10.1038/ng.2764.

[43] F. W. Albert and L. Kruglyak, “The role of regulatory variation in complex traits and disease,” *Nat. Rev. Genet.*, vol. 16, no. 4, pp. 197–212, Apr. 2015, doi: 10.1038/nrg3891.

[44] A. C. Nica and E. T. Dermitzakis, “Expression quantitative trait loci: present and future,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 368, no. 1620, p. 20120362, Jun. 2013, doi: 10.1098/rstb.2012.0362.

[45] S. Villicaña and J. T. Bell, “Genetic impacts on DNA methylation: research findings and future perspectives,” *Genome Biol.*, vol. 22, no. 1, p. 127, Apr. 2021, doi: 10.1186/s13059-021-02347-6.

[46] A. Adan, G. Alizada, Y. Kiraz, Y. Baran, and A. Nalbant, “Flow cytometry: basic principles and applications,” *Crit. Rev. Biotechnol.*, vol. 37, no. 2, pp. 163–176, Mar. 2017, doi: 10.3109/07388551.2015.1128876.

[47] A. Tanay and A. Regev, “Scaling single-cell genomics from phenomenology to mechanism,” *Nature*, vol. 541, no. 7637, Art. no. 7637, Jan. 2017, doi: 10.1038/nature21350.

[48] B. Hwang, J. H. Lee, and D. Bang, “Single-cell RNA sequencing technologies and bioinformatics pipelines,” *Exp. Mol. Med.*, vol. 50, no. 8, Art. no. 8, Aug. 2018, doi: 10.1038/s12276-018-0071-8.

[49] E. Z. Macosko *et al.*, “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets,” *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015, doi: 10.1016/j.cell.2015.05.002.

[50] G. X. Y. Zheng *et al.*, “Massively parallel digital transcriptional profiling of single cells,” *Nat. Commun.*, vol. 8, no. 1, Art. no. 1, Jan. 2017, doi: 10.1038/ncomms14049.

[51] S. Baek and I. Lee, “Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation,” *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1429–1439, Jan. 2020, doi: 10.1016/j.csbj.2020.06.012.

[52] A. A. Galitsyna and M. S. Gelfand, “Single-cell Hi-C data analysis: safety in numbers,” *Brief. Bioinform.*, vol. 22, no. 6, p. bbab316, Nov. 2021, doi: 10.1093/bib/bbab316.

[53] J. Ahn, S. Heo, J. Lee, and D. Bang, “Introduction to Single-Cell DNA Methylation Profiling Methods,” *Biomolecules*, vol. 11, no. 7, p. 1013, Jul. 2021, doi: 10.3390/biom11071013.

[54] R. Jiang, T. Sun, D. Song, and J. J. Li, “Statistics or biology: the zero-inflation controversy about scRNA-seq data,” *Genome Biol.*, vol. 23, no. 1, p. 31, Jan. 2022, doi: 10.1186/s13059-022-02601-5.

[55] T. Kim, I. R. Chen, Y. Lin, A. Y.-Y. Wang, J. Y. H. Yang, and P. Yang, “Impact of similarity metrics on single-cell RNA-seq data clustering,” *Brief. Bioinform.*, vol. 20, no. 6, pp. 2316–2326, Nov. 2019, doi: 10.1093/bib/bby076.

[56] C. Ahlmann-Eltze and W. Huber, “Comparison of transformations for single-cell RNA-seq data,” *Nat. Methods*, vol. 20, no. 5, Art. no. 5, May 2023, doi: 10.1038/s41592-023-01814-1.

[57] O. Ben-Kiki, A. Bercovich, A. Lifshitz, and A. Tanay, “Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis,” *Genome Biol.*, vol. 23, no. 1, p. 100, Apr. 2022, doi: 10.1186/s13059-022-02667-1.

[58] S. Sun, J. Zhu, Y. Ma, and X. Zhou, “Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis,” *Genome Biol.*, vol. 20, no. 1, p. 269, Dec. 2019, doi: 10.1186/s13059-019-1898-6.

[59] Y. Baran *et al.*, “MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions,” *Genome Biol.*, vol. 20, no. 1, p. 206, Oct. 2019, doi: 10.1186/s13059-019-1812-2.

[60] H. Ritchie, F. Spooner, and M. Roser, “Causes of death,” *Our World Data*, Feb. 2018, Accessed: Jun. 05, 2023. [Online]. Available: https://ourworldindata.org/causes-of-death

[61] J. Ma, E. M. Ward, R. L. Siegel, and A. Jemal, “Temporal Trends in Mortality in the United States, 1969-2013,” *JAMA*, vol. 314, no. 16, pp. 1731–1739, Oct. 2015, doi: 10.1001/jama.2015.12319.

[62] C. Mattiuzzi and G. Lippi, “Current Cancer Epidemiology,” *J. Epidemiol. Glob. Health*, vol. 9, no. 4, pp. 217–222, Dec. 2019, doi: 10.2991/jegh.k.191008.001.

[63] M. A. Feitelson *et al.*, “Sustained proliferation in cancer: mechanisms and novel therapeutic targets,” *Semin. Cancer Biol.*, vol. 35, no. Suppl, pp. S25–S54, Dec. 2015, doi: 10.1016/j.semcancer.2015.02.006.

[64] D. Hanahan and R. A. Weinberg, “The hallmarks of cancer,” *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000, doi: 10.1016/s0092-8674(00)81683-9.

[65] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 144, no. 5, pp. 646–674, Mar. 2011, doi: 10.1016/j.cell.2011.02.013.

[66] D. Hanahan, “Hallmarks of Cancer: New Dimensions,” *Cancer Discov.*, vol. 12, no. 1, pp. 31–46, Jan. 2022, doi: 10.1158/2159-8290.CD-21-1059.

[67] E. Martínez, K. Yoshihara, H. Kim, G. M. Mills, V. Treviño, and R. G. Verhaak, “Comparison of gene expression patterns across twelve tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects,” *Oncogene*, vol. 34, no. 21, pp. 2732–2740, May 2015, doi: 10.1038/onc.2014.216.

[68] T. A. Baudino, “Targeted Cancer Therapy: The Next Generation of Cancer Treatment,” *Curr. Drug Discov. Technol.*, vol. 12, no. 1, pp. 3–20, 2015, doi: 10.2174/1570163812666150602144310.

[69] A. M. Tsimberidou, E. Fountzilas, M. Nikanjam, and R. Kurzrock, “Review of precision cancer medicine: Evolution of the treatment paradigm,” *Cancer Treat. Rev.*, vol. 86, p. 102019, Jun. 2020, doi: 10.1016/j.ctrv.2020.102019.

[70] O. Yersal and S. Barutca, “Biological subtypes of breast cancer: Prognostic and therapeutic implications,” *World J. Clin. Oncol.*, vol. 5, no. 3, pp. 412–424, Aug. 2014, doi: 10.5306/wjco.v5.i3.412.

[71] J. J. Jimenez, R. S. Chale, A. C. Abad, and A. V. Schally, “Acute promyelocytic leukemia (APL): a review of the literature,” *Oncotarget*, vol. 11, no. 11, pp. 992–1003, Mar. 2020, doi: 10.18632/oncotarget.27513.

[72] X. An, A. K. Tiwari, Y. Sun, P.-R. Ding, C. R. Ashby, and Z.-S. Chen, “BCR-ABL tyrosine kinase inhibitors in the treatment of Philadelphia chromosome positive chronic myeloid leukemia: A review,” *Leuk. Res.*, vol. 34, no. 10, pp. 1255–1268, Oct. 2010, doi: 10.1016/j.leukres.2010.04.016.

[73] L. J. van ’t Veer *et al.*, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002, doi: 10.1038/415530a.

[74] C. M. Perou *et al.*, “Molecular portraits of human breast tumours,” *Nature*, vol. 406, no. 6797, Art. no. 6797, Aug. 2000, doi: 10.1038/35021093.

[75] D. C. Koboldt *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, Art. no. 7418, Oct. 2012, doi: 10.1038/nature11412.

[76] Cancer Genome Atlas Research Network, “Integrated genomic analyses of ovarian carcinoma,” *Nature*, vol. 474, no. 7353, pp. 609–615, Jun. 2011, doi: 10.1038/nature10166.

[77] Cancer Genome Atlas Network, “Genomic Classification of Cutaneous Melanoma,” *Cell*, vol. 161, no. 7, pp. 1681–1696, Jun. 2015, doi: 10.1016/j.cell.2015.05.044.

[78] Cancer Genome Atlas Network, “Comprehensive molecular characterization of human colon and rectal cancer,” *Nature*, vol. 487, no. 7407, pp. 330–337, Jul. 2012, doi: 10.1038/nature11252.

[79] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, “Clustering cancer gene expression data: a comparative study,” *BMC Bioinformatics*, vol. 9, no. 1, p. 497, Nov. 2008, doi: 10.1186/1471-2105-9-497.

[80] R. Sharan, A. Maron-Katz, and R. Shamir, “CLICK and EXPANDER: a system for clustering and visualizing gene expression data,” *Bioinformatics*, vol. 19, no. 14, pp. 1787–1799, Sep. 2003, doi: 10.1093/bioinformatics/btg232.

[81] E. A. Collisson *et al.*, “Comprehensive molecular profiling of lung adenocarcinoma,” *Nature*, vol. 511, no. 7511, Art. no. 7511, Jul. 2014, doi: 10.1038/nature13385.

[82] K. A. Hoadley *et al.*, “Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer,” *Cell*, vol. 173, no. 2, pp. 291-304.e6, Apr. 2018, doi: 10.1016/j.cell.2018.03.022.

[83] C. J. Vaske *et al.*, “Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM,” *Bioinformatics*, vol. 26, no. 12, pp. i237–i245, Jun. 2010, doi: 10.1093/bioinformatics/btq182.

[84] G. Chao, S. Sun, and J. Bi, “A Survey on Multiview Clustering,” *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 146–168, Apr. 2021, doi: 10.1109/TAI.2021.3065894.

[85] T. Cremer, C. Cremer, T. Schneider, H. Baumann, L. Hens, and M. Kirsch-Volders, “Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments,” *Hum. Genet.*, vol. 62, no. 3, pp. 201–209, 1982, doi: 10.1007/BF00333519.

[86] L. Manuelidis, “Individual interphase chromosome domains revealed by in situ hybridization,” *Hum. Genet.*, vol. 71, no. 4, pp. 288–293, 1985, doi: 10.1007/BF00388453.

[87] M. Cremer *et al.*, “Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes,” *Methods Mol. Biol. Clifton NJ*, vol. 463, pp. 205–239, 2008, doi: 10.1007/978-1-59745-406-3_15.

[88] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, “Capturing Chromosome Conformation,” *Science*, vol. 295, no. 5558, pp. 1306–1311, Feb. 2002, doi: 10.1126/science.1067799.

[89] M. Simonis *et al.*, “Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C),” *Nat. Genet.*, vol. 38, no. 11, pp. 1348–1354, Nov. 2006, doi: 10.1038/ng1896.

[90] E. Lieberman-Aiden *et al.*, “Comprehensive mapping of long range interactions reveals folding principles of the human genome,” *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009, doi: 10.1126/science.1181369.

[91] M. J. Fullwood, C.-L. Wei, E. T. Liu, and Y. Ruan, “Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses,” *Genome Res.*, vol. 19, no. 4, pp. 521–532, Apr. 2009, doi: 10.1101/gr.074906.107.

[92] M. R. Mumbach *et al.*, “HiChIP: efficient and sensitive analysis of protein-directed genome architecture,” *Nat. Methods*, vol. 13, no. 11, Art. no. 11, Nov. 2016, doi: 10.1038/nmeth.3999.

[93] T.-H. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando, “Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C,” *Cell*, vol. 162, no. 1, pp. 108–119, Jul. 2015, doi: 10.1016/j.cell.2015.05.048.

[94] E. Hildebrand and J. Dekker, “Mechanisms and functions of chromosome compartmentalization,” *Trends Biochem. Sci.*, vol. 45, no. 5, pp. 385–396, May 2020, doi: 10.1016/j.tibs.2020.01.002.

[95] T. Misteli, “The Self-Organizing Genome: Principles of Genome Architecture and Function,” *Cell*, vol. 183, no. 1, pp. 28–45, Oct. 2020, doi: 10.1016/j.cell.2020.09.014.

[96] B. D. Pope *et al.*, “Topologically associating domains are stable units of replication-timing regulation,” *Nature*, vol. 515, no. 7527, Art. no. 7527, Nov. 2014, doi: 10.1038/nature13986.

[97] S. S. P. Rao *et al.*, “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping,” *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014, doi: 10.1016/j.cell.2014.11.021.

[98] J. A. Beagan and J. E. Phillips-Cremins, “On the existence and functionality of topologically associating domains,” *Nat. Genet.*, vol. 52, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41588-019-0561-1.

[99] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, “Formation of Chromosomal Domains by Loop Extrusion,” *Cell Rep.*, vol. 15, no. 9, pp. 2038–2049, May 2016, doi: 10.1016/j.celrep.2016.04.085.

[100] M. Ganji *et al.*, “Real-time imaging of DNA loop extrusion by condensin,” *Science*, vol. 360, no. 6384, pp. 102–105, Apr. 2018, doi: 10.1126/science.aar7831.

[101] M. Gabriele *et al.*, “Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging,” *Science*, vol. 376, no. 6592, pp. 496–501, Apr. 2022, doi: 10.1126/science.abn6583.

[102] K. P. Eagen, “Principles of Chromosome Architecture Revealed by Hi-C,” *Trends Biochem. Sci.*, vol. 43, no. 6, pp. 469–478, Jun. 2018, doi: 10.1016/j.tibs.2018.03.006.

[103] A.-L. Valton and J. Dekker, “TAD disruption as oncogenic driver,” *Curr. Opin. Genet. Dev.*, vol. 36, pp. 34–40, Feb. 2016, doi: 10.1016/j.gde.2016.03.008.

[104] M. Spielmann, D. G. Lupiáñez, and S. Mundlos, “Structural variation in the 3D genome,” *Nat. Rev. Genet.*, vol. 19, no. 7, Art. no. 7, Jul. 2018, doi: 10.1038/s41576-018-0007-0.

[105] A. T. L. Lun and G. K. Smyth, “diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data,” *BMC Bioinformatics*, vol. 16, no. 1, p. 258, Aug. 2015, doi: 10.1186/s12859-015-0683-0.

[106] H. Spemann and H. Mangold, “Induction of embryonic primordia by implantation of organizers from a different species. 1923,” *Int. J. Dev. Biol.*, vol. 45, no. 1, pp. 13–38, 2001.

[107] M. Boareto, “Patterning via local cell-cell interactions in developing systems,” *Dev. Biol.*, vol. 460, no. 1, pp. 77–85, Apr. 2020, doi: 10.1016/j.ydbio.2019.12.008.

[108] F. Spitz and E. E. M. Furlong, “Transcription factors: from enhancer binding to developmental control,” *Nat. Rev. Genet.*, vol. 13, no. 9, Art. no. 9, Sep. 2012, doi: 10.1038/nrg3207.

[109] K. W. Rogers and A. F. Schier, “Morphogen Gradients: From Generation to Interpretation,” *Annu. Rev. Cell Dev. Biol.*, vol. 27, no. 1, pp. 377–407, 2011, doi: 10.1146/annurev-cellbio-092910-154148.

[110] K. Takahashi *et al.*, “Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors,” *Cell*, vol. 131, no. 5, pp. 861–872, Nov. 2007, doi: 10.1016/j.cell.2007.11.019.

[111] R. D. Riddle, R. L. Johnson, E. Laufer, and C. Tabin, “Sonic hedgehog mediates the polarizing activity of the ZPA,” *Cell*, vol. 75, no. 7, pp. 1401–1416, Dec. 1993, doi: 10.1016/0092-8674(93)90626-2.

[112] Y. Litingtung and C. Chiang, “Control of Shh activity and signaling in the neural tube,” *Dev. Dyn. Off. Publ. Am. Assoc. Anat.*, vol. 219, no. 2, pp. 143–154, Oct. 2000, doi: 10.1002/1097-0177(2000)9999:9999<::aid-dvdy1050>3.3.co;2-h.

[113] N. Nishioka *et al.*, “The Hippo signaling pathway components Lats and Yap pattern Tead4 activity to distinguish mouse trophectoderm from inner cell mass,” *Dev. Cell*, vol. 16, no. 3, pp. 398–410, Mar. 2009, doi: 10.1016/j.devcel.2009.02.003.

[114] Y. Zeng and T. Chen, “DNA Methylation Reprogramming during Mammalian Development,” *Genes*, vol. 10, no. 4, p. 257, Mar. 2019, doi: 10.3390/genes10040257.

[115] Z. D. Smith and A. Meissner, “DNA methylation: roles in mammalian development,” *Nat. Rev. Genet.*, vol. 14, no. 3, pp. 204–220, Mar. 2013, doi: 10.1038/nrg3354.

[116] A. Jambhekar, A. Dhall, and Y. Shi, “Roles and regulation of histone methylation in animal development,” *Nat. Rev. Mol. Cell Biol.*, vol. 20, no. 10, Art. no. 10, Oct. 2019, doi: 10.1038/s41580-019-0151-1.

[117] A. Piunti and A. Shilatifard, “The roles of Polycomb repressive complexes in mammalian development and cancer,” *Nat. Rev. Mol. Cell Biol.*, vol. 22, no. 5, pp. 326–345, May 2021, doi: 10.1038/s41580-021-00341-1.

[118] M. Mittnenzweig *et al.*, “A single-embryo, single-cell time-resolved model for mouse gastrulation,” *Cell*, vol. 184, no. 11, pp. 2825-2842.e22, May 2021, doi: 10.1016/j.cell.2021.04.004.

[119] B. Pijuan-Sala *et al.*, “A single-cell molecular map of mouse gastrulation and early organogenesis,” *Nature*, vol. 566, no. 7745, Art. no. 7745, Feb. 2019, doi: 10.1038/s41586-019-0933-9.

[120] T. Nagano *et al.*, “Single-cell Hi-C reveals cell-to-cell variability in chromosome structure,” *Nature*, vol. 502, no. 7469, Art. no. 7469, Oct. 2013, doi: 10.1038/nature12593.

[121] T. Nagano *et al.*, “Cell-cycle dynamics of chromosomal organization at single-cell resolution,” *Nature*, vol. 547, no. 7661, Art. no. 7661, Jul. 2017, doi: 10.1038/nature23001.

[122] T. J. Stevens *et al.*, “3D structures of individual mammalian genomes studied by single-cell Hi-C,” *Nature*, vol. 544, no. 7648, Art. no. 7648, Apr. 2017, doi: 10.1038/nature21429.

[123] V. Ramani *et al.*, “Massively multiplex single-cell Hi-C,” *Nat. Methods*, vol. 14, no. 3, Art. no. 3, Mar. 2017, doi: 10.1038/nmeth.4155.

[124] D.-S. Lee *et al.*, “Simultaneous profiling of 3D genome structure and DNA methylation in single human cells,” *Nat. Methods*, vol. 16, no. 10, pp. 999–1006, Oct. 2019, doi: 10.1038/s41592-019-0547-z.

[125] J. Zhou *et al.*, “Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation,” *Proc. Natl. Acad. Sci.*, vol. 116, no. 28, pp. 14011–14018, Jul. 2019, doi: 10.1073/pnas.1901423116.

[126] H.-J. Kim *et al.*, “Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data,” *PLOS Comput. Biol.*, vol. 16, no. 9, p. e1008173, Sep. 2020, doi: 10.1371/journal.pcbi.1008173.

[127] R. Zhang, T. Zhou, and J. Ma, “Multiscale and integrative single-cell Hi-C analysis with Higashi,” *Nat. Biotechnol.*, vol. 40, no. 2, Art. no. 2, Feb. 2022, doi: 10.1038/s41587-021-01034-y.

[128] R. Duan *et al.*, “Evaluation and comparison of multi-omics data integration methods for cancer subtyping,” *PLoS Comput. Biol.*, vol. 17, no. 8, p. e1009224, Aug. 2021, doi: 10.1371/journal.pcbi.1009224.

[129] C. Nießl, S. Hoffmann, T. Ullmann, and A.-L. Boulesteix, “Explaining the optimistic performance evaluation of newly proposed methods: A cross-design validation experiment,” *Biom. J.*, vol. n/a, no. n/a, p. 2200238, doi: 10.1002/bimj.202200238.

[130] H. G, G. M, and S. M, “Exact log-rank tests for unequal follow-up,” *Biometrics*, vol. 59, no. 4, Dec. 2003, doi: 10.1111/j.0006-341x.2003.00132.x.

[131] W. R, L. Sw, and G. Rj, “Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring,” *Biostat. Oxf. Engl.*, vol. 11, no. 4, Oct. 2010, doi: 10.1093/biostatistics/kxq021.

[132] J. Liu, C. Wang, J. Gao, and J. Han, “Multi-View Clustering via Joint Nonnegative Matrix Factorization,” in *Proceedings of the 2013 SIAM International Conference on Data Mining (SDM)*, in Proceedings. Society for Industrial and Applied Mathematics, 2013, pp. 252–260. doi: 10.1137/1.9781611972832.28.

[133] H. Hotelling, “Relations Between Two Sets of Variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936, doi: 10.2307/2333955.

[134] B. Xie, Y. Mu, and D. Tao, “m-SNE: multiview stochastic neighbor embedding,” in *Proceedings of the 17th international conference on Neural information processing: theory and algorithms - Volume Part I*, in ICONIP’10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 338–346.

[135] V. H. Do and S. Canzar, “A generalization of t-SNE and UMAP to single-cell multimodal omics,” *Genome Biol.*, vol. 22, no. 1, p. 130, May 2021, doi: 10.1186/s13059-021-02356-5.

[136] C. Angermueller *et al.*, “Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity,” *Nat. Methods*, vol. 13, no. 3, pp. 229–232, Mar. 2016, doi: 10.1038/nmeth.3728.

[137] S. J. Clark *et al.*, “scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells,” *Nat. Commun.*, vol. 9, no. 1, Art. no. 1, Feb. 2018, doi: 10.1038/s41467-018-03149-4.

[138] Y. Hao *et al.*, “Integrated analysis of multimodal single-cell data,” *Cell*, vol. 184, no. 13, pp. 3573-3587.e29, Jun. 2021, doi: 10.1016/j.cell.2021.04.048.

[139] “Method of the Year 2019: Single-cell multimodal omics,” *Nat. Methods*, vol. 17, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41592-019-0703-5.

5. **Single cell Hi-C identifies plastic chromosome conformations underlying the gastrulation enhancer landscape**

   Nimrod Rappoport*, Elad Chomsky*, Takashi Nagano, Charlie Seibert, Yaniv Lubling, Yael Baran, Aviezer Lifshitz, Wing Leung, Zohar Mukamel, Ron Shamir, Peter Fraser, Amos Tanay; Accepted to Nature Communications.

   * denotes equal contribution.

התפתחות עוברית כרוכה בפרוליפרציה והתמיינות מסיבית של שושלות תאים. תהליכים אלה חייבים להיתמך על ידי שכפול כרומוזומים ותכנות מחדש (reprogramming) אפיגנטי, אבל האופן שבו פרוליפרציה ורכישת גורל התא מאוזנים בתהליך זה אינו מובן היטב. כאן אנו משתמשים ב-Hi-C של תא בודד (single-cell Hi-C) כדי למפות קונפורמציות כרומוזומליות בתאים של עוברי עכברים לאחר גסטרולציה, וכדי לחקור את ההתפלגויות והמתאמים שלהם עם אטלסים תואמים של שעתוק בעוברים. אנו מגלים שכרומוזומים עובריים מראים חתימה חזקה להפליא של מחזור התא. למרות זאת, תזמון השכפול, מבנה הכרומוזומים, דומיינים קשורים-טופולוגית (TADs) וקשרים בין מקדמים ומעצמים (promoter-enhancer) משתנים בין מצבים אפיגנטיים שונים. כ-10% מהגרעינים מזוהים כתאי דם אדומים פרימיטיביים, המראים מבנה תא קומפקטי ומאורגן במיוחד. התאים הנותרים קשורים באופן נרחב לזהות אקטודרם ומזודרם, ומציגים רק התמיינות מתונה של TADs, אך קשרים מקומיים ספציפיים יותר במאות זוגות מקדמים-מעצמים באקטודרם ומזודרם. הנתונים מצביעים על כך שבעוד שושלות עובריות ממוינות יכולות לרכוש במהירות קונפורמציות כרומוזומליות ספציפיות, רוב התאים העובריים מראים חתימות פלסטיות המונעות על ידי מצבי מעצמים מורכבים.

בנתונים חלקיים. בחלק מבדיקות הנתונים החלקיים, PCV, אלגוריתם מרובה תצוגה, הציג ביצועים טובים יותר, אך הוא מוגבל לשני אומיקים ולנתונים חלקיים חיוביים. לבסוף, אנו מדגימים את היתרון של NEMO בניתוח מפורט של נתונים חלקיים של חולי לוקמיה מיאלואידית חריפה (Acute Myeloid Leukemia). האלגוריתם NEMO הוא מהיר ופשוט הרבה יותר מאלגוריתמים קיימים של קיבוץ נתונים מרובי אומיק, ונמנע מאופטימיזציה איטרטיבית.

3. **Inaccuracy of the log-rank approximation in cancer data analysis**
   Nimrod Rappoport, Ron Shamir; Mol Syst Biol. (2019) 15: e8754.

השוואת ההישרדות בין קבוצות שונות של חולים נמצאת בשימוש נרחב בחקר הסרטן, וכאמצעי השוואה בין אלגוריתמי קיבוץ של נתונים מחולי סרטן. המבחן הסטטיסטי הנפוץ ביותר עבור השוואות מסוג זה הוא מבחן ה-log-rank. בעבודה זו אנו מראים שרוב כלי התוכנה משתמשים בגרסה אסימפטוטית של המבחן, שהיא מאוד לא מדויקת בבסיסי נתונים עם מספר החולים שנמצאים בבסיסי נתונים של סרטן. אנו מראים כי ערכי המובהקות המדווחים מפריזים במובהקות התוצאות, מצביעים על תגליות שווא קודמות שנעשו באמצעות בדיקה זו, ומספקים מימוש לבדיקה מדויקת עבור מספר קבוצות.

4. **MONET: Multi-omic module discovery by omic selection**
   Nimrod Rappoport, Roy Safra, Ron Shamir; PLOS Computational Biology 16(9): e1008182.

התקדמויות עכשוויות בביולוגיה ניסויית מאפשרות יצירת מערכי נתונים שבהם נמדדים מספר סוגי נתונים על-פני כל הגנום (הנקראים אומיקים) עבור כל דגימה. ניתוח אינטגרטיבי של מערכי נתונים מרובי אומיק באופן כללי, וקיבוץ דגימות במערכי נתונים כאלה באופן ספציפי, יכולים לשפר את ההבנה שלנו של תהליכים ביולוגיים ולגלות תת-סוגים שונים של מחלות. בעבודה זו אנו מציגים את האלגוריתם MONET, המציג גישה ייחודית לקיבוץ נתונים מרובי אומיק. MONET מגלה מודולים של דגימות דומות, כך שכל מודול רשאי להוות קיבוץ רק עבור תת-קבוצה של האומיקים. גישה זו נבדלת מרוב אלגוריתמי הקיבוץ הקיימים, אשר מניחים מבנה קיבוץ משותף על פני כל האומיקים, וממספר אלגוריתמים עדכניים הממדלים מבני קיבוץ שונים. בדקנו את MONET בהרחבה על נתונים מדומים, על נתוני תמונות ועל עשרה מערכי נתונים מרובי אומיק של סרטן מ-TCGA. הניתוח שלנו מראה כי ל-MONET ביצועים טובים ביחס לשיטות אחרות לקיבוץ נתונים מרובי אומיק. אנו מדגימים את הרלוונטיות הביולוגית והקלינית של MONET על ידי ניתוח תוצאותיו עבור סרטן מסוג ציסט אדנוקרצינומה של השחלות. אנו גם מראים כי MONET עמיד לנתונים חסרים, יכול לקבץ גנים בנתונים מרובי אומיק, ולחשוף מודולים של סוגי תאים בנתונים מרובי אומיק מתאים בודדים. העבודה שלנו מראה כי MONET הוא כלי רב ערך שיכול לספק תוצאות משלימות לאלו המסופקות על ידי אלגוריתמים קיימים לניתוח נתונים מרובי אומיק.

תקציר המאמרים הכלולים בתזה

להלן תקצירי המאמרים עליהם מבוססת עבודה זו:

1. **Multi-omic and multi-view clustering algorithms: review and cancer benchmark**
   Nimrod Rappoport, Ron Shamir; Nucleic Acids Research, Volume 46, Issue 20, 16 November 2018, Pages 10546–10562.

לאחרונה נעשה שימוש בשיטות ניסוייות בתפוקה-גבוהה (high-throughput) לאיסוף בסיסי נתונים ביו-רפואיים גדולים. קיבוץ (clustering) של בסיסי נתונים של אומיק בודד (single-omic) הוכח כבעל ערך רב עבור מחקר ביולוגי ורפואי. העלות היורדת והפיתוח של שיטות נוספות עם תפוקה גבוהה מאפשרים כעת מדידה של נתונים מרובי אומיק (multi-omic). לקיבוץ נתונים מרובי אומיק יש פוטנציאל לחשוף תובנות נוספות ברמה המערכתית, אך הוא מעורר אתגרים חישוביים וביולוגיים. בעבודה זו, אנו סוקרים אלגוריתמים לקיבוץ נתונים מרובי אומיק, ודנים בסוגיות מפתח ביישום האלגוריתמים הללו. הסקירה שלנו מכסה שיטות שפותחו במיוחד עבור נתוני אומיק, כמו גם שיטות מרובות תצוגה (multi-view) גנריות שפותחו בקהילת המחקר של למידת מכונה (machine learning) לקיבוץ משותף של מספר סוגי נתונים. בנוסף, תוך שימוש בנתוני סרטן מ-TCGA, אנו מבצעים השוואת ביצועים (benchmark) נרחב המשתרע על פני עשרה סוגי סרטן שונים, ומספק את ההשוואה השיטתית הראשונה של אלגוריתמים מרובי אומיק ומרובי תצוגה מובילים. התוצאות מדגישות סוגיות מפתח בנוגע לשימוש באומיק יחיד מול רבים, בחירת אסטרטגיית קיבוץ, הכוח של שיטות מרובות תצוגה גנריות ושימוש בערכי מובהקות (p-values) מקורבים למדידת איכות הפתרון. בשל השימוש הגובר בנתוני אומיק מרובים, אנו מצפים כי נושאים אלו יהיו חשובים להתקדמות עתידית בתחום.

2. **NEMO: cancer subtyping by integration of partial multi-omic data**
   Nimrod Rappoport, Ron Shamir; Bioinformatics, Volume 35, Issue 18, September 2019, Pages 3348–3356.

מוטיבציה: תתי סוגים של סרטן הוגדרו בדרך כלל על סמך אפיון מולקולרי של נתוני אומיק בודדים. יותר ויותר, מדידות של מספר פרופילי אומיק עבור אותה עוקבה (cohort) נעשות זמינות. הגדרת תתי-סוגים של סרטן באמצעות נתונים מרובי אומיק עשויה לשפר את ההבנה שלנו לגבי סרטן, ולהציע טיפול מדויק יותר לחולים.

תוצאות: אנו מציגים את NEMO, אלגוריתם חדשני לקיבוץ נתונים מרובי אומיקים. מאפיין חשוב של NEMO הוא שניתן להפעיל אותו על מערכי נתונים חלקיים, בהם לחלק מהמטופלים יש נתונים עבור תת-קבוצה של האומיקים, מבלי לבצע הערכה של הנתונים החסרים (imputation). בבדיקות מקיפות על עשרה מערכי נתונים של סרטן המשתרעים על פני 3168 חולים, NEMO השיג תוצאות דומות לטוב מבין תשעה אלגוריתמים מתקדמים של קיבוץ נתונים מרובי אומיק מלאים, והראה שיפור

תוצאות

בתיזה זו אנו מפתחים אלגוריתמים לניתוח נתונים מרובי אומיק, ונתונים מתאים בודדים. ספציפית, הבעיות שאנו בוחנים הן הגדרת תתי סוגי סרטן מנתונים מרובי אומיק, ואפיון המבנה המרחבי של הגנום בהתפתחות עוברית של עכבר באמצעות נתונים מתאים בודדים.

בפרק 2 אנו מבצעים השוואה בין אלגוריתמים להגדרת תתי סוגים של סרטן באמצעות נתונים מרובי אומיק. לצורך ההשוואה אנחנו בוחנים הן אלגוריתמים שפותחו במיוחד עבור מידע ביולוגי, והן אלגוריתמים שפותחו על ידי קהילת המחקר שחוקרת למידת מכונה שלא בהקשר ביולוגי. אנו מראים שאלגוריתמים שפותחו עבור נתונים ביולוגיים הם לאו דווקא בעלי ביצועים טובים יותר, ושלא תמיד כדאי להשתמש בכל סוגי האומיק הזמינים.

בפרק 3 אנו מציגים אלגוריתם שפיתחנו, NEMO, עבור הגדרת תתי סוגים של סרטן. אנו מראים את ביצועיו הטובים של NEMO ביחס לאלגוריתמים אחרים על-פני עשרה סוגי סרטן שונים. כמו כן, אנו מראים כיצד NEMO מסוגל לעבוד על מאגרי מידע חלקיים, בהם עבור חלק מהדגימות נמדדו רק חלק מהאומיקים. לבסוף, אנו משתמשים ב-NEMO על מנת להגדיר תתי סוגים של סרטן דם מסוג לוקמיה מיאלואידית חריפה.

בפרק 4 אנו מציגים בעיה במבחן סטטיסטי שמשווה בין שרידות של קבוצות. המבחן הוא מבחן אסימפטוטי, ואנו מראים שערכי המובהקות המדווחים הם נמוכים (כלומר, מובהקים יותר) מערכי המובהקות האמיתיים. בנוסף אנו מספקים מימוש עבור גרסה מדויקת של המבחן.

פרק 5 עוסק ב-MONET, אלגוריתם נוסף שפיתחנו עבור הגדרת תתי סוגים של סרטן מנתונים מרובי אומיק. לאור התוצאות שהוצגו בפרק 2, לפיהן לא תמיד כדאי להשתמש בכל סוגי האומיק, MONET מוצא קבוצות של חולי סרטן שדומים זה לזה רק בתת קבוצה של האומיקים הזמינים. אנו משווים את MONET לאלגוריתמים אחרים ומנתחים באמצעותו נתוני סרטן של ציסט אדנוקרצינומה של השחלות. בנוסף, אנחנו מפעילים את MONET על נתונים שהם גם מרובי אומיקים וגם מתאים בודדים בתהליך ההתפתחות העוברית, ומראים כיצד MONET עוזר לאפיין תהליכים אפיגנטיים בהתמיינות.

פרק 6 עוסק בנתונים מתאים בודדים, כאשר הנתון הנמדד מהם הוא המבנה המרחבי של הגנום שלהם. התאים נלקחו מעוברי עכבר בני תשעה ימים. אנו מפתחים שיטות ניתוח ייעודיות עבור סוג נתונים זה, ומראים שתאי הדם של עוברי עכבר הם בעלי מבנה DNA שונה מהותית מתאים אחרים. בנוסף, אנו מוצאים חלוקה של שאר תאי העכבר לקבוצות תאים בעלות מבנה דומה, תוך התחשבות בשינויים המרחביים שמתרחשים בגנום כתוצאה ממחזור התא. לסיום, אנו מראים כיצד מבנה הגנום מתקשר לשכבות בקרה אפיגנטיות אחרות בעת ההתפתחות העוברית.

רקע כללי

המחקר הביולוגי והרפואי התקדם מאז ומעולם לצד פיתוח כלים טכנולוגיים, שאפשרו למדוד ולאפיין מערכות ביולוגיות בדיוק הולך וגובר. טכנולוגיות אלה התאפיינו ברזולוציה הולכת וגדלה, ומעבר למחקר שמתבצע ברמה המולקולרית. לקראת סוף המאה ה-20 החלו להתפתח שיטות שמאפשרות לבצע עשרות אלפי מדידות מולקולריות על קבוצת תאים. עם התפתחות טכנולוגיית ריצוף DNA, ולאור הוזלתה המהירה, שיטות שמסוגלות לבצע מדידות רבות על קבוצת תאים תוך שימוש בריצוף DNA נעשו נפוצות יותר. כל שיטה כזו אוספת מידע מסוג מסוים, אשר נקרא "אומיק" (omic).

השיפורים בטכנולוגיות למדידת מידע אומיק הובילו להיווצרות של שני סוגי מאגרי מידע ייחודיים. הסוג הראשון הוא מאגרי מידע בהם מבצעים לא רק ניסוי אחד על כל דגימה ביולוגית, אלא מספר רב של ניסויי אומיק (multi-omic). כך, במאגר המידע יש דגימות רבות, שעל כל אחת מהן בוצעו מספר ניסויי אומיק, וכל ניסוי בדק רבבות של פרמטרים שונים. הסוג השני קשור לרגישות ההולכת וגדלה של ניסויי האומיק. ניסויים אלה מדדו תחילה מאפיינים מולקולריים של קבוצת תאים, כאשר הערך שנמדד הוא ערך ממוצע על פני כל התאים בניסוי. בשנים האחרונות פותחו שיטות שמאפשרות לבצע ניסויי אומיק כך שהמדידה נעשית ברמת התא הבודד (single-cell), כך שרבבות של מאפיינים מולקולריים נמדדים על מספר רב של תאים בניסוי יחיד.

באותן שנים של התפתחות בשיטות המחקר הביולוגיות, נעשו קפיצות דרך משמעותיות גם במדעי המחשב, וספציפית ביכולות החישוב ובאלגוריתמים לניתוח מידע רחב היקף. לאור כמויות הנתונים הגדולות שנוצרות בניסויי אומיק, ניתוח הנתונים יכול להתבצע רק בכלים חישוביים, והשילוב בין כמויות הנתונים הביולוגיים והאלגוריתמים לניתוחם תרם לצמיחת תחום הביואינפורמטיקה. אולם לא כל אלגוריתם שפותח במדעי המחשב מתאים בהכרח למידע ביולוגי, ויש צורך בפיתוח אלגוריתמים ייעודיים שמתאימים לנתונים ביולוגיים ולשאלות מחקר ביולוגיות. בפרט, קיים צורך באלגוריתמים לניתוח מידע ממספר ניסויי אומיק, ובאלגוריתמים לניתוח מידע מתאים בודדים.

סרטן היא אחת המחלות שעברו את השינוי המהותי ביותר בזכות ניסויי אומיק. במהלך המאה ה-20 התחדדה ההבנה שסרטן היא למעשה קבוצה של מחלות, כאשר ייתכנו הבדלים גדולים אפילו בין סרטנים מאותה הרקמה. ניסויי אומיק אפשרו להגדיר את ההבדלים בין סרטנים שונים בצורה מדויקת, ואף ליצור כלים שמסייעים לרופאים בקבלת החלטות טיפוליות.

תחום ביולוגי נוסף בו נעשה שימוש בניסויי אומיק הוא אפיגנטיקה. תחום זה עוסק בשאלת הזהות התאית – כיצד ייתכן שתאים שונים באותו אורגניזם, שלהם אותו DNA, מראים הבדלים כה גדולים זה מזה. האפיגנטיקה מתארת כיצד תאים "מתמיינים" ומתמחים בביצוע פעולות ביולוגיות מסוימות, ומה המנגנונים המולקולריים שמעורבים בכך. ניסויי אומיק שונים פותחו על מנת לאפיין רמות שונות של בקרה אפיגנטית, כולל ניסויים שמבצעים מדידות ברמת התא הבודד. מאפיין אפיגנטי שחשיבותו הולכת ומתבהרת הוא סידורה המרחבי של מולקולת ה-DNA.

תמצית

שתי מגמות מרכזיות ביצירת נתונים ביו-רפואיים הפכו בולטות בשנים האחרונות. ראשית, שיטות ניסוי יכולות כעת למדוד מספר סוגים שונים של פרמטרים מולקולריים עבור רקמות ביולוגיות. כל סוג כזה נקרא אומיק (omic), ומערכי נתונים מרובי-אומיק - שבהם נמדדים מספר אומיקים עבור כל מדגם - הופכים נפוצים יותר. שנית, ניסויים חדשניים יכולים כעת למדוד נתוני אומיק ברזולוציית תא בודד, במקום למדוד ממוצעים על פני כל התאים ברקמה.

בעוד זמינותם של מערכי נתונים מרובי-אומיק ונתונים מתאים בודדים הולכת וגדלה, אלגוריתמים לניתוח שלהם עדיין חסרים. האלגוריתמים הקיימים אינם מתייחסים למספר מאפיינים של מערכי נתונים מרובי-אומיק, כגון נוכחות של נתונים חלקיים, והמבנה השונה של הנתונים באומיקים שונים. עבור נתונים מתאים בודדים, השיטות הקיימות אינן מספקות מודלים פרמטריים המנצלים את מלוא היתרונות של המדידות הבודדות כדי לחלץ ידע ביולוגי חדש.

בעבודה זו פיתחנו שיטות לניתוח מערכי נתונים מרובי-אומיק ונתונים מתאים בודדים. עבור נתונים מרובי-אומיק, שיטות אלו התמקדו באפיון טוב יותר של תתי סוגים של סרטן. עבור נתונים של תא בודד, הן התמקדו בהבנת מאפיינים אפיגנטיים של תאים, ובמיוחד בארגון הגנום באמצעות נתוני Hi-C מתאים בודדים.

אוניברסיטת תל-אביב TEL AVIV UNIVERSITY

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלבטניק

שיטות לניתוח נתונים מ-"אומיקים" מרובים ומתאים בודדים

חיבור לשם קבלת תואר "דוקטור לפילוסופיה"

מאת **נמרוד רפופורט**

בהנחייתם של **פרופ' רון שמיר** ו**פרופ' עמוס תנאי**

הוגש לסנאט של אוניברסיטת תל אביב

יוני 2023