BMC Bioinformatics

SOFTWARE

Open Access

Pattern recognition in lymphoid malignancies using CytoGPS and Mercator

Zachary B. Abrams[1*], Dwayne G. Tally[2], Lin Zhang[3], Caitlin E. Coombes[1], Philip R. O. Payne[3], Lynne V. Abruzzo[4] and Kevin R. Coombes[1]

*Correspondence:
Zachary.Abrams@osumc.edu
[1] Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA
Full list of author information is available at the end of the article

Abstract

Background: There have been many recent breakthroughs in processing and analyzing large-scale data sets in biomedical informatics. For example, the CytoGPS algorithm has enabled the use of text-based karyotypes by transforming them into a binary model. However, such advances are accompanied by new problems of data sparsity, heterogeneity, and noisiness that are magnified by the large-scale multidimensional nature of the data. To address these problems, we developed the Mercator R package, which processes and visualizes binary biomedical data. We use Mercator to address biomedical questions of cytogenetic patterns relating to lymphoid hematologic malignancies, which include a broad set of leukemias and lymphomas. Karyotype data are one of the most common form of genetic data collected on lymphoid malignancies, because karyotyping is part of the standard of care in these cancers.

Results: In this paper we combine the analytic power of CytoGPS and Mercator to perform a large-scale multidimensional pattern recognition study on 22,741 karyotype samples in 47 different hematologic malignancies obtained from the public Mitelman database.

Conclusion: Our findings indicate that Mercator was able to identify both known and novel cytogenetic patterns across different lymphoid malignancies, furthering our understanding of the genetics of these diseases.

Keywords: Karyotype, Pattern recognition, CytoGPS, Mercator, Lymphoid malignancies

Background

Cytogenetics

As biology and medicine advance, our ability to generate ever-increasing amounts of data also expands [1]. While a boon for biomedical research, this increase in the volume and diversity of data poses challenges to data scientists [2]. Issues of noisiness, dimensionality, and heterogeneity can prove problematic when performing large-scale biomedical analyses [3]. These problems become more common and more severe as larger, higher-dimensional data sets are collected; as a result, some biomedical data remain underused.

For years, technical issues have limited the use of karyotype data in secondary computational analyses. Karyotype data are one of the most common forms of genetic information collected on patients, since cytogenetic karyotyping is part of the standard of care for most hematologic malignancies [4]. The current standard for large-scale cytogenetic analyses is manual classification by cytogenetic pathologists. This is extremely time consuming and can introduce human error into downstream analysis. Thus, these data have not been used in large-scale computational analyses because of the text-based standard format in which they are recorded [4]. In response, we recently developed the CytoGenetic Pattern Sleuth (CytoGPS), a tool that converts karyotypes into binary vectors that can be analyzed using modern computational methods [5]. However, CytoGPS is only a first step in understanding and exploring these data, since it merely enables (but does not carry out) the application of pattern recognition methods. To actually apply such methods systematically, we developed the Mercator R package. Mercator provides a consistent, unified interface to a suite of unsupervised pattern recognition algorithms. Mercator uses 10 distance metrics between binary vectors, selected from 76 metrics described and classified in a review paper by Choi et al. [6], to provide a representative sample of this wide scope of different metrics. Mercator also supports five data visualization methods designed for both standard and high dimensional data analysis; the visualization tools work with arbitrary distance metrics for any data type, not just binary. Mercator makes it easy to produce visualizations with consistent color schemes. More importantly, since cluster labels from different unsupervised algorithms are arbitrary, Mercator includes tools to synchronize and compare these labels. Thus, Mercator enables the exploratory unsupervised analysis of large, high-dimensional data sets, accompanied by clear, easy visualizations.

Research design

In this article, we apply CytoGPS and Mercator to understand the structure of a data set containing more than 22,000 karyotypes from lymphoid malignancies. Lymphoid cells are one of the two most common cell types from which leukemias and lymphomas are derived, the other cell type being myeloid cells [7, 8]. Lymphoid cells include B cells, T cells, and natural killer (NK) cells. The current World Health Organization (WHO) classification of lymphoid malignancies [9] incorporates a variety of factors, including cytogenetics, cell-of-origin, location (bone marrow, blood, lymph node, etc.), clinical findings, immunophenotype, histological patterns (e.g., follicular or diffuse), and mutations or rearrangements of specific genes. The current WHO classification of lymphoid malignancies includes at least 60 subtypes; the historical karyotype data from the Mitelman. Database of Chromosome Aberrations and Gene Fusions in Cancer includes 47 named subtypes [10]. These subtypes include chronic lymphocytic leukemia (CLL), which is known to include at least four prognostic subgroups defined by different cytogenetic abnormalities [11]. Similarly, various cytogenetic abnormalities are known to be prognostic in subsets of acute lymphocytic leukemia (ALL) [12, 13]. One of the strengths of the Mercator approach is its ability to discover, visualize, and interpret large numbers of subtypes in large data sets. Here, our goal is to apply Mercator in order to determine whether lymphoid malignancies can be separated into clusters based on their patterns of cytogenetic abnormalities alone.

Implementation

We first describe the data source, then the software packages and computational algorithms used to perform the analysis.

Data

Cytogenetic data were obtained from the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer, a curated, publicly available database containing all cytogenetic karyotypes published since the early 1970s [10]. When downloaded on 4 April 2019, it contained 22,741 karyotypes from patients with lymphoid malignancies, classified into 47 different disease domains. The karyotypes were written using the International System for Human Cytogenomic Nomenclature (ISCN) [14]. The ISCN is a semi-structured, semi-context-free grammar that produces a textual representation of the complete genetic information seen by the cytogeneticist evaluating the sample. We transformed the ISCN karyotypes representations into a binary representation based on a loss, gain, and fusion (LGF) biological model using CytoGPS (www.cytogps.org) [5].

Algorithms

The Mercator package (version 0.10.0) facilitates the exploration of binary data sets. Mercator (1) removes redundant features, (2) calculates a variety of distance metrics, and (3) generates multiple visualizations using a consistent color scheme and interface. This approach is designed to help users more easily navigate through the process of data analysis and visualization for pattern recognition. Although the Mercator package implements multiple distance metrics (Jaccard, Sokal & Michener, Hamming, Russell-Rao, Pearson, Goodman & Kruskal, Manhattan, Canberra, Binary and Euclidean) between binary vectors [6], in this article we rely primarily on a metric derived from Jaccard similarity [15]. The Jaccard similarity index between two binary vectors is defined as $J = N_{11}/(N_{11} + N_{10} + N_{01})$, where N_{ij} is the number of entries where the first vector contains the value i and the second vector contains the value j. Because it ignores the "insignificant" 0–0 matches, it is particularly well adapted to finding structure in sparse binary data. (For comparison, the Additional files 1 and 2 also investigates the Soakl–Michener and Goodman–Kruskal metrics.)

Mercator relies on the Thresher (version 1.1.2) and PCDimension (version 1.1.11) R packages to remove outliers and to determine the number of clusters [16, 17]. The number of clusters depends on the number of significant principal components, which is determined using the Bayesian method of Auer and Gervini [18]. Next, samples are assigned to clusters using Partitioning Around Medoids (PAM) [19]. Although PAM is the default clustering algorithm, Mercator allows the user to apply their preferred clustering algorithm before using its visualization tools. Finally, Mercator provides an interface to data visualization (with a consistent color scheme) using a variety of techniques including multidimensional scaling (MDS) [20], hierarchical clustering [21], t-distributed Stochastic Neighbor Embedding (t-SNE) [22], and adjacency graphs. To simplify the visualization of graphs with more than 20,000 nodes, we also used Mercator to perform down-sampling. This approach was inspired by Peng Qiu's implementation of the Spanning-tree Progression Analysis of Density-normalized Events (SPADE) clustering

algorithm for mass cytometry data [23]. The main point is to under-sample the densest regions of the data space to make it more likely that rare clusters will still be adequately sampled. Mercator is available as an R package at (https://cran.r-project.org/web/packages/Mercator/index.html) where further information on the packages available.

Results

In this section, we present results using the Jaccard distance metric. For theoretical reasons, we prefer the Jaccard metric, which is directly specifically designed to calculate distances for sparse binary data, where most of the values are zero [15]. Since most of each patient's genome is normal, the cytogenetic data vectors for each patient are sparse. In the Mitelman data set, most cytogenetic features are also sparse, occurring in relatively small subsets of patients. However, analogs of all four figures and the final table from parallel analyses using the Sokal–Michener distance and the Goodman–Kruskal distance are presented in Additional files 1 and 2.

Number of components and clusters

We applied CytoGPS to the lymphoid malignancy samples from the Mitelman database, which generated a binary matrix of 22,741 samples and 2748 binary LGF features. Because large-scale copy number changes such as gains or losses of entire chromosomes affect many cytogenetic bands in the same way, we next removed any redundant features (i.e., features represented by identical vectors across the full data set). This step reduced the size to 1,144 unique features. We then applied Thresher to the features in order to identify "outliers" that do not contribute significantly to the principal components [16]. After this step, 814 unique informative features remained. Transposing the data, we also used the Jaccard metric to compute the distance between samples based on cytogenetic profiles. We used Thresher to find the number, K, of significant components (Fig. 1). We then assigned patient karyotypes to clusters using PAM with K = 134.

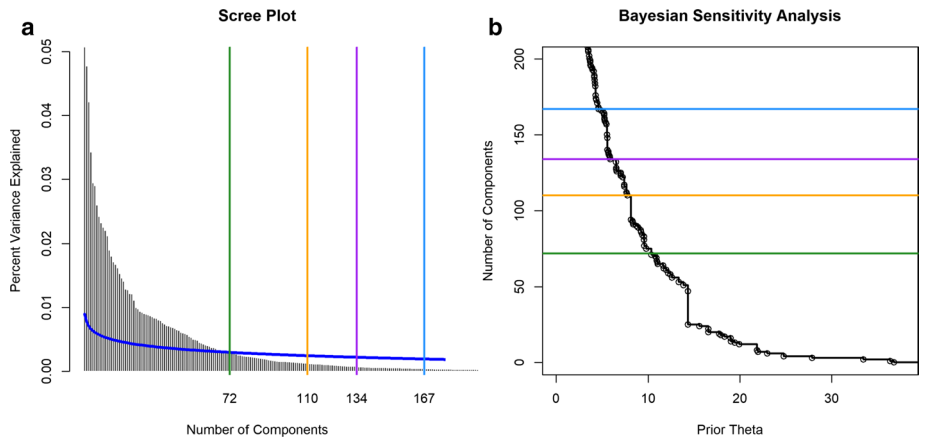

Fig. 1 Number of principal components. **a** The scree plot shows the percent of variance explained. **b** The Auer–Gervini plot shows the maximum posterior number of components as a step function of the parameter theta selecting an exponentially decaying prior. In both panels, the green line is the number (N = 72) selected by the broken-stick model. The orange (N = 110), purple (N = 134), and blue (N = 167) lines mark "long" steps that are potential cutoffs for the number of components. We selected N = 134

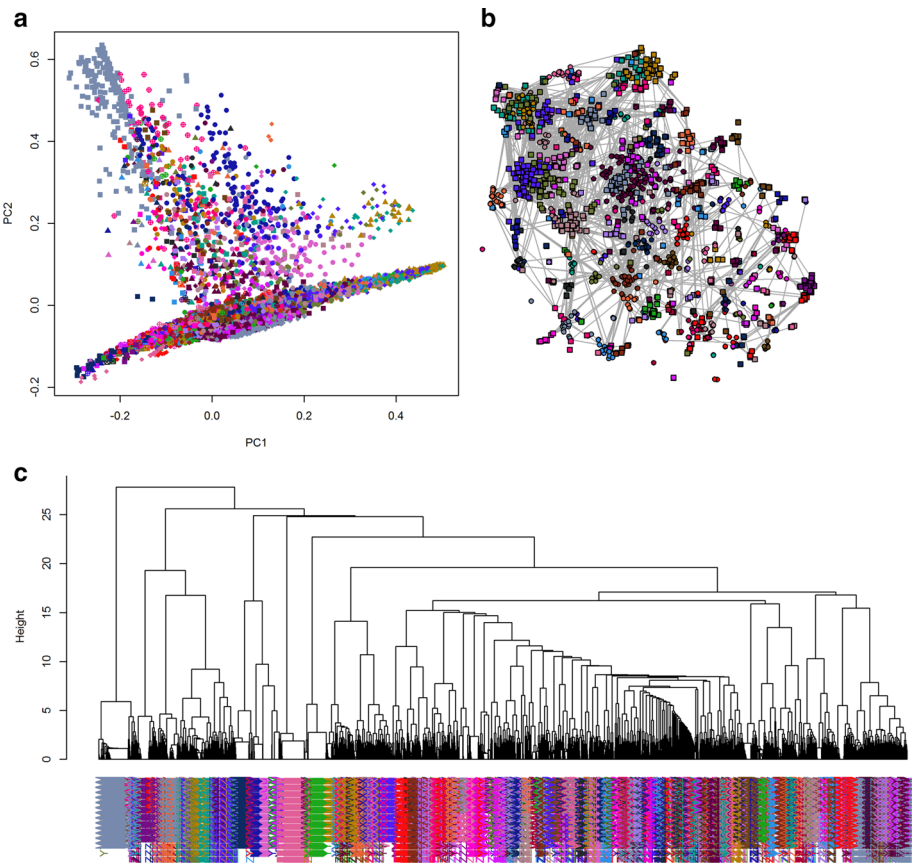

Fig. 2 LGF binary karyotypes color-coded based on PAM clustering based on Jaccard distance were visualized using three methods. **a** Multi-dimensional scaling. **b** Down-sampled adjacency graph. **c** Hierarchical clustering using Ward's linkage rule

Data visualizations

To visualize the results of PAM clustering, we applied a variety of standard methods (Fig. 2). As an initial pass through the data, we performed hierarchical clustering using the Jaccard distance matrix. The first two principal coordinates derived from multi-dimensional scaling were unable to separate the PAM-defined clusters, which is unsurprising since we believe the principal component dimension to be K = 134 (Fig. 2a). In order to visualize the distance matrix as an adjacency graph, we down-sampled the data from 22,000 nodes to 2000 (Fig. 2b). Nodes were connected by an edge if the Jaccard distance was less than 0.6 or, equivalently, if the Jaccard similarity was greater than 0.4. This threshold was determined by identifying an inflection point in the distribution of all Jaccard values (data not shown). This threshold thus removed uninformative edges while preserving biologically informative connections. This would connect nodes if the two corresponding karyotypes shared 40% or more of their cytogenetic abnormalities, indicating a high degree a cytogenetics similarity. Both the adjacency graph and the dendrogram produced by hierarchical clustering (Fig. 2c) gave some visual support to the clusters found by PAM. Finally, we used the non-linear t-SNE algorithm to produce yet another plot of the data (Fig. 3). This visualization technique clearly shows separation between most of the PAM-clusters throughout the entire plot. Some of the tightest

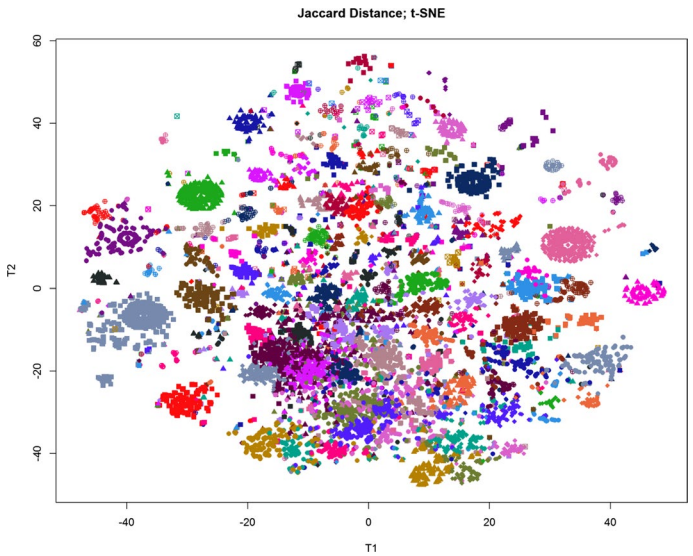

Fig. 3 T-distributed Stochastic Neighbor Embedding (t-SNE) plot of the 134 karyotype clusters. Samples are color coded based on PAM clustering using Jaccard distance

lusters form a distinguishing eye-shape; such clusters form an oval with a single point in the center. These clusters consist of groups of samples with identical or nearly identical karyotypes.

Clustering and interpreting LGF features

We now turned our attention to the 814 unique informative LGF features. Using Thresher and the Auer-Gervini method, we determined that the 814 features could be clustered into 72 groups. We again assigned features to clusters using PAM. In order to interpret these feature clusters, we determined all cytogenetic event types (loss, gain, or fusion) and chromosome bands present in any of the members of each cluster. (See the dendrogram and the labels along the side of Fig. 4.) Of the 72 clusters, 26 were only associated with a loss of a single chromosome or chromosome arm and 22 clusters were only associated with a gain of a single chromosome or chromosome arm. The remaining 24 clusters were associated with fusions and either one (N = 16 clusters) two (n = 7 clusters), or three (N = 1 cluster) chromosomes. Further, 11 of the 24 fusion-associated clusters were also associated with a loss of chromosomal material, 7 were associated with a gain, and 6 were pure fusions. All of the associations are consistent with single cytogenetic events.

Interpreting sample clusters using high frequency cytogenetic aberrations

In order to interpret the sample-clusters, we next computed, for each cluster, the fraction of patients in that cluster who exhibited the well-characterized cytogenetic events defined by each of the 72 feature clusters described in the previous section. These frequency data were used to construct a two-way clustered heatmap based on Pearson correlation and Ward's linkage (Fig. 4). The main split in the feature-dendrogram along the side of the heatmap is between losses (top branch) and gains and fusions (bottom branch). In other words, gains tend to occur along with other gains, and losses tend to occur along with other losses.

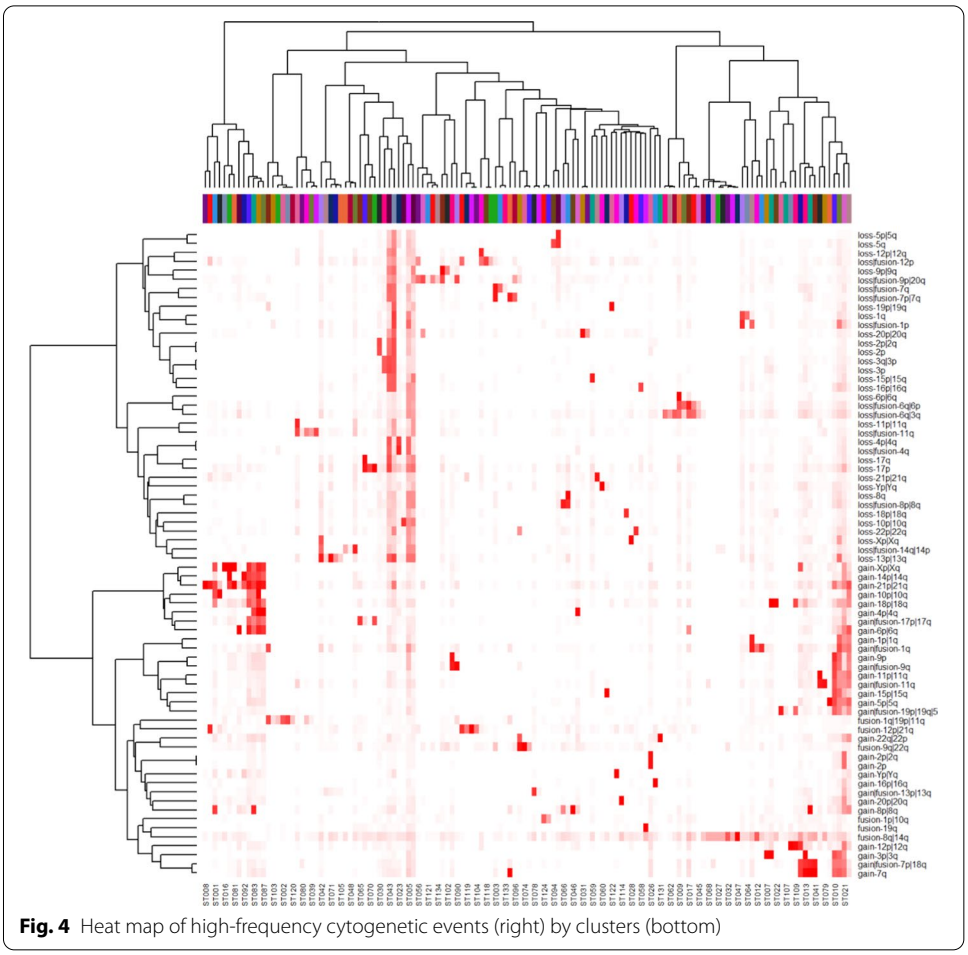

Fig. 4 Heat map of high-frequency cytogenetic events (right) by clusters (bottom)

Finally, we recorded the most frequent events (down to a frequency cutoff of 60%) in each sample-cluster. For 18 of the 134 clusters, at least one cytogenetic abnormality was present in at least 99% of the cases; for 42, at least one abnormality in at least 95% of case; for 50, 90%; for 71, 80%; for 84, 70%, and for 94, 60%. The 40 best characterized sample clusters by this measure are listed in Table 1. For each cluster we also calculated the disease prevalence based on karyotype disease labels. This was performed to with data interpretation, as some clusters contain multiple disease groups. The following diseases are part of at least one of the top 40 best characterized clusters: ALL = acute lymphocytic leukemia, CLL = chronic lymphocytic leukemia, Burkitt = Burkitt's lymphoma, FL = follicular lymphoma, DLBCL = diffuse large B-cell lymphoma and MM = multiple myeloma.

Discussion

Lymphoid karyotype clusters

We have shown that, by combining CytoGPS with Mercator to analyze 22,741 karyotypes obtained from the public Mitelman database, we are able to recover both simple and complex cytogenetic events that are important for understanding and classifying lymphoid malignancies. Using Mercator to cluster the binary LGF features, we found 72 clusters. Of these, 71 clusters represented simple losses restricted to one

Table 1 The top forty well characterized sample clusters

Cluster	Symbol	Karyotype	Relative Frequency	Disease	Prevalence
ST132	⊠	+Y	100	ALL	46
ST131	⊠	+22	100	ALL	57
ST114	⊕	+20	100	ALL	54
ST101	⊕	t(19q)	100	ALL, CLL	40, 36
ST055	▲	+10,+4,+21,+6,+X, +18,-14,+17	100,100,96,87,86, 79,74,63	ALL	100
ST053	▲	t(12p;21q)	100	ALL	100
ST047	■	t(8q;14q)	100	Burkitt	71
ST087	♦	+4,+21,+6,+X,-14	99,92,80,80,63	ALL	96
ST035	■	add(7q),+8, add(7p)t(18q12.1,*)	99,96, 84	FL	23
ST060	▲	-Y	99	CLL, ALL	26, 22
ST081	♦	+X,+21	99,98	ALL	85
ST046	■	+8	99	ALL	65
ST016	●	+X	99	ALL	53
ST075	♦	+X,-14,+21	99,98,75	ALL	90
ST097	⊕	+16	99	ALL	60
ST095	♦	+15	99	ALL	24
ST041	■	add(7q),add(7p)t(18q12.1,*)	99,88	FL	33
ST008	●	+21	99	ALL	88
ST057	▲	+5	98	ALL	41
ST099	⊕	del(17p),+17	98,70	ALL, CLL	34, 27
ST074	♦	t(9q;22q)	98	ALL	98
ST019	●	+6	98	ALL	56
ST011	●	+18,+3	98,98	DLBCL	35
ST040	■	+12	98	CLL	75
ST013	●	+3,add(7q), add(7p)t(18q12.1,*)	98,97, 81	DLBCL	29
ST033	■	add(7q),LF(7p;7q)	98,89	ALL	59
ST089	♦	-9,t(9q;20q)	97,70	ALL	58
ST054	▲	+10	97	ALL	83
ST090	♦	add(9q),t(9q;20q)	97,72	ALL	94
ST028	■	-X	97	ALL	38
ST073	♦	del(8q),-8	97,82	ALL	41
ST022	●	+18	97	FL	39
ST006	●	+11,add(11q)	96,89	ALL	36
ST009	●	-6,LF(6p;6q),t(3q;6q)	96,69,60	ALL	25
ST031	■	-20	96	ALL	69
ST092	♦	-14,+21	96,63	ALL	75
ST071	▲	-13	95	ALL, MM	28, 23
ST007	●	+3	95	DLBCL	11
ST034	■	del(17q),del(17p)	95,91	ALL, CLL	30, 23
ST109	⊕	+12,+18	95,72	CLL	62

chromosome, simple gains restricted to one chromosome, or simple fusions involving at most two chromosomes. The remaining cluster was a fusion event involving three chromosomal arms: 1q, 11q, and 19p. Although more complicated than the others, this cluster represents a known phenomenon of "jumping" translocations involving 1q that has been seen in both lymphoid and myeloid malignancies [24, 25].

The lymphoid karyotypes from the Mitelman database represent 47 disease morphologies. Our analysis with Mercator found 134 clusters based on cytogenetics. We used the 72 elementary cytogenetic events above to characterize the 134 sample clusters. One of the well-known patterns is the t(8q;14q) translocation, which produces a fusion protein by juxtaposing the immunoglobulin heavy chain locus on chromosome 14 with the MYC oncogene on chromosome 8 [26]. This abnormality is the only recurrent event in cluster ST047 of Table 1, and occurs in 100% of the cases in that

cluster. However, it is not unique to that cluster; as shown in Fig. 4, it is present at varying frequencies in the majority of lymphoid malignancy clusters. This finding can be explained by the fact that this translocation does not just occur as the sole abnormality in lymphoid malignancies, but also occurs in concert with many other combinations of abnormalities.

One of the strengths of using Mercator is its ability to uncover more complicated patterns that represent the recurrent co-occurrence of cytogenetic events. The most striking examples in Table 1 are clusters ST087 (with gains of four and loss of one chromosome) and ST055 (with gains of seven and loss of one chromosome). Both of these clusters display complex cytogenetic patterns that could only be uncovered using computational techniques and are unlikely to have been found simply by visual inspection of large sets of complex karyotypes. Looking deeper at these two clusters reveals that they share the events -14 and $+21$. In fact, these two events also co-occur in other clusters, including ST092 (which has only those two recurrent events) and ST075 (which combines them with an extra copy of the X chromosome). To our knowledge, this co-occurrence has not previously been recognized as a separate entity by cytogeneticists or hematopathologists. Preliminary visual inspections of the text-based karyotypes suggests that -14 and $+21$ almost always occur in the context of highly complex karyotypes where picking this pair out as a separate feature would be unlikely without computational assistance.

In general, the co-occurrence of a monosomy with a trisomy is unusual. A primary feature of Fig. 4 is that losses (monosomies) cluster together and gains (trisomies) cluster together, on separate branches in the (side) dendrogram. Hyperdiploidy (having more than the usual number of chromosomes) is a common feature of multiple myeloma [27] and of acute lymphoblastic leukemia [28] and has been reported in diffuse large B-cell lymphoma [29]. Hypodiploidy (having fewer than the normal number of chromosomes) is also common in lymphoid malignancies [30, 31].

A fundamental challenge when using any clustering method to perform unsupervised analysis arises from the difficulty of correctly ascertaining the number of clusters present in the data. We found that 94 (70%) of the 134 clusters have at least one cytogenetic abnormality that is present in at least 60% of the cases, and that many of those clusters have one or more abnormalities present at much higher frequencies. Thus, Mercator is able to identify high fidelity patterns and generates clusters that have a natural biological interpretation. It is possible, however, that the "true" number of clusters lies somewhere between the 134 found by Mercator and Thresher and the 47 known disease morphologies. Ideally, every cytogenetic cluster should be characterized by a unique combination of events.

Distance metrics

We looked at different distance metrics to determine which metric would work best on cytogenetic data. In addition to the Jaccard distance, we performed our analyses using both the Sokal–Michener and Goodman–Kruskal metrics. These results are shown in Additional file 1 (Sokal–Michener) and Additional file 2 (Goodman–Kruskal). Sokal–Michener was not selected due to poor cluster differentiation (Additional file 1: Figure S2). Sokal–Michener did identify complex cytogenetic clusters, so it may be of research benefit for identifying recurrent complex events. Goodman–Kruskal identified weaker

connections between karyotypes than Jaccard (Additional file 2: Figure S2), and thus was not selected. This is likely due to Goodman–Kruskal taking into consideration zero-zero matches when looking at binary data. This is in contrast to Jaccard, which only considers one–one matches to be meaningful. Since the LGF model data is a sparse binary vector it makes sense that a distance metric that only values one–one matches would outperform a metric that considers all matches.

One critically important aspect of Mercator is its use of shared cluster color schemes across different methods. It has been known for many years that humans are better than computers at determining visual patterns [32]. For this reason we designed Mercator to use a shared color scheme when using different methods on the same data set. This enables users to look at plots generated by different algorithms and visually compare them to determine the best algorithm for a given dataset. Keeping the color schemes consistent allows clusters that are based on a similar underlying characteristic can be compared across different clustering algorithms. This allows Mercator to leverage the intelligence of the researcher to help identify the best algorithm for a given dataset.

Conclusions

In the future, it may be possible to address this issue by applying Mercator recursively. That is, we would first remove any cytogenetic event that is used to fully characterize one or more clusters at very high frequencies, and would then remove samples that only present with those abnormalities. We could then apply Mercator to the remaining features and samples to see if the resulting clusters can be characterized by other abnormalities at high frequency. We also intend to examine the associations between cytogenetically defined sample clusters and the known disease morphologies. A cursory examination suggests that the cytogenetic classification may be independent of and orthogonal to the known disease classification. If that observation holds up, then it will also be important to find other karyotype data sets that can be linked to clinical outcomes in order to test whether the cytogenetics can give better insight into an appropriate choice of therapies across disease types.

Mercator, in conjunction with CytoGPS, was able to identify biological patterns of shared elements within the cytogenetic profiles of different diseases. Data heterogeneity remains a very common problem in karyotype data analysis due to the innate linkage of cytogenetic features with one another due to colocating on the same chromosome. Mercator solves this problem by identifying unique feature sets and combining features to reduce the dimensionality while still preserving all relevant information. Mercator solves the related problem of data sparsity by selecting the proper measurement of distance. By utilizing the Jaccard distance, we were able to address the high levels of sparsity within our data set by focusing solely on 1 to 1 matches across our binary vectors. This elegant solution enabled both clustering and large-scale visualizations to be performed on an otherwise highly sparse and noisy high-dimensional data set.

Although we highlighted the usage of the Mercator package on binary cytogenetic data in this paper, it is important to note that Mercator is "data-type agnostic". Many other forms of biomedical data could be easily processed and visualized using the Mercator methodology. This is particularly relevant in many omics fields where the large feature

space requires clever feature reduction techniques, such as Thresher, to improve the overall computational analysis of the data. The standard visualizations used by Mercator will also aid these omics experiments, providing a clear visualization of the underlying data and thus a better understanding of the structure of omics data sets.

Availability and requirements

Project name: Mercator
Project home page: https://CRAN.R-project.org/package=Mercator
Operating system(s): Windows/macOS
Programming language: R
Other requirements: R (≥ 3.5), Thresher (≥ 1.1)
License: Apache License ($==2.0$)
Any restrictions to use by non-academics: No restrictions, only acknowledge authors contribution

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-03992-1.

Additional file 1. Sokal Michener experiments.

Additional file 2. Goodman Kruskal experiments.

Acknowledgements
We thank the Summer Internship Program at the Ohio State University Department of Biomedical Informatics for their support.

Authors' contributions
ZA: Helped design software, experiments and was the lead in writing the paper. DT: Help run experiments and edit and write the paper. LZ: Helped edit and write the paper. CC: Helped design and develop software, edit and write the paper. PP: Helped edit and write the paper. LA: Helped with experimental design, edit and write the paper. KC: Oversaw all aspects of the project e.g. Software design, software development, experimental design, paper writing and editing. All authors read and approved the final manuscript.

Funding
This work was supported by the National Library of Medicine (NLM) Grant Number T15 LM011270, the National Cancer Institute Grant Number R03 CA235101, and by Pelotonia Intramural Research Funds from the James Cancer Center, Columbus Ohio. Non of the funding bodies played any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
All data generated or analyzed during this study are included in this published article, and its supplementary information files, or upon a reasonable request. The public data set Mitelman was used in this paper and can be found at: https://mitelmandatabase.isb-cgc.org/

Ethics approval and consent to participants
Not applicable.

Consent for publication
Not applicable.

Competing interests
Kevin R. Coombes is also a member of the editorial board (Associate Editor) of the BMC Bioinformatics journal.

Author details
[1] Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA. [2] The Center for Genomic Advocacy At Indiana State University, Terre Haute, IN 47809, USA. [3] Institute for Informatics, Washington University School of Medicine in St. Louis, St. Louis, MO 63108, USA. [4] Department of Pathology, The Ohio State University, Columbus, OH 43210, USA.

Received: 7 February 2020 Accepted: 2 February 2021
Published online: 01 March 2021

References

1. Andreu-Perez J, et al. Big data for health. IEEE J Biomed Health Inform. 2015;19(4):1193–208.
2. Margolis R, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21(6):957–8.
3. Miotto R, et al. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2017;19(6):1236–46.
4. Stevens-Kroef M, et al. Cytogenetic nomenclature and reporting. Methods Mol Biol. 2017;1541:303–9.
5. Abrams ZB, et al. CytoGPS: a web-enabled karyotype analysis tool for cytogenetics. Bioinformatics. 2019;216:1037.
6. Choi SS, Cha SH, Tappert CC. A survey of binary similarity and distance measures. Syst Cybern Inf. 2010;8910:43–8.
7. Alizadeh AA, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503.
8. Collins SJ, Gallo RC, Gallagher RE. Continuous growth and differentiation of human myeloid leukaemic cells in suspension culture. Nature. 1977;270(5635):347.
9. Swerdlow SH, et al. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood. 2016;127(20):2375–90.
10. Mitelman F, Johansson B, Mertens F. Catalog of chromosome aberrations in cancer, vol. 1. New York: Wiley-Liss; 1991.
11. Dohner H, et al. Chromosome aberrations in B-cell chronic lymphocytic leukemia: reassessment based on molecular cytogenetic analysis. J Mol Med (Berl). 1999;77(2):266–81.
12. Seol CA, et al. Prognostic significance of recurrent additional chromosomal abnormalities in adult patients with Philadelphia chromosome-positive acute lymphoblastic leukemia. Cancer Genet. 2017;216–217:29–36.
13. Shago M. Recurrent cytogenetic abnormalities in acute lymphoblastic leukemia. Methods Mol Biol. 2017;1541:257–78.
14. McGowan-Jordan J, Simons A, Schmid M, editors. ISCN 2016: An International System for Human Cytogenomic Nomenclature (2016). Basel: Karger Medical and Scientific Publishers; 2016.
15. Jaccard P. The distribution of the flora in the alpine zone. 1. New Phytol. 1912;11(2):37–50.
16. Wang M, et al. Thresher: determining the number of clusters while removing outliers. BMC Bioinform. 2018;19(1):9.
17. Wang M, Kornblau SM, Coombes KR. Decomposing the apoptosis pathway into biologically interpretable principal components. Cancer Inform. 2018;17:1176935118771082.
18. Auer P, Gervini D. Choosing principal components: a new graphical method based on Bayesian model selection. Commun Stat Simul Comput. 2008;37(5):962–77.
19. Van der Laan M, Pollard K, Bryan J. A new partitioning around medoids algorithm. J Stat Comput Simul. 2003;73(8):575–84.
20. Borg I, Groenen P. Modern multidimensional scaling: theory and applications. J Educ Meas. 2003;40(3):277–80.
21. Johnson SC. Hierarchical clustering schemes. Psychometrika. 1967;32(3):241–54.
22. Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(9):2579–605.
23. Peng QIU, et al. Spanning-tree progression analysis of density-normalized events (spade). 2013, Google Patents.
24. Couture T, et al. Jumping translocations of 1q in myelodysplastic syndrome and acute myeloid leukemia: report of three cases and review of literature. Case Rep Genet. 2018;2018:8296478.
25. Sawyer JR, et al. Jumping translocations of chromosome 1q in multiple myeloma: evidence for a mechanism involving decondensation of pericentromeric heterochromatin. Blood. 1998;91(5):1732–41.
26. Haluska FG, et al. The t(8; 14) chromosomal translocation occurring in B-cell malignancies results from mistakes in V-D-J joining. Nature. 1986;324(6093):158–61.
27. Manier S, et al. Genomic complexity of multiple myeloma and its clinical implications. Nat Rev Clin Oncol. 2017;14(2):100–13.
28. de Smith AJ, et al. Clonal and microclonal mutational heterogeneity in high hyperdiploid acute lymphoblastic leukemia. Oncotarget. 2016;7(45):72733–45.
29. Nanjangud G, et al. Spectral karyotyping identifies new rearrangements, translocations, and clinical associations in diffuse large B-cell lymphoma. Blood. 2002;99(7):2554–61.
30. Holmfeldt L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013;45(3):242–52.
31. Van Wier S, et al. Hypodiploid multiple myeloma is characterized by more aggressive molecular markers than non-hyperdiploid multiple myeloma. Haematologica. 2013;98(10):1586–92.
32. Schur AI, Tappert CC. Speed and accuracy improvements in visual pattern recognition tasks by employing human assistance. In: Advances in human factors and system interactions. Springer; 2017. p. 293–300.

Publisher's Note