Short Papers

APPLES: Fast Distance-Based Phylogenetic Placement

Metin Balaban[1], Shahab Sarmashghi[2], and Siavash Mirarab[2(✉)]

[1] Bioinformatics and Systems Biology, UC San Diego, La Jolla, CA 92093, USA
[2] Electrical and Computer Engineering, UC San Diego, La Jolla, CA 92093, USA
smirarab@ucsd.edu

Keywords: Phylogenetic placement · Distance-based methods · Genome-skimming

Extended Abstract

Methods for inferring phylogenetic trees from very large datasets exist, yet, large-scale tree reconstructions still require significant resources. New species are continually being sequenced, and as a result, even large trees can become outdated. Reconstructing the tree *de novo* each time new sequences become available is not practical. An alternative approach is phylogenetic placement where new sequence(s) are simply added to an existing *backbone* tree. Phylogenetic placement has applications other than updating trees, including sample identification, where the goal is to detect the identity of given *query* sequences of unknown origins. This problem arises [3] in the study of mixed environmental samples that make up much of the microbiome literature. Sample identification is also the essence of barcoding and meta-barcoding, methods used often in biodiversity studies.

Maximum Likelihood (ML) methods of phylogenetic placement are now available and in wide use (e.g., [4] and EPA(-ng) [2]). The ML approach is computationally demanding, and in particular requires large amounts of memory, and therefore, is limited in the size of the backbone tree it can use. More fundamentally, existing placement tools take as input alignments of assembled sequences for the backbone set, even when queries allowed to be unassembled reads. This reliance on assembled sequences makes them unsuitable for alignment and assembly-free scenarios. For example, sample identification using genome-skimming is fast becoming cost-effective. Methods like Skmer [5] (introduced in RECOMB 2018) can be used to infer k-mer-based estimates of phylogenetic distance from genome skims, and these distances can potentially be used for placement on phylogenetic trees. However, existing methods cannot be used for this purpose.

Distance-based phylogenetics has a rich methodological history, and yet, there are no existing tools for distance-based phylogenetic placement. Such methods, if developed, can be scalable to ultra-large backbone trees. Moreover,

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 287–288, 2019.
https://doi.org/10.1007/978-3-030-17083-7

distance-based methods only need distances, not assembled sequences, and therefore, can be used for sample identification from reads in an assembly-free and alignment-free fashion.

We have developed a new method for distance-based phylogenetic placement called APPLES (Accurate Phylogenetic Placement using LEast Squares). APPLES finds the placement of a query sequence that minimizes the least square error of phylogenetic distances with respect to sequence distances. It can also operate on the minimum evolution principle, or a hybrid of minimum evolution and least square error. Using dynamic programming, APPLES is able to perform placement in time and memory that both scale linearly with the size of the backbone tree.

We have performed extensive studies on simulated and real datasets to evaluate APPLES. Our results show that in the alignment-based scenario, APPLES is much faster than ML tools, uses much less memory, and is very close to ML in the accuracy. Moreover, APPLES can handle much larger backbone trees (we have tested up to 200,000 leaves), and has *increased* accuracy when the backbone trees become larger and more densely sampled. In contrast, ML methods cannot handle backbones with several thousand species. For assembly-free scenarios, we study three genome skimming datasets of insects and show that APPLES applied to Skmer distances can accurately identify genome skim samples using coverage below 1X [1]. APPLES is open-source and freely available at https://github.com/balabanmetin/apples.

References

1. Balaban, M., Sarmashghi, S., Mirarab, S.: Apples: Fast distance-based phylogenetic placement. bioRxiv (2018). https://doi.org/10.1101/475566. https://www.biorxiv.org/content/early/2018/11/23/475566
2. Barbera, P., et al.: EPA-ng: massively parallel evolutionary placement of genetic sequences. BioRxiv, 291658 (2018)
3. Janssen, S., et al.: Phylogenetic placement of exact amplicon sequences improves associations with clinical information. mSystems **3**(3), 00021–18 (2018). https://doi.org/10.1128/mSystems.00021-18. http://msystems.asm.org/lookup/doi/10.1128/mSystems.00021-18
4. Matsen, F.A., Kodner, R.B., Armbrust, E.V.: pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinf. **11**(1), 538 (2010)
5. Sarmashghi, S., Bohmann, K., Gilbert, M.T.P., Bafna, V., Mirarab, S.: Assembly-free and alignment-free sample identification using genome skims. Genome Biology (abstract appeared at RECOMB 2018) (2018, in press). https://doi.org/10.1101/230409. https://www.biorxiv.org/content/early/2018/04/02/230409

De Novo Peptide Sequencing Reveals a Vast Cyclopeptidome in Human Gut and Other Environments

Bahar Behsaz[1], Hosein Mohimani[2], Alexey Gurevich[3], Andrey Prjibelski[3], Mark F. Fisher[4], Larry Smarr[2,5,6], Pieter C. Dorrestein[6,7], Joshua S. Mylne[4], and Pavel A. Pevzner[2,3,6](✉)

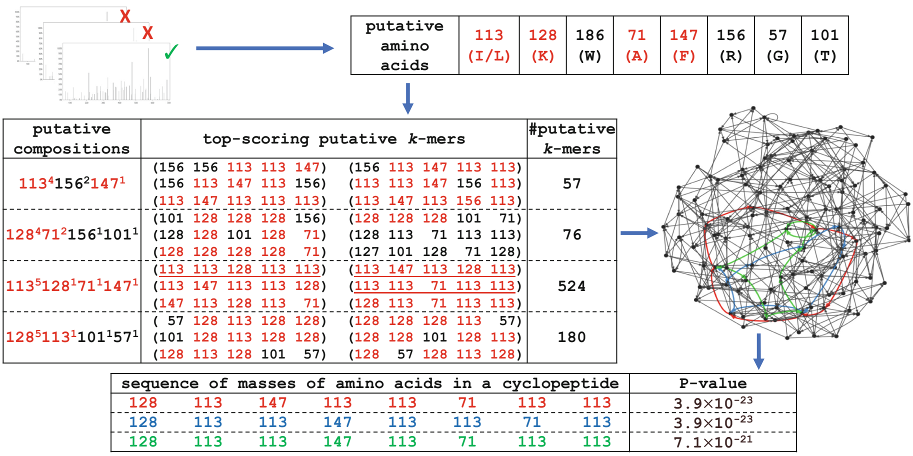
[1] Bioinformatics and Systems Biology Program,
University of California San Diego, La Jolla, USA
[2] Department of Computer Science and Engineering,
University of California San Diego, La Jolla, USA
ppevzner@ucsd.edu
[3] Center for Algorithmic Biotechnology, Institute of Translational Biomedicine,
St. Petersburg State University, St. Petersburg, Russia
[4] School of Molecular Sciences and The ARC Centre of Excellence in Plant
Energy Biology, The University of Western Australia, Crawley, Australia
[5] California Institute for Telecommunications and Information Technology,
University of California San Diego, La Jolla, USA
[6] Center for Microbiome Innovation,
University of California at San Diego, La Jolla, USA
[7] Department of Pharmacology,
University of California at San Diego, La Jolla, USA

Extended Abstract

Cyclic and branch cyclic peptides (cyclopeptides) represent an important class of bioactive natural products that include many antibiotics and anti-tumor compounds. However, little is known about cyclopeptides in the human gut, despite the fact that humans are constantly exposed to them. To address this bottleneck, we developed CycloNovo algorithm [1] for *de novo* cyclopeptide sequencing that employs de Bruijn graphs, the workhorse of DNA sequencing algorithms. Figure 1 illustrates the CycloNovo pipeline. CycloNovo reconstructed many new cyclopeptides that we validated with transcriptome, metagenome, and genome mining analyses.

We applied CycloNovo to high-resolution spectral dataset generated from daisy seeds (*Senecio vulgaris*), human microbiome (HUMANSTOOL), and a large dataset of 40 high-resolution spectra from GNPS (GNPS). CycloNovo reconstructed ten cyclopeptides in *S. vulgaris* including 4 known and 6 novel cyclopeptides that were further validated using assembled RNA-seq transcripts. Our analysis revealed 703 cyclospectra in HUMANSTOOL dataset corresponding to 79 unique putative cyclopeptides (identified by MS-Cluster) forming 69 spectral families (identified by molecular networking). Dereplicator search yielded only nine PSMs with 0% FDR and

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 289–291, 2019.
https://doi.org/10.1007/978-3-030-17083-7



putative amino acids	113 (I/L)	128 (K)	186 (W)	71 (A)	147 (F)	156 (R)	57 (G)	101 (T)

putative compositions	top-scoring putative *k*-mers		#putative *k*-mers
$113^4 156^2 147^1$	(156 156 113 113 147) (156 113 147 113 156) (113 147 113 113 113)	(156 113 147 113 113) (113 113 147 156 113) (113 147 113 156 113)	57
$128^4 71^2 156^1 101^1$	(101 128 128 128 156) (128 128 101 128 71) (128 128 128 128 71)	(128 128 128 101 71) (128 113 71 113 113) (127 101 128 71 128)	76
$113^5 128^1 71^1 147^1$	(113 113 128 113 113) (113 147 113 113 128) (147 113 128 113 71)	(113 147 113 128 113) (113 113 71 113 113) (128 113 71 113 113)	524
$128^5 113^1 101^1 57^1$	(57 128 113 128 128) (101 128 113 128 128) (128 113 128 101 57)	(128 128 128 113 57) (128 128 101 128 113) (128 57 128 113 128)	180

sequence of masses of amino acids in a cyclopeptide								P-value
128	113	147	113	113	71	113	113	3.9×10^{-23}
128	113	113	147	113	113	71	113	3.9×10^{-23}
128	113	113	147	113	71	113	113	7.1×10^{-21}

Fig. 1. CycloNovo outline illustrated using *Spectrum*$_{Surugamide}$.CycloNovo includes six steps: (*i*) recognizing cyclospectra using their spectral-convolution [2], (*ii*) predicting amino acids in a cyclopeptide, (*iii*) predicting amino acid composition of a cyclopeptide, (*iv*), predicting *k*-mers in a cyclopeptide, (*v*) constructing the de Bruijn graph of a spectrum, and (*vi*) generating cyclopeptide reconstructions and calculating P-values [3, 4]. Only six top-scoring putative *k*-mers for each putative amino acid composition are shown. Masses of amino acids occurring in surugamide are shown in red and *k*-mers occurring in surugamide are underlined. To simplify the de Bruijn graph (corresponding to the composition $71^1 113^5 128^1 147^1$), all tips and isolated edges in the graph were removed. Red, blue and green feasible cycles in the graph spell out three cyclopeptides shown in the bottom table along with their P-values. The red cycle spells out surugamide.

P-value < 10^{-15}, seven that originated from Flax cyclolinopeptides A [5], B [6], C [7], D [7], H [7], E [7], and P [8] as well as Citrusin V and Massetolide F. Cyclolinopeptides belong to the family of flaxseed orbitides that are present in the seeds of *Linum usitatissimum*. We confirmed that the diet of the individual who provided the HUMANSTOOL sample (L.S., co-author) contained flaxseed eaten frequently as an ingredient in his cooking. Citrusin V belong to the citrusin family of antimicrobial orbitides found in the extracts of various species from the *Citrus* genus [9]. Massetolides are non-ribosomal lipopeptides produced by *Pseudomonas fluoresences*, an indigenous member of human and plant microbiota [10, 11]. Analysis of the metagenome assembly of reads paired with the HUMANSTOOL dataset confirmed that *P. fluoresences* is present in the stool samples where massetolide F was detected.

In addition to the nine identified cyclopeptides, CycloNovo reconstructed 32 cyclopeptides in the HUMANSTOOL dataset with P-values below 10^{-15} forming 26 cyclofamilies. Finding many bioactive cyclopeptides in our study that remain stable in the proteolytic environment of the human gut raises the question of how these bioactive antimicrobial cyclopeptides affect the bacterial composition of the human microbiota.

We analyzed cyclopeptide spectra identified in the GNPS dataset with the goal of estimating the number of still unknown cyclopeptides from spectra already deposited in GNPS. Dereplicator search of the entire GNPS dataset identified 80 unique known cyclopeptides containing 41 cyclofamilies. CycloNovo predicted a total of 12,004 cyclopeptide spectra representing 512 putative cyclopeptides forming 213 cyclofamilies. These putative cyclopeptides include 67 (37 cyclofamilies) of the 80 known cyclopeptides. We showed that even in the case of the phyla with extensively analyzed cyclopeptides (*Cyanobacteria*, *Pseudomonas*, and *Actinomyces*), only less than 30% of the predicted cyclopeptides are already known.

Link to preprint version: https://www.biorxiv.org/content/10.1101/521872v2

References

1. Behsaz, B., et al.: De novo peptide sequencing reveals a vast cyclopeptidome in human gut and other environments. bioRxiv (2019)
2. Ng, J., et al.: A: Dereplication and de novo sequencing of nonribosomal peptides. Nat Methods **6**, 596–599 (2009)
3. Mohimani, H., Kim, S., Pevzner, P.A.: A new approach to evaluating statistical significance of spectral identifications. J. Proteome Res. **12**, 1560–1568 (2013)
4. Mohimani, H., et al.: Dereplication of peptidic natural products through database search of mass spectra. Nat. Chem. Biol. **13**, 30–37 (2017)
5. Kaufmann, H.P., Tobschirbel, A.: Über ein oligopeptid aus leinsamen. Eur. J. Inorg. Chem. **92**, 2805–2809 (1959)
6. Morita, H., et al.: A new immunosuppressive cyclic nonapeptide, cycloinopeptide B from linum usitatissimum. Bioorg. Med. Chem. Letts. **7**, 1269–1272 (1997)
7. Morita, H., Shishido, A., Matsumoto, T., Itokawa, H., Takeya, K.: Cyclolinopeptides B-E, new cyclic peptides from Linum usitatissimum. Tetrahedron. **55**, 967–976 (1999)
8. Okinyo-Owiti, D.P., Young, L., Burnett, P.G.G., Reaney, M.J.T.: New flaxseed orbitides: Detection, sequencing, and15N incorporation. Biopolymers - Peptide Science Section. **102**, 168–175 (2014)
9. Noh, H.J., et al.: Anti-inflammatory activity of a new cyclic peptide, citrusin XI, isolated from the fruits of Citrus unshiu. J. Ethnopharmacol. **163**, 106–112 (2015)
10. Scales, B.S., Dickson, R.P., Lipuma, J.J., Huffnagle, G.B.: Microbiology, genomics, and clinical significance of the pseudomonas fluorescens species complex, an unappreciated colonizer of humans. Clin. Microbiol. Rev. **27**, 927–948 (2014)
11. O'Sullivan, D.J., O'Gara, F.: Traits of fluorescent Pseudomonas spp. involved in suppression of plant root pathogens. Microbiol. Rev. **56**, 662–676 (1992)

Biological Sequence Modeling with Convolutional Kernel Networks

Dexiong Chen[1](✉), Laurent Jacob[2], and Julien Mairal[1]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
{dexiong.chen,julien.mairal}@inria.fr
[2] Univ. Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR 5558, Lyon, France
laurent.jacob@univ-lyon1.fr

Understanding the relationship between biological sequences and the associated phenotypes is a fundamental problem in molecular biology. Accordingly, machine learning techniques have been developed to exploit the growing number of phenotypic sequences in automatic annotation tools. Typical applications include classifying protein domains into superfamilies [6, 9], predicting whether a DNA or RNA sequence binds to a protein [1], its splicing outcome [3], or its chromatin accessibility [4], predicting the resistance of a bacterial strain to a drug [2], or denoising a ChIP-seq signal [5]. Choosing how to represent biological sequences is a critical part of methods that predict phenotypes from genotypes. Kernel-based methods [6, 9, 8] have often been used for this task. They have been proven efficient to represent biological sequences in various tasks but only construct fixed representations and lack scalability to large amount of data. By contrast, convolutional neural networks (CNN) [1] have recently shown scalable and able to optimize data representations for specific tasks. However, they typically lack interpretability and require large amounts of annotated data, which motivates us to introduce more data-efficient approaches.

In this work we introduce CKN-seq, a strategy combining kernel methods and deep neural networks for sequence modeling, by adapting the convolutional kernel network (CKN) model originally developed for image data [7]. CKN-seq relies on a convolutional kernel, a continuous relaxation of the mismatch kernel [6], and the Nyström approximation. The relaxation makes it possible to learn the kernel from data, and we provide an unsupervised and a supervised algorithm to do so—the latter leading to a special case of CNNs.

On a transcription factor binding prediction task and a protein remote homology detection task, both approaches show better performance than DeepBind, another existing CNN [1], especially when the amount of training data is small. On the other hand, the supervised algorithm produces task-specific and small-dimensional sequence representations while the unsupervised version dominates all other methods on small-scale problems but leads to higher dimensional representations. Consequently, we introduce a hybrid approach which enjoys the benefits of both supervised and unsupervised variants, namely the ability of learning low-dimensional models with good prediction performance in all data size regimes. Finally, the kernel point of view of our method provides us simple ways to visualize and interpret our models, and obtain sequence logos. On some

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 292–293, 2019.
https://doi.org/10.1007/978-3-030-17083-7

simulated data, the logos given by CKN-seq are more informative and match better with the ground truth in terms of any probabilistic distance measures. We provide a free implementation of CKN-seq for learning from biological sequences, which can easily be adapted to other sequence prediction tasks and is available at https://gitlab.inria.fr/dchen/CKN-seq.

The fact that CKNs retain the ability of CNNs to learn feature spaces from large training sets of data while enjoying a reproducing kernel Hilbert space structure has other uncharted applications which we would like to explore in future work. First, it will allow us to leverage the existing literature on kernels for biological sequences to define the bottom kernel instead of the mismatch kernel, possibly capturing other aspects than sequence motifs. More generally, it provides a straightforward way to build models for non-vector objects such as graphs, taking as input molecules or protein structures. Finally, it paves the way for making deep networks amenable to statistical analysis, in particular to hypothesis testing. This important step would be complementary to the interpretability aspect, and necessary to make deep networks a powerful tool for molecular biology beyond prediction.

A full version of the paper is available at https://doi.org/10.1101/217257.

References

1. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat. Biotechnol. **33**(8), 831–838 (2015)
2. Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V.G., Bourgault, A.M., Laviolette, F., Corbeil, J.: Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. BMC Genomics **17**(1), 754 (2016)
3. Jha, A., Gazzara, M.R., Barash, Y.: Integrative deep models for alternative splicing. Bioinformatics **33**(14), 274–282 (2017)
4. Kelley, D.R., Snoek, J., Rinn, J.L.: Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. **26**(7), 990–999 (2016)
5. Koh, P.W., Pierson, E., Kundaje, A.: Denoising genome-wide histone chip-seq with convolutional neural networks. Bioinformatics **33**(14), i225–i233 (2017)
6. Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S.: Mismatch string kernels for discriminative protein classification. Bioinformatics **20**(4), 467–476 (2004)
7. Mairal, J.: End-to-end kernel learning with supervised convolutional kernel networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1399–1407 (2016)
8. Rangwala, H., Karypis, G.: Profile-based direct kernels for remote homology detection and fold recognition. Bioinformatics **21**(23), 4239–4247 (2005)
9. Saigo, H., Vert, J.P., Ueda, N., Akutsu, T.: Protein homology detection using string alignment kernels. Bioinformatics **20**(11), 1682–1689 (2004)

Dynamic Pseudo-time Warping of Complex Single-Cell Trajectories

Van Hoan Do[4], Mislav Blažević[1], Pablo Monteagudo[4], Luka Borozan[1],
Khaled Elbassioni[2], Sören Laue[3], Francisca Rojas Ringeling[4],
Domagoj Matijević[1], and Stefan Canzar[4](✉)

[1] Department of Mathematics, University of Osijek, Osijek, Croatia
[2] Khalifa University of Science and Technology, Abu Dhabi, UAE
[3] Friedrich-Schiller-Universität Jena, Jena, Germany
[4] Gene Center, Ludwig-Maximilians-Universität München, Munich, Germany
canzar@genzentrum.lmu.de

1 Introduction

Single-cell RNA sequencing enables the construction of trajectories [1] describing the dynamic changes in gene expression underlying biological processes such as cell differentiation and development. The comparison of single-cell trajectories under two distinct conditions can illuminate the differences and similarities between the two and can thus be a powerful tool for analysis [2]. Recently developed methods for the comparison of trajectories [2, 3] rely on the concept of dynamic time warping (dtw), originally proposed for the comparison of two time series and consequently restricted to simple, linear trajectories. Here, we adopt and theoretically link arboreal matchings to dtw and implement a suite of exact and heuristic algorithms suitable for the comparison of complex trajectories of different characteristics in our tool Trajan (Fig. 1). Trajan's alignment enables the meaningful comparison of gene expression dynamics along a common pseudo-time scale. Trajan is available at https://github.com/canzarlab/Trajan.

2 Methods

Dynamic time warping (dtw) is the algorithmic workhorse underlying current methods that compare linear single-cell trajectories. We develop Trajan, the first method to compare and align complex trajectories (trees) with multiple branch points. Trajan aligns each path in one tree to at most one path in the second tree and vice versa and, similar to dtw, preserves the order of nodes along the paths. In [4] we have introduced *arboreal matchings* that formalize such a consistent path-by-path alignment of trees.

We devise scoring schemes for arboreal matchings that yield (guaranteed) similar distance measures between linear trajectories as dtw, but naturally

The full version of this paper is available as preprint at bioRxiv 522672.
Van Hoan Do and Mislav Blažević—equal contribution.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 294–296, 2019.
https://doi.org/10.1007/978-3-030-17083-7

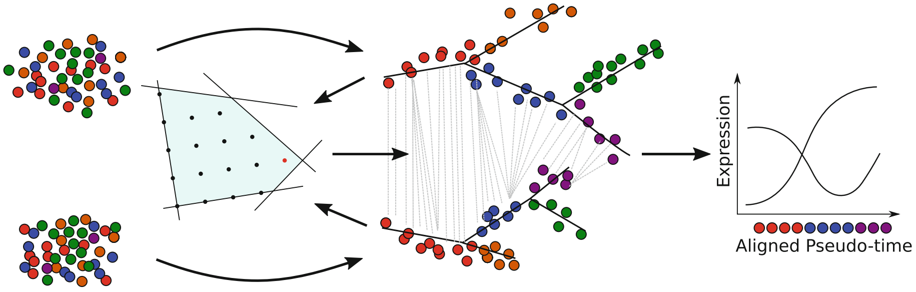


Fig. 1. Complex trajectories, reconstructed from single-cell RNA measurements using, e.g., Monocle 2, are aligned by Trajan based on arboreal matchings. The matching warps individual pseudo-time scales into a shared one along which expression kinetics can be compared.

extend to complex trajectories. Trajan implements a thoroughly engineered branch-and-cut algorithm that allows to practically compare complex single-cell trajectories. It repeatedly determines cutting planes that strengthen the LP relaxation in [4] in polynomial-time and uses an in-house developed, non-commercial, non-linear solver for all continuous optimization problems. For trajectories with a small number of cell fates k we employ a fixed-parameter tractable algorithm, parameterized by k, that applies a dynamic program similar to [5] to align them optimally.

3 Results

Adopting a strategy similar to [2], we re-analyzed two public single-cell datasets: human skeletal muscle myoblast (HSMM) differentiation and human fibroblasts undergoing MYOD-mediated myogenic reprogramming (hFib-MyoD). Trajan is able to align the core paths of each complex trajectory, without any previous knowledge of myoblast differentiation markers. From Trajan's alignment, we construct gene expression kinetics for a set of genes that were assessed in [2] and are able to reproduce their key findings, including the molecular barriers identified in [2] that hinder the efficient reprogramming of fibroblasts to myotubes.

In a perturbation experiment we demonstrate the benefits in terms of robustness and accuracy of our model which compares entire trajectories at once, as opposed to a pairwise application of dtw.

Acknowledgments. Sören Laue has been funded by Deutsche Forschungsgemeinschaft (DFG) under grant LA 2971/1-1. Mislav Blažević was supported in part by BAYHOST. Francisca Rojas Ringeling was supported by the Bavarian Gender Equality Grant (BGF).

References

1. Qiu, X.: Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods **14**(10), 979–982 (2017)
2. Cacchiarelli, D.: Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. Cell Syst. 1–18 (2016)
3. Alpert, A.: Alignment of single-cell trajectories to compare cellular expression dynamics. Nat. Methods **15**(4), 267–270 (2018)
4. Böcker, S., Canzar, S., Klau, G.W.: The generalized robinson-foulds metric. In: Darling, A., Stoye, J. (eds.) WABI 2013. LNCS, vol. 8126, pp. 156–169. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40453-5_13
5. Zhang, K.: Simple fast algorithms for the editing distance between trees and related problems. SIAM J. Comput. **18**, 1245–1262 (1989)

netNMF-sc: A Network Regularization Algorithm for Dimensionality Reduction and Imputation of Single-Cell Expression Data

Rebecca Elyanow[1,2], Bianca Dumitrascu[3], Barbara E. Engelhardt[2,4], and Benjamin J. Raphael[2](✉)

[1] Center for Computational Molecular Biology, Brown University,
Providence, RI 029012, USA

[2] Department of Computer Science, Princeton University,
Princeton, NJ 08540, USA
braphael@princeton.edu

[3] Lewis Sigler Institute for Integrative Genomics,
Princeton University, Princeton, NJ 029012, USA

[4] Center for Statistics and Machine Learning,
Princeton University, Princeton, NJ 08540, USA

Abstract

Motivation. Single-cell RNA-sequencing (scRNA-seq) enables high throughput measurement of RNA expression in individual cells. Due to technical and financial limitations, scRNA-seq datasets often contain zero counts for many transcripts in individual cells. These zero counts, or *dropout events*, complicate the analysis of scRNA-seq data using standard analysis methods developed for bulk RNA-seq data. Current methods for analysis of scRNA-seq data typically overcome dropout by combining information across cells, leveraging the observation that the cells measured in any scRNA-seq experiment generally occupy a small number of RNA expression states.

Results. We describe an algorithm to overcome dropout by combining information across *both* cells and genes. Our algorithm, netNMF-sc, combines network-regularized non-negative matrix factorization with a specialized procedure to handle the large fraction of zero entries in the transcript count matrix. The matrix factorization results in a low-dimensional representation of the transcript count matrix, while the network regularization encourages two genes connected in the network to be close in the low-dimensional representation. In addition, the two matrix factors can be used to cluster cells and to impute values for dropout events. While our netNMF-sc algorithm may use any type of network as prior information, a particularly promising approach is to leverage tissue-specific gene-coexpression networks derived from the vast repository of RNA-seq/microarray studies of bulk tissue.

We show that netNMF-sc outperforms existing methods in both clustering cells and imputing transcript counts on simulated data. netNMF-sc's advantages were especially pronounced at high dropout rates e.g. above 60%. Such high

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 297–298, 2019.
https://doi.org/10.1007/978-3-030-17083-7

dropout rates are common in newer scRNA-seq technologies, such as from 10X Genomics, that measure large number of cells with low sequence coverage per cell. We also show that netNMF-sc outperforms existing methods on real scRNA-seq datasets, including the clustering of mouse embryonic stem cells into cell-cycle states and the clustering of mouse embryonic brain cells into known cell types. Finally, we show that gene-gene correlations computed from the netNMF-sc imputed data are more biologically meaningful than the gene-gene correlations obtained from existing algorithms.

Availability. netNMF-sc is available at https://github.com/raphael-group/netNMF-sc. The preprint is available at https://www.biorxiv.org/content/10.1101/544346v1.

Geometric Sketching of Single-Cell Data Preserves Transcriptional Structure

Brian Hie[1(✉)], Hyunghoon Cho[1], Benjamin DeMeo[2,4], Bryan Bryson[3], and Bonnie Berger[1,4(✉)]

[1] Computer Science and Artificial Intelligence Laboratory, MIT,
Cambridge, MA 02139, USA
bab@mit.edu

[2] Department of Biomedical Informatics, Harvard University,
Cambridge, MA 02138, USA

[3] Department of Biological Engineering, MIT, Cambridge, MA 02139, USA

[4] Department of Mathematics, MIT, Cambridge, MA 02139, USA

1 Introduction

Single-cell RNA-sequencing (scRNA-seq) experiments that profile hundreds of thousands of cells or more are becoming increasingly common. These large-scale data sets present a key computational bottleneck for conventional scRNA-seq analysis pipelines [1]. Standard methods of reducing the size of data sets, such as uniform downsampling, frequently remove rare transcriptional states, mitigating the advantage that large-scale experiments provide. Here we present *geometric sketching*, an efficient downsampling method that newly preserves the transcriptional heterogeneity of single-cell data sets by sampling evenly across transcriptomic space, thinning out dense clusters of common cells and preferentially selecting cells from sparser regions.

We empirically demonstrate that geometric sketches represent the geometry rather than the density of the original data set. We show that our sketches enhance and accelerate downstream analyses by: preserving rare cell types, producing visualizations that capture the full transcriptomic heterogeneity, and facilitating the identification of cell types via clustering. Geometric sketching downsamples from data sets with millions of cells in a matter of minutes, with an asymptotic runtime nearly linear in the size of the data set. As the size of single-cell data grows, geometric sketching will become increasingly crucial for broadening access to single-cell omics experiments even for researchers without expensive computational resources. The full version of this paper can be found at https://www.biorxiv.org/content/10.1101/536730v2.

2 Methods

Geometric sketching is based on the key insight that common cell types form dense clusters in transcriptomic space, while rare cell types may occupy larger

B. Hie and H. Cho—Contributed equally to this work.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 299–301, 2019.
https://doi.org/10.1007/978-3-030-17083-7

regions with much greater sparsity. To accurately summarize the transcriptomic landscape, geometric sketching first obtains a geometric approximation of the data set with equal-sized, non-overlapping, axis-aligned boxes (hypercubes), which we refer to as a plaid covering (Fig. 1). Once the geometry of the data is approximated with a set of covering boxes, we sample cells by uniformly sampling a covering box, then choosing a cell in the box also uniformly at random. The samples therefore more evenly cover the gene expression landscape, naturally diminishing the influence of densely populated regions and increasing the representation of rare transcriptional states.

The plaid covering generalizes grid-based approximation while maintaining computational efficiency in assigning points to their respective covering box. To obtain a plaid covering, we fix an interval length ℓ, and for each coordinate construct a minimal covering of the projected data with intervals of length ℓ. The Cartesian product of these coordinate-wise coverings yields a plaid covering of the original data set by axis-aligned boxes of side length ℓ. Note that after an $O(n \log(n))$ sorting operation, points can be assigned to boxes by rounding up or down, yielding an overall $O(n \log(n))$ runtime in each dimension. In practical scenarios where each coordinate requires only a small constant number of intervals to cover, we achieve $O(n)$ time complexity by using linear scans to find the next interval without sorting. We perform a binary search to find the value of ℓ that produces the number of covering boxes that match the number of samples to be taken. In addition, we use a fast random projection-based PCA to project the data to a relatively low-dimensional space (100 dimensions in the experiments below) before applying the sketching algorithm.

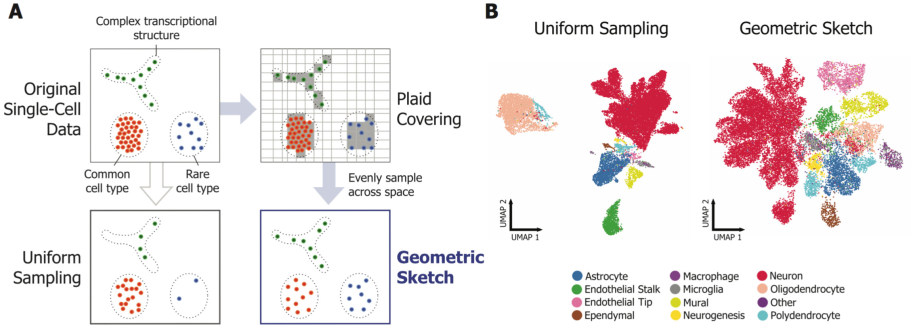


Fig. 1. Geometric sketches capture transcriptional heterogeneity. (**A**) An illustration of the geometric sketching algorithm. (**B**) Geometric sketches more evenly represent the transcriptomic landscape of a data set. Shown are the sketches of 20k cells sampled from a mouse brain data set with 666k cells.

3 Results

Visualizations of geometric sketches reflect the geometric "map" of the transcriptional variability within a data set, allowing researchers to more easily gain insight into rarer transcriptional states (Fig. 1). On data sets with three clusters of similar volumes but different densities, our algorithm samples each cluster with near equal probability (KL divergence = 0.063 versus $\geq$ 0.85 for other sampling methods). Our algorithm also detects rare cell types in a variety of settings: 293T cells mixed with Jurkat cells at a concentration of 0.66%, CD14+ monocytes at a concentration of 1.2%, and macrophages in a mouse brain data set at a concentration of 0.27%. In all cases, rare cell types are substantially better represented in geometric sketches than in subsamples made with other methods, which include spatial random sampling [2] and k-means++ [3], which have not been previously considered for the problem of subsampling scRNA-seq data. Finally, Louvain clustering on data subsampled via geometric sketching resulted in comparable or better agreement with known cell labels across a range of Louvain resolution parameters.

References

1. Angerer, P., et al.: Single cells make big data: new challenges and opportunities in transcriptomics. Curr. Opin. Syst. Biol. (2017)
2. Rahmani, M., Atia, G.K.: Spatial random sampling: a structure-preserving data sketching tool. IEEE Signal Process. Lett. (2017). 1705.03566
3. Arthur, D., Vassilvitskii, S.: K-Means++: the advantages of careful seeding. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (2007). 1212.1121

Sketching Algorithms for Genomic Data Analysis and Querying in a Secure Enclave

Can Kockan[1], Kaiyuan Zhu[1], Natnatee Dokmai[1], Nikolai Karpov[1], M. Oguzhan Kulekci[2], David P. Woodruff[3], and S. Cenk Sahinalp[1](✉)

[1] Department of Computer Science, Indiana University, Bloomington, IN, USA
cenksahi@indiana.edu
[2] Informatics Institute, Istanbul Technical University, Istanbul, Turkey
[3] Department of Computer Science, Carnegie Mellon University, Pittsburgh, USA

Extended Abstract

Current practices in collaborative genomic data analysis (e.g. PCAWG [1]) necessitate all involved parties to exchange individual patient data and perform all analysis locally, or use a trusted server for maintaining all data to perform analysis in a single site (e.g. the Cancer Genome Collaboratory [2]). Since both approaches involve sharing genomic sequence data - which is typically not feasible due to privacy issues, collaborative data analysis remains to be a rarity in genomic medicine.

In order to facilitate efficient and effective collaborative or remote genomic computation we introduce SkSES (Sketching algorithms for Secure Enclave based genomic data analysiS), a computational framework for performing data analysis and querying on multiple, individually encrypted genomes from several institutions in an untrusted cloud environment. Unlike other techniques for secure/privacy preserving genomic data analysis, which typically rely on sophisticated cryptographic techniques with prohibitively large computational overheads, SkSES utilizes the secure enclaves supported by current generation microprocessor architectures such as Intel's SGX. The key conceptual contribution of SkSES is its use of *sketching* data structures that can fit in the limited memory available in a secure enclave.

While streaming/sketching algorithms have been developed for many applications, their feasibility in genomics has remained largely unexplored. On the other hand, even though privacy and security issues are becoming critical in genomic medicine, available cryptographic techniques based on, e.g. homomorphic encryption, secure multi-party computing or garbled circuits, can not always address the performance demands of this rapidly growing field [3–6]. The alternative offered by Intel's SGX, a combination of hardware and software solutions for secure data analysis, is severely limited by the relatively small size of a secure enclave, a private region of the memory protected from other processes [7]. SkSES addresses this limitation through the use of sketching data structures to support efficient secure and privacy preserving SNP analysis across individually

C. Kockan, K. Zhu and N. Dokmai—Joint first authors.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 302–304, 2019.
https://doi.org/10.1007/978-3-030-17083-7

encrypted VCF files from multiple institutions. In particular SkSES provides the users the ability to query for the "k most significant SNPs" among any set of user specified SNPs and any value of k - even when the total number of SNPs to be maintained is far beyond the memory capacity of the secure enclave.

SkSES processes individual genomic data presented as VCF files from participating parties who aim to perform collective statistical tests. For compacting the input VCF files, SkSES uses a simple scheme to filter out non-essential components of a VCF file and encode essential components efficiently - reducing the storage and communication needs and speeding up encryption/decryption within the framework. SkSES then builds a sketch of the compacted VCF files, based on either the count-min sketch [8] or the count sketch [9] structures in order to approximate the actual allele count distribution with respect to L_1 measure (the difference between case and control) - as a proxy to the χ^2 statistic.

Results: We tested SkSES on the extended iDASH-2017 competition data set comprised of 1000 case and 1000 control samples related to an unknown phenotype. SkSES was able to identify the top SNPs with respect to the χ^2 statistic, among any user specified subset of SNPs across this data set of 2000 individually encrypted complete human genomes quickly and accurately - significantly improving our iDASH-2017 (http://www.humangenomeprivacy.org/2017/) runner-up software for secure GWAS - demonstrating the feasibility of secure and privacy preserving computation at human genome scale via Intel's SGX.

Availability: https://github.com/ndokmai/sgx-genome-variants-search
Full Text: https://www.biorxiv.org/content/early/2018/11/12/468355

Acknowledgements. S.C.S. was supported in part by grant NSF CCF-1619081, NIH GM108348 and the Indiana University Grand Challenges Program, Precision Health Initiative; M.O.K. was partially supported by the TUBITAK grant number 114E293; Part of the work was done while D.P.W. was visiting the Simons Institute for the Theory of Computing.

References

1. Campbell, P.J., et al.: Pan-cancer analysis of whole genomes. bioRxiv (2017)
2. Yung, C.K., et al.: Abstract 378: the cancer genome collaboratory. Cancer Res. **77**(13 Supplement), 378–378 (2017)
3. Constable, S.D., et al.: Privacy-preserving gwas analysis on federated genomic datasets. BMC Med. Inform. Decis. Mak. **15**(5), S2 (2015)
4. Zhang, Y., et al.: Secure distributed genome analysis for gwas & sequence comparison computation. BMC Med. Inform. Decis. Mak. **15**(5), S4 (2015)
5. Xie, W., et al.: Securema: protecting participant privacy in genetic association meta-analysis. Bioinformatics **30**(23), 3334–3341 (2014)
6. Cho, H., et al.: Secure genome-wide association analysis using multiparty computation. Nat. Biotechnol. **36**(6), 547 (2018)
7. Chen, F., et al.: Princess: privacy-protecting rare disease international network collaboration via encryption through software guard extensions. Bioinformatics **33**(6), 871–878 (2017)

8. Cormode, G., et al.: An improved data stream summary: the count-min sketch and its applications. J. Algorithms **55**(1), 58–75 (2005)
9. Charikar, M., Chen, K., Farach-Colton, M.: Finding frequent items in data streams. In: Widmayer, P., Eidenbenz, S., Triguero, F., Morales, R., Conejo, R., Hennessy, M. (eds.) ICALP 2002. LNCS, vol. 2380, pp. 693–703. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45465-9_59

Mitigating Data Scarcity in Protein Binding Prediction Using Meta-Learning

Yunan Luo[1], Jianzhu Ma[2], Xiaoming Zhao[1], Yufeng Su[3], Yang Liu[1],
Trey Ideker[2], and Jian Peng[1](✉)

[1] University of Illinois at Urbana-Champaign, Champaign, USA
jianpeng@illinois.edu
[2] University of California San Diego, San Diego, USA
[3] Shanghai Jiao Tong University, Shanghai, China

1 Introduction

A plethora of biological functions are performed through various types of protein-peptide binding, e.g., protein kinase phosphorylation on peptide substrates. Understanding the specificity of protein-peptide interactions is critical for unraveling the architectures of functional pathways and the mechanisms of cellular processes in human cells. A line of computational prediction methods has been recently proposed to predict protein-peptide bindings which efficiently provide rich functional annotations on a large scale. To achieve a high prediction accuracy, these computational methods require a sufficient amount of data to build the prediction model. However, the number of experimentally verified protein-peptide bindings is often limited in real cases. These methods are thus limited to building accurate prediction models for only well-characterized proteins with a large volume of known binding peptides and cannot be extended to predict new binding peptides for less-studied proteins.

2 Methods

We propose a new two phases *meta-learning* framework, named MetaKinase, for the prediction of kinase phosphorylation sites. In phase one, using multiple training kinase families, we train a model which can generate more adaptable representations which are broadly suitable for every kinase family (called meta-learning). In phase two, using a few (e.g., <10) known phosphorylation sites from a new target kinase family, we fine-tune the model on this target family to capture its specificity. With the general patterns captured in phase one, the adaption to the target family in phase two is very sample-efficient: we can tweak the model by only using a few data points to make it family-specific and accurately predict the specificity of the target family (called few-shot learning). With its transferability and fast adaptability, our framework can thus be applied to mitigate the data scarcity issue in characterizing specificities of less-studied kinases. Even with only a few known phosphorylation sites, the model is still able to accurately characterize the specificity of the target kinase family.

Y. Luo and J. Ma—Equal contribution.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 305–307, 2019.
https://doi.org/10.1007/978-3-030-17083-7

3 Results

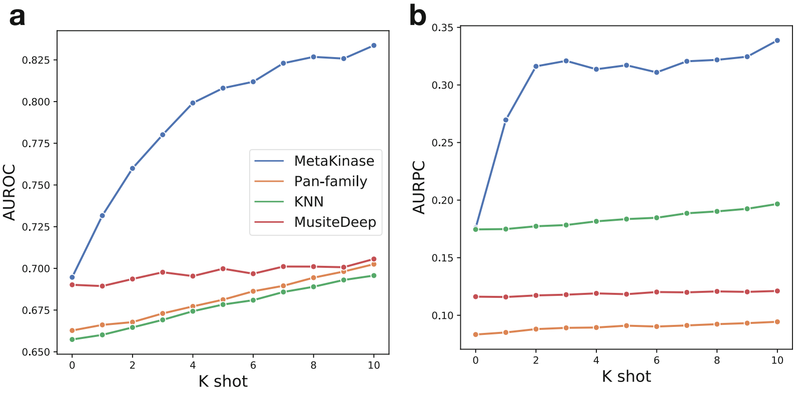


Fig. 1. Evaluation of few-shot learning. MetaKinase was trained with data of multiple kinase families in the meta-learning phase and fine-tuned in the few-shot learning phase using k samples of the test family for $k = 1, 2, \ldots, 10$.

We compared our framework with three baseline methods: pan-family approach (one prediction model for all kinase families), K-nearest neighbor, and MusiteDeep [1]. We varied the value of k-shot from 0 to 10 (0-shot means the model was trained on training family only), and for each value of k, we randomly sampled k samples from the target family and used the remaining samples as test data. The process was repeated for 50 times for each value of k. We used the AUROC and AUPRC scores as the evaluation metrics and showed the results in Fig. 1. We first observed that MetaKinase outperformed other methods for each value of k in terms of both AUROC and AUPRC scores. In addition, while other methods had relatively similar prediction performance as the number of k-shot increased, we observed that the improvement was clear for MetaKinase when more k-shot samples were provided. Our framework also achieved fast adaption to a target family. For example, the predictor had a 0.316 AUPRC score when using 2-shot samples in the few-shot learning phase, which was closed to AUPRC score achieved with 10-shot (0.338). These results demonstrated the transferability and fast-learning ability of MetaKinase. The full paper describing MetaKinase is available at [2].

Acknowledgements. This work was supported in part by the NSF CAREER Award, the Sloan Research Fellowship, the PhRMA Foundation Award in Informatics, and the CompGen Fellowship.

References

1. Wang, D., et al.: Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics **33**, 3909–3916 (2017)
2. Luo, Y. et al.: Mitigating data scarcity in protein binding prediction using meta-learning. bioRxiv (2019). https://doi.org/10.1101/519413

Efficient Estimation and Applications of Cross-Validated Genetic Predictions

Joel Mefford[1(✉)], Danny Park[1], Zhili Zheng[2], Arthur Ko[3],
Mika Ala-Korpela[4,5], Markku Laakso[6], Paivi Pajukanta[3], Jian Yang[2],
John Witte[7], and Noah Zaitlen[8]

[1] Program in Pharmaceutical Science and Pharmacogenomics, UCSF,
San Francisco, USA
joelmefford@gmail.com
[2] Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia
[3] Department of Human Genetics, UCLA, Los Angeles, USA
[4] Systems Epidemiology, Baker Heart and Diabetes Institute,
Melbourne, Victoria, Australia
[5] Computational Medicine, Faculty of Medicine,
University of Oulu and Biocenter Oulu, Oulu, Finland
[6] School of Medicine, University of Eastern Finland, Kuopio, Finland
[7] Departments of Epidemiology and Biostatistics, and Urology, UCSF,
San Francisco, USA
[8] Department of Neurology, UCLA, Los Angeles, USA

While complex traits are highly heritable, individual genetic polymorphisms typically explain only a small proportion of the heritability [1]. Polygenic scores (PS), also known as polygenic risk scores for disease phenotypes, aggregate the contributions of multiple genetic variants to a phenotype [2]. These scores can be calculated using routinely recorded genotypes [1, 2], are strongly associated with heritable traits [1], and are independent of environmental exposures or other factors that are uncorrelated with germ line genetic variants. These properties have motivated a rapidly expanding list of applications from basic science (e.g. causal inference and Mendelian randomization [3], hierarchical disease models and identification of pleiotropy [4]) to translation (e.g. estimating disease risk [5], identifying patients who are likely to respond well to a particular therapy [6], or flagging subjects for modified screening [7]).

Polygenic scores are calculated as a weighted sum of genotypes. This may include all genotyped SNPs, but often only a small set is given nonzero weight – such as a genome spanning but uncorrelated (LD-pruned) set or SNPs with independent evidence of association with the phenotype of interest. Gene-specific polygenic score are also generated using selected sets of SNPs within a region of the genome, such as a window around the coding region of a particular gene [8]. The weights on the SNPs included in a polygenic score are often derived from the regression coefficients of an external GWAS [9, 10], but they may instead be based on predictive models using all SNPs. Joint predictive models include LMMs and their sparse extensions and other regularized regression models such as the lasso or elastic net [8, 11]. The predictions from these joint analyses using genome wide variation are also approximated by post-processing of GWAS summary statistics [8, 11].

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 308–310, 2019.
https://doi.org/10.1007/978-3-030-17083-7

For these SNP-weights to accurately reflect the SNPs' joint association with the phenotype and to generate informative and interpretable polygenic scores, the reference data set must match the target data set in many ways: the populations must have similar ancestry; the trait of interest must be measured; and identical genotypes must be assayed or imputed. Further, the reference data must be large enough to accurately learn the PS weights. An alternative approach is to use the studied data set to build a reference-free PS. This eliminates the need for an external reference data set with matched genotypes, phenotypes, and populations. However, as we show below, naive approaches can easily overfit genetic effects. This overfitting results in PS correlated with non-genetic components of phenotype, that will induce bias or other errors in downstream applications.

Here we report an efficient method to generate PS by using the out-of-sample predictions from a cross-validated linear mixed model (LMM). Our approach generates leave-one-out (LOO) polygenic scores, which we call *cvBLUPs* after a single LMM fit, with computational complexity linear in sample size. In addition to eliminating the reliance on external data and guaranteeing the PS are generated from a relevant population and phenotype, we describe several applications that are only feasible with cvBLUPs. We first demonstrate several desirable statistical properties of cvBLUPs and then consider applications including evidence of polygenicity across metabolic phenotypes, estimation of the shrink term in linear mixed models, a novel formulation of mixed model association studies, and selection of relevant principal components for downstream analyses. To make the results of this work accessible to the community, we have implemented them in the GCTA software package [12].

Full paper on bioRxiv at: https://doi.org/10.1101/517821

References

1. Nolte, I.M., et al.: Missing heritability: is the gap closing? an analysis of 32 complex traits in the lifelines cohort study. Eur. J. Hum. Genet. **25**, 877 (2017)
2. Torkamani, A., Wineinger, N.E., Topol, E.J.: The personal and clinical utility of polygenic risk scores. Nature Rev. Genet. 1 (2018)
3. Burgess, S., Thompson, S.G.: Use of allele scores as instrumental variables for mendelian randomization. Int. J. Epidemiol. **42**, 1134–1144 (2013)
4. Cortes, A., et al.: Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. Nature Genet. **49**, 1311 (2017)
5. Maas, P., et al.: Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. JAMA Oncol. **2**, 1295–1302 (2016)
6. Natarajan, P., et al.: Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. Circulation **135**, 2091–2101 (2017)
7. Seibert, T.M., et al.: Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. bmj **360**, j5757 (2018)
8. Gusev, A., et al.: Integrative approaches for large-scale transcriptome-wide association studies. Nature Genet. **48**, 245 (2016)
9. Dudbridge, F.: Polygenic epidemiology. Genet. Epidemiol. **40**, 268–272 (2016)

10. Wray, N.R., Goddard, M.E., Visscher, P.M.: Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. **17** (2007)
11. Vilhjálmsson, B.J., et al.: Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. **97**, 576–592 (2015)
12. Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M.: GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. **88**, 76–82 (2011)

Inferring Tumor Evolution from Longitudinal Samples

Matthew A. Myers[1], Gryte Satas[1,2], and Benjamin J. Raphael[1(✉)]

[1] Department of Computer Science, Princeton University, Princeton, NJ 08540, USA
braphael@princeton.edu
[2] Department of Computer Science, Brown University, Providence, RI 02912, USA

Abstract

Background: Determining the clonal composition and somatic evolution of a tumor greatly aids in accurate prognosis and effective treatment for cancer. In order to understand how a tumor evolves over time and/or in response to treatment, multiple recent studies have performed longitudinal DNA sequencing of tumor samples from the same patient at several different time points. However, none of the existing algorithms that infer clonal composition and phylogeny using several bulk tumor samples from the same patient integrate the information that these samples were obtained from longitudinal observations.

Results: We introduce a model for a longitudinally-observed phylogeny and derive constraints that longitudinal samples impose on the reconstruction of a phylogeny from bulk samples. These constraints form the basis for a new algorithm, Cancer Analysis of Longitudinal Data through Evolutionary Reconstruction (CALDER), which infers phylogenetic trees from longitudinal bulk DNA sequencing data. We show on simulated data that constraints from longitudinal sampling can substantially reduce ambiguity when deriving a phylogeny from multiple bulk tumor samples, each a mixture of tumor clones. On real data, where there is often considerable uncertainty in the clonal composition of a sample, longitudinal constraints yield more parsimonious phylogenies with fewer tumor clones per sample. We demonstrate that CALDER reconstructs more plausible phylogenies than existing methods on two longitudinal DNA sequencing datasets from chronic lymphocytic leukemia patients. These findings show the advantages of directly incorporating temporal information from longitudinal sampling into tumor evolution studies.

Availability: CALDER is available at https://github.com/raphael-group.

Preprint: Preprint version of the full manuscript is available at https://www.biorxiv.org/content/10.1101/526814v1.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, p. 311, 2019.
https://doi.org/10.1007/978-3-030-17083-7

Scalable Multi-component Linear Mixed Models with Application to SNP Heritability Estimation

Ali Pazokitoroudi[1], Yue Wu[1], Kathryn S. Burch[2], Kangcheng Hou[3,4], Bogdan Pasaniuc[3,5,6], and Sriram Sankararaman[1,5,6](✉)

[1] Department of Computer Science, UCLA, Los Angeles, CA, USA
[2] Bioinformatics Interdepartmental Program, UCLA, Los Angeles, CA, USA
[3] Department of Pathology and Laboratory Medicine,
David Geffen School of Medicine, UCLA, Los Angeles, CA, USA
[4] College of Computer Science and Technology, Zhejiang University,
Hangzhou, Zhejiang, China
[5] Department of Human Genetics, David Geffen School of Medicine,
UCLA, Los Angeles, CA, USA
[6] Department of Computational Medicine, David Geffen School of Medicine,
UCLA, Los Angeles, CA, USA
sriram@cs.ucla.edu

A central question in human genetics is to find the proportion of variation in a trait that can be explained by genetic variation [1]. A number of methods have been developed to estimate this quantity, termed narrow-sense heritability, from genome-wide SNP data [2–6]. Recently, it has become clear that estimates of narrow-sense heritability are sensitive to modeling assumptions that relate the effect sizes of a SNP to its minor allele frequency (MAF) and linkage disequilibrium (LD) patterns [6, 7]. A principled approach to estimate heritability while accounting for variation in SNP effect sizes involves the application of linear Mixed Models (LMMs) [8] with multiple variance components where each variance component represents the fraction of genetic variance explained by SNPs that belong to a given range of MAF and LD values. Beyond their importance in accurately estimating genome-wide SNP heritability, multiple variance component LMMs are useful in partitioning the contribution of genomic annotations to trait heritability which, in turn, can provide insights into biological processes that are associated with the trait.

Existing methods for fitting multi-component LMMs rely on maximizing the likelihood of the variance components. These methods pose major computational bottlenecks that makes it challenging to apply them to large-scale genomic datasets such as the UK Biobank which contains half a million individuals genotyped at tens of millions of SNPs.

We propose a scalable algorithm, RHE-reg-mc, to jointly estimate multiple variance components in LMMs. RHE-reg-mc is a randomized method-of-moments estimator with a runtime that is observed to scale as $\mathcal{O}(\frac{NMB}{\max(\log_3(N),\log_3(M))} + k^3)$ for N individuals, M SNPs, k variance components, and $B \approx 10$, a parameter that controls the number of random matrix-vector multiplication. RHE-reg-mc also efficiently computes asymptotic and jackknife standard errors. We evaluate the accuracy and scalability of RHE-reg-mc for estimating the total heritability as well as in partitioning heritability.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 312–313, 2019.
https://doi.org/10.1007/978-3-030-17083-7

The ability to fit multiple variance components to SNPs partitioned according to their MAF and local LD allows RHE-reg-mc to obtain relatively unbiased estimates of SNP heritability. On the UK Biobank dataset consisting of $\approx 300,000$ individuals and $\approx 500,000$ SNPs, RHE-reg-mc can fit 250 variance components, corresponding to genetic variance explained by 10 MB blocks, in ≈ 40 minutes on standard hardware. The full version of the paper is available at: http://biorxiv.org/cgi/content/short/522003v2.

References

1. Visscher, P.M., Hill, W.G., Wray, N.R.: Heritability in the genomics era: concepts and misconceptions. Nat. Rev. Genet. **9**(4), 255 (2008)
2. Yang, J., et al.: Common snps explain a large proportion of the heritability for human height. Nat. Genet. **42**(7), 565 (2010)
3. Zhou, X.: A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann. Appl. Stat. **11**(4), 2027 (2017)
4. Lee, S.H., Wray, N.R., Goddard, M.E., Visscher, P.M.: Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. **88**(3), 294–305 (2011)
5. Golan, D., Lander, E.S., Rosset, S.: Measuring missing heritability: inferring the contribution of common variants. Proc. Nat. Acad. Sci. **111**(49), E5272–E5281 (2014)
6. Speed, D., Hemani, G., Johnson, M.R., Balding, D.J.: Improved heritability estimation from genome-wide snps. Am. J. Hum. Genet. **91**(6), 1011–1021 (2012)
7. Evans, L.M., et al.: Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. Nat. Genet. **50**(5), 737 (2018)
8. McCulloch, C.E., Searle, S.R.: Generalized, linear, and mixed models. John Wiley & Sons, (2004)

A Note on Computing Interval Overlap Statistics

Shahab Sarmashghi[1] and Vineet Bafna[2](✉)

[1] Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA
ssarmash@ucsd.edu

[2] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA
vbafna@cs.ucsd.edu

Extended Abstract

We consider the following problem: Let I and I_f each describe a collection of n and m non-overlapping intervals on a line segment of finite length. Suppose that k of the m intervals of I_f are intersected by some interval(s) in I. Under the null hypothesis that intervals in I are randomly arranged w.r.t I_f, what is the significance of this overlap? This is a natural abstraction of statistical questions that are ubiquitous in the post-genomic era. The interval collections represent annotations that reveal structural or functional regions of the genome, and overlap statistics can provide insight into the correlation between different structural and functional regions. However, the statistics of interval overlaps have not been systematically explored. We propose a combinatorial algorithm for a constrained interval overlap problem that can accurately compute very small p-values. Specifically, we define $N(i,h,k,a)$ as the number of randomized arrangements of the first i intervals in I such that the i-th interval ends at genomic location h, and k intervals in I_f are hit by the first i intervals in I (a is an auxiliary binary variable). Assuming that the order of intervals in I is retained, $N(i,h,k,a)$ is computed using a dynamic programming algorithm in pseudo-polynomial time $\mathcal{O}(ngm)$ [1], where n and m are the number of intervals in I and I_f, and g is the genome length. The p-value of the overlap is then given by

$$P\text{-value}(k) = \frac{\sum_{\kappa=k}^{m} N_1(n,g,\kappa,0)}{\sum_{\kappa=0}^{m} N_1(n,g,\kappa,0)}.$$

We have also provided a fast approximate method based on Poisson binomial distribution to facilitate problems consisted of very large number of intervals, and have introduced parameter η as a measure of the spread of intervals to estimate the closeness of approximated p-values.

We tested our tool, iStat, on simulated interval data to obtain precise estimates of low p-values, and characterize the performance of our methods. We also applied iStat to four cases of interval overlap problem from previous studies, and showed that iStat can estimate very small p-values, considering the length

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 314–315, 2019.
https://doi.org/10.1007/978-3-030-17083-7

and structure of intervals, while avoiding inflated p-values reported from basic permutation or parametric tests. The iStat software is made publicly available on Github (https://github.com/shahab-sarmashghi/ISTAT.git).

Reference

1. Sarmashghi, S., Bafna, V.: A Note on Computing Interval Overlap Statistics. bioRxiv, p. 517987, January 2019. https://doi.org/10.1101/517987, https://www.biorxiv.org/content/early/2019/01/11/517987.full.pdf+html

Distinguishing Biological from Technical Sources of Variation Using a Combination of Methylation Datasets

Mike Thompson[1](✉), Zeyuan Johnson Chen[1], Elior Rahmani[1],
and Eran Halperin[1,2,3,4](✉)

[1] Department of Computer Science, University of California Los Angeles,
Los Angeles, CA, USA
mjthompson@ucla.edu, ehalperin@cs.ucla.edu
[2] Department of Human Genetics, University of California Los Angeles,
Los Angeles, CA, USA
[3] Department of Anesthesiology and Perioperative Medicine, University of California
Los Angeles, Los Angeles, CA, USA
[4] Department of Biomathematics, University of California Los Angeles,
Los Angeles, CA, USA

DNA methylation remains one of the most widely studied epigenetic markers. One of the major challenges in population studies of methylation is the presence of global methylation effects that may mask local signals [1, 2]. Such global effects may be due to either technical effects (e.g., batch effects) or biological effects (e.g., cell type composition, genetics). Many methods have been developed for the detection of such global effects, typically in the context of Epigenome-wide association studies [3–9]. However, current unsupervised methods do not distinguish between biological and technical effects, resulting in a loss of highly relevant information. Though supervised methods can be used to estimate known biological effects, it remains difficult to identify and estimate unknown biological effects that globally affect the methylome.

Here, we propose CONFINED (CCA ON Features for INter- dataset Effect Detection), a reference-free method based on sparse canonical correlation analysis (CCA) that captures replicable sources of variation across multiple methylation datasets such as age, sex, and cell-type composition and distinguishes them from dataset-specific sources of variability (e.g., technical effects). Our method is based on the observation that the same biological sources of variation typically affect different studies that are performed under the same conditions (e.g., on the same tissue type), while technical variability is study-specific. Thus, unlike previous unsupervised methods that utilize single-matrix decomposition techniques to account for covariates in methylation data, we propose the use of canonical correlation analysis, which captures shared signal across multiple datasets. Nonetheless, there are two substantial differences between CONFINED and traditional uses of CCA in genomic studies. First, CONFINED looks for shared structure of one methylation profile across two sets of individuals rather than looking for shared structure in one set of individuals across two sets of genomic measurements. Second, CONFINED performs a feature selection procedure that is critical to detect the shared sources of variability across the different datasets.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 316–317, 2019.
https://doi.org/10.1007/978-3-030-17083-7

Across several datasets we demonstrate that CONFINED accurately captures global biological sources of variability. Specifically, we shrow through simulated and real data that our approach captures replicable sources of biological variation such as age, sex, and cell-type composition better than the state-of-the-art methods and is considerably more robust to technical noise than previous reference-free methods. Additionally, we demonstrate that the features selected by CONFINED recapitulate biological functionality inherent to both datasets. For example, when pairing two whole-blood datasets together, the sites best ranked by CONFINED were significantly enriched for immune cell function.

CONFINED is available at https://github.com/cozygene/CONFINED as an R package. The calculations in the R package were optimized with C++ code using Rcpp and RcppArmadillo. Also included in the package is an ultra-fast function for performing CCA. The preprint of the manuscript can be found at https://www.biorxiv.org/content/early/2019/01/16/521146.

References

1. Schmidt, F., List, M., Cukuroglu, E., Köhler, S., Göke, J., Schulz, M.H.: An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. Bioinformatics **34**(17), i908–i916 (2018)
2. Maksimovic, J., Gagnon-Bartsch, J.A., Speed, T.P., Oshlack, A.: Removing unwanted variation in a differential methylation analysis of illumina humanmethylation450 array data. Nucleic Acids Res. **43**, e106–e106 (2015)
3. Rahmani, E., et al.: Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat. Methods **13**, 443 (2016)
4. Zou, J., Lippert, C., Heckerman, D., Aryee, M., Listgarten, J.: Epigenome-wide association studies without the need for cell-type composition. Nat. Methods **11**, 309 (2014)
5. Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T., Marsit, C.J.: Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinf. **17**, 259 (2016)
6. Lutsik, P., Slawski, M., Gasparoni, G., Vedeneev, N., Hein, M., Walter, J.: Medecom: discovery and quantification of latent components of heterogeneous methylomes. Genome Biol. **18**, 55 (2017)
7. Rahmani, E., et al.: Bayescce: a bayesian framework for estimating cell-type composition from dna methylation without the need for methylation reference. Genome Biol. **19**, 141 (2018)
8. Houseman, E.A., Molitor, J., Marsit, C.J.: Reference-free cell mixture adjustments in analysis of dna methylation data. Bioinformatics **30**(10), 1431–1439 (2014)
9. Rahmani, E., et al.: Correcting for cell-type heterogeneity in DNA methylation: a comprehensive evaluation. Nat. Methods **14**, 218 (2017)

GRep: Gene Set Representation via Gaussian Embedding

Sheng Wang[2,3], Emily Flynn[1], and Russ B. Altman[1,2,3](✉)

[1] Biomedical Informatics Training Program, Stanford University,
Stanford, CA 94035, USA
[2] Department of Bioengineering, Stanford University, Stanford, CA 94035, USA
[3] Department of Genetics, Stanford University, Stanford, CA 94035, USA
russ.altman@stanford.edu

1 Introduction

Molecular interaction networks are our basis for understanding functional interdependencies among genes. Network embedding approaches analyze these complicated networks by representing genes as low-dimensional vectors based on the network topology. These low-dimensional vectors have recently become the building blocks for a larger number of systems biology applications. Despite the success of embedding genes in this way, it remains unclear how to effectively represent gene sets, such as protein complexes and signaling pathways. The direct adaptation of existing gene embedding approaches to gene sets cannot model the diverse functions of genes in a set. Here, we propose GRep, a novel gene set embedding approach, which represents each gene set as a multivariate Gaussian distribution rather than a single point in the low-dimensional space. The diversity of genes in a set, or the uncertainty of their contribution to a particular function, is modeled by the covariance matrix of the multivariate Gaussian distribution. By doing so, GRep produces a highly informative and compact gene set representation. Using our representation, we analyze two major pharmacogenomics studies and observe substantial improvement in drug target identification from expression-derived gene sets. Overall, the GRep framework provides a novel representation of gene sets that can be used as input features to off-the-shelf machine learning classifiers for gene set analysis. A full version of the paper can be found on bioRxiv https://www.biorxiv.org/content/early/2019/01/13/519033.

2 Methods

Biologically meaningful gene sets, such as signaling pathways and protein complexes, aggregate gene level information into higher level patterns. A key observation behind our approach is that gene sets can have diverse molecular functions and/or biological processes. GRep explicitly models this diversity as a low-dimensional Gaussian distribution which summarizes both location and uncertainty of each dimension. To summarize, GRep takes a network and a collection

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 318–319, 2019.
https://doi.org/10.1007/978-3-030-17083-7

of gene sets as input. It first calculates the diffusion states of each gene and gene set to characterize their topological information in the network. GRep then finds the low-dimensional representations for genes and gene sets according to these diffusion states. Each gene is represented as a single point in the low-dimensional space. Each gene set is represented as a multivariate Gaussian distribution which is parameterized by a mean vector and a covariance matrix. In this paper, we present GRep (Gene set Representation), a novel computational method that represents each gene set as a highly informative and compact multivariate Gaussian distribution. GRep takes a biological network and a collection of gene sets as input. It represents each gene as a single point and each gene set as a multivariate Gaussian distribution parameterized by a low-dimensional mean vector and a low-dimensional covariance matrix. The mean vector of each gene set describes the joint contribution of genes in this gene set, and the covariance matrix characterizes the agreement among individual genes in each dimension. By using this representation, GRep is able to differentiate between gene sets that would be considered equivalent by average embedding. The key idea of GRep is to use the prior knowledge in gene sets and group genes in the same set closely as a multivariate Gaussian distribution in the low-dimensional space. To achieve this, GRep solves an optimization problem to preserve the network topology according to diffusion states. We evaluate GRep on a collection of drug response correlated gene sets derived from Genomics of Drug Sensitivity in Cancer (GDSC) and The Cancer Therapeutic Portal (CTRP). We demonstrate that representing those gene sets using GRep substantially outperforms comparison approaches on drug-target identification in both datasets.

3 Results

To evaluate GRep, we performed large-scale drug target identification on two pharmacogenomics studies, GDSC and CTRP. Our approach significantly outperforms comparison approaches on both datasets. In CTRP, our method achieved 0.8667 AUROC, which is much higher than 0.7102 AUROC of plain average embedding, 0.7104 of weighted gene set average embedding and 0.7319 AUROC of weighted average embedding. The same improvement was observed on GDSC where our method achieved 0.8890 AUROC, which is again substantially higher than 0.6870 AUROC of plain average embedding, 0.7325 of weighted average embedding and 0.6870 AUROC of weighted gene set average embedding. All improvements were statistically significant ($P < 0.05$; paired Wilcoxon signed-rank test). The above results suggest that representing a gene set through simple averaging is not able to modeling uncertainty, leading to worse performance. By incorporating prior knowledge about gene sets and jointly optimizing the gene and gene set representations, our method substantially improved drug target identification.

Accurate Sub-population Detection and Mapping Across Single Cell Experiments with PopCorn

Yijie Wang, Jan Hoinka, and Teresa M. Przytycka(✉)

National Center of Biotechnology Information, National Library of Medicine, NIH,
Bethesda, MD 20894, USA
przytyck@ncbi.nlm.nih.gov

Extended Abstract

Recent technological advances have facilitated unprecedented opportunities for studying biological systems at single-cell level resolution. One notable example is single-cell RNA sequencing (scRNA-seq), which enables the measurement of transcriptomic information of thousands of individual cells in one experiment. Single cell measurements open the ability of capturing the heterogeneity of a population of cells and thus provide information that is not accessible using bulk sequencing. Among its many applications, scRNA-seq is more prominently employed in the identification of sub-populations of cells present in a sample, and for comparative analysis of such sub-populations across samples [3–6, 8–11].

We report PopCorn (single cell Populations Comparison)– a new method allowing for the identification of sub-populations of cells present within individual experiments and their mapping across experiments. PopCorn uses several innovative ideas to perform this task accurately. First, in contrast to previous approaches, PopCorn performs the two tasks (sub-population identification and mapping) simultaneously by optimizing a function that combines both objectives. This allows for integrating information across experiments and reducing noise. The second key innovation consists of a new approach to identify sub-populations of cells within a given experiment. Specifically, PopCorn utilizes Personalized PageRank vectors [1] and a quality measure of cohesiveness of a cell population to perform this task. Finally, the simultaneous identification of sub-populations within each experiment and their mapping across experiments uses a graph theoretical approach.

We tested the performance of PopCorn in two distinct settings. We demonstrated its potential in identifying and aligning sub-populations informed by single cell data from human and mouse pancreatic singe cell data [2]. In addition, we applied PopCorn to the task of aligning biological replicates of mouse kidney single cell data [7]. In both scenarios PopCorn achieved a striking improvement over alternative tools.

Taken together, our results demonstrate that PopCorn's novel approach provides a powerful tool for comparative analysis of single-cells sub-populations.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 320–321, 2019.
https://doi.org/10.1007/978-3-030-17083-7

The preprint of the manuscript is available at https://www.biorxiv.org/content/early/2018/12/28/485979.article-metrics.

References

1. Andersen, R., Chung, F., Lang, K.: Local graph partitioning using pagerank vectors. In: FOCS, pp. 475–486 (2006)
2. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R.: Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. **36**(5), 411–420 (2018)
3. Byrnes, L.E., et al.: Lineage dynamics of murine pancreatic development at single-cell resolution. Nat. Commun. **9**(1), 3922 (2018)
4. Duan, L., et al.: PDGFRβ Cells Rapidly Relay Inflammatory Signal from the Circulatory System to Neurons via Chemokine CCL2. Neuron **100**(1), 183–200 (2018)
5. Mayer, C., et al.: Developmental diversification of cortical inhibitory interneurons. Nature **555**(7697), 457–462 (2018)
6. Ordovas-Montanes, J., et al.: Allergic inflammatory memory in human respiratory epithelial progenitor cells. Nature **560**(7720), 649–654 (2018)
7. Park, J., et al.: Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science **360**(6390), 758–763 (2018)
8. Paulson, K.G., et al.: Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. Nat. Commun. **9**(1), 3868 (2018)
9. Shrestha, B.R., Chia, C., Wu, L., Kujawa, S.G., Liberman, M.C., Goodrich, L.V.: Sensory neuron diversity in the inner ear is shaped by activity. Cell **174**(5), 1229–1246 (2018)
10. Sun, S., et al.: Hair cell mechanotransduction regulates spontaneous activity and spiral ganglion subtype specification in the auditory system. Cell **174**(5), 1247–1263 (2018)
11. Verma, M., et al.: Muscle satellite cell cross-talk with a vascular niche maintains quiescence via VEGF and notch signaling. Cell Stem Cell **23**(4), 530–543 (2018)

Fast Estimation of Genetic Correlation for Biobank-Scale Data

Yue Wu[1], Anna Yaschenko[3], Mohammadreza Hajy Heydary[4],
and Sriram Sankararaman[1,2](✉)

[1] Department of Computer Science, UCLA, Los Angeles, USA
[2] Department of Human Genetics, UCLA, Los Angeles, USA
sriram@cs.ucla.edu
[3] Department of Computer Science and Electrical Engineering,
University of Maryland, Baltimore County, Baltimore, USA
[4] Department of Computer Science, California State University, Fullerton,
Fullerton, USA

Genetic correlation, *i.e.*, the proportion of phenotypic correlation across a pair of traits that can be explained by genetic variation, is an important parameter in efforts to understand the relationships among complex traits [1]. The observation of substantial genetic correlation across a pair of traits, can provide insights into shared genetic pathways as well as providing a starting point to investigate causal relationships. Attempts to estimate genetic correlations among complex phenotypes attributable to genome-wide SNP variation data have motivated the analysis of large datasets as well as the development of sophisticated methods.

Bi-variate Linear Mixed Models (LMMs) have emerged as a key tool to estimate genetic correlation from datasets where individual genotypes and traits are measured [2]. The bi-variate LMM jointly models the effect sizes of a given SNP on each of the pair of traits being analyzed. The parameters of the bi-variate LMM, *i.e.*, the variance components, are related to the heritability of each trait as well as correlation across traits attributable to genotyped SNPs. The most commonly used method for estimating genetic correlation as well as trait heritabilities in a bi-variate LMM relies on the restricted maximize likelihood method, termed genomic restricted maximum likelihood (GREML) [3–6] However, GREML poses serious computational burdens. GREML is a non-convex optimization problem that relies on an iterative optimization algorithm.

Another state-of-the-art method, LD-score regression (LDSC), requires only summary statistics from genome-wide association studies (GWAS) to estimate genetic correlations [1]. As LD-score preserves privacy and has substantially reduced computational requirements (assuming that the summary statistics have been computed), LDSC has some drawbacks: its estimates tend to have large standard errors and is prone to bias in some settings [7].

We propose, RG-Cor, a scalable randomized Method-of-Moments (MoM) estimator of genetic correlations in bi-variate LMMs. RG-Cor leverages the structure of genotype data to obtain runtimes that scale sub-linearly with the number of individuals in the input dataset (assuming the number of SNPs is held constant). We perform extensive simulations to validate the accuracy and scalability of RG-Cor. Compared to GREML estimators, we show that the loss in

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 322–323, 2019.
https://doi.org/10.1007/978-3-030-17083-7

statistical inefficiency of RG-Cor is fairly modest. On the other hand, RG-Cor is several orders of magnitude faster than other methods. RG-Cor can compute the genetic correlations on the UK biobank dataset consisting of 430,000 individuals and 460,000 SNPs in 3 hours on a stand-alone compute machine.

Link to the full paper: https://www.biorxiv.org/content/early/2019/01/20/525055

References

1. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., et al.: An atlas of genetic correlations across human diseases and traits. Nat. Genet. **47**(11), 1236 (2015)
2. Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M.: GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. (2010)
3. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., Wray, N.R.: Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics **28**(19), 2540–2542 (2012)
4. Chen, G.-B.: Estimating heritability of complex traits from genome-wide association studies using ibs-based haseman-elston regression. Front. Genet. **5**, 107 (2014)
5. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., et al.: Efficient bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. **47**(3), 284 (2015)
6. Loh, P.-R., Bhatia, G., Gusev, A., et al.: Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet. **47**(12), 1385 (2015)
7. Ni, Guiyan, Moser, Gerhard, Ripke, Stephan, et al.: Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. Am. J. Hum. Genet. (2018)

Distance-Based Protein Folding Powered by Deep Learning

Jinbo Xu(✉)

Toyota Technological Institute at Chicago, Chicago, USA
jinboxu@gmail.com

Accurate description of protein structure and function is a fundamental step towards understanding biological life and highly relevant in the development of therapeutics. Although greatly improved, experimental protein structure determination is still low-throughput and costly, especially for membrane proteins. Predicting the structure of a protein with a new fold is very challenging and usually needs a large amount of computing power. We show that we can accurately predict the distance matrix of a protein by deep learning (DL), even for proteins with few sequence homologs. Using only the geometric constraints given by the resulting distance matrix we may construct 3D models without involving any folding simulation.

This work is an extension of our previous CASP-winning deep learning method RaptorX-Contact [1] that uses deep and global (or fully) convolutional residual neural network (ResNet) to predict protein contacts. ResNet is one type of DCNN (deep convolutional neural network), but much more powerful than the traditional DCNN. RaptorX-Contact is the first DL method that greatly outperforms DCA (direct coupling analysis) and shallow learning methods such as the CASP11 winner MetaPSICOV. The accuracy of RaptorX-Contact decreases much more slowly than DCA when more predicted contacts are evaluated even when the protein under study has thousands of sequence homologs (see Table 1 in the paper [1]). As reported in [1, 2], without folding simulation, RaptorX-Contact may produce much better 3D models than DCA methods such as CCMpred and shallow methods such as MetaPSICOV. RaptorX-Contact also works well for membrane proteins even trained by soluble proteins [2] and for complex contact prediction even trained by single-chain proteins [3]. Inspired by the success of RaptorX-Contact, many CASP13 participants have adopted global ResNet or DCNN into their prediction pipeline, as shown in the CASP13 abstract book, and made very good progress. As a result, CASP13 has achieved the largest progress in the history of CASP.

Instead of contact prediction, here we study distance prediction. The distance matrix contains finer-grained information than contact matrix and provides more physical constraints of a protein structure, e.g., distance is metric while contact is not. A distance matrix can determine a protein structure (except mirror image) much more accurately than a contact matrix. Different from DCA that aims to predict only a small number of contacts and then use them to assist folding simulation, we predict the whole distance matrix and then directly construct protein 3D models without invoking any folding simulation at all. This significantly reduces running time needed for protein folding, especially for a large protein. Distance prediction is not totally new. In addition

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 324–325, 2019.
https://doi.org/10.1007/978-3-030-17083-7

to few previous studies, my group employed a probabilistic neural network to predict inter-residue distance and then derived protein- and position-specific statistical potential from predicted distance distribution [4]. We have also studied folding simulation using this distance-based statistical potential [5]. Recently, we showed that protein-specific distance potential derived from deep ResNet may improve by a large margin protein threading with weakly similar templates [6].

We feed our predicted distance into CNS to generate 3D models for a protein under prediction. Our method successfully folded 21 of the 37 CASP12 hard targets with a median family size of 58 effective sequence homologs within 4 h on a Linux computer of 20 CPUs. In contrast, DCA cannot fold any of these hard targets in the absence of folding simulation, and the best CASP12 group folded only 11 of them by integrating DCA-predicted contacts into complex, fragment-based folding simulation. Rigorous experimental validation in CASP13 shows that our distance-based folding server successfully folded 17 of 32 hard targets (with a median family size of 36 sequence homologs) and obtained 70% precision on top L/5 long-range predicted contacts. In CASP13, our method was officially ranked first in terms of contact prediction accuracy among all CASP13 groups and our server was ranked second among all CASP13-participating servers in terms of tertiary structure prediction.

An extended version of this abstract is available at https://www.biorxiv.org/content/early/2018/12/20/465955 and https://arxiv.org/abs/1811.03481.

References

1. Wang, S., Sun, S., Li, Z., Zhang, R., Xu, J.: Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput. Biol. **13**, e1005324 (2017)
2. Wang, S., Li, Z., Yu, Y., Xu, J.: Folding membrane proteins by deep transfer learning. Cell Syst. **5**, 202–211 (2017). e203
3. Zeng, H., et al.: ComplexContact: a web server for inter-protein contact prediction using deep learning. Nucleic Acids Res. **46**, W432–W437 (2018)
4. Zhao, F., Xu, J.: A position-specific distance-dependent statistical potential for protein structure and functional study. Structure **20**, 1118–1126 (2012)
5. Wang, Z.: Knowledge-based machine learning methods for macromolecular 3D structure prediction. Ph.D. thesis (2016)
6. Zhu, J.W., Wang, S., Bu, D.B., Xu, J.B.: Protein threading using residue co-variation and deep learning. Bioinformatics **34**, 263–273 (2018)

Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF

Yang Yang[1], Yang Zhang[1], Bing Ren[2], Jesse Dixon[3], and Jian Ma[1](✉)

[1] Computational Biology Department, School of Computer Science,
Carnegie Mellon University, Pittsburgh, USA
jianma@cs.cmu.edu

[2] Ludwig Institute for Cancer Research, Department of Cellular and Molecular Medicine, Moores Cancer Center and Institute of Genomic Medicine,
UCSD School of Medicine, San Diego, USA

[3] Salk Institute for Biological Studies, San Diego, USA

Recent developments in whole-genome mapping approaches for the chromatin interactome (such as Hi-C) have facilitated the identification of genome-wide three-dimensional (3D) chromatin organizations comprehensively, and offered new insights into 3D genome architecture. However, our knowledge of the evolutionary patterns of 3D genome structures in mammalian species remains surprisingly limited. In particular, there are no existing phylogenetic-model based methods to analyze chromatin interactions as continuous features across different species to uncover evolutionary patterns of 3D genome organization.

Here we develop a new probabilistic model, named phylogenetic hidden Markov random field (Phylo-HMRF), to identify evolutionary patterns of 3D genome structures based on multi-species Hi-C data by jointly utilizing spatial constraints among genomic loci and continuous-trait evolutionary models. Specifically, Phylo-HMRF integrates the continuous-trait evolutionary constraints (based on Ornstein-Uhlenbeck process in this work) with the hidden Markov random field (HMRF) model, enabling the joint modeling of general types of spatial dependencies among genomic loci and evolutionary temporal dependencies among species. The overview of Phylo-HMRF is shown in Fig. 1. The effectiveness of Phylo-HMRF is demonstrated in both simulation evaluation and application to real Hi-C data. We used Phylo-HMRF to uncover cross-species 3D genome patterns based on Hi-C data from the same cell type in four primate species (human, chimpanzee, bonobo, and gorilla). Phylo-HMRF identified genome-wide evolutionary patterns of Hi-C contact frequency across the four species, including conserved patterns and lineage-specific patterns. The identified evolutionary patterns of 3D genome organization correlate with other features of genome structure and function, including long-range interactions, topologically-associating domains (TADs), and replication timing patterns.

This work provides a new framework that utilizes general types of spatial constraints to identify evolutionary patterns of continuous genomic features and has the potential to reveal the evolutionary principles of 3D genome organization.

Link to the bioRxiv preprint: doi: http://doi.org/10.1101/552505.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 326–327, 2019.
https://doi.org/10.1007/978-3-030-17083-7

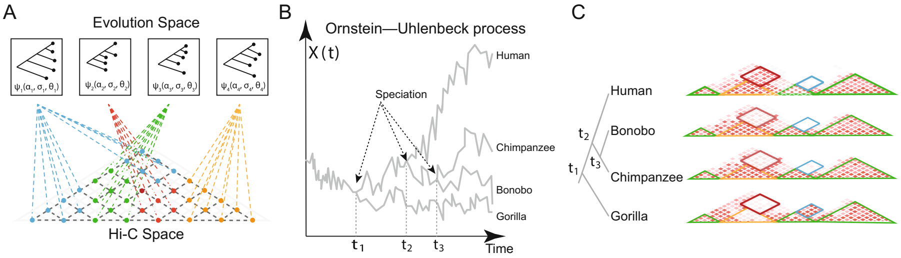


Fig. 1. Overview of Phylo-HMRF. (A) Illustration of the possible evolutionary patterns of chromatin interaction. The Hi-C space is a combined multi-species Hi-C contact map, which integrates aligned Hi-C contact maps of each species. Each node represents the multi-species observations of Hi-C contact frequency between a pair of genomic loci, with a hidden state assigned. Nodes with the same color have the same hidden state and are associated with the same type of evolutionary pattern represented by a parameterized phylogenetic tree ψ_i. The parameters of ψ_i include the selection strengths α_i, Brownian motion intensities σ_i, and the optimal values θ_i based on the Ornstein-Uhlenback (OU) process assumption. (B) Illustration of the OU process over a phylogenetic tree with four observed species. Time axis represents the evolution history. $X(t)$ represents the trait at time t. The trajectories reflect the evolution of the continuous-trait features in different lineages, where the time points t_1, t_2, t_3 represent the speciation events. (C) A cartoon example of the possible evolutionary patterns (partitioned with different colors). Phylo-HMRF aims to identify evolutionary Hi-C contact patterns among four primate species in this work. The four Hi-C contact maps represent the observations from the four species, which are combined into one multi-species Hi-C map as the input to Phylo-HMRF, as shown in (A). The phylogenetic tree of the four species in this study is on the left. The partitions with green borders are conserved Hi-C contact patterns. The partitions with red or blue borders represent lineage-specific Hi-C contact patterns.

Towards a Post-clustering Test for Differential Expression

Jesse M. Zhang, Govinda M. Kamath, and David N. Tse(✉)

Department of Electrical Engineering, Stanford University, Palo Alto 94304, USA
jessez@stanford.edu, gkamath@stanford.edu, dntse@stanford.edu

Extended Abstract

Single-cell technologies have seen widespread adoption in recent years. The datasets generated by these technologies provide information on up to millions or more individual cells; however, the identities of the cells are often only determined computationally. Single-cell computational pipelines involve two critical steps: organizing the cells in a biologically meaningful way (clustering) and identifying the markers driving this organization (differential expression analysis). Because clustering algorithms *force* separation, performing differential expression analysis after clustering on the same dataset will generate artificially low p-values, potentially resulting in false discoveries.

While several differential expression methods exist, as a motivating example we consider the classic Student's t-test introduced in 1908 [2]. The t-test was devised for controlled experiments where the hypothesis to be tested was defined before the experiments were carried out. For example, to test the efficacy of a drug, the researcher would randomly assign individuals to case and control groups, administer the placebo or the drug, and take a set of measurements. Because the populations were clearly defined a priori, so was the null hypothesis. Therefore, under the null hypothesis where no effect exists, the mean measurement should be the same across the two populations, and the p-value should be uniformly distributed between 0 and 1.

For single-cell analysis, however, the populations are often obtained *after* the measurements are taken, via clustering, and therefore we can expect the t-test to return significant p-values even if the null hypothesis was true. Figure 1 shows how a measurement, such as expression of a gene, is deemed significantly different between two clusters even though all samples came from the same normal distribution. The clustering introduces a **selection bias** [1, 3] that would result in several false discoveries if uncorrected.

In this work, we introduce the truncated normal (TN) test, an approximate test based on the truncated normal distribution that corrects for a significant portion of the selection bias generated by clustering. We condition on the clustering event using the hyperplane that separates the clusters. By incorporating this

Full paper available at https://www.biorxiv.org/content/early/2018/11/05/463265. Code provided at https://github.com/jessemzhang/tn_test.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 328–329, 2019.
https://doi.org/10.1007/978-3-030-17083-7

hyperplane into our null model, we can obtain a uniformly distributed p-value even in the presence of clustering (Fig. 1). To our knowledge, the TN test is the first test to correct for clustering bias while addressing the differential expression question: *is this feature significantly different between the two clusters?* Based on the TN test, we provide a data-splitting based framework that allows us to generate valid p-values for differential expression of genes for clusters obtained from any clustering algorithm. We validate the method using both synthetic and real data, such as the peripheral blood mononuclear cell (PBMC) dataset generated using recent techniques developed by 10x Genomics [4], and we compare the method to several existing differential expression methods.

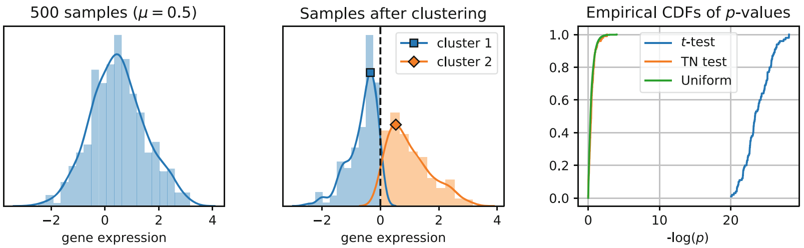


Fig. 1. Artificially low p-values due to clustering. Although the 500 samples are drawn from the same $\mathcal{N}(\mu, 1)$ distribution, our simple clustering approach will always generate two clusters that seem significantly different under the t-test. In this work, we explore an approach for correcting the selection bias due to clustering. In other words, we attempt to close the gap between the blue and green curves in the rightmost plot. We introduce the TN test, which generates significantly more reasonable p-values.

Acknowledgements. We thank Jonathan Taylor, Martin Zhang, and Vasilis Ntranos of Stanford University and Aaron Lun of the CRUK Cambridge Institute for helpful discussions about selective inference and applications of the method. GMK and JMZ are supported by the Center for Science of Information, an NSF Science and Technology Center, under grant agreement CCF-0939370. JMZ and DNT are supported in part by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG008164.

References

1. Fithian, W., Sun, D., Taylor, J.: Optimal inference after model selection (2014). arXiv preprint, http://arxiv.org/abs/1410.2597
2. Student: The probable error of a mean. Biometrika pp. 1–25 (1908)
3. Taylor, J., Tibshirani, R.J.: Statistical learning and selective inference. Proc. Nat. Acad. Sci. **112**(25), 7629–7634 (2015)
4. Zheng, G.X., et al.: Massively parallel digital transcriptional profiling of single cells. Nat. Commun. **8**, 14049 (2017)

AdaFDR: A Fast, Powerful and Covariate-Adaptive Approach to Multiple Hypothesis Testing

Martin J. Zhang[1], Fei Xia[1], and James Zou[1,2,3(✉)]

[1] Department of Electrical Engineering, Stanford University, Palo Alto 94304, USA
{jinye,feixia,jamesz}@stanford.edu

[2] Department of Biomedical Data Science, Stanford University,
Palo Alto 94304, USA

[3] Chan-Zuckerberg Biohub, San Francisco 94158, USA

Introduction. Multiple hypothesis testing is an essential component in many modern data analysis workflows. A very common objective is to maximize the number of discoveries while controlling the fraction of false discoveries. For example, we may want to identify as many genes as possible that are differentially expressed between two populations such that less than, say, 10% of these identified genes are false positives.

In the standard setting, the data for each hypothesis is summarized by a p-value, with a smaller value presenting stronger evidence against the null hypothesis that there is no association. Commonly-used procedures such as the Benjamini-Hochberg procedure (BH) [1] works solely with this list of p-values [3, 7]. Despite being widely used, these multiple testing procedures fail to utilize additional information that is often available in modern applications that are not directly captured by the p-value.

For example, in expression quantitative trait loci (eQTL) mapping or genome-wide association studies (GWAS), single nucleotide polymorphism (SNP) in active chromatin state are more likely to be significantly associated with the phenotype [2]. Such chromatin information is readily available in public databases, but is not used by standard multiple hypothesis testing procedures—it is sometimes used for post-hoc biological interpretation. Similarly, the location of the SNP, its conservation score, etc., can alter the likelihood for the SNP to be an eQTL. Together such additional information, called covariates, forms a feature representation of the hypothesis; this feature vector is ignored by the standard multiple hypothesis testing procedures.

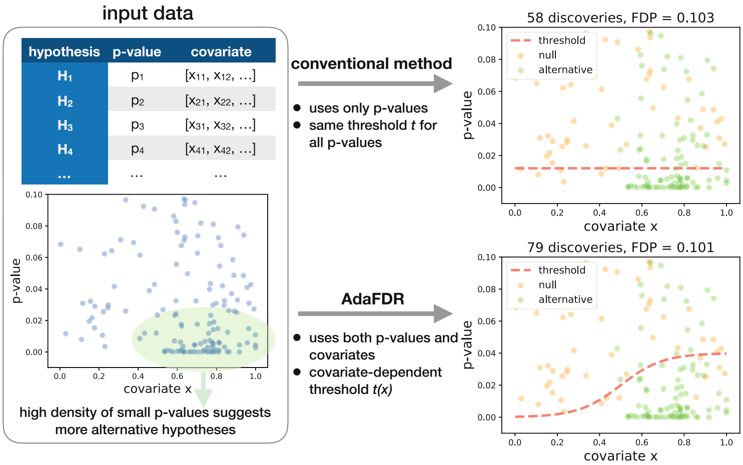
In this paper, we present AdaFDR, a fast and flexible method that adaptively learns the decision threshold from covariates to significantly improve the detection power while having the false discovery proportion (FDP) controlled at a user-specified level. A schematic diagram for AdaFDR is shown in Fig. 1.

Full paper available at https://www.biorxiv.org/content/early/2018/12/13/496372.

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 330–333, 2019.
https://doi.org/10.1007/978-3-030-17083-7

AdaFDR takes as input a list of hypotheses, each with a p-value and a covariate vector. Conventional methods like BH use only p-values and have the same p-value threshold for all hypotheses (Fig. 1 top right). However, as illustrated in the bottom-left panel, the data may have an enrichment of small p-values for certain values of the covariate, which suggests an enrichment of alternative hypotheses around these covariate values. Intuitively, allocating more FDR budget to hypothesis with such covariates could increase the detection power. AdaFDR adaptively learns such pattern using both p-values and covariates, resulting in a covariate-dependent threshold that makes more discoveries under the same FDP constraint (Fig. 1 bottom right).

Methods. AdaFDR extends conventional procedures like BH and Storey-BH [1, 7] by considering multiple hypothesis testing with side information on the hypotheses. The input of AdaFDR is a set of hypotheses each with a p-value and a vector of covariates, whereas the output is a set of selected (also called rejected) hypotheses. For eQTL analysis, each hypothesis is one pair of SNP and gene, and the p-value tests for association between their values across samples. The covariate can be the location, conservation, and chromatin status at the SNP and the gene. The standard assumption of AdaFDR and all the related methods is that the covariates should not affect the p-values under the null hypothesis. AdaFDR learns the covariate-dependent p-value selection threshold by first fit-



hypothesis	p-value	covariate
H_1	p_1	[x_{11}, x_{12}, …]
H_2	p_2	[x_{21}, x_{22}, …]
H_3	p_3	[x_{31}, x_{32}, …]
H_4	p_4	[x_{41}, x_{42}, …]
…	…	…

Fig. 1. Intuition of AdaFDR. Top-left: As input, AdaFDR takes a list of hypotheses, each with a p-value and a covariate that may be multi-dimensional. Bottom-left: A toy example with a univariate covariate. The enrichment of small p-values in the bottom right corner suggests more alternative hypotheses there. Leveraging this structure can lead to more discoveries. Top-right: Conventional method uses only p-values and has the same threshold for all hypotheses. Bottom-right: AdaFDR adaptively learns the uneven distribution of the alternative hypotheses, and makes more discoveries while controlling the false discovery proportion (FDP) at the desired level (0.1 in this case).

ting a mixture model using expectation maximization (EM) algorithm, where the mixture model is a combination of a generalized linear model (GLM) and Gaussian mixtures. Then it makes local adjustments to the p-value threshold by optimizing for more discoveries. We prove that `AdaFDR` controls FDP under standard statistical assumptions. `AdaFDR` is designed to be fast and flexible — it can simultaneously process more than 100 million hypotheses within an hour and allows multi-dimensional covariates with both numeric and categorical values. In addition, `AdaFDR` provides exploratory plots visualizing how each covariate is related to the significance of the hypotheses, allowing users to interpret their findings.

Results. We systematically evaluate the performance of `AdaFDR` across multiple datasets. We first consider the problem of eQTL discovery using the data from the Genotype-Tissue Expression (GTEx) project [2]. As covariates, we consider the distance between the SNP and the gene, the gene expression level, the alternative allele frequency as well as the chromatin states of the SNP. Across all 17 tissues considered in the study, `AdaFDR` has an improvement of 32% over BH and 27% over the state-of-art covariate-adaptive method independent hypothesis weighting (IHW) [4]. We next consider other applications, including three RNA-Seq datasets with the gene expression level as the covariate, two microbiome datasets with ubiquity (proportion of samples where the feature is detected) and the mean nonzero abundance as covariates, a proteomics dataset with the peptides level as the covariate, and two fMRI datasets with the Brodmann area label as the covariate that represents different functional regions of human brain. In all experiments, `AdaFDR` shows a similar improvement. Finally, we perform extensive simulations, including ones from a very recent benchmark paper [5], to demonstrate that `AdaFDR` has the highest detection power while controlling the false discovery proportion in various cases where the p-values may be either independent or dependent. The default parameters of `AdaFDR` are used for every experiment in this paper, both real data analysis and simulations, without any tuning. In addition to the experiments, we theoretically prove that `AdaFDR` controls FDP with high probability when the null p-values, conditional on the covariates, are independently distributed and stochastically greater than the uniform distribution, a standard assumption also made by related literature [1, 6].

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. B (Methodol.) 289–300 (1995)
2. GTEx Consortium: Genetic effects on gene expression across human tissues. Nature **550**(7675), 204 (2017)
3. Dunn, O.J.: Multiple comparisons among means. J. Am. Stat. Assoc. **56**(293), 52–64 (1961)
4. Ignatiadis, N., Klaus, B., Zaugg, J.B., Huber, W.: Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat. Methods **13**(7), 577–580 (2016)

5. Korthauer, K., et. al.: A practical guide to methods controlling false discoveries in computational biology. bioRxiv (2018). https://doi.org/10.1101/458786, https://www.biorxiv.org/content/early/2018/10/31/458786
6. Lei, L., Fithian, W.: Adapt: an interactive procedure for multiple testing with side information. J. Roy. Stat. Soc. B (Stat. Methodol.) **80**(4), 649–679 (2018)
7. Storey, J.D.: A direct approach to false discovery rates. J. Roy. Stat. Soc. B (Stat. Methodol.) **64**(3), 479–498 (2002)

Author Index