Sackler Faculty of Exact Sciences,
Blavatnik School of Computer Science

Statistical and computational methods for studying genomic spatial structure and properties

THESIS SUBMITTED FOR THE DEGREE OF
"DOCTOR OF PHILOSOPHY"
by

Shay Ben-Elazar

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

Prof. Ben Zion Chor
Prof. Zohar Yakhini

Submitted to the Senate of Tel Aviv University
November 2019

Acknowledgements

To my academic advisors and mentors, Zohar Yakhini and Benny Chor – it has been my distinct privilege to work with and learn from you. I am lucky to have had two sharp-witted and ego-free supervisors guiding my journey and consider you true role models. To my managers as an applied researcher at Microsoft, Daniel Sitton, Noam Koenigstein, Royi Ronen – Thank you for enriching my toolset with ideas and perspectives of different domains, ideas that have undoubtedly carried over and improved my academic work and capabilities as a researcher. To my colleagues from Chor group, Yakhini group, Microsoft Recommendations team, Microsoft Video Indexer and Microsoft Education Analytics – it was a pleasure working alongside each of you and I hope to continue collaborating.

I would like to also thank all the sources who generously funded my research: Blavatnik Computer Science Research Fund, Agilent Technologies University Relations grant, Interdisciplinary Center Tuition grant, Microsoft tuition assistance program.

Finally, to my dear family, thank you for your love, support and understanding throughout the years. To my parents, Lili and David, thank you for your encouragement during more frustrating times. To my beloved wife, Yael, and daughter, Abigail – thank you for providing the drive to all that I do.

- *"If we don't mark the milestones, we're just passing with the time"– Lara Axelrod*

Preface

This thesis is based on the following three articles. At the time of writing, two of the three papers below were published in scientific journals, one of which was presented in a leading conference and the third paper was submitted and is awaiting peer review.

1. **Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data**
   Shay Ben-Elazar, Benny Chor, Zohar Yakhini
   Published in *Bioinformatics 2016,* Presented as poster and orally at *ECCB 2016*

2. **The functional 3D organization of unicellular genomes**
   Shay Ben-Elazar, Benny Chor, Zohar Yakhini
   Published in *Nature Scientific Reports 2019*

3. **miRNA normalization enables joint analysis of several datasets to increase sensitivity and to reveal novel miRNAs differentially expressed in breast cancer**
   Shay Ben-Elazar, Miriam Ragle Aure, Kristin Jonsdottir, Suvi-Katri Leivonen, Vessela N. Kristensen, Emiel A.M. Janssen, Kristine Kleivi Sahlberg, Ole Christian Lingjærde and Zohar Yakhini
   Submitted to *PLOS Computational Biology 2019*

Abstract

Genomes store and encode complex instruction sets for the production and regulation of genes. In turn, genes interact with each other and with their environment to dictate the phenotype and function of biological cells. Elucidating genomic mechanisms of operation can potentially deliver innovation in healthcare, agriculture, ecology and even in encoding digital information and computing. Technological advancements and emergent experimental procedures continuously produce new sets of challenges in efficiently scaling and correctly interpreting genomic observational data. This thesis addresses data analysis aspects related to two molecular biology measurement techniques, focusing on specific challenges that have emerged in the context of interpreting their results. We present three separate research projects that share a common goal – deciphering genomic and epigenomic properties from measurement data. We provide novel algorithms, motivate them with simulations, apply them on real data and provide statistical evidence and biological interpretation of the analysis findings. The approaches discussed in this thesis advance the state-of-the-art and provide new insights on genomic and epigenomic characteristics of cells and their functional roles. In particular, our contribution includes:

- An approach for using Hi-C data to infer full haplotypes from partially phased genotypes.
- A statistical approach to characterizing the functional 3D organization of unicellular genomes using Hi-C data.
- A novel normalization approach to miRNA data, that enables the integration of several datasets, leading to increased statistical power.

Contents

List of Figures

List of Supplementary Figures

List of Tables

Chapter 1:

Introduction

Chromosome conformation capture (3C), and related methods (e.g. High-throughput 3C, or Hi-C), are a set of experimental biology protocols based on DNA sequencing technology that produce a (sparse) map of paired read counts across chromosomes. These read counts are (approximately) proportional to spatial proximities between pairs of chromosomal loci (Nynke L. van Berkum *et al.*, 2010a). A myriad of approaches embed Hi-C read counts into (qualitative) 3D models in order to smooth out sampling noise and offer an intuitive glimpse into the underlying genome structure. 3C and related techniques have paved the way for experimentally charting 3-dimensional structural properties of genomes in living cells at detail currently unavailable to volumetric microscopy. Key discoveries attributable in part to 3C include: Functional-organizational unit of TADs (topologically associating domains) as a structural epigenetic mechanism enforcing promoter-enhancer contacts and enabling neighborhood insulation. Systematic evidence for the “transcription factory” hypothesis. Namely, regulatory co-factors and transcription machinery co-localize to sub-compartments in the nucleus (Eukaryotes) / nucleoid (Prokaryote) along with their genomic targets.

In chapter 1.1 we introduce an approach that leverages information obtained by genome conformation capture to address the “last-mile” sequence assembly problem of Haplotyping. Haplotyping is the process of assigning nucleotide sequence variants and aberrations to one corresponding homologous chromosome copy. In our work (S. Ben-Elazar *et al.*, 2016) we demonstrate methods that are useful for both 1) de-multiplexing “traditional” averaged pairwise-chromosome Hi-C proximity maps into maps containing pairwise-homologous copy information and 2) “Phasing” (un-shuffling) homologous Hi-C maps to the correct homologous block structure. Such de-multiplexed and phased Hi-C maps are important to improve the precision and applicability of further interpretation of Hi-C data for other downstream tasks.

In chapter 1.2 we revisit a problem related to the downstream analysis of 3D models derived from Hi-C data. In this work (Ben-Elazar *et al.*, 2019) we develop an algorithmic and statistical framework to identifying 3D spherical compartments in which genomic elements with some common biological property significantly co-localize. This approach overcomes a limitation of our previous work where candidate enrichment spheres were centered on the 1D genome. We provide rigorous analysis of this method and illustrate its benefit in detecting novel patterns with plausible biological interpretation. We describe findings in several organisms.

MicroRNAs (miRNAs) are short RNA molecules that are typically functional although they do not undergo translation to protein. miRNA has evolved to play a regulatory role in gene expression as well as in immune system activity and modulation. miRNAs have been implicated in both malignant tumor suppression and development depending on various conditions (Peng and Croce, 2016). Of particular interest is the precise characterization of their relation to cancer subtypes as potential biomarkers for driving personalized clinical care.

In chapter 1.3 we present a method for the normalization and integrative analysis of miRNA expression data. Our methods mitigate batch-related effects. We apply this approach to jointly analyze four cohorts of miRNA expression in breast cancer, present potential novel miRNA biomarkers and discuss the statistical advantages of our approach. We also discuss some specific observations that would not have emerged without normalization.

In the sections of this chapter we provide more background and a more detailed overview of each of the aforementioned questions. In later chapters we present computational methods to address these problems and discuss potential novel findings surfaced by our techniques along with their detailed analysis. Finally, we conclude by discussing our work and offering directions for future investigation.

1.1 Hi-C and Phasing

Haplotyping, or phasing, is the process of determining the physical co-occurrence of genomic variations along intact maternal or paternal homologous chromosomes in diploid or polyploid

organisms. Several methods aim to determine haplotype, ranging from population-genetics based approaches that require some knowledge on the parental genomes, to methods based on time consuming and elaborate molecular isolation of chromosomes prior to high-throughput sequencing.

We developed a computational pipeline that combines Hi-C data with partial phasing data to infer full haplotypes as well as the fully phased Hi-C proximity map (Shay Ben-Elazar, Chor, Yakhini, *et al.*, 2016). In this work we embedded genomic locations based on corrected Hi-C maps and used this representation to apply a greedy partitioning method in order to decode the correct homologous assignment of different alleles observed from a partially phased genotype assay. We prove that our solution yields the global maximum likelihood configuration in a statistical formalization of the problem. We show that embedding Hi-C data offers a better proximity measure for haplotype decoding suggesting that embedding is an essential step in smoothing Hi-C data which can be sparse and noisy. This is possibly true for other Hi-C analysis pipelines as well. Additionally, we process reads that overlap mono-allelic (indiscernible between maternal and paternal copies) loci and not only reads that overlap loci that are bi-allelic for the measured individual. Such mono-allelic reads are softly assigned under a uniform prior over its potential homolog copies of origin. Traditionally, such reads are ignored by many state-of-the-art Hi-C data analysis approaches.

In chapter 2 we explore the problem of recovering phased Hi-C and full haplotype data using un-phased Hi-C data and partial haplotype data. The proposed solution is exemplified by analyzing its accuracy on human diploid Hi-C and ground-truth haplotype data available via Trio-phasing (Auton et al., 2015). Our results show that the proposed method results in haplotypes that have 98% agreement with ground truth data (averaged across chromosomes). We show potential added value in correctly interpreting diploid Hi-C data by applying a co-localization analysis that shows patterns in which single copies of genes on different homologous chromosomes reside in a proposed transcription factory. For completeness, we provide a more rigorous mathematical definition to the underlying geometric problem addressed in the paper and available as an additional supplementary chapter directly following chapter 2.

1.2 Hi-C, Spatial Enrichment and Transcription Factories

Chapter 3 discusses algorithmics and statistics for assessing spatial enrichment of a binary property on a given spatially organized dataset (coordinates in R^2, R^3). We developed a method to identify locations within a 2D-3D Euclidean space, around which a specific subset of elements is localized with significantly high density. We studied the validity and efficiency of this method on simulated data and applied it to 3D embeddings of Hi-C data from multiple unicellular organisms and across multiple genomic annotation sets (Ben-Elazar *et al.*, 2019). We compare this method to directly studying raw paired read counts and discuss its advantages.

Previous studies, both by us and by others, have suggested heuristics for performing spatial co-localization analysis on Hi-C data. In this work we explored a rigorous formal definition of the spatial co-localization problem. We present compelling evidence to support our methodology compared to those used previously and apply our method to obtain statistically significant results suggesting potential novel biology. Applying spatial co-localization for obtaining a more precise characterization and means of identifying transcription factories is of particular interest. Transcription factories are a regulatory mechanism manifested as confined spaces within the nucleus, where transcription machinery recruits relevant cofactors and genomic stretches such as to regulate the activation of specific cellular functions (F. J. Iborra *et al.*, 1996; Sutherland and Bickmore, 2009a). Previous studies have attempted to statistically assess the existence of transcription factories. The authors of (Dai and Dai, 2012) compared the number of interactions in different functionally-related gene sets and observed statistical enrichment under the hypergeometric null model for interactions among transcription factor (TF) targets. However, a follow-up study (Witten and Noble, 2012) argued that edges in the 3C interaction graph are not statistically independent, as was assumed under the model used by Dai and Dai, and that co-localization events would therefore be over-counted. To correct for this issue, Witten and Noble applied a re-sampling methodology under which no signal for TF target co-localization was detected. Our approach, applied in both (Ben-Elazar *et al.*, 2013a) and (Shay Ben-Elazar, Chor, Yakhini, *et al.*, 2016)), avoids comparing between populations of proximities altogether, and so avoids any statistical dependence issues which arise in former methods. Instead, we

focused on the distances to a single pivot locus – a reference point around which we measure co-localization statistical significance, as described below.

In our previous study, we identified transcription factory candidates by developing a statistical model based on the minimum Hypergeometric (mHG) statistical framework (Eden *et al.*, 2007, 2009a). In more detail, consider a genomic locus, l. Rank all other genomic loci $l_1, \dots, l_N$ by some distance function to l, $d(l_i, l)$. Consider a transcription factor (TF) and its set of targets, T. We consider T to represent the genomic locations at which transcription is driven by the TF. Define a binary vector, λ, of length N, where $\lambda(i) = 1 \; iff \; l_i \in T$. For $1 \leq n \leq N$ we define $\Lambda_n = [\lambda(1), \dots, \lambda(n)]$ as the prefix of length n of the binary vector λ. Let $b_n = \Sigma\Lambda_n$, $B = \Sigma\Lambda_N$. The mHG score is defined by the threshold, n, that minimizes the right tail of the hypergeometric CDF of observed b_n values. That is,

$$mHG(\lambda) = \min_{1 \leq n \leq N} \sum_{i=b_n}^{\min(n,B)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}}$$

The null hypothesis in mHG is that given the number of B 1's, they are uniformly distributed in the binary vector of length N. In our context, rejecting the null hypothesis suggests that TF targets are localized in significantly close proximity to the pivot locus. We repeat this experiment for all loci and TFs, Bonferroni correcting for multiple testing.

The mHG statistic was used to measure the probability with which an observed ranking of genes by their distance from a certain pivot point would surface an 'unlikely' number of genes that are known targets of a specific TF to the 'top' of the ranked list. The observed value of this statistic is assessed against a background model of random permutations, to which we add appropriate controls for the 1D gene order, to isolate the effect of actual 3D spatial localization. See an illustration of our previously developed method in Figure 1 taken from (Ben-Elazar *et al.*, 2013a).

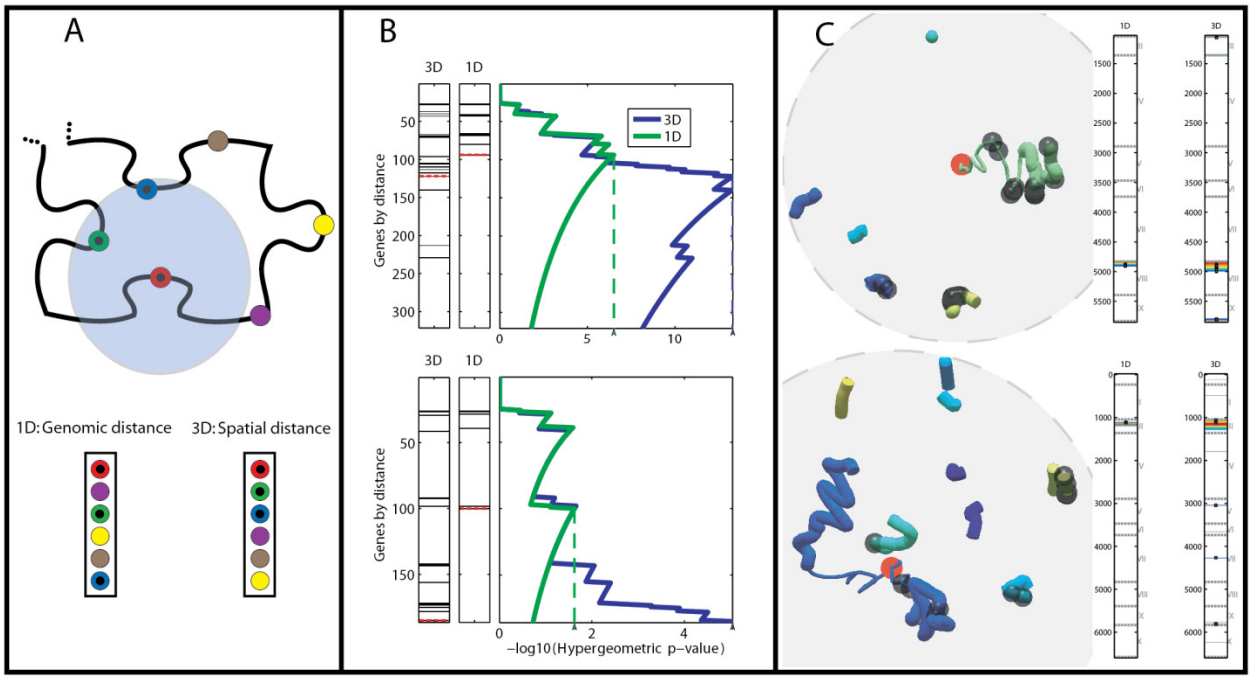


Figure 1. Comparing functional enrichment between the genomic and spatial regions of the genome. (A) Two genomic distances. The schematic shows the gene neighborhood surrounding a particular gene (red). The neighboring genes may be ranked by their genomic proximity (left) or their spatial proximity (right). (B) Detecting areas of enrichment for TF-cohorts. In ranked gene lists, generated by either genomic or spatial proximity, the genes annotated as targets of a particular TF are indicated as black lines. The p-value of the enrichment of the targets for each threshold is indicated on the right. The threshold with the best p-value is indicated by the dashed line (see Methods). This analysis is shown for two genomic loci surrounding genes YCL012C and YHL050C respectively and querying for targets of GLN3. (C) Local structures of the two loci examined in B. Colors indicate distinct yeast chromosomes. The red circles indicate the center gene around which co-localization was tested. The center genes shown are YCL012C (top) and both YHL050C and YHL050W-A (bottom). The content shown in each sphere is the environment which corresponds to the mHG threshold, dictated by the most enriched spatial environment for GLN3 targets. Bars on the right mark the loci along the linear genome which participate in the most enriched environment by both the genomic and spatial rankings. Black dots, both in the bars and the visualized structure, indicate gene targets of GLN3.

In the approach described above, we scanned genes along the 1-dimensional genome as pivots to identify potential transcription factories, measuring whether TF targets are enriched in 3D space around each such pivot. However, transcription factories need not be centered around a gene and not even around a pivot along the 1D genome, as we show in the paper. In this work swe expand the applicability of the 3D enrichment method described above and develop approaches that relax the limits of considering pivots only along the 1D genome.

Venturing out of the discrete space of possible pivots along the 1D genome in order to cover all possible pivots in 3D is generally intractable as there are infinitely many possible pivots.

However, as our enrichment analysis is based on rank orders and not on actual distance values, not all possible points in space need to be considered as pivots. We show that only a polynomial number of sets of pivots can induce different rank orders and can yield different mHG values. In our work we characterize the underlying combinatorial space precisely and provide an online branch and bound approach to scan for co-localization in arbitrarily pivots. Our algorithmics rely on properties of the hypergeometric distribution to efficiently discard candidate regions and recursively refine potential enrichments. In continuation to our previous work we applied this method to analyze Hi-C data and identify points in space as transcription factory candidates based on genomic 3D configurations. We evaluated this method on multiple datasets including simulated 3D data. More importantly – we applied it to Hi-C data from unicellular organisms and discuss several interesting biological results: Spatially co-localized peri-telomeric copy number increase in Rad21 knockout mutant, alluding to a deep connection between a functional cohesion complex and peri-telomeric integrity presented in Figure 2 taken from our paper. Co-localized genome replication genes partitioned to two copies that are in close proximity to the *ori* and *ter* (origin and terminus of DNA replication, accordingly) providing evidence for an evolved "backup" template useful for recovering stalled replication, etc.

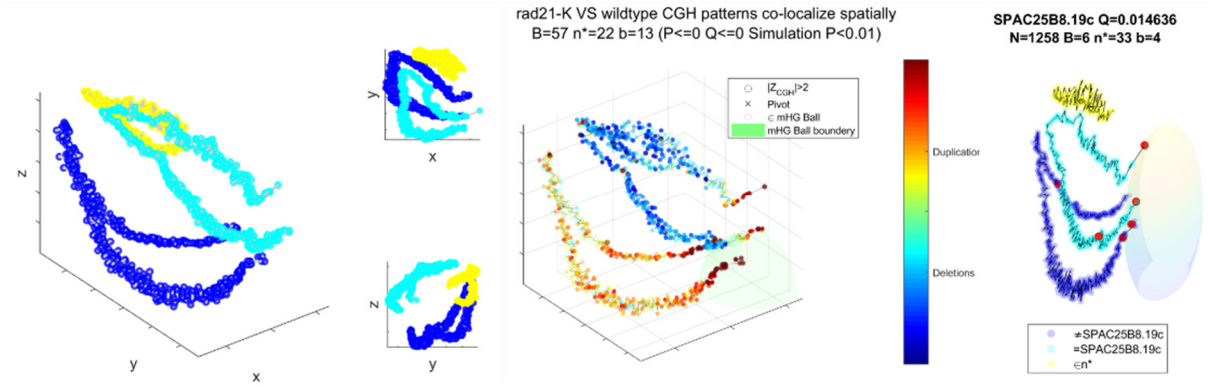


Figure 2. Example of spatial co-localization identified by our method.Left: sNMDS embedding for S. pombe *with colour coded chromosomes. Middle (animation available as Supplementary Video 5): Bins are colour coded by average aCGH value, with marked outliers (opaque red for Z>2 and blue for Z<-2). We can observe a weak duplication signal on ChrII, and deletion on ChrI, ChrIII. Strongest duplication is evident at the telomeres. Right (animation available as Supplementary Video 6): Red bins contain Loz1 transcription factor targets. The resulting* $smHG$ *pivot and corresponding ball are visible containing 4/6 TF targets.*

1.3 Integrative analysis of miRNA expression data

Chapter 4 describes an adjusted quantile-normalization approach (AQN) for the integrative analysis of breast-cancer miRNA expression data from multiple experiments sampled using different technologies. microRNAs (miRNAs) are endogenous, small non-coding RNAs (~22 nucleotides) that bind to target-specific sites most often found in the 3'-untranslated regions (UTRs) of target messenger RNAs (mRNAs). By this binding miRNAs regulate gene expression, inhibiting mRNA translation or marking the mRNA molecules for degradation. miRNA expression profiling is an important tool for studying tumor biology and classification and has shown to be important with respect to diagnostic and prognostic assessments.

The approaches to jointly analyze expression data from multiple sources (with source-specific biases) can be split to two families: meta-analysis and integrative analysis. In meta-analysis we study each dataset independently and combine results to make more robust conclusions. Meta-analysis is considered to benefit less from the added statistical power of an increased sample size when compared to integrative analysis. In contrast integrative analysis attempts to overcome batch effects by shifting the distributions of expression values from different experiments such that they are comparable, under different and specific considerations.

We develop a quantized and jittered variant of quantile normalization, denoted AQN, that reduces batch related clustering effects. We show that when coupled with appropriate downstream statistics our method is able to surface more differentially expressed miRNAs between estrogen receptor (ER) positive and negative patients. In particular, using the combined dataset, we implicate hsa-miR-193b-5p as a potential tumor suppressor. Our approach yields expression values that correlate better with known miRNA targets and increases GO (gene ontology) enrichment score for terms that are consistent with observational studies. We compare our method to commonly used normalization schemes and provide different lines of evidence in its favor.

1.4 Summary of articles included in this Thesis

1. **Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data**
   Shay Ben-Elazar, Benny Chor, Zohar Yakhini
   Published in *Bioinformatics 2016,* Presented as poster and orally at *ECCB 2016*

   *Motivation***:** Complex interactions among alleles often drive differences in inherited properties including disease predisposition. Isolating the effects of these interactions requires phasing information that is difficult to measure or infer. Furthermore, prevalent sequencing technologies used in the essential first step of determining a haplotype limit the range of that step to the span of reads, namely hundreds of bases. With the advent of pseudo-long read technologies, observable partial haplotypes can span several orders of magnitude more. Yet, measuring whole-genome-single-individual haplotypes remains a challenge. A different view of whole genome measurement addresses the 3D structure of the genome – with great development of Hi-C techniques in recent years. A shortcoming of current Hi-C, however, is the difficulty in inferring information that is specific to each of a pair of homologous chromosomes.
   *Results***:** In this work we develop a robust algorithmic framework that takes two measurement derived datasets: raw Hi-C and partial short-range haplotypes, and constructs the full-genome haplotype as well as phased diploid Hi-C maps. By analyzing both data sets together we thus bridge important gaps in both technologies – from short to long haplotypes and from un-phased to phased Hi-C. We demonstrate that our method can recover ground truth haplotypes with high accuracy, using measured biological data as well as simulated data. We analyze the impact of noise, Hi-C sequencing depth and measured haplotype lengths on performance. Finally, we use the inferred 3D structure of a human genome to point at transcription factor targets nuclear co-localization.

2. **The functional 3D organization of unicellular genomes**
   Shay Ben-Elazar, Benny Chor, Zohar Yakhini
   Published in *Nature Scientific Reports 2019*

Genome conformation capture techniques permit a systematic investigation into the functional spatial organization of genomes, including functional aspects like assessing the co-localization of sets of genomic elements. For example, the co-localization of genes targeted by a transcription factor (TF) within a transcription factory. We quantify spatial co-localization using a rigorous statistical model that measures the enrichment of a subset of elements in neighbourhoods inferred from Hi-C data. We also control for co-localization that can be attributed to genomic order.

We systematically apply our open-sourced framework, *spatial-mHG*, to search for spatial co-localization phenomena in multiple unicellular Hi-C datasets with corresponding genomic annotations. Our biological findings shed new light on the functional spatial organization of genomes, including: In *C. crescentus,* DNA replication genes reside in two genomic clusters that are spatially co-localized. Furthermore, these clusters contain similar gene copies and lay in genomic vicinity to the *ori* and *ter* sequences. In *S. cerevisae,* Ty5 retrotransposon family element spatially co-localize at a spatially adjacent subset of telomeres. In *N. crassa*, both Proteasome lid subcomplex genes and protein refolding genes jointly spatially co-localize at a shared location. An implementation of our algorithms is available online.

4. **miRNA normalization enables joint analysis of several datasets to increase sensitivity and to reveal novel miRNAs differentially expressed in breast cancer**
Shay Ben-Elazar, Miriam Ragle Aure, Kristin Jonsdottir, Suvi-Katri Leivonen, Vessela N. Kristensen, Emiel A.M. Janssen, Kristine Kleivi Sahlberg, Ole Christian Lingjærde and Zohar Yakhini
Submitted to *PLOS Computational Biology 2019*

Different miRNA profiling protocols and technologies introduce differences in the resulting quantitative expression profiles. These include differences in the presence (and measurability) of certain miRNAs. We present and examine a method based on quantile normalization, Adjusted Quantile Normalization (AQN), to combine miRNA expression data from multiple studies in breast cancer to a single joint dataset for integrative

analysis. By pooling multiple datasets, we obtain increased statistical power, surfacing patterns that do not emerge as statistically significant when separately analyzing these datasets. To merge several datasets, as we do here, one needs to overcome both technical and batch differences between these datasets. We compare several approaches to merging and jointly analyzing miRNA datasets. We investigate the statistical confidence for known results and highlight potential new findings that resulted from the joint analysis using AQN. In particular, we detect several previously associated breast-cancer miRNAs to be differentially expressed in estrogen receptor (ER) positive versus ER negative, thereby identifying new potential biomarkers and therapeutic targets for both categories. More specifically, using the AQN-derived dataset we detect hsa-miR-193b-5p to have statistically significant higher expression in ER positive samples, a phenomenon that was not previously reported. Furthermore, overexpression of hsa-miR-193b-5p in breast cancer cell lines resulted in decreased cell viability and expression of cancer-relevant proteins in addition to induced apoptosis, suggesting a novel functional role for this miRNA in breast cancer. Packages implementing AQN are provided for Python, Matlab and R.

1.5 Summary of contributions

The contributions of our work include:

- An approach for using Hi-C data to infer full haplotypes from partially phased genotypes. At the time of writing this thesis, state-of-the-art haplotyping (and similarly, metagenomic analysis) is typically accomplished using a hybrid of short range and long range (e.g. Hi-C) sequencing technologies. We provide an advanced algorithmic approach that better utilizes Hi-C data for improved performance, as we show in a direct comparison. We demonstrate an interesting implication of haplotype information via a downstream co-localization analysis on human (diploid) Hi-C data. Namely: we observe genomic co-localization patterns in which a single copy of a homologous gene pair appears to co-localize into a potential transcription factory.

- A statistical approach to characterizing the functional 3D organization of unicellular genomes using Hi-C data. We apply the approach to discover evidence of functional 3D organization across multiple organisms and multiple functional annotation sets. We present novel biological potential findings based on our analyses. An additional side-product (and prerequisite) from this work is the 3D embedding of several Hi-C datasets.
- A novel normalization approach to miRNA data that enables the integration of several datasets, leading to increased statistical power. We present statistically significant differentially expressed miRNA in estrogen receptor (ER) positive compared to ER negative breast-cancer patients, including a newly identified tumor suppressor miRNA that could potentially aid with future prognosis and treatment.

Finally, packages and code implementing our work presented herein is available as open source software for the community.

Chapter 2:

Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data

2.1. Introduction

Chromosome conformation capture (3C), and derived high-throughput methods (Hi-C), are a experimental protocols that yield a sparse map of read counts that are proportionally related to spatial proximities between pairs of genomic loci (Nynke L van Berkum *et al.*, 2010). Hi-C and related methods have been used to assess structural properties of genomes (Ay and Noble, 2015). Haplotyping is the process of determining the physical co-occurrence of genomic variations along intact maternal or paternal homolog chromosomes in diploid or polyploid organisms. Co-localization and linkage of such variations are key factors in determining the complex nature of some phenotypes and as essential tools in understanding genetics (Tewhey *et al.*, 2011).

Both haplotype information and genome conformation are typically investigated using “next generation” DNA sequencing technology, which suffers from an inherent ambiguity in its interpretability. In current state-of-the-art DNA sequencing methods, high-throughput short reads are computationally aligned or assembled to recover contiguous regions of the genome. The ambiguity in sequencing becomes evident when considering diploid genomes. Diploid genomes, by definition, contain at least two copies of almost every genomic region (up to CNVs and other genetic variations between homologous regions and sex chromosomes). Thus, when two long genomically identical regions are flanked by bi-allelic genotypes (such as in single nucleotide polymorphisms, SNPs) for a measured genome, we are unable to determine whether said variations reside on the same copy of the chromosome homologs, or belong to different homologs. The problem becomes practically intractable when considering the combinatorics involved in accurate pairwise assignment of variations along an entire genome.

While many methods attempt to address the problem of haplotype phasing it is difficult to achieve a satisfactory balance between the scalability and reliability required in practice. Methods vary from population-based methods, to methods requiring vast sequencing depth with multiple insert sizes, complex manual isolation or cutting-edge non-mainstream technology. A good overview of these methods including their advantages and disadvantages is presented in (Snyder *et al.*, 2015; Glusman *et al.*, 2014). In recent years technological progress

enabled cost-effective long-read or pseudo-long-read sequencing such as Pacific Biosciences SMRT and Oxford Nanopore and 10X (McCoy *et al.*, 2014; Patterson *et al.*, 2015; Pirola *et al.*, 2016; Eisenstein, 2015). With such technology, experimentally-derived short-range haplotypes are becoming viable, and the algorithmic challenges, common to all combinatorial methods for haplotype assembly, are shifting towards extending short-range phasing to cover longer genomic stretches with hybrid approaches (Glusman *et al.*, 2014; Auton *et al.*, 2015; Kuleshov *et al.*, 2014). On the other hand, Hi-C and related technologies are becoming more accurate, cheaper, and with higher coverage (Suhas S P Rao *et al.*, 2014). Information about distances in the genome can be used in the context of haplotyping, as shown in the pioneering work from Selvaraj, Bafna and colleagues (Selvaraj *et al.*, 2013). The authors developed 'HaploSeq', an adjusted version of 'HapCut' (Bansal and Bafna, 2008) that was shown to recover haplotypes with high coverage and quality. In their study, the authors utilize a Monte Carlo scheme to maximize the agreement of a sampled phasing with the observed read counts, and iteratively solve Min-Cut instances to identify and replace discordant phase assignments in a way that guarantees convergence to a local minimum. By rerunning this process $O(n)$ times, where n is the number of variation sites, they expect to find a single haplotype assignment that minimizes an overall discordance score.

In this work we introduce an algorithmic approach to using Hi-C data to extend haplotype information to longer genomic stretches. Several aspects of the approach presented in (Selvaraj *et al.*, 2013) can potentially be addressed to improve performance. First – distances measured at any two loci can be used to more robustly infer a distance between these two loci. We address this by computing the embedding of local similarities that induces a global proximity measure among loci by using all available data to establish coordinate locations in the ambient space. Embedding helps correct observed connections, which can be sparse and noisy. We also process reads that overlap mono-allelic loci and not only reads that overlap loci that are bi-allelic for the measured individual. Our algorithm is deterministic and therefore consistent in its output.

The premise of this work relies on a fundamental duality between diploid genome structural inference and haplotyping. In Hi-C data, one cannot determine an assignment of a read pair to

the physical chromosomal copy directly from sequencing. Naïve approaches average reads from both chromosomal copies into the same co-occurrence matrix. Some biological phenomena can survive this distortion and still be captured. However, this structure based on averages is unlikely to capture the original structure of chromosome copies (Figure S1). Therefore, studies based on distance averages become questionable and are likely to miss much of the information related to the actual structure (Figure S6). With haplotype information, reads spanning bi-allelic instances of SNPs would be uniquely-mappable to their chromosome copy of origin, and a partial reconstruction of the manifold would potentially be achievable with a sufficient amount of reads. On the dual side, if one had SNP allele specific positions along the geometrical structure of the chromosomal manifold – it would potentially be possible to interpolate a reasonable manifold structure and infer the haplotype by partitioning SNP alleles based on their geometric relationships.

Recent studies (Suhas S P Rao *et al.*, 2014; Servant *et al.*, 2015) make use of fully phased Hi-C data in a straight-forward way. These approaches produce a partial Hi-C co-occurrence matrix in the resolution of single nucleotides and focus only on reads overlapping specific alleles in bi-allelic loci. These can be uniquely mapped and then phased using measured haplotype data. This enables inference from phased Hi-C data. One possible shortcoming is, however, the focus on bi-allelic loci.

Our contribution consists of:

(1) An algorithmic approach that takes short range haplotype blocks (as can be inferred from current and near future NGS techniques) and Hi-C data and produces much longer blocks and phased Hi-C data.

(2) Performance analysis of the above and comparison to other approaches.

(3) An algorithmic approach to computing a distance matrix between haplotype blocks, using Hi-C data. In particular, we use mono-allelic as well as bi-allelic loci.

(4) A component of 1 above uses an embedding of the haplotype blocks into an inferred 3-dimensional configuration.

(5) An example of how the phased Hi-C data can be used to better understand genome 3-dimensional structure. Specifically – we address the spatial co-occurrence of TF (transcription factor) targets, using inferred phased Hi-C data. We show that using genomic order or averaged unphased Hi-C data is not sufficiently strong to identify this co-occurrence.

2.2 Methods

We present an algorithmic framework to computationally extend partial short-range haplotypes based on Hi-C data. Our algorithm relies in its core on embedding of a Hi-C-based $n \times n$ similarity data matrix, S, to a set of n coordinates, $\mathrm{E} = \{c_i | c_i \in \mathbb{R}^3\}_1^n$. That is – we seek a conformation of points that maintains the similarity values in S with minimal error. Because of measurement noise values in the matrix S do not behave as a metric. Since E is forced to a Euclidean space, the embedding approximates transitive relations possibly violated in S, elucidating a viable geometrical interpretation of the similarity data. Based on an abstraction of the resulting geometry, we partition bi-allelic coordinates and recover a haplotype. To overcome several issues affecting our embedding strategy (elaborated in The Supplementary section – 5.1) we introduce additional preliminary steps of computing dot product and performing connected component analysis. Our algorithm for determining the complete phasing is composed of five steps:

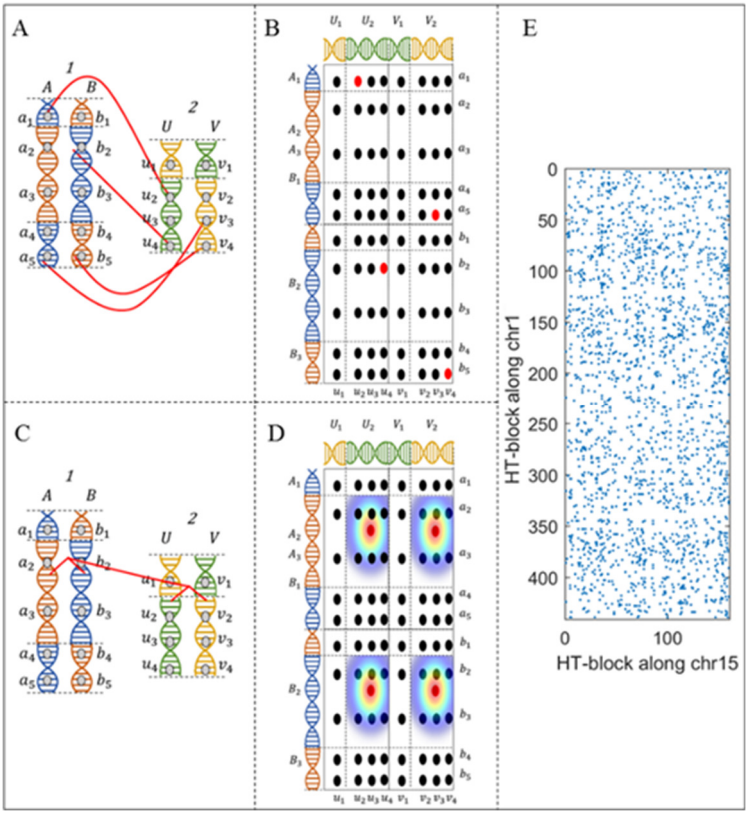


Figure 3. Illustration of steps applied to parse Hi-C data into a similarity matrix between HT-blocks (A) Contiguous HT-blocks (with known partial phasing data) along two pairs of homologous chromosomes, chromosome 1 and 2 with homolog pairs (A, B) and (U, V) are illustrated. Following our notation, HT-block $A_2 = (a_2, a_3)$ is depicted as an orange section of the chromosome and thus the correct underlying phasing for this chromosome should be (A_1, B_2, A_3) and (B_1, A_2, B_3). Some gray dots are connected with a red line representing a paired-end Hi-C read mapped to bi-allelic genomic loci on both ends. E.g., $e = (a_1, u_2)$ is one such read at the top of the illustration. (B) Reads depicted in Figure 3A, map to their corresponding pair of bi-allelic loci shown as red dots. E.g., $e = (a_1, u_2)$ is the top-left-most red dot. (C) Illustrated Hi-C paired-read that overlaps only mono-allelic loci on both ends. (D) Showing the 2D Gaussian interpolation within the corresponding HT-block pair for the read illustrated in Figure 3C. Since the read can potentially map to either block of the matrix, it is split proportionally according to phased bi-allelic-overlapping reads in its genomic neighborhood in the bin (intersecting dashed lines). (E) Spy-plot of a resulting enriched map with simulated HT-blocks (Supp. Methods) fully ordered by ground-truth phasing data.

Algorithm Overview – SPECTRALPHASING

Input: Aligned Hi-C data and partial short-range phasing data.

Do:

(1) Aggregate filtered sequencing reads to yield a matrix of similarities among haplotype blocks (Figure 3).

For each pair of homologous chromosomes:

(2) Compute dot product similarities for haplotype blocks (Figure 4).

(3) Identify connected components and partition accordingly.

(4) Embed HT-blocks to a 3D model using Multidimensional Scaling (MDS), per connected component.

(5) Compute Trellis phasing using the 3D Euclidean distances of the embedded representation of points.

Output: Report the extended phasing and a phased Hi-C map.

We define the following notation: for a chromosome with homologs *A* and *B* we denote phased alleles as ordered sets,

$A = \left((a_1, \dots a_j), (a_{j+1}, \dots, a_k), \dots, (a_l, \dots, a_n) \right)$ and

$B = \left((b_1, \dots b_j), (b_{j+1}, \dots, b_k), \dots, (b_l, \dots, b_n) \right)$. $\{a_j, b_j\}$ are alleles belonging to bi-allelic locus j. Denote $|A| = |B| = m$ as the number of sets of phased alleles in the chromosome. We index haplotype blocks (HT-blocks) in the homolog copies following the above notation. That is, for example: $A_2 = (a_{j+1}, \dots, a_k)$ is the 2nd (according to genomic order) HT-block of homolog A. An HT-block is defined as the genomic region demarcated by a set of phased alleles assumed or measured to be on the same homolog, according to partial phasing data. Finally, an ordered set of HT-blocks, e.g. $\mathcal{H} = (A_1, B_2, A_3, \dots A_n) \leftrightarrow \overline{\mathcal{H}} = (B_1, A_2, B_3, \dots B_n)$, is the information that extends a partial phasing to a complete one.

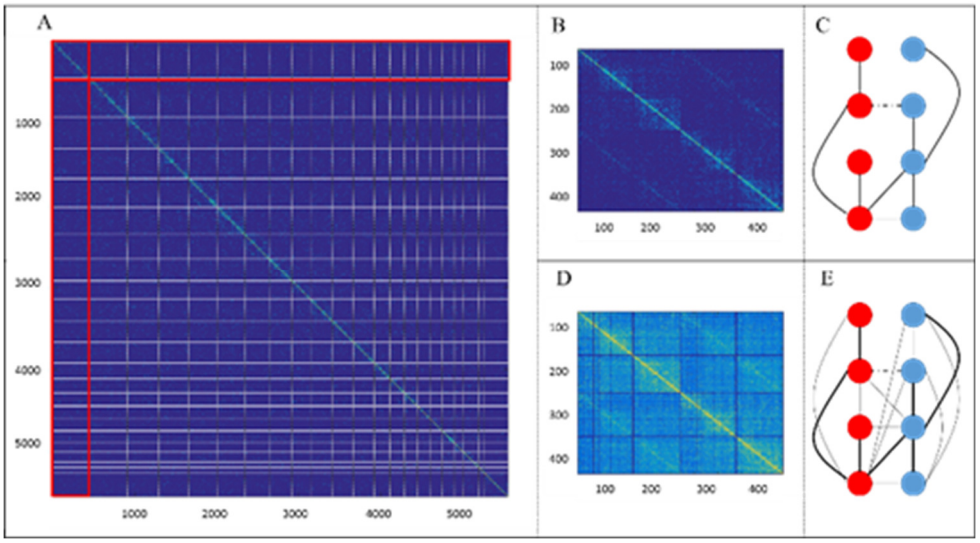


Figure 4. Computing dot products on a chromosome's genome-wide map enriches intra-chromosomal maps. (A) Figure 3E shows a single inter-chromosomal enriched contact map, while in fact, there are 23^2 maps for each pair of chromosomes, shown here separated by a white grid. Multiplying the sub-matrices belonging to the first block-row with the first block-column (highlighted in red), that correspond to all inter-chromosomal Hi-C data for Chromosome 1, yields the dot product of Chromosome 1 (B). Intra-chromosomal map of Chromosome 1, before computing the dot product. (C) Illustration shows a bipartite graph representation of an intra-chromosomal contact map. Nodes belong to HT-blocks and are colored by the homolog of origin of the block. Edges represent observed contacts between HT-blocks in the Hi-C contact map. Dashed edge belongs to observations that are eliminated by the algorithm. (D) Result of the dot product computation for Chromosome 1. (E) Illustration of the impact of the dot product computation on edges in the underlying graph. In real data the graph is also enriched with edges spanning different chromosomes, not shown here.

2.2.1 Haplotype-block binned Hi-C contact maps

The first step of our algorithm aims to prepare diploid Hi-C similarity matrices by utilizing all reads. Our algorithm is parallelized across pairs of both homologs of each two chromosomes. We produce similarity matrices for each two chromosomes. A total of $\binom{23}{2} + 23$ similarity matrices are produced. For this step we begin by binning sufficiently high quality Hi-C reads that overlap bi-allelic loci (Figure 3A, Figure 3B) on both ends into their HT-block pair to a read count matrix, C. I.e., for a pair of HT-blocks K_1, K_2 with homologous blocks $\overline{K}_1, \overline{K}_2$ and bi-allelic loci pair b_1, b_2:

(1) $C_{K_1,K_2}(b_1, b_2) = \#reads\ overlapping\ b_1 \in K_1, b_2 \in K_2$

Once all such reads are mapped, we compute the ratio of observed reads for each HT-block pair in the bi-allelic loci pair:

(2) $R_{K_1,K_2}(b_1, b_2) =$

$$C_{K_1,K_2}(b_1, b_2) / \left(C_{K_1,K_2}(b_1, b_2) + C_{K_1,\overline{K}_2}(b_1, b_2) + C_{\overline{K}_1,K_2}(b_1, b_2) + C_{\overline{K}_1,\overline{K}_2}(b_1, b_2) \right)$$

Next, we map each of the ambiguous reads, reads that have at least one end in a mono-allelic region (Figure 3C) to the four corresponding bins in the ratio matrix R (for each HT-block pair) and interpolate the ratio at its chromosomal loci along a chromosome with a 2D Gaussian kernel (Figure 3D). The interpolant is given by

(3) $f(x, y) = f_{K_1,K_2}(x, y) =$

$$\sum_{\substack{b_1,b_2\ are\ bi-allelic; \\ x,b_1 \in K_1 \\ y,b_2 \in K_2}} \left[e^{-\frac{((x-b_1)^2+(y-b_2)^2)}{2\sigma^2}} \times R_{K_1,K_2}(b_1, b_2) \right]$$

The interpolated ratio for each bin is added to the final read-count matrix. Finally, since bins are not equally-sized, all read counts are averaged by the product of the number of nucleotide in K_1 and K_2, i.e.

(4) $Q(K_1, K_2) =$

$$(|K_1| \cdot |K_2|)^{-1} \cdot \sum_{\substack{b_1 \in K_1 \\ b_2 \in K_2}} C(b_1, b_2) + \sum_{(x,y) \in (K_1, K_2)} f(x, y)$$

The final block-matrix (Figure 3E) represents a more robust picture of a similarity measure between HT-blocks, in their genomic locations across both homologous chromosome pairs, for all chromosomes. Note that utilizing the Gaussian interpolant enabled utilizing all mappable reads to obtain the resulting similarity matrix. The choice of a Gaussian interpolant is further elaborated in the Discussion section.

2.2.2 Mitigating noise and sparsity: dot-product similarities

Our algorithm is founded on the basis of constructing a global similarity measure that integrates over observed local similarities in the partially phased Hi-C map. In the latent 3D structure underlying our data, similarities are inherently transitive, a property that we aim to exploit. Specifically, to determine for a certain HT-block, A_i, whether it should be phased to the same homolog with A_{i+1} or with B_{i+1} we would like to infer a robust measure of which homolog is a more likely pairing based on spatially adjacency.

Embedding discovers the latent structure, however, it can be unfeasible for large matrices (see Supplementary for more on embedding and spectral theory). With this in mind, we devise a "divide and conquer" strategy, solving for each homologous chromosome pair separately. The downside of dividing to sub problems is that informative inter-chromosomal similarities are lost. To alleviate this loss, we introduce a step of computing the whole-genome dot product for each homologous chromosome pair. This calculation is described in Figure 4.

Since both Hi-C and phasing data can have potential errors and biases, we perform a seemingly heuristic step of removing all cross-homolog edges referencing the same HT-block, i.e. edges of the form $\{(A_i, B_i)|\forall i\}$ shown in Figure 4B as a dashed edge. This is a noise-reduction step used to avoid sequencing biases, as previously described (Suhas S P Rao *et al.*, 2014). This type of error appears to be prevalent in Hi-C data (see the light-colored secondary diagonal in Figure 4B in inter-homolog block matrix) and cleaning it is essential to recovering a partitionable embedding, as we show in Figure 7. We further justify this step in the discussion. Finally, the diagonal, $S(K, K)$ for every HT-block K is set to be 1.

2.2.3 Connected components

The dot product matrix described above is not guaranteed to recover estimated weights on all edges. In some cases, partially phased Hi-C can give rise to blocks that are completely unreachable to one another by traversing graph edges. We therefore apply connected component analysis (Dulmage and Mendelsohn, 1958) and perform the embedding and phasing analysis per (non-trivial) component. This issue reduces the coverage of a possible complete phasing that utilizes Hi-C, as discussed in the results section.

2.2.4 Embedding of HT-blocks with multidimensional scaling

In Figure 5 we show several iterations during the convergence process of a single embedding from the ensemble, that contains most of Chromosome 1's HT-blocks. The process is initialized by setting coordinates to the top eigenvectors from the Classical Multidimensional Scaling (Mead, 1992a) on the dot product matrix, that includes explicit zeros. This initialization is a heuristic that helps converge to a local minimum that is more likely to treat a zero value as dissimilar, rather than as a missing value. We then apply non-classical multidimensional scaling (Kruskal, 1964a) where zeroes in the dot product matrix are masked as missing values and are ignored in the optimization. Non-classical MDS attempts to minimize the mistakes between the order of Euclidean distances in the embedding and (non-missing) distances in the input matrix, D, or in our case $D = 1 - \sqrt{S}$. We observe that the quality of phasing increases as the stress criterion for embedding diminishes, while the embedding is agnostic to phasing quality.

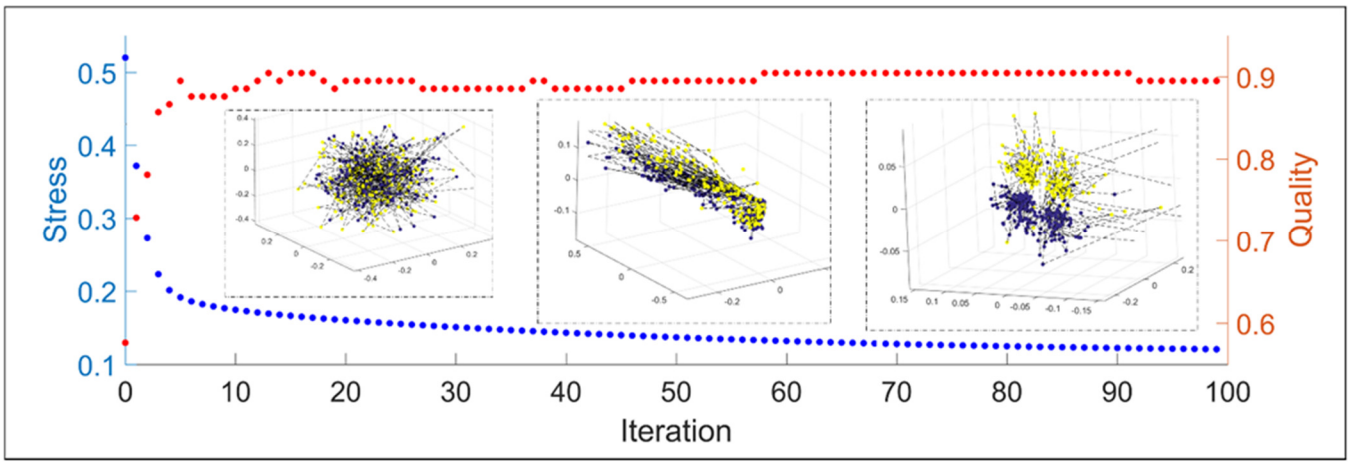


Figure 5. Embedding convergence. Showing the progression of the optimization for the embedding of a connected component in Chromosome 1. Insets from left to right: Random initialization, Classical Multidimensional Scaling, Optimization convergence. HT-blocks of different homologs (according to ground-truth) are colored in yellow and blue, accordingly. Dashed lines correspond to phasing assignments according to the algorithm. The figure shows the stress optimization target function value in blue and the phasing quality (unsupervised) in red. After 6 iterations of optimization the phasing already yields better quality when relying on the embedding rather than relying on local Hi-C similarity. Note that quality is not guaranteed to monotonically increase with embedding steps but is highly correlated. Animations showing convergence progression are available in Online Materials.

2.2.5 Trellis recovery of phasing

Distances in an embedding can be used as estimators for the likelihood of HT-blocks to phase to the same homolog. We apply a simple decision rule along consecutive homologous HT-blocks to compute their best haplotype assignments. For HT-blocks $A_i, B_i\ \forall i \in \{1, \dots m\}$ let,

(1) $$S_{A_i} = \frac{d(A_i,A_{i+1})}{d(A_i,A_{i+1})+d(A_i,B_{i+1})}\ ; W_A = 1 - S_A$$

(2) $$S_{B_i} = \frac{d(B_i,B_{i+1})}{d(B_i,B_{i+1})+d(B_i,A_{i+1})}\ ; W_B = 1 - S_B$$

Where $d(\cdot)$ is the 3D Euclidean distance between HT-block coordinates in the embedding latent space. We define

$$\delta_i = \begin{cases} 1 & S_{A_i} + S_{B_i} \geq 1 \\ 0 & otherwise \end{cases}$$

If $\delta_j = 1$ we call this a 'stay' transition, as we keep the order induced by the arbitrary HT-block ordering, and if $\delta_j = 0$ we call it a 'switch' transition. The set of assignments $\delta_i, i = 1 \dots m-1$

defines the full haplotypes $\mathcal{H}, \overline{\mathcal{H}}$ for the $1..m$ HT-blocks. To compute an optimal haplotype, we maximize

$$\sum_{i=1}^{m-1}\left(S_{A_i}+S_{B_i}\right)^{\delta_i}\left(W_{A_i}+W_{B_i}\right)^{1-\delta_i} \quad (3)$$

We call attention to the fact that a "greedy" solution to this equation, or, taking the maximal P_j assignments $\forall j$, yields the global maximum for Equation (3) by its definition. We visualize the set of HT-blocks and relevant embedding distances in graph form using the Trellis graph (Figure 6).

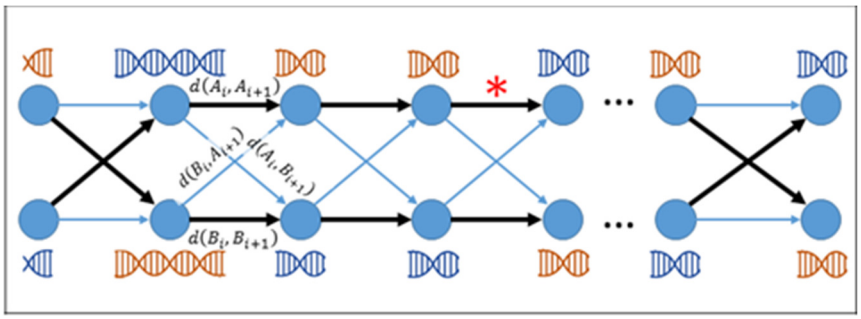


Figure 6. Trellis diagram. Illustration of Trellis graph with selected transitions by 'SpectraPh' highlighted in bold. Nodes represent HT-blocks with homologs along the graph in orange and blue. HT-blocks are randomly permuted between homologs to illustrate the arbitrary order given by the partial phasing data. Edges are weighted by the Euclidean distance between the HT-blocks' corresponding coordinates in latent space. The red asterix shows an erroneous selected a $\delta_* = 1$, 'stay' transition, i.e. the algorithm chose to traverse edges that are not validated by the ground truth phasing data. See an example of selected transitions on real data for Chromosome 1 in Figure S4.

2.3 Results

To investigate the applicability of our algorithm we simulated partial short-range phasing data (see Supp. Methods) at different HT-block lengths. We used the "gold standard" trio-phased GM12878 genome (Auton *et al.*, 2015) as our baseline and show that the algorithm is able to recover the haplotype with high quality using experimentally available GM12878 Hi-C data from (Selvaraj *et al.*, 2013). To investigate the robustness of our method to noise we defined a generative model to sample Hi-C-like data from a Log-Normal distribution. We inspected the effect of noise on our algorithm and compared to 'HapCut'. We define natural quality, confidence and coverage scores, to be computed for each chromosome. Given κ connected

components for a chromosome, and each underlying Trellis ordered according to ground truth phasing:

- Coverage is the percent of remaining Trellis transitions when factoring over all components. Namely, $1 - \frac{\kappa - 1}{m - 1}$.
- Confidence is computed per transition as the difference between 'stay' and 'switch' transition probabilities (E.g. Figure S4). Namely, $S_{A_i} + S_{B_i} - W_{A_i} - W_{B_i}$.
- Quality is computed as the fraction of 'stay' out of all transitions. Namely, $\frac{\Sigma_i^{m-1} \delta_i}{m-1}$.

We emphasize that ground truth order is assumed when computing Confidence and Quality only for performance assessment when a ground-truth phasing is available.

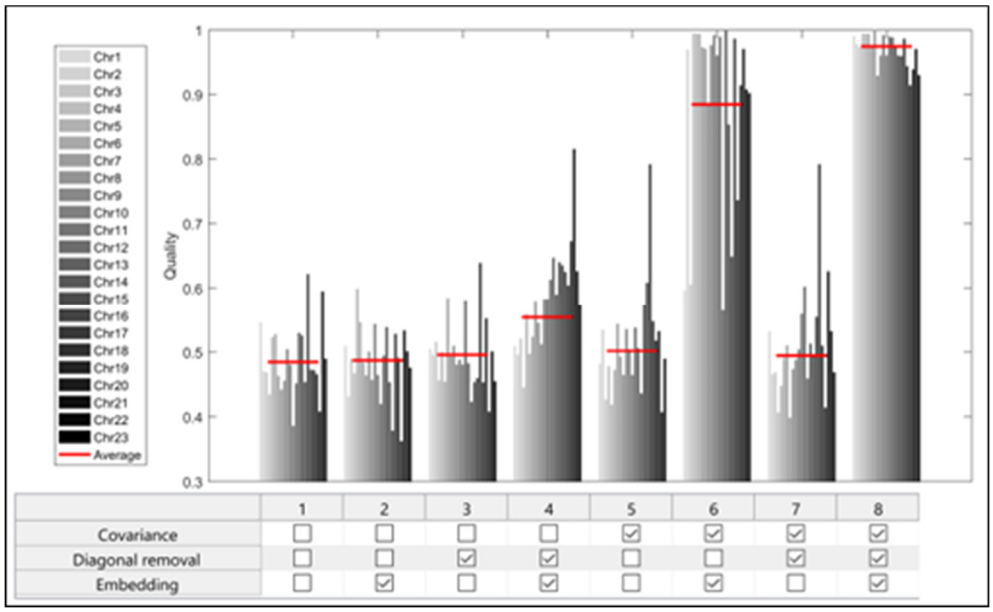


Figure 7. Showcasing impact of applying combinations of the algorithm on quality of phasing. Each bar group contains 23 columns corresponding to the different chromosomes, and a red horizontal line representing the average of the group (weighted by number of transitions per chromosome). Columns below each bar group show which configurations of the algorithm were applied. Bar group #8, that corresponds to applying all algorithm steps, shows near-perfect phasing quality compared to ground truth. Covariance refers to computing the dot-product similarities of two homologs.

2.3.1 Extending partial haplotype in humans with Hi-C data

GM12878 has trio-based phasing data available. To emulate experimentally unavailable short-range phasing, we scan each chromosome for SNP loci, adding ground truth phasing as long as the resulting HT-block length is below a certain threshold. Quality of phasing for all

chromosomes at ≤1Mb HT-block length is shown in Figure 7, including a breakdown of the impact of different non-trivial steps in our algorithm. Figure S2 shows our results for different simulated partial short-range phasing HT-block lengths. Collectively, these show that the algorithm is able to completely recover the ground-truth data for most chromosomes, reaching an average quality of 0.98. We observe that short chromosomes tend to yield poorer results. This effect is amplified when we do not apply the "diagonal removal" heuristic, suggesting that errors in mapping Hi-C reads or in ground-truth phasing data are more easily corrected by taking into account more similarity observations. We have verified the quality results for 100Kb, 500Kb, 800Kb, 2Mb thresholds as well (Figure S2).

2.3.2 Simulated Hi-C data

To investigate the effect of noise we have simulated Hi-C-like data. We begin by generating two 3D curves by iteratively appending random unit vectors within a fixed angle range. Each curve is normalized to its center of mass, and rotated in a random direction. We then sample values from a log-normal distribution based on a transformation of the pairwise Euclidean distances among resulting curves. $\mu_{i,j} = \log(c/d_{i,j})$, where c is a normalizing constant used to control the number of simulated 'reads' in the experiment reflecting a fixed sequencing depth. σ is set as a function of the coefficient of variation $CV = \sigma/\mu$, to control the level of noise in the simulation.

To compare robustness between our algorithm and HapCut, we have implemented a non-optimized version of HapCut in Matlab that can accommodate the format of our simulated data. This simplified version is able to handle the scale of our simulated data and was run for $O(n)$ iterations, as suggested by HapCut authors. Results for the analysis are shown in Figure 8 and indicate that HapCut suffers from inclusion of noise in simulation while our algorithm can reach quality of ~0.87 for $CV = 0.5$.

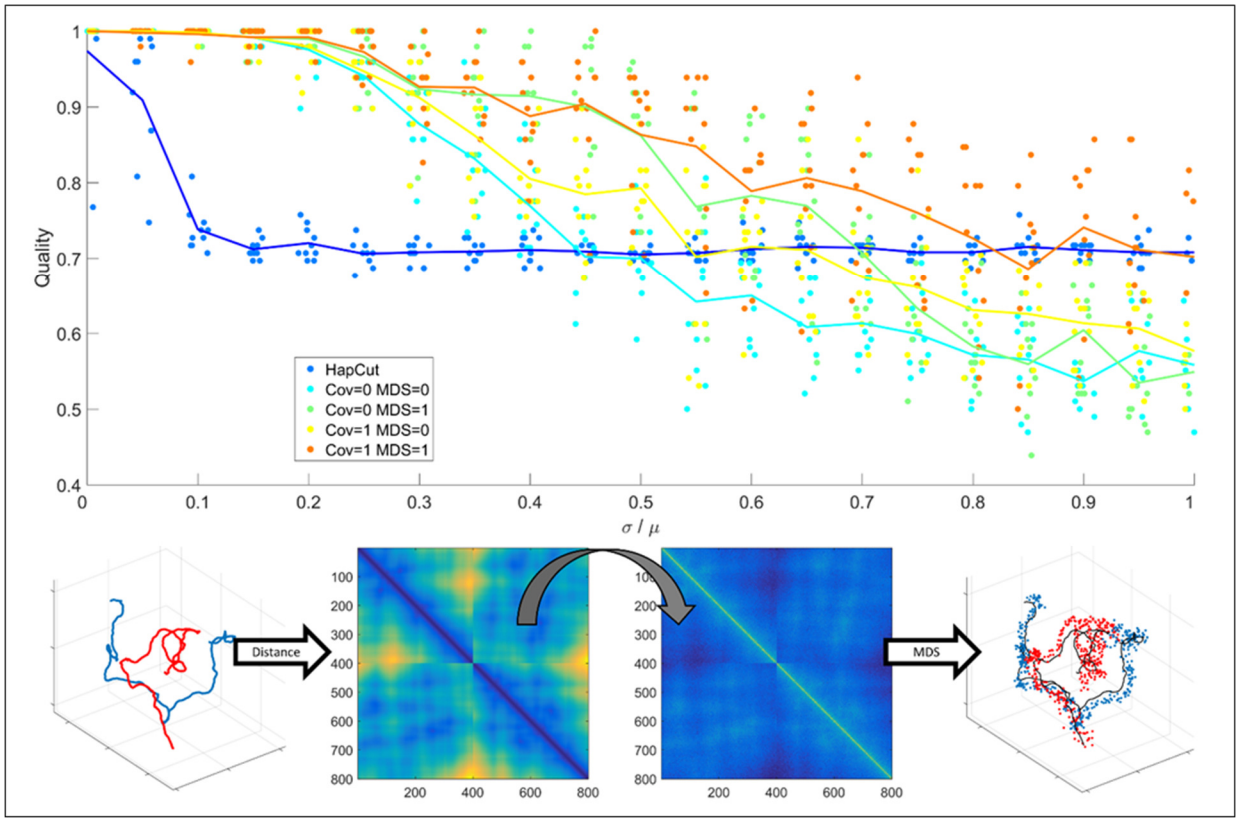


Figure 8. Simulated data signal-to-noise analysis. (Top) Each line at the top represents the average quality computed over 10 instances of simulated Hi-C data for a pair of simulated homologs. We compare our algorithm with HapCut (Blue). (Bottom) Left to right: 1. Example of a generated pair of homologs. 2. Pairwise distances. 3. Transformation to noisy Hi-C data with a log normal sampling with CV=0.2. 4. Embedding result overlaid on top of originally generated homologs.

2.3.3 Enrichment analysis on a diploid genome structure

Once a complete phasing is known we can utilize Hi-C to investigate co-localization of genomic functions. In (Ben-Elazar *et al.*, 2013a) we describe co-localization (in a haploid genome) of yeast transcription factor (TF) targets. Such co-localization supports the existence of Transcription Factories, regions in the nucleus where transcription machinery operates in concert to regulate transcription activity. We now apply a similar enrichment analysis of TF targets (Bovolenta *et al.*, 2012) to demonstrate such analysis on phased Hi-C. Analysis of co-localization in averaged Hi-C data for diploid genomes is also addressed in (Diament *et al.*, 2014). Our results indicate that diploid Hi-C maps provide insights into the distribution of genes in the nucleus that current, averaging based Hi-C analysis approaches cannot identify. An example for the TFAP2C transcription factor is shown in Figure 9. In our analysis we compute

the mHG (Eden *et al.*, 2007, 2009a) enrichment score of TF targets ordered by proximity to a pivot locus. In this example we clearly see patterns that would not emerge from a naïve interpretation of Hi-C data. We can see that often only one homolog of each TF target is within the suggested transcription factory according to the mHG threshold. In more detail, consider a genomic locus, l. Rank all other genomic loci $l_1, \ldots, l_N$ by the distance to l, $d(l_i, l)$. Consider a TF and its set of targets, T. Define a binary vector of length N, $\lambda(i) = 1 \; iff \; l_i \in T$. For n we define $\Lambda_n = [\lambda(1), \ldots, \lambda(n)]$ as the prefix of length n of the binary vector. Let $b_n = \Sigma\Lambda_n$, $B = \Sigma\Lambda_N$. The mHG score is defined by the threshold, n, that minimizes the right tail of the hypergeometric CDF. I.e.,

$$mHG(\lambda) = \min_{1 \le n \le N} \sum_{i=b_n}^{\min(n,B)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}}$$

The null hypothesis in the mHG statistical framework is that all binary vectors of length N with exactly B 1's are equi-probable. In our context, rejecting the null hypothesis suggests that TF targets are localized in significantly close proximity to the pivot locus. We repeat this experiment for all loci and TF, correcting for multiple hypotheses with Bonferroni correction.

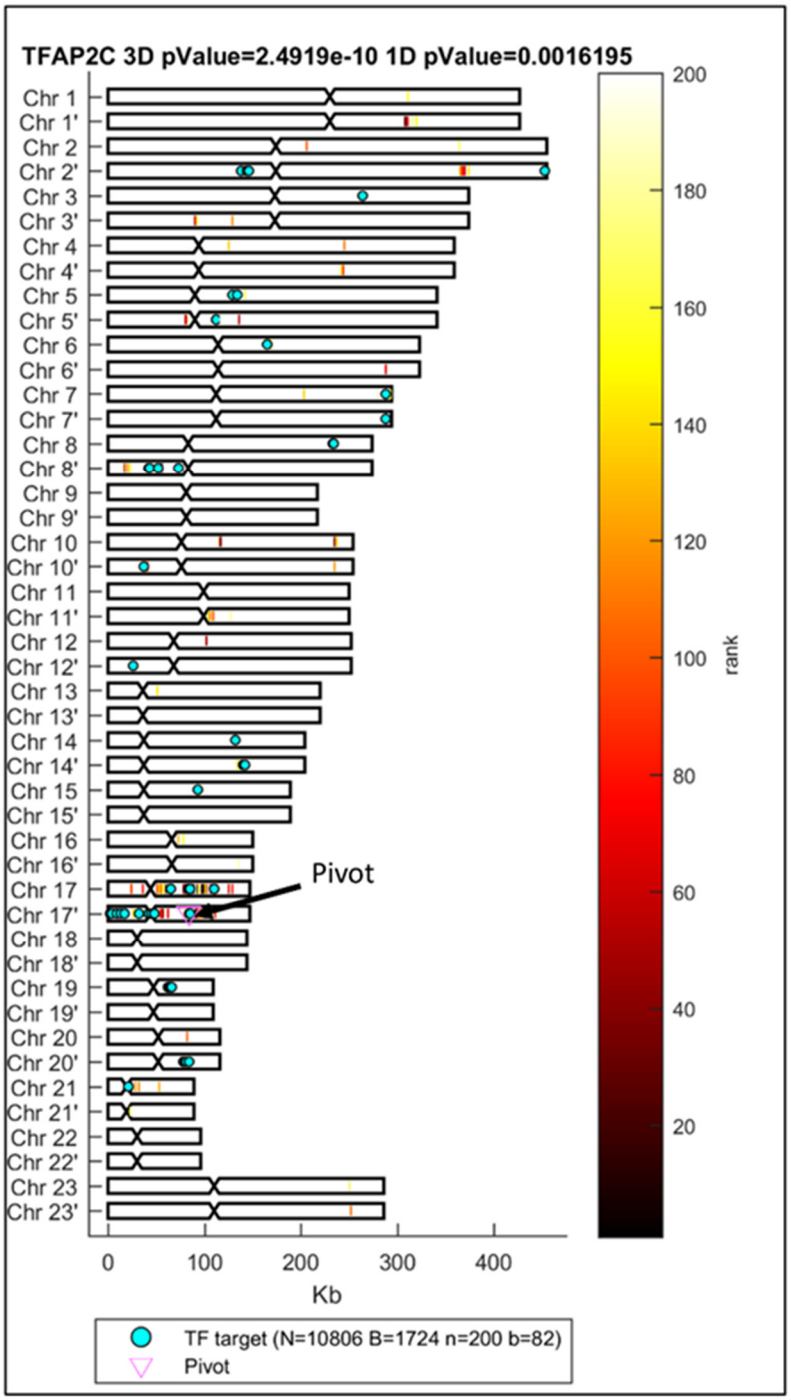


Figure 9. TFAP2C target 3D co-localization pattern. Human genome chromosomes in pairs of homologs. We mark the closest HT-blocks to a pivot HT-block (200 HT-blocks, as determined by the optimal mHG threshold, $p<10^{-10}$, Bonferroni), colored by the rank in the phased dot product similarity. The pivot HT-block is marked as a magenta triangle and arrow pointing to its position on Chr 17'. Targets of the TFAP2C transcription factor that are positioned within the mHG threshold are marked as teal dots. The co-localization pattern evident in the figure illustrates the importance of phasing homologs in Hi-C data, as mostly distinct copies of each TF target inhabit the suggested transcription factory.

2.4 Discussion

The method presented in this paper can refine the haplotype signal in Hi-C data without assuming a complex prior on the experimental setup.

We note that the quality achieved by our method is highly dependent on the genomic size of partially phased HT-blocks (Figure S2). Short HT-blocks characteristically yield sparser maps as the same number of Hi-C reads are binned to a quadratically larger contact matrix. This will become less of a problem as sequencing depth improves with technology. More surprisingly, perhaps, is that long HT-blocks also yield lower quality results. We observe that this phenomenon is related to the skew of the distribution of similarity values used in the embedding. In large HT-blocks the underlying structural signal is averaged over significant portions of the chromosome, and the embedded structure no longer contains the information required for phasing. This issue can be bypassed by preliminarily subdividing HT-blocks in the known partial phasing to produce better Hi-C data for embedding.

Another issue plaguing short HT-blocks is the runtime complexity of the embedding algorithm. To investigate the applicability of our algorithm on HT-block sizes $\leq 100Kb$ would require a higher-performance implementation of the algorithm, or a different algorithmic approach.

One seemingly heuristic step performed in our algorithm is the removal of cross-homolog edges referencing the same HT-block, i.e. secondary diagonal. We argue that while this indeed has impact on the resulting embedding, keeping these edges can only lower the quality of phasing as they can only increase the ratio $(W_A + W_B)/(S_A + S_B)$ when the trellis is ordered according to ground-truth phasing. This notion relates to another important distinction that we would like to stress – while our method is completely reliant on embedding for phasing, it is by no means suggesting that the recovered structures represent actual chromosomal conformation. Specifically, embedding is used as a tool to integrate global similarities into Trellis edge weights to facilitate phasing.

Another point worth discussing is the application of the Gaussian kernel as the interpolant used on the ratio matrix. By utilizing an interpolant that includes a variance parameter we can guarantee a small effect of genomically distant reads, as would be expected by the mechanical

properties of DNA structure. This is especially beneficial when interpolating on large HT-blocks to reduce the effect of genomically distant loci in the block on the interpolated value. In analyses we performed we observe that inclusion of an interpolant is beneficial to the quality of phasing.

Finally, in this work we only briefly address the problem of enrichment analysis on diploid genomes to illustrate the potential advantages of correctly interpreting diploid Hi-C data. We showed how Hi-C data assists haplotyping and the relevance of haplotyping to co-localization of TF targets. It is of interest to further explore TF binding sites and to expand the analysis to other genomic markers.

Acknowledgements

We thank Erez Lieberman Aiden for references and for critical feedback on some of the results discussed in this section.

2.5 Chapter Supplementary Materials

Online materials: Implementation available at https://github.com/YakhiniGroup/SpectraPh

Embedding convergence animations are available at http://imgur.com/a/fwzBD.

Supplementary Methods

Embedding and spectral theory

In the context of matrix theory our embedding approach is a simplified version of more general Spectral methods (Chung, 1994; Spielman, 2007). Embedding theorem can be naturally interpreted by treating values in our similarity matrix as probabilities of a random-walk operator in a Markov Process. In certain conditions, an infinite random walk traversing edges in the similarity graph converges to a stationary distribution which can be applied to compute edge values which capture all transitive relations. Eigenvalue methods (Kruskal, 1964b; Ham *et al.*, 2004) solve this but tend to break down in simulations when we introduce missing edges (Figure S3) as these are treated as explicit zeros in the linear equation solver (Van Der Maaten *et al.*, 2009).

We have previously (Ben-Elazar *et al.*, 2013a) applied an optimization method (Kruskal, 1964a) which avoids missing data in this context, however it does not scale very well and difficult to apply to the entire human genome. To overcome this, we would like to distribute the workload per homologous chromosome pair, without losing all inter-chromosomal transitivity. To this end, we apply Step 2 of the algorithm (Figure 4), and compute the empirical dot product of the full genome matrix which captures two-hop transitive relations in the graph.

Detailed formulation of the haplotyping problem

Let $f(t), g(t)$: $[0,1] \to \mathbb{R}^3$ be smooth and differentiable maps into two arbitrary curves in $\mathbb{R}^3$of unit velocity, i.e. $\forall t \in \mathbb{R},\ f(t), g(t) \in \mathbb{R}^3$ and $f(t), g(t)$ are of differentiability class C^∞.

We are given (approximate, noisy, partial) pairwise distances, $d(\cdot)$, between n consecutively sampled coordinates $t_1 < t_2 < \cdots < t_n$ along each curve ($2n$ in total). The distances between

all pairs of points, $f(t_i), g(t_j), i, j \in [1..2n]$, are given by the $2n \times 2n$ distance matrix, $\widetilde{D}$, as follows:

$$\widetilde{D}(i,j) \equiv \begin{cases} d\left(f(t_i), g(t_j)\right) & i \leq n, j > n \\ d\left(f(t_i), f(t_j)\right) & i \leq n, j \leq n \\ d\left(g(t_i), g(t_j)\right) & i > n, j > n \\ d\left(g(t_i), f(t_j)\right) & i > n, j \leq n \end{cases}$$

Let I_i be n independent coin tosses with $p = 0.5$, for $i \in [1..n]$.

We define $D = \hat{I} \cdot \widetilde{D}$, where $\hat{I}$ is the following, $2n \times 2n$, permutation block-matrix:

$$\hat{I} = \begin{bmatrix} I_{intra} & I_{inter} \\ I_{inter} & I_{intra} \end{bmatrix}$$

With the corresponding blocks,

$$I_{intra} = \begin{bmatrix} I_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I_n \end{bmatrix};\ I_{inter} = \begin{bmatrix} 1 - I_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 - I_n \end{bmatrix}$$

More intuitively, $\hat{I}$ "tosses a coin" to decide if it switches the distance values between coordinate i and $n + i$ (the ith coordinate of curve f and the ith coordinate of curve g). This permutation causes the identity of the curve from which each coordinate was sampled (previously encoded as the corresponding block in which its distance values appear in the matrix $\widetilde{D}$) to be lost.

The geometric task we address is as follows. Given D, assumed to be constructed by such process, we would like to recover the most likely partitioning of the $2n$ coordinates into two corresponding curves, representing $f(t), g(t)$.

Supplementary Figures

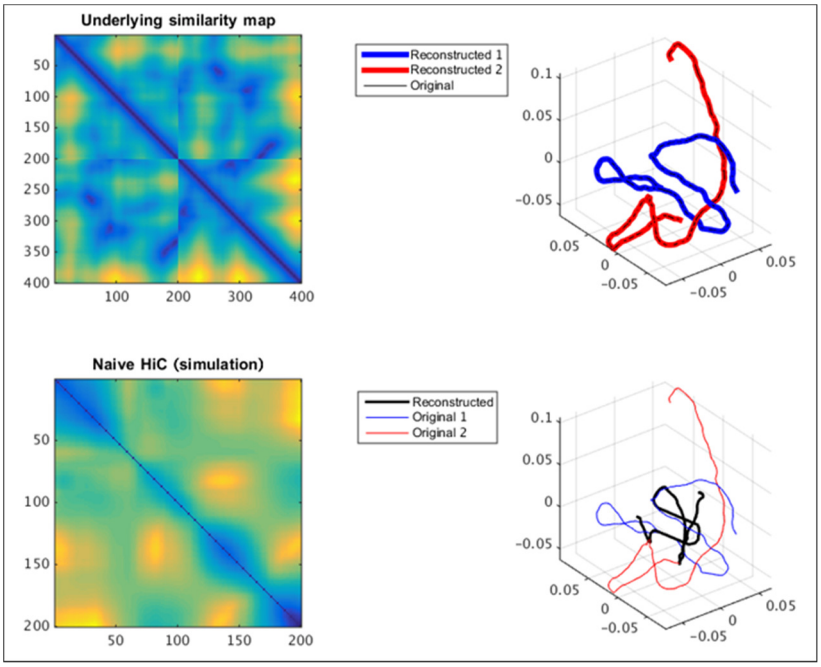


Figure S1. Simulated naïve Hi-C on diploid genome. (Top) Showing a simulated pair of homolog chromosome structures and their underlying distances as a proxy to Hi-C. (Bottom) When naïve approaches multiplex homolog data, the structure does not resemble the original.

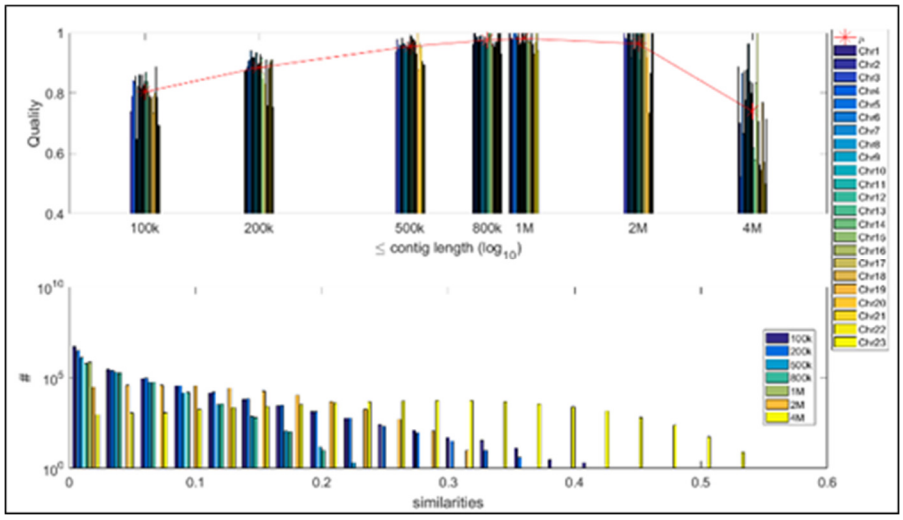


Figure S2. Effect of HT-block size on phasing quality. (Top) We repeated our phasing analysis pipeline for multiple simulated HT-block sizes. We see a pattern where large HT-blocks have lower phasing quality. (Bottom) Distribution of similarity values in our normalized dot product matrix showing a more uniform distribution for large HT-block sizes (in log scale). This shows that at some HT-block length bin size the structural signal is averaged over too many bases to a point where the algorithm cannot phase accurately.

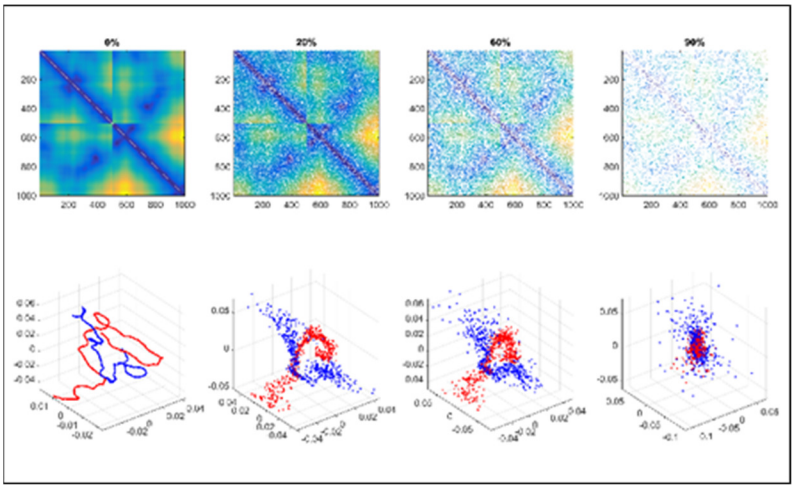


Figure S3. Effect of sparsity on Eigenvalue based methods. (Top) We sample missing data uniformly over a simulated similarity matrix of two homologs. As reference, the Hi-C dataset we use, binned at 500Kb resolution has >98% missing values. (Bottom) Results of classical multidimensional scaling on the corresponding similarities above.

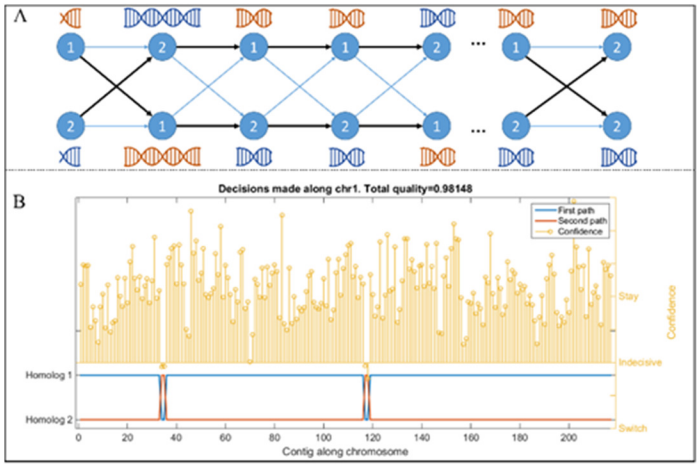


Figure S4. Trellis graph edge weight example for real data. (A) Illustration of Trellis graph with selected transitions. Nodes represent known partial phasing. Homologs in each column along the graph are randomly permuted to illustrate the arbitrary order given in the partially phased Hi-C maps (when no ground truth is known). Edges represent transitions along consecutive phased regions and are weighted by transition probability which is proportional to Euclidean distance between corresponding coordinates in latent space. Edges colored black illustrate the 'switch' and 'stay' decisions the algorithm makes to complete the phasing. Nodes are marked by predicted homolog number (numbered arbitrarily). (B) Showing decisions chosen by the algorithm for chromosome 1. On the primary Y axis and bottom of the graph we see the decisions along the Trellis for a ground-truth-ordered (non-permuted) Trellis. We see the algorithm makes 4 mistakes (chooses to cross from one homolog to the other) along chromosome 1 in this case. On the secondary Y axis and related stem-plot we see the confidence for the choice calculated as the difference between the sum of parallel edge weights and sum of cross edge weights. When the difference is zero the algorithm is 'Indecisive', when the difference is positive the algorithm identifies the likely phasing is the same as ground truth and when negative to 'Switch' from the ground-truth.

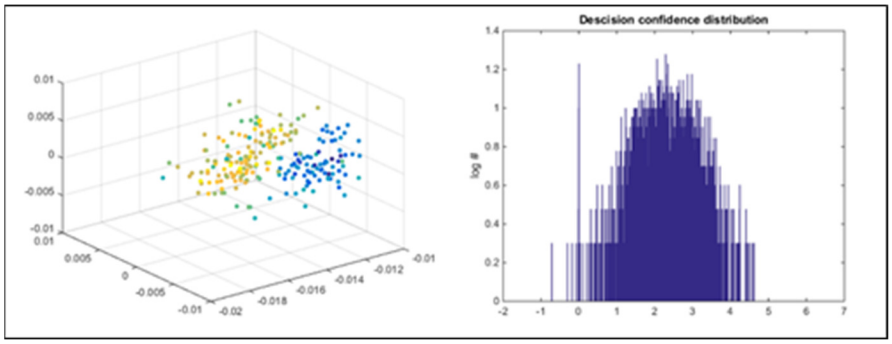


Figure S5. Distribution of confidence, coverage and quality of phased haplotypes. (Left) Embedding of results from Figure S4 color coded by Confidence (see Results) shows correlation between 3D separation and Confidence. (Right) Distribution of all confidence values across all ensembles of all chromosomes. Note that confidence is a signed quantity where positive values are haplotyping decisions which correlate with the ground truth.

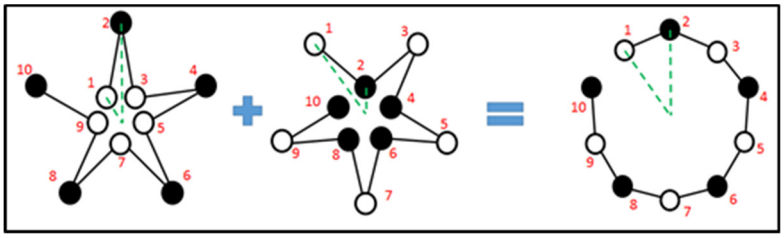


Figure S6. Presenting the under-determinism in naïve multiplexing of Hi-C data. An example of co-localization analysis. (Left) We illustrate a simplified diploid genome composed of two star-shaped 2D homologs, and 10 genes. In this toy example, white colored circles represent genes that share a common function, and black circles genes that do not. For this specific configuration, pivot genes 1,3,5,7,9 on the left-most chromosome homolog display a significant co-localization pattern ($P < 0.005$). (Right) When averaging the pairwise distances of homologous genes to other genes, the resulting naïve "haploid view" representing the underlying diploid genome is presented, and the intricate details of the conformation are lost. Green dashed lines illustrate the averaging effect on distances, wherein genes 1,2 alternate between short and long distances to the center of the homolog in the true, diploid structure, and are averaged to be equally distant in the haploid view. Using the resulting model directly will not yield significant co-localization patterns. Furthermore, the problem of recovering the details of the original, diploid, conformation directly from this view is under-determined. I.e. There are an infinite number of equally valid diploid models that yield the same resulting haploid view, some of which will not express an enriched co-localization pattern.

Chapter 3:

The Functional 3D Organization of Unicellular Genomes

3.1 Introduction

Studying the co-localization of elements along the genome (Kanduri *et al.*, 2018) is used for providing evidence of evolutionary or mechanistic relationships between genomic elements and genomic organization. There are well established functional mechanisms that are known to interact in cis via genomic proximity, such as genes along an operon, promotors and their associated coding sequence, nucleosome modifications and proximal chromatin accessibility, etc. Studying trans interactions has remained elusive until recent technological breakthroughs that have enabled the assessment of the 3D structural properties of genomes. Chromosome conformation capture (3C) and methods derived therefrom (Hi-C) (Ay and Noble, 2015; Lin *et al.*, 2018) are, generally speaking, experimental protocols that yield a sparse map of paired sequencing read counts. These counts correlate with 3D spatial proximities between pairs of genomic loci (Nynke L. van Berkum *et al.*, 2010b). These methods allow for a methodical examination of how the genome folds (Lieberman-Aiden *et al.*, 2009; Suhas S.P. Rao *et al.*, 2014; Sanborn *et al.*, 2015) and how genomic elements co-localize to potentially interact in three-dimensional space (Varoquaux *et al.*, 2015; Sanyal *et al.*, 2012; Thévenin *et al.*, 2014; Nurick *et al.*, 2018), opening the door to studying trans interaction systematically.

Hi-C has established a prominent and noteworthy contribution to our understanding of cis chromatin order and epigenetics with progress in the study and characterization of topologically associated domains (TADs) (Dixon *et al.*, 2012; Nora *et al.*, 2012; de Laat and Duboule, 2013). Such domains are typically presented as local triangle-shapes in a triangular view of the Hi-C interaction matrix, corresponding to local clusters of high intra-cluster, low inter-cluster read density. Studies pertaining to the underlying mechanism of TAD formation have implicated the contribution of CTCF and cohesin, key contributors to cell-type-specific genome conformation (Junier *et al.*, 2012). TADs are believed to form higher-order insulated intra-chromosomal neighbourhoods, regulating gene-enhancer interactions, and their disruption has been shown to cause disease (Denker and de Laat, 2016).

Imaging and Hi-C data, as well as data collected from related techniques, have been used to demonstrate co-localization of active genes in specific conditions and in a handful of organisms. The authors of (Mahy *et al.*, 2002) were among the first to experimentally assess the nuclear localization of active genes. They applied FISH (fluorescence in situ hybridization) to provide evidence contrary to the hypothesis that active genes co-localize at the periphery of chromosome territories. A later study (Osborne *et al.*, 2004), followed with a systematic analysis using independent 3C (chromosome conformation capture) and 3D-FISH experiments. Their results provided early evidence to the dynamic nature of co-localization of active genes. One purpose of this current work is to expand this investigation of co-localization in a more systematic manner. To achieve this, we developed streamlined algorithmic and statistical approaches as described herein.

Transcription factories (Cook, 2010) are an example of an established regulatory mechanism manifested as confined compartments within the nucleus, wherein transcription machinery recruits both cis or trans cofactors and genomic elements to regulate specific cellular functions (A. Iborra *et al.*, 1996; Sutherland and Bickmore, 2009b; Junier *et al.*, 2010). Previous studies have attempted to address the task of statistically assessing the existence of transcription factories. The authors of (Dai and Dai, 2012) compared the number of inter-chromosomal interactions in different functionally-related gene sets and observed statistical enrichment under the hypergeometric null model for interactions among transcription factor (TF) targets. However, a follow-up study (Witten and Noble, 2012) argued that edges in the inter-chromosomal 3C interaction graph are not statistically independent, as was assumed under the model used by (Dai and Dai, 2012), and that co-localization events would therefore be over-counted. To correct for this issue, some studies (Witten and Noble, 2012) applied a re-sampling procedure under which no signal for TF target co-localization was detected. Another study (Paulsen *et al.*, 2013) developed an extended approach that includes intra-chromosomal interactions along with a more elaborate sampling methodology which controls for local genomic structural features and applied this method to discover 3D co-localization of mutations in cancer and chromatin states. Studies from our group (Ben-Elazar *et al.*, 2013a; Shay Ben-Elazar, Chor, and Yakhini, 2016) took a different approach to statistically assess transcription factories (Witten and Noble, 2012; Dai and Dai, 2012) that avoids comparing between populations of pairwise proximities altogether, and so circumvents any statistical dependence issues that fail some earlier methods. Specifically, in the aforementioned work (Ben-Elazar *et al.*, 2013b; Shay Ben-Elazar, Chor, and Yakhini, 2016) we compute our statistics independently on each genomic bin – a pivot point centered at some locus along the genome around which we measure the statistical significance of co-localization. Since this approach is only concerned with distances measured from a single fixed point, it avoids dependence issues related to working with all interaction pairs. For example, this approach never considers a triplet of significantly interacting genomic bin pairs $(i,j),(j,k),(i,k)$ and therefore avoids dependence arising from transitivity, which was correctly pointed out by (Witten and Noble, 2012). We rank all genes according to the number of interactions recorded between them and the pivot point under consideration. Using the ranked list of genes, we applied a statistical model to quantify whether targets from the functional set are significantly localized close to that pivot. We then apply additional safeguards to control for multiple hypotheses evaluated across different genomic bins and for events confounded by genomic proximity. The approach of (Ben-Elazar *et al.*, 2013b; Shay Ben-Elazar, Chor, and Yakhini, 2016) is flexible in its inherent ability to detect partial co-localization of only a subset of the query set of TF targets, where approaches based on averaged Hi-C signal would require exponentially enumerating all possibilities. In addition to producing this subset, our method also produces the set of all genomic bins that geometrically reside within the convex subset of co-localized TF targets, but are not labelled as belonging to the query set. These bins could potentially hold elements that are functionally related to group in questions. A shortcoming of the above is that, in reality, co-localization needs not be geometrically restricted to a 3D point positioned precisely on a genomic locus but can be arbitrarily

centered in space. Thus, events of significant colocalization may remain undetected by this method, as shown by the synthetic construction in (Figure 10, Left). We later report a conceptually similar result on actual biological data for *Caulobacter crescentus*, further illustrating the need for a method that can overcome the shortcoming of such an approach. In both synthetic and real-data examples, none of the genomic bins yield a statistically significant co-localization result and such phenomena would be inadvertently ignored by methods that are limited to genomic bins as pivots.

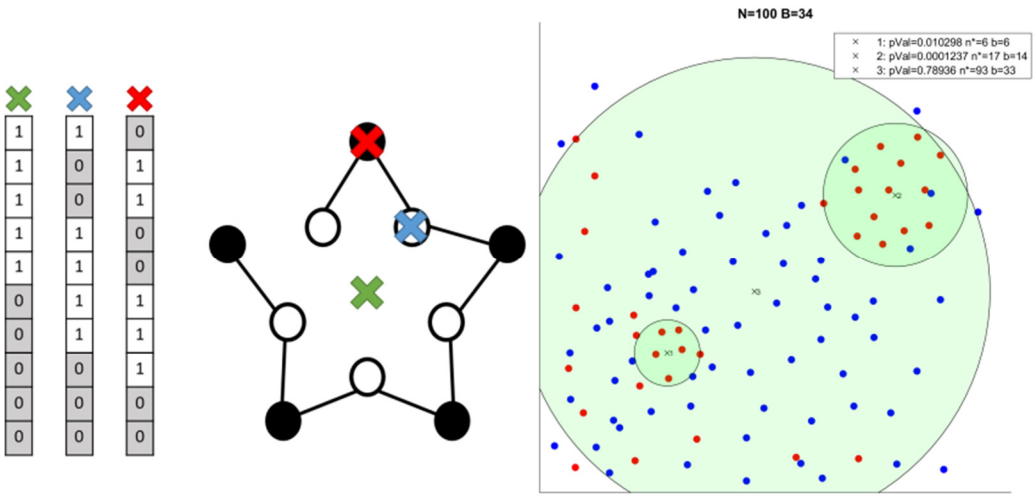


Figure 10. Synthetic examples of co-localization. Left: A construct showing that (2D) spatial co-localization might not be identified by selecting positions along a 1D curve. Circles represent genomic bins. White circles contain TF targets; black circles are bins without TF targets. Red and blue 'X' represent both possible distinct pivots due to symmetry. On the left side we show the corresponding binary vectors reflecting the 2D (Euclidean) distance from each possible pivot. Green 'X' marks the optimal position (yielding the most significant mHG p-Value, see methods) and would not be identified with previous methods. Right: Showcasing three example pivots in a synthetic example. Three green discs representing three pivots (center of disc) with corresponding mHG p-values (in legend) and thresholds are reported. Red points are treated as binary '1' in the corresponding λ vectors. x_3 represents the center of mass of red points, illustrating its sensitivity to the distribution of red and blue points. x_1, x_2 show that the method can adjust to different densities in the data.

In this work, we aim to extend our previous studies by removing the requirements for the pivot to reside on the genome. Our approach, as reported here, enables the study of co-localization of a set of genomic elements centered at arbitrary points in 3D space representations of Hi-C data. Investigating cis driven chromatin order, such as TADs, relies on the 1D topology of genomic order. Clearly, studying trans chromatin order, as in transcription factories, benefits from understanding the embedding of measured proximity data. We provide insights into the difficulty of solving this problem exactly and suggest several heuristics to approach it. We provide code and software implementing these approaches efficiently. In the discussion section, we compare our statistical enrichment approach to co-localization with a more simplistic sampling-based assessment. While a sampling-based approach will find some of the co-localization events, it will, as we show, miss several significant ones. Finally, we apply our method to multiple publicly available datasets across several species. Our analysis is able to uncover previously unreported cases of various genomic elements that appear significantly spatially co-localized. Co-localization alone cannot be used as

direct evidence of an underlying mechanism due to potential confounding linkage. Although requiring additional experimental validation, these results shed new light on the genomic 3D organization of unicellular organisms.

3.2 Methods

We present a statistical-algorithmic framework, referred to as *Spatial-mHG (*$smHG$*, in short)*, that can quantify patterns of spatial co-localization of binary-labelled elements.

Intuitively, our method scans an input set of 3D locations (for example, genomic bins in a 3D embedding of Hi-C data) labelled by some binary property, looking for 'hotspots'. These are regions in which we observe an enrichment of '1'-labelled and a depletion of '0'-labelled genomic bins. Our method identifies hotspots as specified by 3D balls centered at pivot points. These events are statistically quantified for each pivot under a null model. We specifically use the, previously developed (Eden *et al.*, 2007, 2009a), minimum hypergeometric null model. In the next two subsections we provide detailed formal definitions and analyze the computational complexity of providing exact solutions. We consider different algorithmic and heuristic strategies as well as statistical controls. This formal mathematical exposition can be skipped by readers who are not interested in such details of the methodology. The results section uses graphical representations that explain the nature of the results without relying on the mathematical details.

In the second part of this section, we list several Hi-C datasets as well as functional annotation sets explored in this study. We conclude this section by presenting a novel smoothed embedding approach that we applied for generating 3D configurations based on Hi-C data as input for $smHG$.

3.2.1 Spatial-mHG: statistics

Consider a set of points in 3D with binary labels:

$$\mathcal{D} = \{x_i, y_i | \{x_i \in \mathbb{R}^3\},\ \ y_i \in \{0,1\}\}_1^N$$

We define $B = \Sigma_1^N y_i$ to represent the number of '1' labelled points in the data.

Let $p \in \mathbb{R}^3$ be some arbitrary point, also referred to as the 'pivot'.

Define $\lambda_p = (y_{r_1}, y_{r_2}, \dots, y_{r_N})$, the binary vector that satisfies $\|p - x_{r_1}\|_2 \leq \|p - x_{r_2}\|_2 \leq \cdots \leq \|p - x_{r_N}\|_2$.

That is λ_p is the binary vector induced by ranking points x_i according to their Euclidean distance from p.

Further consider

$$\phi(p) = mHG(\lambda_p) = \min_{1 \leq n \leq N} \sum_{i=b_n}^{\min(n,B)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}}$$

where $b_n = \Sigma_i^n \lambda_p(i)$.

mHG is a, previously published (Eden *et al.*, 2007, 2009a; Ben-Elazar *et al.*, 2013b; Shay Ben-Elazar, Chor, and Yakhini, 2016), statistical framework that inspects prefixes of a binary vector, such as λ_p, for

overabundance of '1' under a hypergeometric null model. Intuitively, the likelihood of an overabundance of '1's is compared against a uniform distribution of such labels along λ_p.

Since any two prefixes are statistically dependent, the resulting score requires a correction scheme to be applicable as a p-value. mHG corrects for multiple hypotheses by explicitly, and efficiently, computing the cumulative probability distribution function (CDF) for a given configuration of N, B. Querying the CDF at the resulting score yields a corrected p-value (Eden *et al.*, 2007).

In $smHG$, $\phi(p)$ would be small when '1' labelled points co-localize around p (Figure 10, Right).

Recall that we are interested in points that minimize $\phi(p)$, formally

$$(*) \qquad \boxed{\arg smHG = argmin_p\{mHG(\lambda_p)\}}$$

The $smHG$ framework is therefore seeking pivots where a statistically significant mHG is obtained for the data, $\mathcal{D}$. As stated, solving $(*)$ naively requires searching through all 3D space - a continuum of pivots. A relatively simple observation shows that the number of pivots that needs to be considered is actually finite. For every pair of points such that one is labelled as '1' and the other as '0' we can divide $\mathbb{R}^3$ using a plane that is perpendicular to their connecting line segment, and crosses in its middle. The arrangement of such (perpendicular bisecting) planes, or 'bisectors', tessellates the space into convex polygonal compartments, or 'cells'. It is easy to see that given a single pivot from each cell (e.g. its centroid) we can cover all distinct binary vectors, λ_p, for a given dataset. In Supplementary 10 we provide an exact polynomial bound on the number of pivots that produce distinct λ_p vectors as $\Theta\binom{B(N-B)}{3}$, leading to a worst case bound of $O(N^6)$, as previously described in (Yaglom and Yaglom, 1987).

Unfortunately, from a practical perspective, this number of cells quickly becomes intractable even for moderately sized datasets, leading to statistical as well as algorithmic challenges. For a single cell (pivot) we can report precise *p*-values using the exact distribution of the mHG statistic (Eden *et al.*, 2007), however, there is a vast number of multiple hypotheses, namely cells, investigated in a single spatial-mHG instance as in $(*)$. Characterizing a precise probability distribution for spatial-mHG remains a difficult task and so we apply FDR correction and report *q*-values. We also apply statistical assessment based on simulations as described below.

3.2.2 Spatial-mHG: algorithmics and heuristics

An approach to evaluate spatial enrichment for a given set of labelled 3D data is a function $\mathcal{F}: \mathcal{D} \to [0,1]$. As indicated in the above discussion, the fast growth of the number of cells leads to algorithmic issues. Specifically, a naïve exhaustive approach for large *N*, although possible in principle, is practically infeasible due to the $O(N^6)$ complexity. In our analysis, we compare several heuristic approaches that aim to deal with this challenge. These approaches, denoted by $smHG^{Grid}$ and $smHG^{Sample}$ correspondingly, provide an upper bound on $smHG$. As described, our methods are designed to detect significant results but cannot guarantee a recall of all significant results.

See Supplementary 1,2 for discussion of the performance and trade-offs of the heuristics tested here and See Supplementary 3 for more technical notes on our experimental set up. An illustration summarizing the key differences between both approaches is available in Figure 11.

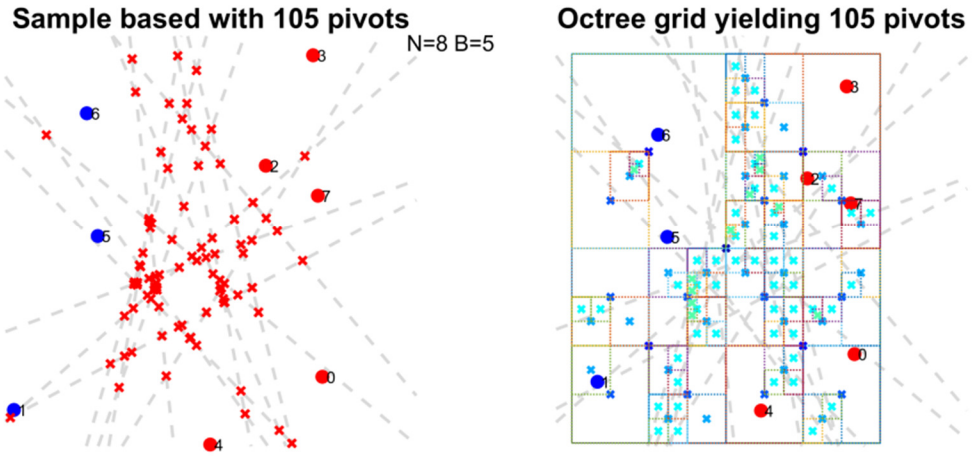


Figure 11. Illustration comparing implemented heuristics. Original points shown as red/teal and numbered from 0 to 7 where $B = 4$. 16 Bisectors are drawn as dashed gray lines, yielding 120 (closed) cells. Left (animation available as Supplementary Video 1): pivots generated in $smHG^{Sample}$ are red x's. In this example our sampling algorithm is run to exhaustion Right (animation available as Supplementary Video 2): pivots generated in $smHG^{Grid}$ are teal 'x's and corresponding dynamic grid structure colour coded by BFS depth in quad-tree. Here we stop the algorithm after yielding 120 pivots, illustrating the difference in behaviour to $smHG^{sample}$.

Grid approach: $smHG^{Grid}$. We recursively iterate over a uniform 3D-grid. Namely, we partition space into eight disjoint, nested, cubes where the center of each cube is to be used as a pivot. This uses a common underlying data structure called octree (Meagher, 1982), and a branch-and-bound algorithmic approach. Let C_{t+1} be the t+1st - cube evaluated. C_0 is the root node in the tree referring to a cube bounding our input data (with some slack to allow pivots outside the convex set to be considered). We dynamically build the octree while traversing it in a breadth-first manner by maintaining a priority queue. Let $OPT(t)$ be the best observed $smHG$ after t cubes are evaluated, and set $Bi_{C_{t+1}}$ = {bisectors that intersect with C_{t+1} | bisectors that intersected C_{t+1}'s parent cube}. Denote $smHG(P_{C_{t+1}})$ the $smHG$ score given by using the center of C_{t+1}, $P_{C_{t+1}}$, as a pivot. We observe that at this point we have enough information available to compute a lower bound on the best theoretically-achievable *p*-value for all cells contained by the cube C_{t+1}. If this lower bound is $> OPT(t)$ we stop the recursion at C_{t+1} since no sub-cube can possibly improve on $OPT(t)$.

Assume there exists a hypothetical pivot, $p^{hyp} \in C_{t+1}$, for which every bisector $bi \in Bi_{C_{t+1}}$ is 'satisfied': Let $\{x_1, 1\}, \{x_2, 0\}$ (W.L.O.G.) be the data points and labels which induced the bisector bi, p^{hyp} 'satisfies' bi if $\|p^{hyp} - x_1\|_2 < \|p^{hyp} - x_2\|_2$. Let k be the number of bisectors in $Bi_{C_{t+1}}$ that are not satisfied by $P_{C_{t+1}}$. We can compute $smHG(P^{hyp})$ by exploiting the data structure used to compute $smHG(P_{C_{t+1}})$. Intuitively, we

append k '1's after every valid prefix of $\lambda_{P_{C_{t+1}}}$ (such that B does not increase) and evaluate the resulting mHG *p*-value.

We note that this method guarantees a finite number of pivots, but each cell may be visited more than once. Details on this and more caveats are available in Supplementary 3.

Sampling approach: $smHG^{Sample}$. Every three bisecting planes in general position (bisectors $B_i \triangleq a_i X + b_i Y + c_i Z + d_i = 0$) intersect at a point, $p_B = (x, y, z)$. We take an ϵ-step along the gradient of each of the three bisectors and average the resulting points to yield a pivot inside a cell p_c. Formally,

$$p_c = \left(\frac{1}{3} \sum_{i=1}^{3} \frac{-(b_i * (y + \epsilon) + c_i * z + d_i)}{a_i}, y + \epsilon, z \right)$$

This procedure defines a one-to-one mapping for every bisector-point-intersection to cells such that every such pivot point is "bottom-most" (w.r.t. dimension y) of some cell, as illustrated in Supplementary Figure S**14**. With this in mind, we iterate over bisectors to yield combinations of three distinct bisectors and by doing so recover all "bottom-most" pivots exactly once.

Given an actual data instance, $\mathcal{D}$, we are interested in benchmarking the enrichment evaluated by any of the above approaches against adequate controls. To do so, we apply the following controls:

'Bead' pivot control, denoted $Bead\ Control$. Uses every original x_i ('beads' along genome) as a candidate pivot, and only those. This is used to compare results with our previously published method (Ben-Elazar *et al.*, 2013b; Shay Ben-Elazar, Chor, and Yakhini, 2016).

Genomic order control, denoted $1D\ Control$. Uses every original x_i as a candidate pivot, but ranks according to 1D genomic distance (i.e. for x_i, x_j, rank by $(i - j)$), rather than, 3D, Euclidean distance. We restrict this analysis per chromosome where applicable, as genomic inter-chromosomal distance is undefined. This analysis is used to filter out results driven entirely by genomic enrichment, rather than spatial enrichment, as these are not the focus of this paper and can be identified without the need of Hi-C data or $smHG$.

Simulations control, denoted P_{sim}. Runs 100X shuffles on the label vector, y, running both $smHG^{grid}$ and $smHG^{sample}$. P_{sim} is then reported as the empirical CDF where the population is comprised of $100 \times \min\{smHG^{Grid}, smHG^{sample}\}$ values. This evaluation is used as an additional approach of computing an empirically determined corrected *p*-value, since, as previously mentioned, $smHG$ conducts multiple hypothesis testing (many dependent cells are treated independently) without an exact correction scheme.

3.2.3 Hi-C datasets and annotation sets

We investigated several unicellular genomes and functional annotation sets, as follows:

- Bacteria: *C. crescentus.* Le et al. (Le *et al.*, 2013) investigate expression of genes in chromosome interacting domains and their organization under a plectonemic model.

- Bacteria: *B. subtilis.* Marbouty et al. (Marbouty *et al.*, 2015) focus on the 3D architecture of the origin domain and its dynamics during the cell cycle.
- Yeast: *S. pombe.* Mizuguchi et al. (Mizuguchi *et al.*, 2014) experiment with Cohesin mutants illustrating its globule-formation function and discuss the role of heterochromatin in facilitating inter-chromosomal interactions.
- Yeast: *S. cerevisiae.* Duan et al. (Duan *et al.*, 2010) early work on structure reconstruction and the study of transcription factories.
- Fungi: *N. crassa.* Klocko et al. (Klocko *et al.*, 2016) study sub-telomeric facultative heterochromatin and the impact of various histone modifications wildtype chromatin conformation.

Given an annotation dataset, namely one that induces binary labelling on genomic loci, we map annotation elements to genomic bins at the resolution, N, as provided in the aforementioned published Hi-C datasets. We filter out resulting annotation sets that map to less than four '1' labelled bins ($B < 4$). We used several types of annotations, as applicable, for the different organisms.

Common annotation sets.

- Gene Ontologies (GO) are acquired from (Ashburner *et al.*, 2000; The Gene Ontology Consortium, 2017) for all five organisms.
- COGs/KOGs are acquired from (Galperin *et al.*, 2015; Koonin *et al.*, 2004) for bacteria and yeast.
- Transcription factor target cohorts are acquired from (Novichkov *et al.*, 2013) for bacteria and from (Teixeira *et al.*, 2018) for yeast.

Differential annotation sets.

We show how one can turn various types of genomic measurements into binary annotations that can be studied using our proposed framework. To illustrate this capability, we use the data published in *S. pombe* (Mizuguchi *et al.*, 2014) which includes the following datasets for both wild-type and mutants:

- CGH: Do copy number variations co-localize to some spatial locations?
  CGH data was binned to the same resolution as Hi-C, averaged by $\sqrt{\#mapped\ probes}$ in bin. Bins with less than 20 probes were removed. Resulting values, $V = \{v_i\}$ were binarized such that $b_i = \begin{cases} 1 & v_i > \mu + 2\sigma \\ 0 & else \end{cases}$ where μ, σ are the mean and standard deviation of V, accordingly.
- Hi-C Data: Do genomic structural changes occur in spatial clusters?
  To evaluate differential Hi-C structures we compute Z scores from the Hi-C datasets of reference (REF) and variant (VAR). Then, per chromosome, we mask out (set as '0') values in location i, j where $abs(i - j) > 5$ and compute the pairwise Euclidean distance between the masked vectors for locus i in REF and locus i in VAR and compute the Z scores on the results. Next, we binarize

when $|Z| > 1.96$ to produce y_i for $smHG$. Intuitively, these are loci that have changed substantially in (local structure) curvature between REF and VAR. We use x_i from the embedding of REF.

3.2.4 sNMDS smoothing of embedded Hi-C data

Embedding Hi-C data attempts to recover a 3D conformation, or ensemble of, that explains the observed data, with mounting qualitative evidence to support its reliability in capturing biological-structural phenomena (Ay and Noble, 2015; Liu *et al.*, 2018; Varoquaux *et al.*, 2014; Ay *et al.*, 2014; Mercy *et al.*, 2017; Treut *et al.*, 2018). We have previously (Shay Ben-Elazar, Chor, and Yakhini, 2016) demonstrated a quantitative advantage of using embedding distances over Hi-C read counts for the task of phasing haplotypes in a human genome, reinforcing its importance for denoising raw Hi-C read counts. We note that such embeddings cannot necessarily be conceived as representing an actual 3D genomic structure (see Discussion).

NMDS (Nonmetric Multidimensional Scaling) (Ahrens, 2007; Mead, 1992b) is a well-established embedding algorithm that iteratively minimizes a loss function measuring the *violations of ordinality* between the embedding and the input distances. Meaning, it attempts to find a conformation where the two closest points in the input will remain so in the embedding, and so forth. This property is desirable for $smHG$ as it implies the embedding will directly optimize λ_p vectors for $p \in \{x_1, \dots x_N\}$, to reflect the ordinality of observations as much as possible. Applying NMDS to Hi-C data often leads to unlikely discontinuities in the resulting configuration. Such discontinuities are especially evident in degenerate mapping of low-genomic-sequence-complexity regions and biased Hi-C measurements. For example, we may get consecutive genomic bins from the same chromosome that are unreasonably distant in space when compared to any other consecutive pair.

sNMDS (smoothed NMDS) iteratively corrects outliers in the embedding, enforcing smoothness for 1D genomic neighbours. Outliers are defined according to the distribution of distances between all genomically-consecutive bins (the discrete derivative) along the same chromosome. We compute Z-scores and provide thresholds as parameters that determine outliers (genomic discontinuities) for each iteration of the correction. These outliers are then corrected using linear interpolation. We demonstrate that this process results in qualitatively superior embedding configuration in Supplementary 5.

3.3 Results

Using the method described herein we found evidence of functional 3D organization across multiple organisms and multiple functional annotation sets, illustrating the prevalence of structure-function relationship at a genomic scale, in unicellular organisms. Below we describe selected results chosen according to their statistical significance as well as according to their potential biological implications. We provide a supplementary table with more details for all results. as well as some descriptive meta-analysis

is available in Supplementary 8. To further highlight the advantage of the grid method in identifying particular cases of spatial enrichment we performed an additional meta-analysis directly comparing the results among suggested heuristics in Supplementary 13. Finally, a discussion on several noteworthy negative results where functionally related elements did not appear to co-localize is available in Supplementary 9, for completeness.

3.3.1 sNMDS results for Hi-C data of unicellular genomes

The first step of our approach is to apply sNMDS to Hi-C data and produce a 3D embedding configuration that is used to represent denoised distances from noisy measured population Hi-C read counts. We base our enrichment analysis on these configurations. These embeddings should not necessarily be considered as representing actual genomic 3D structure as further considered in the Discussion section. We apply sNMDS and $smHG$ to elucidate distinct spatial enrichment patterns across multiple organisms and provide insights into the variability and prevalence of genomic functional organization across phyla. In the next subsections we list our key findings for each organism and discuss previously unreported phenomena detected as significant by $smHG$, as related to the functional 3D organization of the organisms studied.

3.3.2 Caulobacter crescents

In Figure 12 we present the sNMDS embedding of Hi-C measurements in *C. crescentus* (at synchronized cell cycle t=0, (Le *et al.*, 2013)), displaying a saddle-like, crescent structure, similar to its bacterial cell shape. A recently published (Yildirim and Feig, 2018) high resolution structural study provided qualitatively similar models with experimental validation.

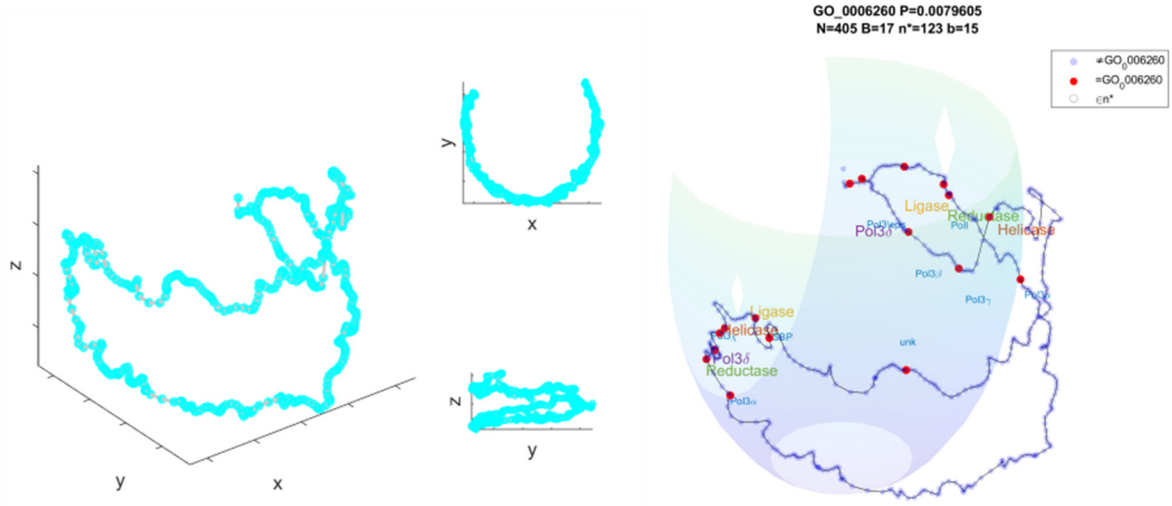


Figure 12. C. crescentus results. Left: sNMDS embedding of C. crescentus from three viewing angles. Right (animation: available as Supplementary Video 3): red spots are genomic bins which contain genes labelled as DNA replication genes under GO:0006260. The floating 'x' is the $smHG$ optimal observed pivot. Translucent semi-sphere represents the ball induced by the $smHG$ threshold. Gray circles indicate bins within the threshold and corresponding ball. Simplified gene labels in GO:0006260. Reductase in green, Helicase in red, Ligase in orange.

Genes annotated as elements of DNA replication (GO:0006260) appear polarized in two distinct sets along the replication axis ($smHG^{Grid}$: $[P < 6e^{-6};\ Q < 8e^{-3};\ P_{sim} < 0.01]$, $Bead\ Control$: $[P < 0.02;\ Q < 0.32]$, $1D\ Control$: $[P < 0.03;\ Q < 0.85]$, Figure 12, middle). Note that this is a real data example resembling the synthetic construction used in Figure 10 in the sense that $smHG$ finds an enrichment centered around a non-genomic pivot that is not evident under the bead pivot nor under the 1D genomic based approaches. Focusing on the individual gene families the observed dichotomy coincides with *ori* and *ter* locations (origin and terminus of DNA replication, accordingly), alluding to evolutionary pressure for duplicated machinery templates possibly related to the replication mechanism. A possible explanation of this observation can come from having a fall-back template for critical elements in the replication machinery in case of a stalled replisome blocking RNAP access (Yeeles *et al.*, 2013). We also observe more subunits from the DNA pol III family available near the Ori, which may relate to the fact that the cell exists longer in a state where these regions are replicated before meiosis.

The observed behavior of polarity along the replication axis appears to be a property of *C. crescentus*. We performed a meta-analysis of our results (Details in Supplementary 6) that illustrate that this property is consistent across available annotation sets and is significant ($P = 0.01$) under an appropriate statistical model.

3.3.3 Bacillus subtilis

In Figure 13 we present four sNMDS embeddings of Hi-C data from available time-course Hi-C measurements in *B. subtilis* (Marbouty *et al.*, 2015).

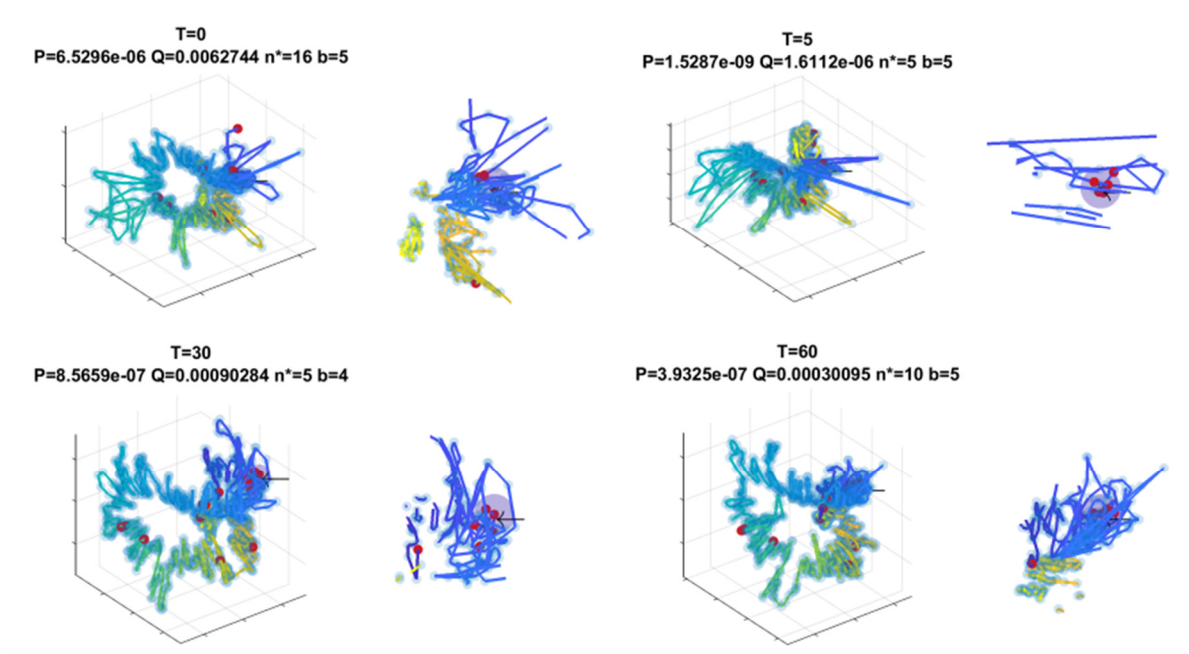


Figure 13. B. subtilis results. Left-to-right, Top-to-bottom (animation available as Supplementary Video 4): Embeddings of time-course Hi-C of B. subtilis at t={0,5,30,60} minutes after release from synchronized G1 into S-phase. Embeddings are aligned with Procrustes analysis. Color gradient along the chromosome is genomic position (showcases the circular nature of the chromosome). Red circles indicate genomic bins that contain gene(s) targeted by BSU00470 (Purine biosynthesis operon repressor). A single translucent ball in each subplot represents the $smHG$ result (pivot and threshold mapped to radius). A black arrow points to the location of the ball. Figure depicts the dynamic nature of co-localization of the targets of the above TF. Next to each subplot we show a zoomed-in plot of the sites of detected co-localization.

Targets of transcription factor BSU00470 (Purine biosynthesis operon repressor) co-localization signal shifts and changes during cell cycle. We observe a substantial colocalization increase in $T = 5$ minutes after release from G1 into S-phase, as defined by the original report (Marbouty *et al.*, 2015). Results are summarized in Table 1 and visualized in Figure 13, top right.

T	$smHG^{Grid}$	$Bead\ Control$	$1D\ Control$
0	$P < 2e^{-6};\ Q < 0.04$	$P < 1e^{-5};\ Q < 0.007$	$P < 1e^{-8};\ Q < 1e^{-5}$
5	$\boldsymbol{P < 1e^{-8};\ Q < 1e^{-3};\ P_{sim} < 0.01}$	$\boldsymbol{P < 1e^{-8};\ Q < 1e^{-5}}$	$P < 1e^{-8};\ Q < 1e^{-5}$
30	$P < 1e^{-6};\ Q < 0.02$	$P < 1e^{-6};\ Q < 1e^{-3}$	$P < 1e^{-8};\ Q < 1e^{-5}$
60	$P < 1e^{-7};\ Q < 0.02$	$P < 1e^{-6};\ Q < 1e^{-3}$	$P < 1e^{-8};\ Q < 1e^{-5}$

Table 1. TF target co-localization dynamics during cell cycle. B. subtilis *BSU00470 (Purine biosynthesis).*

Purine synthesis and salvage gene expression has been observed to fluctuate substantially during the cell cycle and is known to respond quickly to changes in pool availability (Fridman *et al.*, 2013; Nygaard

and Saxild, 2005; Ye *et al.*, 2009). We therefore observe a co-localization of purine biosynthesis targets in the cell cycle period when they are indeed observed as active. Gram positive bacteria, such as *B. subtilis*, have been demonstrated to have a strong strand-specific purine asymmetry, skewed positively to the leading strand and related to the mechanism of DNA replication (Hu *et al.*, 2007). The work by Nouri et al. (Nouri *et al.*, 2018) showed that carbon metabolism in *B. subtilis* affects DNA replication rates. This may relate to our observation as purine biosynthesis requires the fusion of a pyrimidine ring with an imidazole ring and therefore has a higher carbon demand. We propose that there may exist a regulatory link between these phenomena, owing to the differences in strand replication progression that is mastered by the metabolism of purine and pyrimidines. The observed co-localization signal is facilitated via 1D as targets share an operon that appears to be spatially invaded by confounding genomic elements when $T \neq 5$. Our analysis of the temporal dynamics of several TFs (further details in Supplementary 7) provides compelling evidence for the transcription factory model where genes can dynamically co-localize in or out of sites of transcription (Rieder *et al.*, 2012).

3.3.4 Schizosaccharomyces pombe

In Figure 14 we present the sNMDS embedding of Hi-C measurements in *S. pombe* (Mizuguchi *et al.*, 2014), displaying a six-pronged claw shape. The authors of (Tanizawa *et al.*, 2017) predicted a similar mitotic configuration in their proposed model.

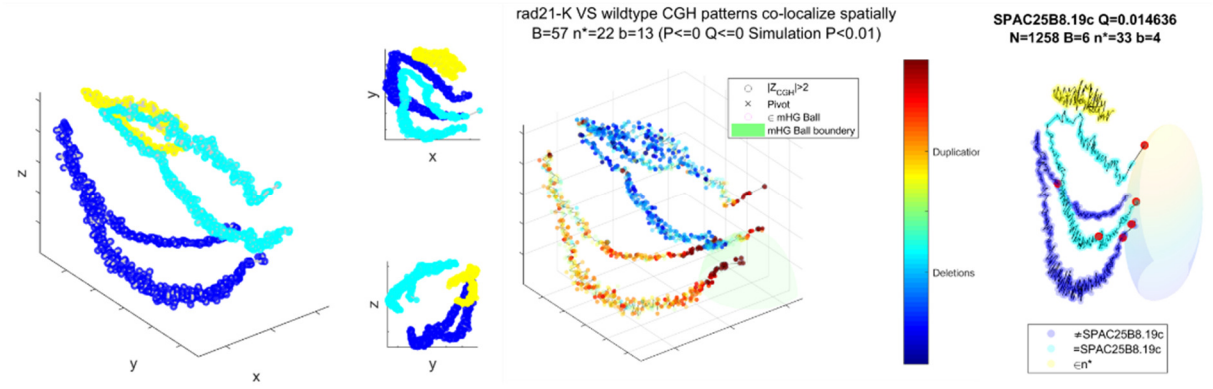


Figure 14. S. pombe results. Left: sNMDS embedding for S. pombe with colour coded chromosomes. Middle (animation available as Supplementary Video 5): Bins are colour coded by average aCGH value, with marked outliers (opaque red for Z>2 and blue for Z<-2). We can observe a weak duplication signal on ChrII, and deletion on ChrI, ChrIII. Strongest duplication is evident at the telomeres. Right (animation available as Supplementary Video 6): Red bins contain Loz1 transcription factor targets. The resulting $smHG$ pivot and corresponding ball are visible containing 4/6 TF targets.

Chromosomal rearrangement of rad21-K1 mutant (compared to Wild Type, based on aCGH data) are spatially co-localized near the telomeres ($smHG^{Grid}$: $[P < 1e^{-300};\ Q < 1e^{-300};\ P_{sim} <$

$0.01], Bead\ Control: [P < 1e^{-300};\ Q < 1e^{-300}], 1D\ Control: [P < 1e^{-8};\ Q < 1e^{-5}]$, Figure 14, middle). rad21-K1 is a mutant selected for partial loss of function in a Cohesin subunit (Tatebayashi *et al.*, 1998). Cohesin is a protein complex implicated in being involved in the determination of chromatin architecture and mitotic domain organization (Mizuguchi *et al.*, 2014; Tanizawa *et al.*, 2017; Sofueva *et al.*, 2013; Lazar-Stefanita *et al.*, 2017). Active chromosomal rearrangement near telomeres have been previously reported using Cohesin mutants in mice and molecular evolution studies in primates (Adelfalk *et al.*, 2009; Trask *et al.*, 2005). In a related observation we see that the transcription factor Loz1 has its targets spatially confined near the telomeres ($smHG^{Grid}: [P < 1.4e^{-5};\ Q < 0.02;\ P_{sim} < 0.02], smHG^{Pivot}: [P < 1e^{-3}; Q < 0.1], mHG^{1D}: [P < 1e^{-2};\ Q < 0.4]$, Figure 14, Right). Two of its targets are SPBC1348.06c and SPAC977.05c, both known to be involved in telomeric duplication. Together, our results indicate a strong relation between a functional Cohesin complex and peri-telomeric integrity, which may be facilitated by DNA repair mechanisms operating during meiotic recombination.

To further inspect the structural conformation changes in rad21-k1, we performed a differential Hi-C analysis (details provided in Methods). Our results show that the major changes in structure are localized and manifested primarily at the middle of each chromosome arm ($smHG^{Grid}: [P < 1e^{-12};\ Q < 1e^{-7};\ P_{sim} < 0.01], Bead\ Control: [P < 1e^{-6}; Q < 1e^{-3}], 1D\ Control: [P < 1e^{-7};\ Q < 1e^{-5}]$, Figure 15). The authors of (Tanizawa *et al.*, 2010) present qualitatively similar interphase models.

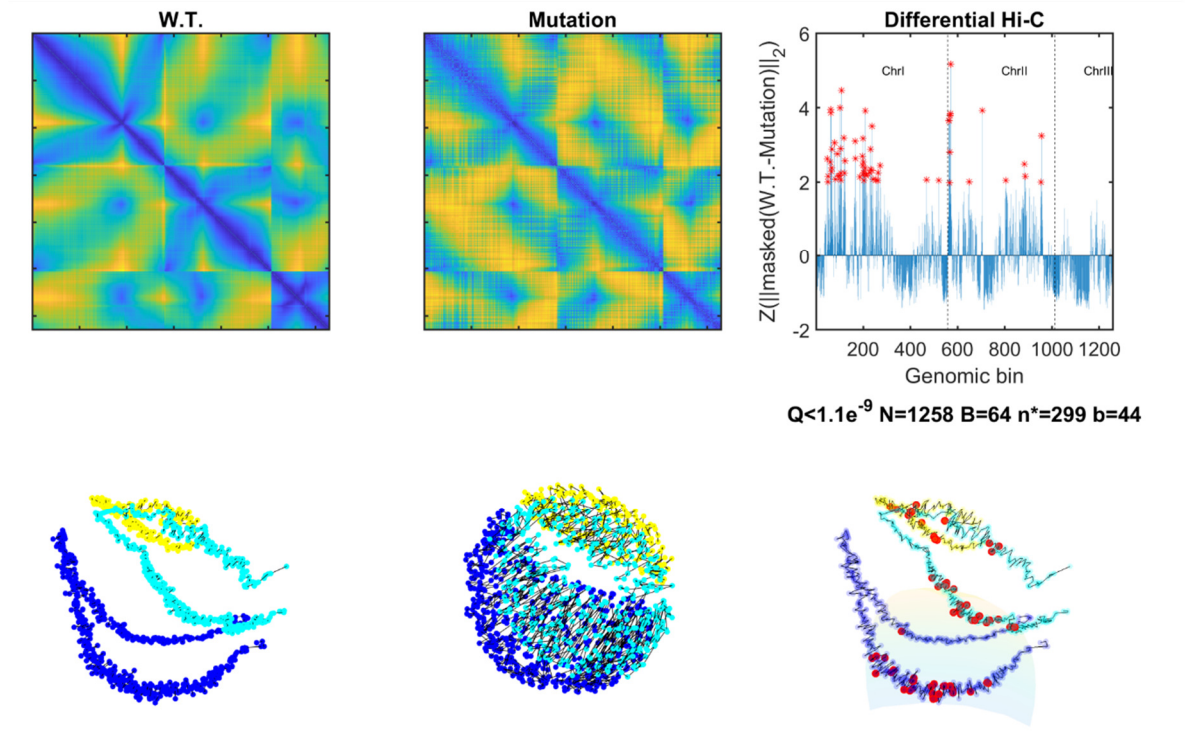


Figure 15. S. pombe mutant structural modifications (animation available as Supplementary Video 7). Left: Top – raw Hi-C read matrix for wildtype. Bottom – resulting sNMDS embedding. Middle: Top – Hi-C data for rad21-k1 mutant. Bottom – resulting sNMDS embedding. Right: Top – ΔZ-scores between both (masked) Hi-C datasets. Red asterix mark loci of Z>1.96 change. Bottom – wildtype sNMDS embedding. Red bins indicate bins that substantially changed in their local structure according to our differential Hi-C analysis (detailed in Methods).

3.3.5 Saccharomyces cerevisiae

In Figure 16 we present the sNMDS embedding of Hi-C measurements in *S. cerevisiae* (Duan *et al.*, 2010), displaying a Rabl (Taddei *et al.*, 2010), Water-lily conformation. This result is qualitatively consistent with previously published models (Ben-Elazar *et al.*, 2013b; Lazar-Stefanita *et al.*, 2017; Capurso *et al.*, 2016).

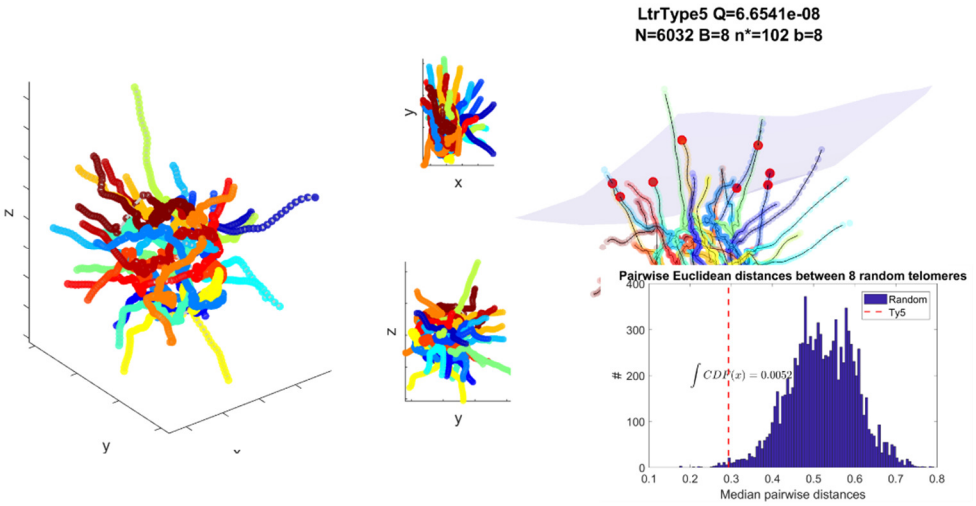


Figure 16. S. cerevisiae results. Left: sNMDS embedding for S. cerevisiae with 16 color-coded chromosomes Right (animation available as Supplementary Video 8): Opaque red colored bins contain Ty5 family LTRs. Inset shows the distribution of mean pairwise Euclidean distances for $\binom{32}{8}$ *telomeres. Red dashed vertical line indicates mean pairwise Euclidean distances for the 8 Ty5 bins. An empirically determined cumulative distribution function evaluated at this point yields* $p < 0.007$.

S. cerevisiae long terminal repeats (LTRs) have been categorized to five distinct families, each with different properties (Kim *et al.*, 1998; Mita and Boeke, 2016). We observe a previously known preference of family Ty5 to associate to peri-telomeric regions ($smHG^{Sample}$: $[P < 1e^{-13};\ Q < 1e^{-7};\ P_{sim} < 0.01]$, $Bead\ Control$: $[P < 1e^{-7}; Q < 1e^{-3}]$, $1D\ Control$: $[P < 1e^{-3};\ Q < 0.04]$, Figure 16). While this association was already known, we offer a refinement in such that the 8 annotated Ty5 LTR elements tend to co-localize at a specific hemisphere of the nucleus, on chromosomes III (3 instances), V (2 instances), VII, VIII and XI. We present the likelihood of such an event to be random in Figure 16, Right inset. We shuffle (10,000 times) the assignment of Ty5 elements to different telomeres and compute the

median of their pairwise Euclidean distances. The resulting empirical CDF at the unpermuted (observed) point yields $p < 0.007$. We propose that this co-localization phenomenon occurs due to the mechanism by which retrotransposons propagate. The probability of a transposing element to integrate in a potential target site is inversely proportional to the distance it needs to travel from its source.

3.3.6 Neurospora crassa

In Figure 17 we present the sNMDS embedding of Hi-C measurements in *N. crassa* (Klocko *et al.*, 2016), displaying a balloon-like shape.

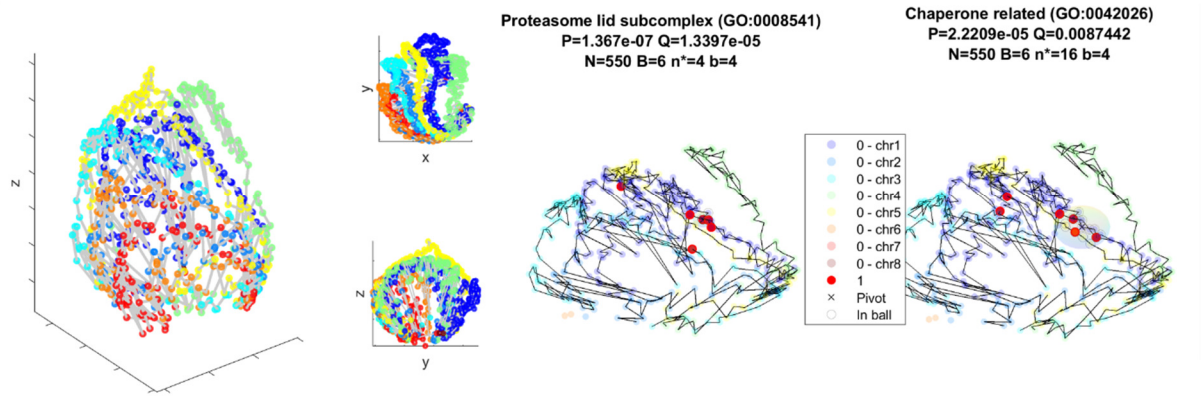


Figure 17. N. crassa results. Left (animation available as Supplementary Video 9): sNMDS embedding of N. crassa. Middle & Right (animations available as Supplementary Video 11 and Supplementary Video 12): Only subset of bins containing mappable genes with GO terms are shown. Red coloured bins contain genes with GO (gene ontology) annotation GO:0008541 and GO:0042026, "Proteasome lid subcomplex" and "Protein refolding" (Chaperone related), accordingly. A black 'x' and translucent sphere depict the resulting $smHG$ position and radius (recovered by mapping mHG threshold back to distance from 'x') for each figure.

Protein folding genes and Proteasome lid subcomplex genes are poised to collaborate by genomic co-localization. In our analysis we observe both gene ontology terms (8541, 42026) to individually co-localize spatially ($smHG^{Grid}$: $[P < 1e^{-9};\ Q < 1e^{-3};\ P_{sim} < 0.01]$, $Bead\ Control$: $[P < 1e^{-6}; Q < 1e^{-3}]$, $1D\ Control$: $[P < 1e^{-6};\ Q < 1e^{-4}]$ and $smHG^{Grid}$: $[P < 1e^{-5};\ Q < 0.02;\ P_{sim} < 0.01]$, $Bead\ Control$: $[P < 1e^{-4}; Q < 1e^{-2}]$, $1D\ Control$: $[P < 1e^{-3};\ Q < 0.02]$ accordingly, Figure 17, Right). Upon inspecting the resulting pivot locations and the sizes of enrichment balls they appear similar to one another. To further validate this result, we compute $smHG$ *on* the union of both GO term targets resulting in $B_{\cup} = 10$, indicating 2 bins overlap. we run $smHG$ on the union without providing an exact statistical model to treat these overlaps, providing an upper bound on the *p*-value ($smHG^{Grid}$: $[P < 1e^{-8};\ Q < 1e^{-4};\ P_{sim} < 0.01]$, $Bead\ Control$: $[P < 1e^{-7}; Q < 1e^{-4}]$, $1D\ Control$: $[P < 1e^{-6};\ Q < 1e^{-4}]$). Additionally, we fixed the 6 target bins of GO: 0042026 and randomly picked 6 targets, computing the mean pairwise distances between both sets of points. The tail of the empirical distribution yielded $CDF < 1e^{-300}$ when

evaluated at the pairwise distances between GO: 0042026 targets and GO: 0008541. These validations further illustrate that these are independent genomic sites with overlapping spatial co-localizations. In summary, we observe a significant co-localization of Proteasome genes as well as of Chaperone genes and furthermore, these two putative transcription factories are spatially close to each other. It has been previously observed that both machineries are intertwined, where chaperones mark for degradation by ubiquitination, physically deliver and interact directly or via coefficients with the proteasome machinery (Imai *et al.*; Carlisle *et al.*, 2017). Our observation suggests that both mechanisms are tightly coupled on the genomic level thereby offering an increased linkage and co-regulation.

3.4 Discussion

In this work we have developed and implemented methods for assessing the statistical significance of spatial co-localization in binary data specified for 3D co-ordinates which overcomes the limitation of being constrained to 'Bead' pivots. Our code is available to the community. We have applied our methods to analyse several Hi-C datasets from unicellular genomes and report statistically significant results detailed above.

Our analyses are performed on previously published "population Hi-C" datasets. That is, Hi-C read counts correspond to evidence of proximity events sampled from millions of independent genomes of distinct biological cells. In this work, as well as in some other Hi-C literature, results are based on analysing such population data. The underlying biology may therefore be obscured by the non-homogeneous character of the data. To mitigate the underlying variability, we focus on analysing datasets of monoclonal single-celled organisms under shared environmental conditions. Furthermore, the bacteria datasets, *C. crescentus* and *B. subtilis* were collected from colonies synchronized to the same cell cycle stages. We therefore expect reduced effects coming from genetic, functional and environmental non-homogeneities. Nonetheless, other factors that contribute to variability remain, and enrichment results should only be interpreted as statistical observations derived from 3D configurations based on sampled population measurements. Applying our methodology on more complex organisms, such as Humans, will require several adjustments: First, methods that sample homogeneous cell populations, or single-cell methods. Next, correctly embedding a polyploid genome. Third, adjustments to the statistical model of mHG to better reflect the availability of gene copies in a gene set. Finally, mitigating the complexity issues discussed above at larger genome scales by developing more advanced heuristics.

Furthermore, we base our analysis on 3D configurations derived from population data as above. sNMDS embeddings probably do not represent the genome structure of any individual biological cell or population member. The spatial manifold in which elements are embedded cannot necessarily be directly interpreted as physical 3D space. Instead, it serves as an abstract 'latent' space, primarily useful for mapping Hi-C data to the geometry required for our statistical 3D enrichment methods, while smoothing out the noisy

character of Hi-C read counts. The approach here could be re-interpreted not as identifying "colocalization" of sets of genomic elements from a spatial model of a genome, but simply testing for statistical enrichment at the level of bulk contact frequency, which hints at some cases of colocalization. We view the fact that resulting embeddings visually correlate with our expectations of polymer behavior without being strictly enforced in the embedding process along with the observed statistically significant $smHG$ results as added qualitative evidence of a population-driven structural signal of genome organization. A quantitative quality control analysis of the embedding process, reinforcing the selection of embedding algorithm and parameters, is displayed in Supplementary 11.

The algorithmic approach we take here is heuristic since the exact calculation of the best $smHG$ pivots in the data corrected for multiple testing is complex. It is clearly a low polynomial search problem as indicated by the combinatorics of the bisector tessellation (see Methods), but still, for thousands of points (as in small genomes), this becomes an unacceptably long calculation. One may consider the use of a Voronoi tessellation. The latter has a far lower computational complexity. However, points in the same Voronoi cell can induce dramatically different rankings on the '0's and '1's, as we illustrate in Supplementary 12. Furthermore – the added complexity of correctly computing a statistically valid result by many repeats to correct for multiple testing, requires even greater time efficiency. We do analyze performance properties of our proposed heuristics, illustrating pros and cons of each.

Further investigation into heuristics may yield improved runtime performance for spatial enrichment methodologies. *Data reduction* methods (Ehrenberg, 1982) may prove useful for filtering or replacing objects of interest (such as input points or tessellation cells) by applying clustering and selecting representatives. A specific noteworthy data reduction approach is to replace objects by fitting them with a density function (Parzen, 1962; Davis *et al.*, 2011). A multiscale density-based representation (Xia *et al.*, 2018) could provide an efficient means of sampling candidate pivots from areas of interest. *Discrete non-convex optimization methods* (Floudas, 1995; Jain and Kar, 2017) such as applying local descent (Snyman and Wilke, 2018) on the mHG p-value of neighboring cells, may offer a mechanism to traverse between cells towards local minima, thereby enabling faster candidate elimination.

A simplistic approach to statistically assessing co-localization for a given set of genomic loci, S, would be to compare the average Hi-C read counts within S to averages obtained over a big number of randomly drawn samples of genomic loci with the same size, $|S|$. In Supplementary Figure S19 we show an analysis comparing this approach with $smHG$ on *B. subtilis* Hi-C data for targets of TF BSU29740 (ccpA), a LacI family transcriptional regulator. Our results in this analysis demonstrate the advantage of using $smHG$ compared to a sampling-based approach which would not report this significant co-localization event. In general, from an algorithmic perspective, applying the sampling approach in a systematic way to find within a moderately enriched functional set (such as a TF cohort) the subsets that are more

significantly enriched, is intractable. Specifically, for a TF cohort S, this is equivalent to enumerating all $2^{|S|}$ subsets.

We applied our statistical methods to several organisms across phyla. To summarize our observations: When analyzing data from TF cohorts we find some of them to be spatially enriched, with evidence that functionally related cohorts can share a common transcription factory. We observe changes in co-localization patterns along cell cycle using time course data, providing evidence for transcription factory dynamics. We further show co-localized retrotransposon telomeric preference, potentially shedding new light on its mechanism of propagation. We observe an axial partitioning of replication machinery genes reinforcing evidence of a deep connection between genome replication and genome organisation.

Overall, we provide distinct lines of evidence for the role of spatial organization in unicellular organisms, illustrating $smHG$'s applicability to studying both cis and trans functional-structural relationships in genomes. Finally, our results and interpretation can benefit from follow-up studies and need to be experimentally validated.

ACKNOWLEDGEMENT

We would like to thank Prof. Dan Halperin for invaluable discussions on line-arrangements and efficient implementations in CGAL. We thank Dr Roi Avraham and Dr Noa Ben-Moshe for useful comments. Thanks to the Yakhini research group for important insights and comments throughout the research process.

3.5 Chapter Supplementary Materials

DATA AVAILABILITY

Spatial-mHG code is open source and available in the Yakhini Group GitHub repository (https://github.com/YakhiniGroup/SpatialEnrichment) along with animated 3D configurations and figures.

SUPPLEMENTARY DATA

Supplementary Data are available at Nature Communications online:
https://www.nature.com/articles/s41598-019-48798-7#Sec22

Supplementary Figures

1. **Empirical comparison of $smHG^{grid}$ and $smHG^{sample}$ on synthetic data:** To allow some degree of control on the optimal enrichment in a synthetically generated instance we provide the following protocol. Pick $\{x_i\}^N, c$ from a multivariate uniform, $\mathcal{U}(0,1)$, and desired minimal enrichment *p*-value, p. Assume x_i is ranked by Euclidean distance to c, i.e. $\|x_i - c\|_2 \leq \|x_{i+1} - c\|_2$. Enumerate all entries in the HGT table (Supplementary Figure S9) that are $\leq p$. Weight each entry with the number of possible non-decreasing paths that cross it (used in mHG multiple hypothesis correction, see (Eden *et al.*, 2007, 2009a) for details) and apply importance sampling to select an entry proportionally to its weight. The entry corresponds to the underlying b, n^* parameters. Generate λ_c by creating a shuffled prefix vector with b '1's and $n^* - b$ '0's, and ap pending a shuffled suffix vector with $B - b$ and $N - n^*$ '1's and '0's accordingly.
   We emphasize that this process only guarantees an *upper bound* on the OPT $smHG$ in this generated instance.
   We run the process described above 10 times each for $N \in \{20,40,60,80,100\}$ with desired $p < 1e^{-3}$, and bound the runtime duration at 2 minutes. During the evaluation we record the best observed p after $iter$ pivot evaluations. The results, presented in Supplementary Figure S7 show an advantage for $smHG^{grid}$ over $smHG^{sample}$ in both convergence time and magnitude (most evident for larger instances) of detected enrichment.

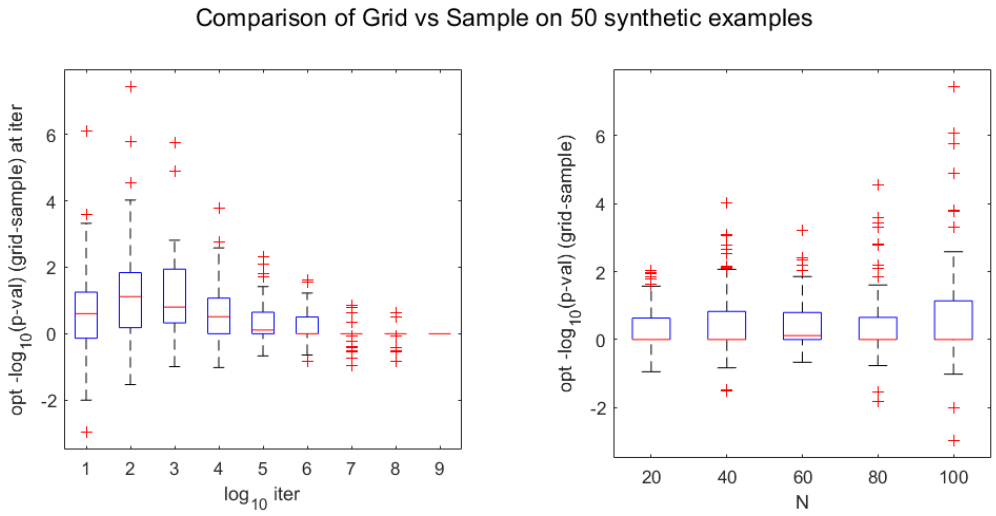


***Figure S7**. Synthetic data comparison of $smHG^{grid}$ and $smHG^{sample}$.showing an advantage for $smHG^{grid}$ over in both convergence time and magnitude (most evident for larger instances) of detected enrichment.*

2. **Additional notes on $smHG^{sample}$ and $smHG^{grid}$ differences:** During our work we have evaluated different approaches by simulating datasets with different tractable parameters where we could compare convergence times to optimal results. Evidently, while $smHG^{sample}$ provides a guarantee to exhaustively cover the exact number of cells, it appears to suffer from one major drawback by its hyper-sensitivity to the distribution of the data. In our simulations we sampled $x_i \propto \mathcal{U}(0,1)$, i.e. from a multivariate (2D or 3D) uniform distribution. Given x_i, x_j the midpoint of their connecting line segment (where the bisector lays) is an average of two uniform random variables, which, thus, in itself, is distributed under a special case of the (normalized) Irwin-Hall triangular distribution (n=2). This implies that uniformly generated data would have a substantially high concentration of bisectors at the center of the ambient space (0.5,0.5 for 2D and 0.5,0.5,0.5 in 3D). In turn, bisectors intersecting each other would yield a significant concentration of cells around that region. Since $smHG^{sample}$ picks cells uniformly, it would adopt this skew and over-represent this specific region of space. In a time-limited / truncated evaluation, we would miss evidence of co-localization in the periphery.
$smHG^{grid}$ adopts a multi-resolution approach, forfeiting on theoretical benefits (that have little practical implications on large scale data) in order to inspect the input for possible co-localizations with increased granularity of over time.

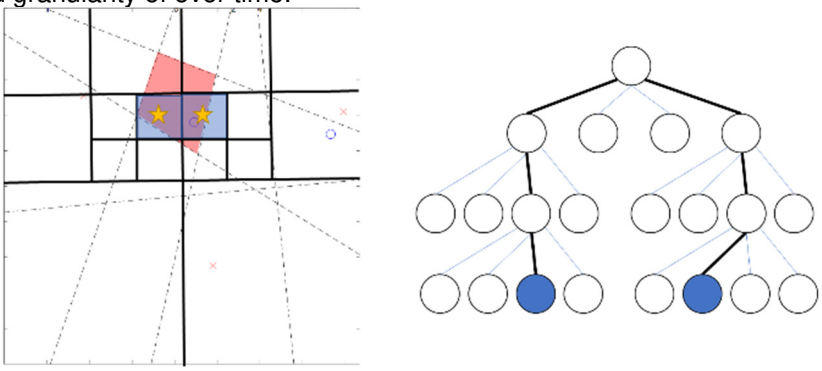


***Figure S8**. Showing how a single cell can be visited more than once by $smHG^{grid}$. different branches of the Octree (2D shows quadtree) yield cubes that intersect it. (Left) 2D instance, bisectors shown as dashed lines. Cell of interest filled with red. Observe that there are bisector intersections that fall outside the axis limits and are not accounted for by this method (Right) corresponding tree graph of the resulting partitioning of running $smHG^{grid}$.*

3. **Spatial-mHG technical notes:** We note that in our experiments we run the Grid and Sample heuristics for a bounded duration of 5 minutes each on a dedicated Azure StandardA8v2 machine.
W.L.O.G. all input data is initially normalized to the unit ball and jittered to guarantee that the bisectors are in general position (with high probability). We also add a virtual sphere containing the normalized inputs, in order to make sure all cells are bounded.
A simplifying observation is that there are no intersections of more than three planes in the same point due to the following argument: Assume by contradiction that four planes intersect in a point. Choose one plane. Each of the three planes intersecting with it forms a line, and since they all intersect in a point. Since every line is a perpendicular bisector for points $x_i \in D$, the point of intersection is a circumcentre of a triangle where the triangle vertices are the bisected points in D. Since we only employ bisectors from pairs that are differently labelled (a '1' and a '0'), this means that every pair of vertices in the triangle is differently labelled (or differently coloured). Since there is obviously no way to 2-colour a clique of size 3 (the triangle vertices), we contradict the original statement.

4. **$smHG^{grid}$ recursion stopping criterion:** λ_p can be illustrated as a non-decreasing path in an $B \times (N-B)$ matrix where each entry corresponds to a Hypergeometric CDF tail score and to a prefix of some possible binary vectors. For this entry, its Manhattan distance from the bottom left corner reflects the "number of draws" its row reflects the number of successes, and its column the

number of failures (and implicitly the population size). The mHG score for a vector is the minimum value in the cells visited by its corresponding non-decreasing path. During an $smHG$ evaluation we track the optimal score observed, p_t^*. p_t^* can be used to estimate the minimal number of cell traversals from p that are necessary for obtaining a score that is better than p_t^*. This number is used as a stopping condition for the octree construction by comparing it to the number of bisectors crossing the cube for which our pivot is close to the '0' coordinate than the '1'. If A visual representation is given in Supplementary Figure S*9*.

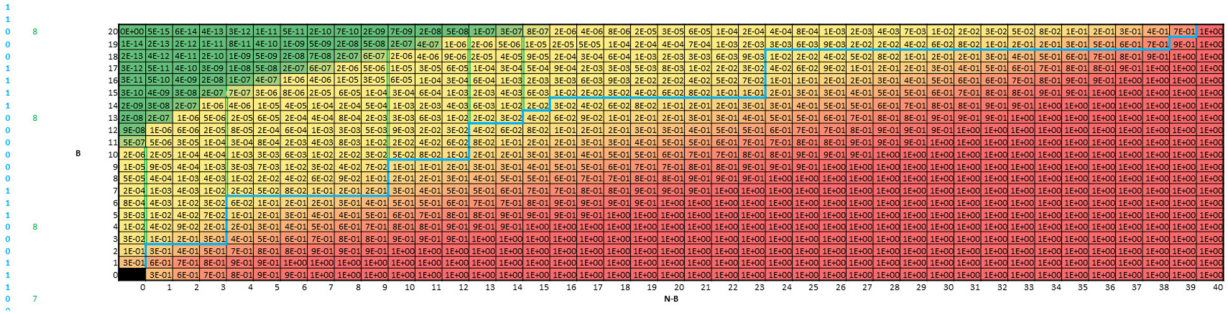
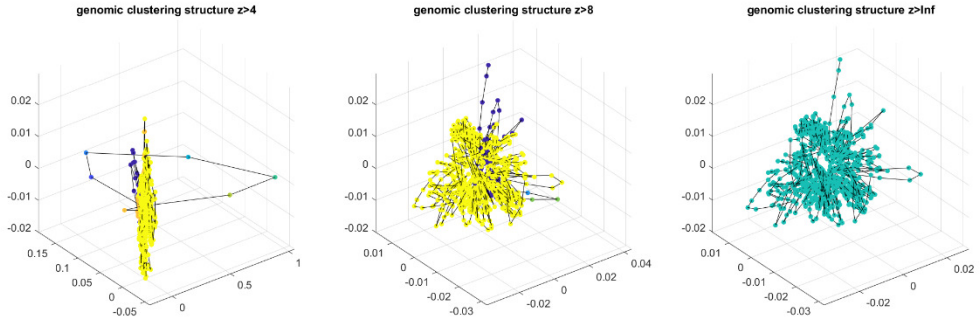


Figure S9*. An example HGT matrix depicting all possible binary vectors of size $N = 60$ with $B = 20$ '1's. Every entry is colored proportionally to the hypergeometric CDF upper tail p-value. Blue path corresponds to some binary vector, λ, the prefix of which is displayed on the left. Vertical green lines emanating from this path towards the greed region of the table correspond to "minimal distance to $p_t^* = 1e^{-6}$" at different thresholds in λ. The overall minimal distance in this case is* 5.

5. **sNMDS outlier correction scheme:** we present one of the NMDS resulting embeddings for B. subtilis time-course (at the 5-minute mark). We applied 2 iterations of smoothing with Z>4, Z>8, top row, visualized from left to right. Clusters identified in each iteration are and colour coded. We see that after each iteration the resulting genomic structure appears smoother and more coherent, unravelling more elaborate detail. We manually tune these hyperparameters to avoid having long stretches of the genome collapse to a line (example of a bad choice of parameters is presented in the bottom line).



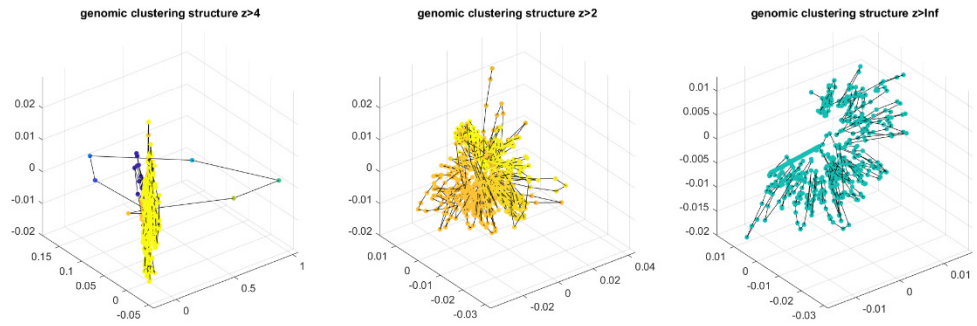


Figure S10. sNMDS outlier correction. (Top) showing two smoothing iterations of sNMDS on B. subtilis Hi-C data with Z>4 and Z>8. Genomic bins are color-coded by clustering them according to the Euclidean distances of consecutive bins. (Bottom) Same example with different parameters (Z>4, Z>2) showing the formation of an undesirable linear segment artifact. Note that this also impacts axis scaling as part of the manifold flattens.

6. **Principal directions of enrichment localization:** We weigh every resulting $smHG$ pivot across the investigated annotation sets with its corresponding $-\log q$-value. Next, we performed PCA on these weighted pivots, yielding the principal directions to explain the main variation in spatial enrichment in C. crescentus. The results, shown in Supplementary Figure S*11*, left, illustrate a primary axis along the direction of the replication axis which explains 58% of the variance in enrichment directions. To quantify the significance of this observation we ran a simulation analysis, shuffling the *q*-value weights across pivots. The distribution of the resulting PC1 and PC2 explained variance are shown in Supplementary Figure S*11*, right. We fit a multivariant gaussian and compute the density of the CDF at the empirically determined point (58%, 34%) showing that our observation is at the tail of the distribution, around 1% of simulated observations.

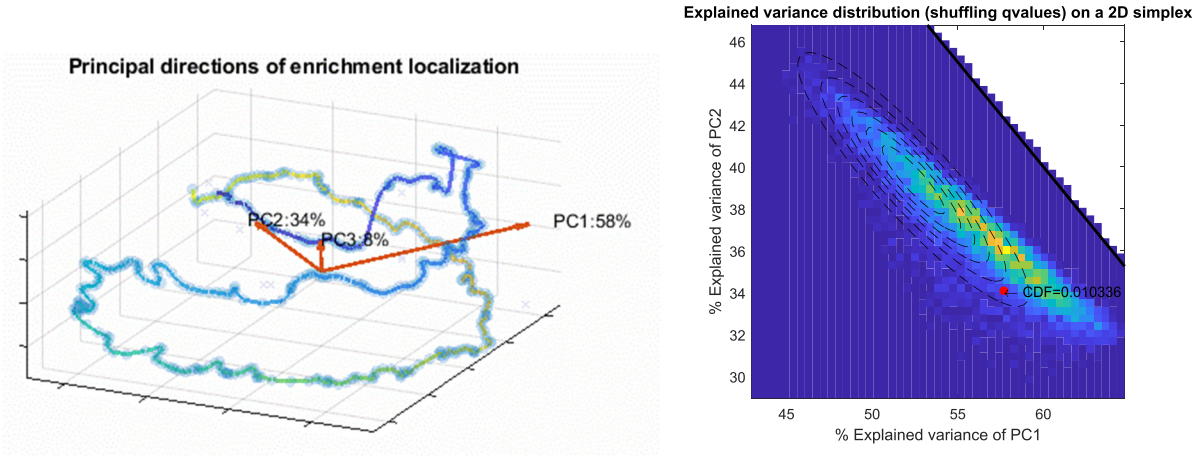


Figure S11. Principal directions of enrichment (Left animation available as Supplementary Video 1s) Principal directions of enrichment as detected by our analysis. (Right) Permutation analysis showing there is a strong bias towards a single dominant axis of enrichment.

7. **Investigating temporal dynamics in time-course Hi-C data:** We plot each transcription factor's $smHG$ -log *q*-value for each of the four available time-course Hi-C datasets. We map each enrichment to the set of genomic bins within the corresponding enrichment ball. Next for each temporally-consecutive pair of sets, we compute the Jaccard similarity to quantify the overlap between their targets. We then manually inspected TFs with temporal dynamics in both Jaccard and *q*-values.

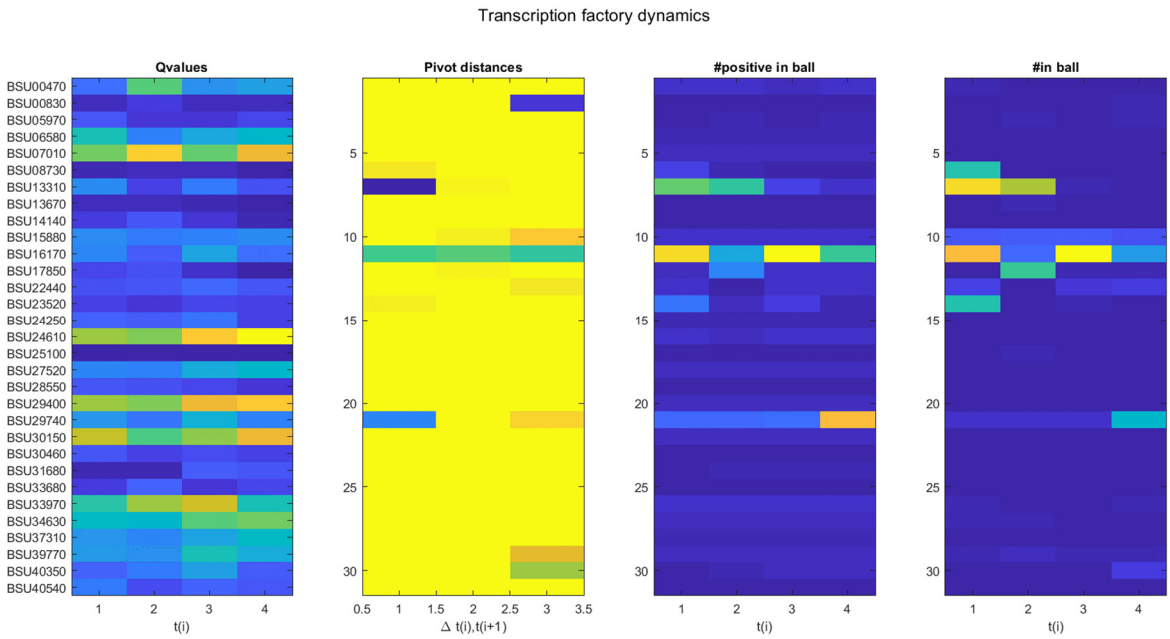


***Figure S12**. Temporal dynamics in B. subtilis TF target smHG results. (Left) Enrichment Q values. (Middle-left) Overlap between bins inside the detected smHG enrichment ball for consecutive Hi-C datasets in the time course. (Middle-right) # of '1's in the enrichment ball. (Right) number of genomic bins in the ball.*

8. **Main results table:**
Supplementary tables available in separate files online. Supplementary Figure S*13* shows a meta-analysis of all *smHG* runs computed in this study.

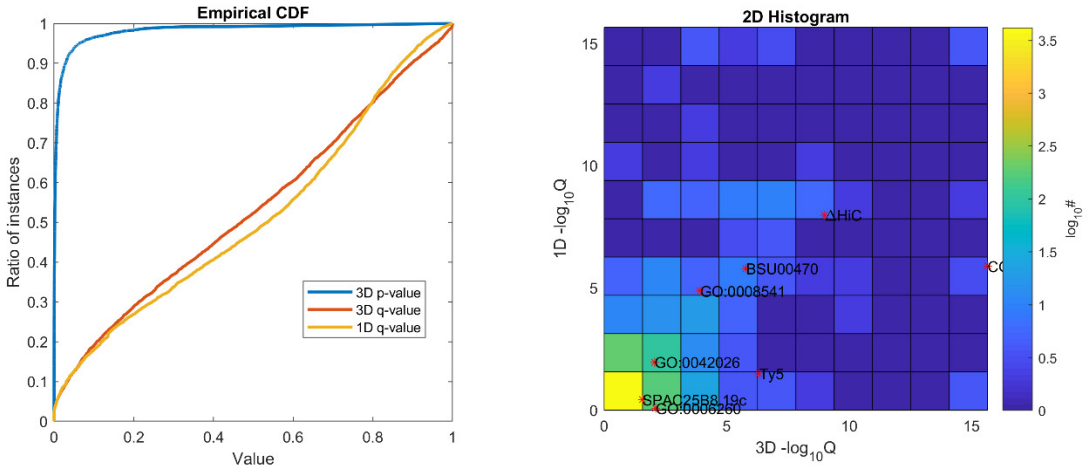


***Figure S13**. Summary of 1D vs 3D Q values on all evaluated smHG instances. (Left) Empirical cumulative distribution plots of of 3D p and q values, and 1D q-value. We see FDR correction yields an empirical distribution that is approximately uniform, as needed, illustrating our sensitivity/specificity as a probabilistic model. (Right) a 2D histogram of 3D vs 1D q values. Results discussed in the main paper are overlaid and marked with a red asterix.*

9. **Negative results:** in this section we detail a few noteworthy efforts that yielded no significant co-localization with the goal of illustrating diverse hypotheses that can be evaluated with our proposed framework.

Differential expression: We evaluated differential expression in two cases, and neither appeared to yield significant co-localization.

- The authors in (Mizuguchi *et al.*, 2014) published a tiling array expression for Chr II in Pombe WT vs Rad21-K1. We average genes per bin, remove bins ≤90th percentile in #mapped probes and with high variance within the bin. We compute the Z-score for differential expression on bins, and binarize with threshold $Z > 1.96$. When limiting our analysis to this subset of bins we observe no spatial co-localization.
- The authors in (Castells-Roca *et al.*, 2011) provide a Heat shock gene expression time course for S. cerevisae. We binarized the relative abundance values by averaging genes in bin and thresholding for $RA \geq 5$, and we do not observe significant spatial co-localization under our model.

Other genomic element annotations:

- Pombe origins of replication (Ori) do not appear to spatially co-localization under our model.

10. **Exact bound on the number of cells induced by the intersection of planes:**

Theorem I: k lines partition the plane, $\mathbb{R}^2$, to at most $\binom{k}{2} + \binom{k}{1} + 1$ distinct 2D cells.

Corollary II: n points in $\mathbb{R}^2$ induce a partitioning of the plane to at most $\binom{\binom{n}{2}}{2} + \binom{\binom{n}{2}}{1} + 1$ distinct 2D cells, when considering all PBHP (perpendicular bisecting lines, in the 2D case) between the points.

Theorem III: k planes in $\mathbb{R}^3$ induce a partitioning to at most $\binom{k}{3} + \binom{k}{2} + \binom{k}{1} + 1$ 3D cells when considering the cells formed by the intersection of all PBHPs between pairs of points.

Corollary IV: n points in $\mathbb{R}^3$, 3D Euclidean space, induce a partitioning to at most $\binom{\binom{n}{2}}{3} + \binom{\binom{n}{2}}{2} + \binom{\binom{n}{2}}{1} + 1$ cells when considering the cells formed by the intersection of all PBHPs (perpendicular bisecting planes, in the 3D case) between the points.

Theorem I Proof. We denote points of intersection between two or more lines as 'vertices'. Note that cells are equally defined by the lines that contain their edges (their boundary set), and by the vertices formed by the intersection of these lines. We define a one to one correspondence between cells and their bottom- most vertex (W.L.O.G. there is always such a vertex, otherwise we can tie-break arbitrarily, e.g. bottom-left vertex first). Assuming at most two lines intersect in any point, every point of intersection of lines serves as the lowest vertex of exactly one cell, thus there are $\binom{k}{2}$ such cells. We now observe that some cells do not have a lowest vertex (they may be non-finite sets). To count them we "hallucinate" a k+1$^{\text{th}}$ horizontal line below any intersection of our original k lines (see Supplementary Figure S14). To count the number of cells formed by the new line we assign each such region to the original line intersected to create it, arbitrarily on the vertex to its left. This process would end after k assignments with one region to spare. Thus, in total we have $\binom{k}{2} + \binom{k}{1} + 1$ cells. Q.E.D. ■

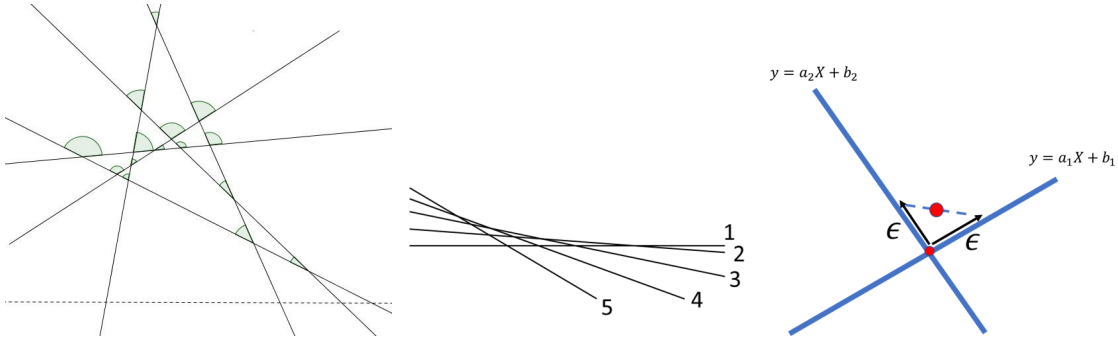


Figure S14**. Number of cells proof* ***– (Left) Number of 2d cells created by k lines. *Bottom vertex of each cell is assigned to it in a one to one correspondence. The angle between the cell and the bottom vertex is colored in green. A horizontal line is added to count non-finite cells on the bottom. Total count of cells is* $\binom{k}{2} + \binom{k}{1} + 1$. ***(Middle) Constructive process to illustrate tightness of result.*** *Adding lines iteratively we can carefully place them in such a way as to ensure that their intersection generates exactly* $\binom{k}{2} + \binom{k}{1} + 1$ *cells.* ***(Right) implementation of*** $smHG^{sample}$ ***illustrated in 2D.*** *Compute plane intersection of 3 planes. In general position these intersect in a point. For each plane, traverse from the intersection a distance of* ϵ *along the gradient of the plane in the direction of* y *dimension. Average the 3 resulting points to yield a pivot inside the cell.*

Theorem III Proof. By induction –

Basis – For k=1 $\binom{k}{3} + \binom{k}{2} + \binom{k}{1} + 1 = 2$ and indeed, a single plane (half-space) divides the space into two halves.

Inductive step – Suppose that k planes have already been added and that the induction hypothesis holds, adding the $k+1$ plane intersects with the first k planes, forming k "new" lines on the $k+1^{\text{th}}$ plane. From Theorem 1 these lines divide the $k+1^{\text{th}}$ plane to $\binom{k}{2} + \binom{k}{1} + 1$ 2D cells. Consequently, each such 2D cell splits a 3D cell in two and

adds this amount to the total cell count. Let R_{k+1} be the recurrence relation defining the maximum number of cells formed by k planes,

$$R_{(k+1)} \leq R_{(k)} + \binom{k}{2} + \binom{k}{1} + 1 \leq \binom{k}{3} + \binom{k}{2} + \binom{k}{1} + 1 + \binom{k}{2} + \binom{k}{1} + 1$$
$$= \left(\binom{k}{3} + \binom{k}{2}\right) + \left(\binom{k}{2} + \binom{k}{1}\right) + \left(\binom{k}{1} + 1\right) + 1 = \binom{k+1}{3} + \binom{k+1}{2} + \binom{k+1}{1} + 1$$

QED. ■

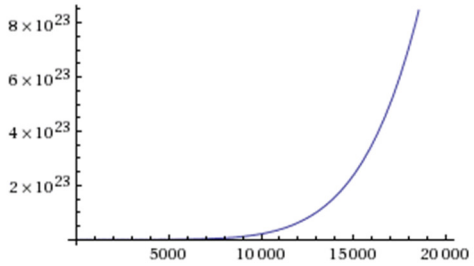


Figure S15. Number of potential different order inducing cells in 3D as a function of number of points. *Showing the intractability of the polynomial problem. X axis denotes the number of points (Genes), and Y-axis denotes the maximum number of cells formed by the bisecting planes between them i.e.* $\left(\binom{n}{2}\right) + \left(\binom{n}{2}\right) + \left(\binom{n}{2}\right) + 1$.

11. ***C. crescentus* NMDS quality controls:** in Supplementary Figure S16 we provide further detail on the quality controls performed during NMDS linear embedding on a sample dataset.

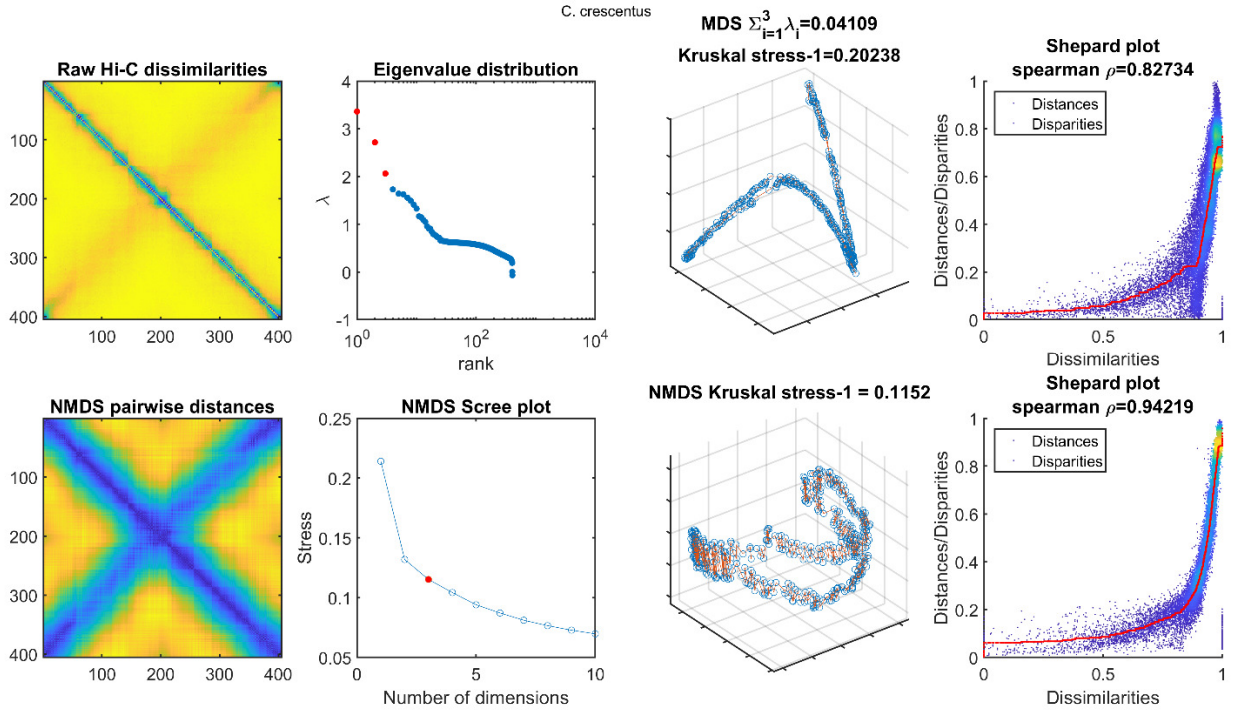


Figure S16. NMDS quality control – Subplots are numbered top-to-bottom, left-to-right. (Top) MDS dimensionality reduction steps: 1) Raw Hi-C input matrix is transformed to represent a dissimilarity matrix. 2) Hi-C matrix is projected to a Euclidean space with the centralized Gram matrix. Eigenvalues of the resulting matrix are shown ranked by magnitude. Top 3 eigenvalues, corresponding to a 3D linear projection into a Euclidean space of Hi-C data, are colored in red and show a 'knee', hinting at an intrinsic manifold dimensionality in this dataset. 3) MDS 3D embedding result represented by the first 3 eigenvectors with corresponding largest eigenvalues. Normalized sum of eigenvalues (measure of captured variance from linear projection) and Kruskal stress-1 criterion (measure of violations of monotonicity between distances and dissimilarities) values are displayed in title. 4) Shepard plot showing correlation between dissimilarities in the Hi-C data, and distances in the resulting MDS 3D embedding. Disparity line indicates deviations from monotonicity, in resulting embedding. Point cloud is overlaid with a density plot. Spearman rank correlation between dissimilarities and distances, ρ, is displayed in title. (Bottom) NMDS dimensionality reduction steps initialized from the MDS solution: 5) pairwise distances in the resulting NMDS embedding (later visualized in subplot 7), a distinct 'cross' pattern emerges that was less visible in the raw Hi-C dissimilarities. 6) A Scree plot showing the impact of selected target dimensionality on the resulting Kruskal stress values shows a 'knee' at 3 dimensions. 7) the resulting NMDS embedding and corresponding Kruskal stress. 8) Shepard plot showing correlation between dissimilarities in the Hi-C data, and distances in the resulting 3D NMDS embedding. We see a clear improvement on the spearman ρ compared to the MDS embedding.

12. **A comparison between bisector and Voronoi tessellations:**

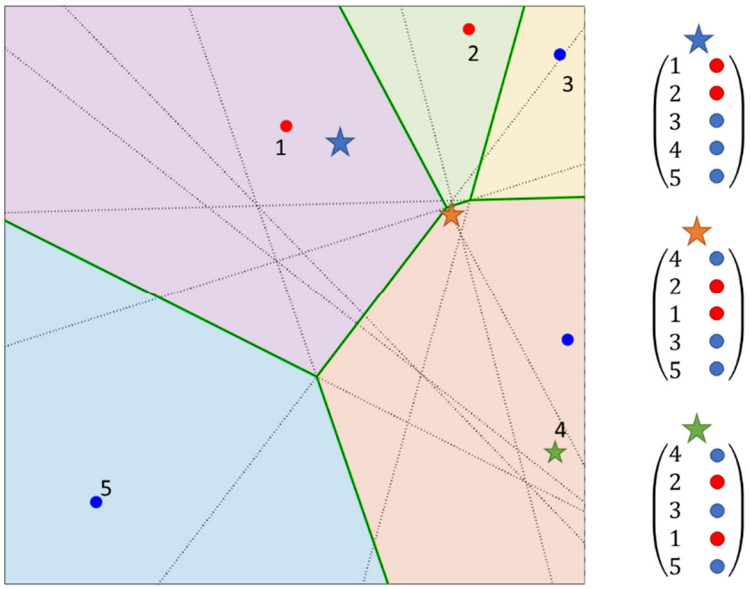


Figure S17*. An illustration of bisector tessellation from five points in 2D. Bisectors are presented as dashed black lines. A Voronoi tessellation is induced by an intersection of a subset of bisectors, highlighted in green. Voronoi cells, represented as differently colored polygons, are cells that induce different rankings on the original points such that neighboring cells have a different point at the top of the ranked vector. An example of three pivots inducing different rankings are shown as blue, green and yellow stars. The induced ranked point ids and corresponding labels are shown on the right. Running* $smHG$ *at the granularity of Voronoi cells would require deciding on a specific pivot for each cell. This example shows that selection of different pivots within a Voronoi cell can have dramatic impact on the ranking and corresponding mHG enrichment results.*

13. **Pairwise empirical comparisons among methods:** This figure provides a breakdown of the main results (Supplementary 8) comparing these across evaluated methods. This result empirically confirms that the bead approach is insufficient for detecting some co-localization events that are detected by the grid method.

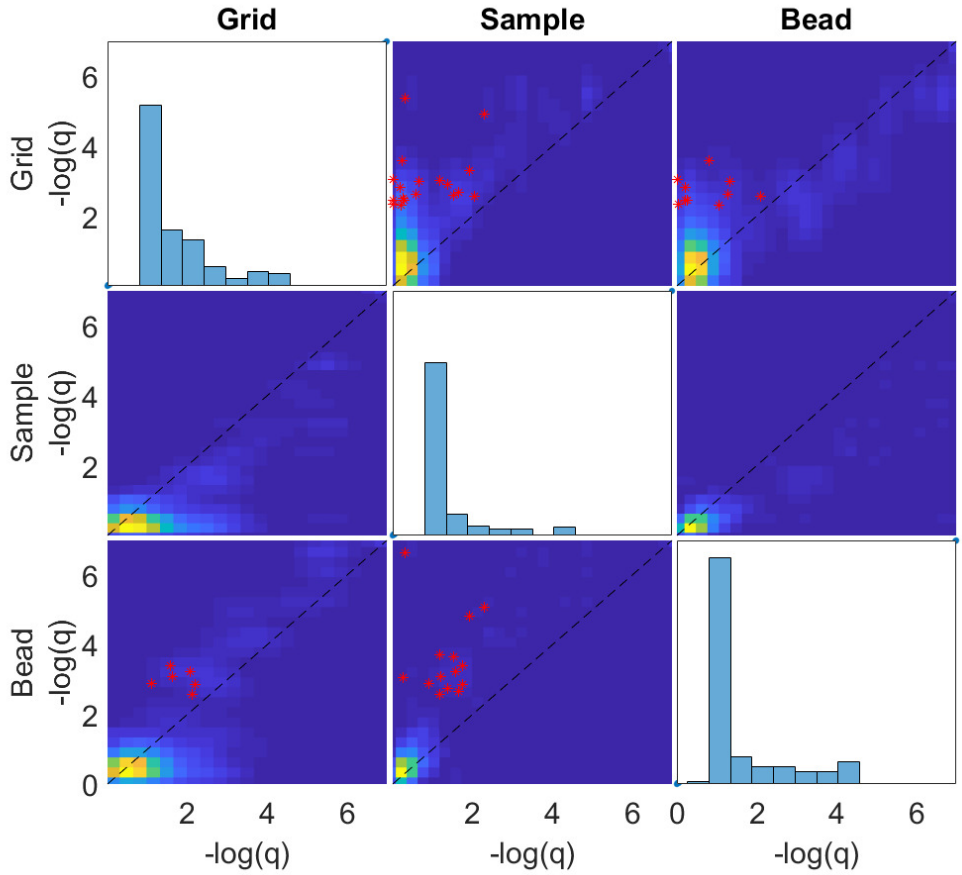


***Figure S18**. A density plot of every experiment's resulting Q values in a pair of methods (indicated by the row and column labels). Red asterisks represent a single evaluation where the method labeled by the row detected (Q<0.1) a potential discovery that the method indicated by column did not. Importantly, we observe potential discoveries that would remain undetected by the bead-based method and vice-versa, showcasing that these methods complement one another.*

14. Comparison between smHG and permutation test on raw Hi-C read counts.

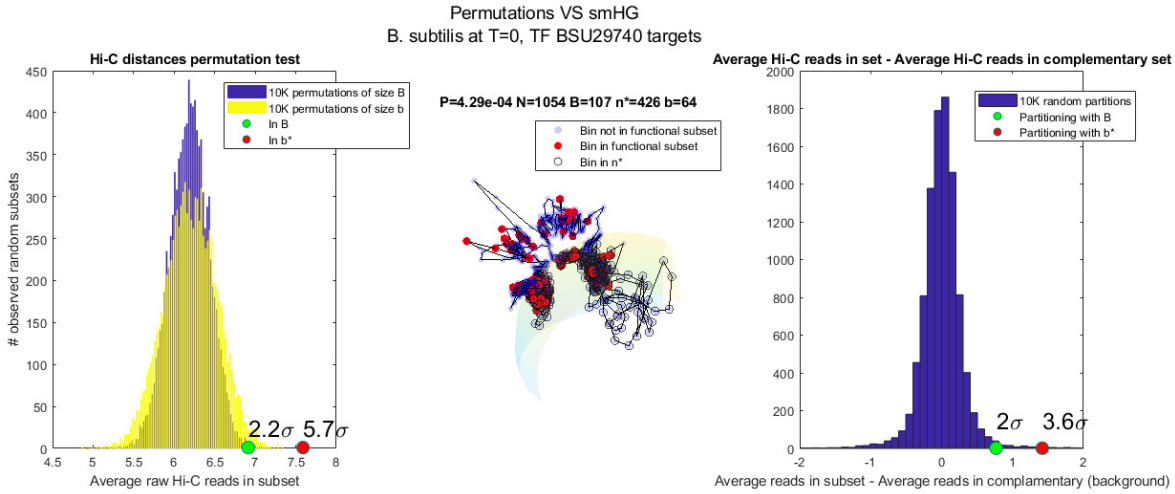


Figure S19. Comparing permutation test with smHG result in B. subtilis (t=0 in timecourse) for a functional group of TF targets (BSU29740): Left) distributions of mean pairwise distances between groups of different sizes are shown in blue and yellow histograms. Correspondingly, the mean pairwise distances between bins in the aforementioned functional group (of size B) is in green. While these are relatively co-localized and are within the 999th-quantile, smHG was able to uncover a substantially more co-localized subset (of size b). While both results appear significant, the green result would not be reported when correcting for hundreds of multiple hypotheses. Middle) a 3D embedding of the Hi-C dataset, bins in B are labled in red, bins in b are the ones that also have a dark circle around them, and are within the translucent sphere to the right. We see that this subset is substantially more clustered. Right) We plot a distribution of 10K random partitions of the genome into two, complementary sets, and compute the difference between the average number of reads in both sets.

Chapter 4:

miRNA normalization enables joint analysis of several datasets to increase sensitivity and to reveal novel miRNA differential expression in breast cancer

4.1 Introduction

microRNAs (miRNAs) are endogenous, small non-coding RNAs (~22 nucleotides) that bind to target-specific sites most often found in the 3'-untranslated regions (UTRs) of target messenger RNAs (mRNAs). By this binding, miRNAs regulate gene expression by conferring inhibition of mRNA translation or mRNA degradation (Bartel, 2009). miRNA expression profiling is an important tool for studying tumor biology and classification and has shown to be important with respect to diagnostic and prognostic assessments. Increasing technological and economic viability of expression sampling methods has enabled the systematic study of miRNA expression in cohorts of hundreds of patients (Aure *et al.*, 2017; Cancer Genome Atlas Network, 2012; Dvinge *et al.*, 2013). On the other hand, inherent measurement noise coupled with complex causes of biological variability affect the statistical confidence in ascertaining consistent differences of low magnitude between populations with small sample sizes. Absolute expression differences are not necessarily linearly correlated with downstream effects of the expressed miRNA, therefore subtle but consistent differences may be of biological importance.

Abnormal miRNA expression in breast cancer has been repeatedly associated with cancer proteins (Aure *et al.*, 2015), molecular subtypes (Enerly *et al.*, 2011), progression (Lesurf *et al.*, 2016; Haakensen *et al.*, 2016; Tahiri *et al.*, 2014) and prognosis (Aure *et al.*, 2017). For example, in one of the first genome-wide characterizations of miRNA expression in breast cancer we identified 63 miRNAs differentially expressed between the two main clinically diverse groups of breast cancer, the estrogen receptor (ER) positive and the ER negative tumors (Enerly *et al.*, 2011).

Combining experimentally measured data from multiple sources is both a challenging and a worthwhile endeavor. Statistical estimation theory formulates a relation between sample size and variance of estimate via the Fisher information that follows the chain rule for independent samples. The ability of statistical hypothesis tests to detect subtle, yet consistent and possibly genuine, differences between populations is directly related to sample size and is quantified as a test's power (Wang and Xu, 2019; Hong and Park, 2012). Increasingly larger power and statistical significance is hindered by sampling costs that can prohibit large sample sizes. This, in turn, leads to the incremental funding of repeated studies aiming to measure the same phenomenon. Follow-up studies tend to vary from their former with newer or alternative experimental protocols, reagents and technologies used for conducting the measurements, introducing batch differences between samples. Such a 'batching' design, inadvertently, introduces distinctions (batch effects) between samples that correlate with the batch and may overshadow subpopulation differences in their magnitude. Blindly testing for hypotheses on batch-collected dataset without taking such effects into account can lead to spurious and erroneous conclusions and can hide significant effects behind batch differences. In this work we address joint analysis of data batched using different miRNA profiling technologies that have been shown to have systematic differences (Git *et al.*, 2010; Mestdagh *et al.*, 2014).

There are various approaches commonly used in practice to address the analysis of combined data containing batch effects. The authors of earlier work (Nygaard *et al.*, 2016; Sims *et al.*, 2008) show that applying standard, parametric, batch correction approaches may introduce bias from uneven sample sizes of the different groups and data idiosyncrasies. A recent study (Gibbons *et al.*, 2018) applied a non-parametric approach for correcting case-control microbiome studies and have shown it compares favorably with former methods. Their method resembles ours, as we further illustrate below.

In this work we apply a non-parametric, quantile-based, batch normalization approach. We use this method for jointly analyzing miRNA expression data in four breast cancer cohorts to obtain increased statistical confidence and power. We demonstrate that, coupled with appropriate non-parametric statistics, our normalization approach mitigates batch effects. We observe stronger statistical evidence of differential expression between ER positive and ER negative samples in multiple miRNA when compared to individually analyzing the cohorts. Moreover, our approach provides interpretable results and is advantageous to direct interpretation of the data conducive to individual examination of findings, as demonstrated herein. Our differential expression analysis surfaces known cancer-related miRNAs, as well as potential new ones.

4.2 Data and Methods

We used miRNA expression data from three previously published breast cancer datasets along with a newly released, fourth, miRNA dataset. These datasets were acquired from frozen material with different minimal amount of tumor cells, using different technologies and experimental protocols as overviewed in Table 2. In addition, we utilized mRNA expression for supporting evidence of the normalization results using one of the cohorts.

We examine miRNA normalization also in the context of jointly analyzing these measurements. Below we elaborate our considerations in the selections made during the normalization process and our means of providing evidence for validating these results.

Dataset	Manufacturer	Technology	Version	Accession number
DBCG (Myhre *et al.*, 2010) – miRNA	Agilent	Human miRNA Microarray Kit	(V2 G4470B) design id 019118	GSE46934
Oslo2 (Aure *et al.*, 2017) – miRNA	Agilent	Human miRNA Microarray Kit (V2)	v14 Rev.2 design id 029297	GSE81000
Oslo2 (Aure *et al.*, 2017) – mRNA	Agilent	SurePrint G3 Human GE 8x60K Microarray	(Probe Name Version) 028004	GSE80999

Micma (Enerly *et al.*, 2011) – miRNA	Agilent	Human miRNA Microarray Kit	(V2 G4470B) design id 019118	GSE19536
Stavanger – miRNA	Exiqon	miRCURY LNA Array	v.11.0	

Table 2. Technical details of platforms used for expression measurements for the four different cohorts. Datasets are color coded consistently throughout the paper. miRNA expression colors are highlighted compared to mRNA measurements.

4.2.1 Dataset pre-processing and coverage

Each miRNA dataset is read from a single-channel image analysis output file acquired from their corresponding GEO repositories (referenced in Table 2) and preprocessed in R using the Limma (Ritchie *et al.*, 2015) package. We note that while Stavanger (Exiqon) data contains a pooled-reference second channel, this measurement is not utilized in our analysis (further discussed in Supplementary 1). Initially, control probes are removed, and the data is corrected by background intensity normalization . Same-probe replicates are replaced by their median value. Probe ids are mapped to their corresponding miRbase v22 accession using miRBaseConverter (Xu *et al.*, 2018). Missing or deleted accession IDs are discarded. Multiple probes that map to the same miRNAs are again replaced by their median value. Next, we apply arrayQualityMetrics (Kauffmann *et al.*, 2009) (resulting Quality Control reports are available in the Supplementary materials) and filter out samples that are marked as outliers by all three outlier detection criteria (L_1-Distance between arrays, Boxplot, MA plot). We thereby filtered out 6, 30, 12 and 2 outliers from DBCG, Oslo2, Micma and Stavanger, respectively. Next, we apply minimum subtraction to avoid log scaling issues with negative numbers where applicable. The joint dataset table is then compiled by applying a “full outer-join” relational operation on the miRbase accession IDs as key. The resulting miRNA cross-dataset table is visualized in Figure 18 (and available in the corresponding online Supplementary materials).

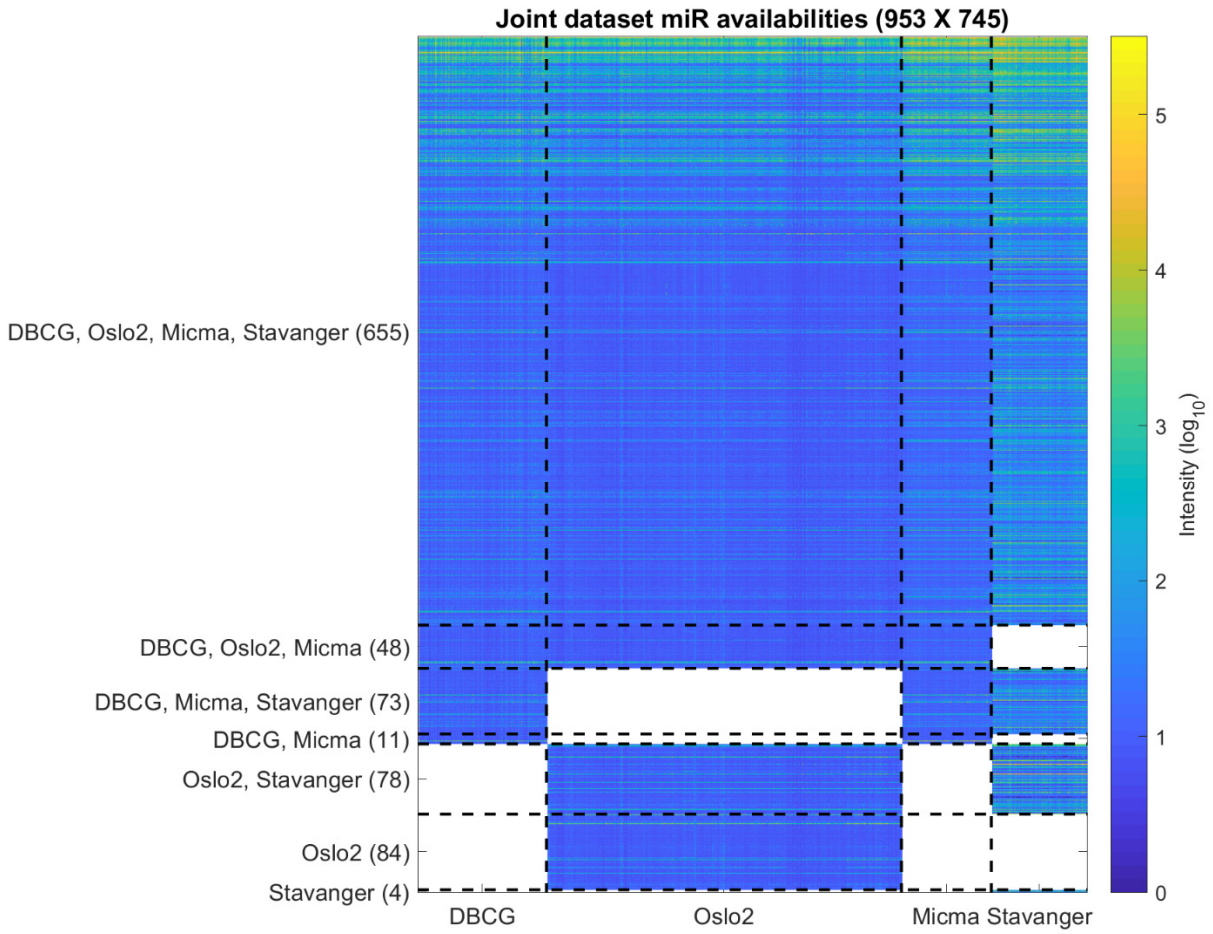


Figure 18. Overview of the miRNA coverage in the dataset. Each row represents one miRNA. Each entry represents the intensity (log_{10}) in a specific sample. Dashed vertical lines separate between samples from the four datasets. Dashed horizontal lines separate between groups of miRNAs by their dataset availability. Blank (white) entries correspond to miRNAs that are missing from a dataset.

4.2.2 Batch effects in joint data

We tested for rank-order consistency of miRNA among pairs of datasets (Figure 19). To do so, we compute for each miRNA the average quantile across all samples belonging in each dataset. We display the resulting value for each pair of datasets in a scatterplot matrix considering the miRNAs (n=655) present in all four cohorts. This analysis shows that Stavanger appears to behave differently than other datasets, presumably due to its fundamentally different measurement technology (Exiqon LNA (Bartel, 2009) vs Agilent Microarray).

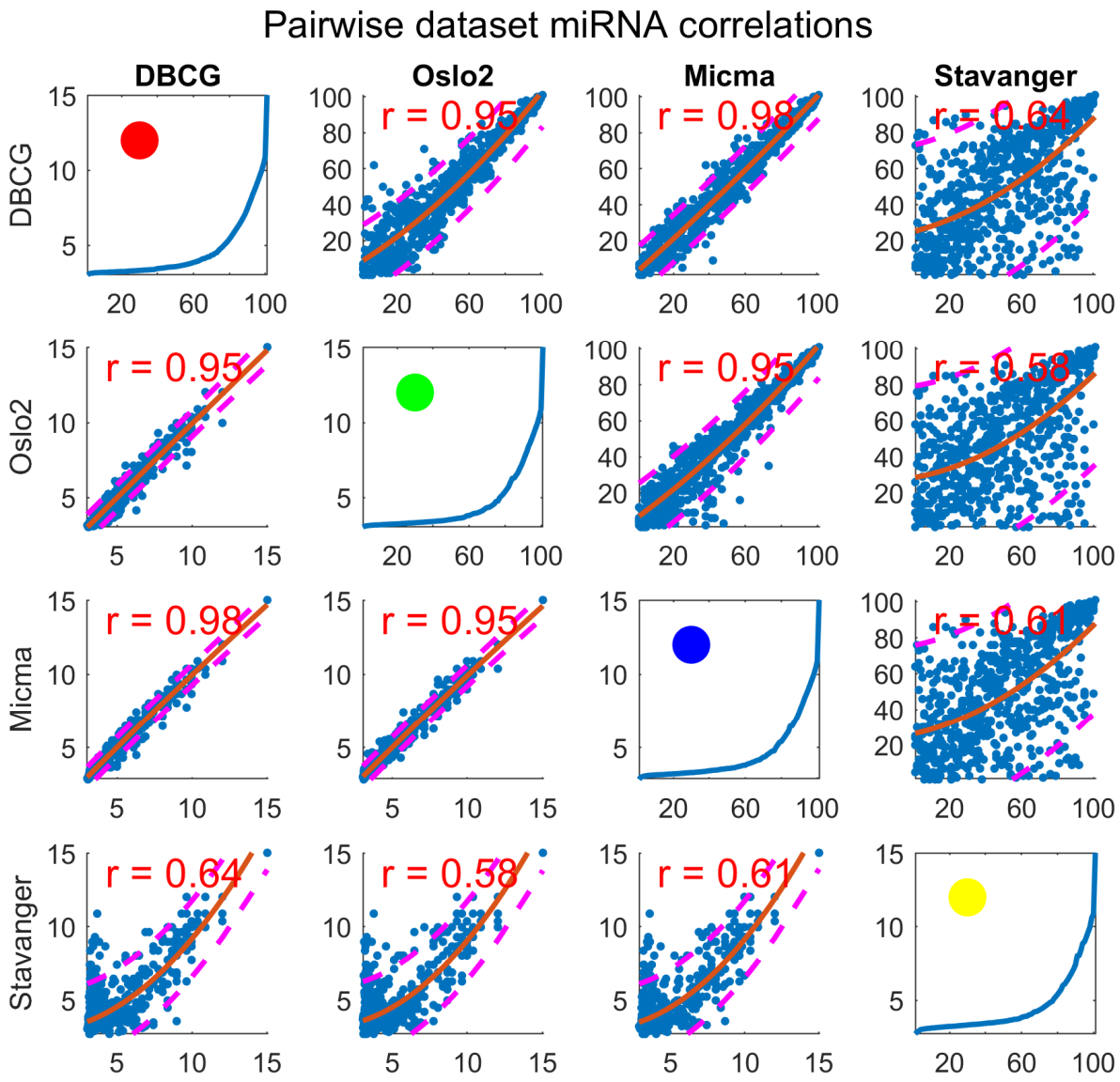


Figure 19. Showing quantile normalized data miRNA expression reproducibility across dataset pairs. Each subplot depicts the miRNA median expression across samples for a pair of datasets. The upper-diagonal-subplots show percentiles, and bottom-diagonal shows log2 expression. A second degree polynomial curve is fitted and prediction intervals at confidence level 0.8 are plotted as dashed lines. Spearman correlation is given for each subplot. Figures at the diagonal show percentile plotted against expression and a circle represents the dataset colorcode as related to other figures in the paper.

We further visualize the batch-clustering behavior of the unnormalized joint dataset in Figure 20. On the left subplot we present hierarchical clustering of the data. Edges of sub-trees in the dendrogram are color-coded by the dataset when all leaves belong to samples from the same original dataset. We observe a visual clustering of colors, especially evident for yellow (Stavanger) being clustered as an outgroup. In the middle subplot we show a silhouette plot, depicting the clustering consistency according to dataset. We can see how a substantial portion

of samples are well assigned to their cluster with large silhouette values, and only a small portion are mis-assigned, again showcasing how batch effects dominate sample behavior. Finally, on the right subplot we present a visualization of the sample-wise pairwise Euclidean distance matrix with dashed lines separating between samples of the same dataset. The block structure that evidently results from coloring according to distances corresponds well to the dashed lines separating samples from different datasets. This analysis demonstrates the prevalence of batch effects in the joint datasets.

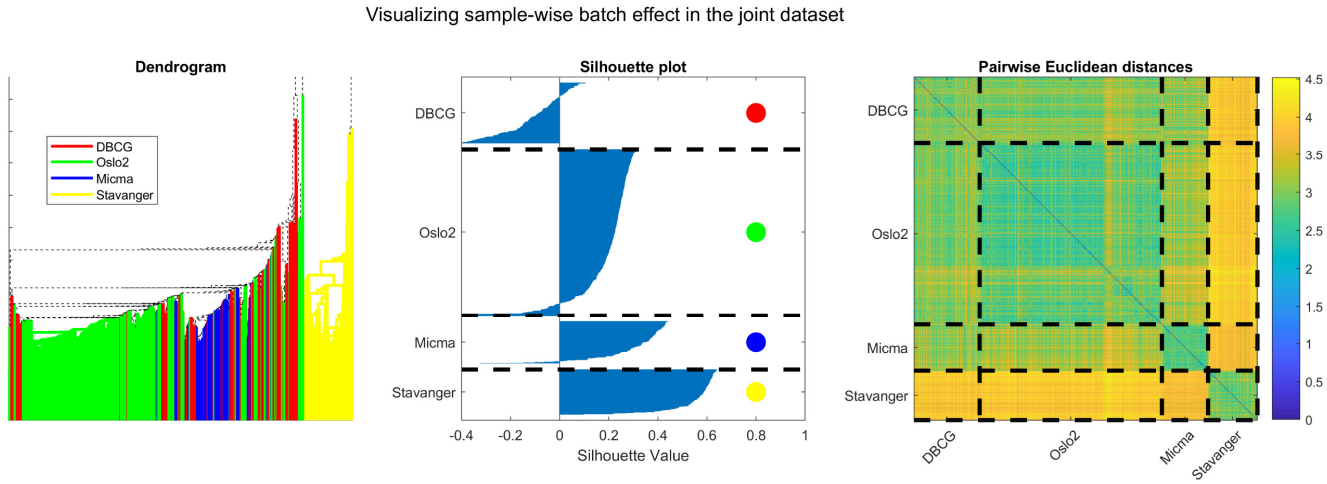


Figure 20. Batch effects in the combined cross-tech miRNA dataset considering the unnormalized data. (Left) Dendrogram with edges colored by dataset. Note that the tree root is not shown. (Middle) Silhouette plot (Rousseeuw, 1987) *showing that most samples cluster according to the dataset they originate from. (Right) Pairwise Euclidean distances showing a block structure that agrees with the sample dataset of origin.*

4.2.3 Adjusted Quantile Normalization (AQN)

In this section we describe our quantile-normalization-based strategy for analyzing combined cross-technology miRNA datasets.

Let X be a batch collected, joint dataset. $X \in R^{n \times m}$ where $X(i,j)$ is the log measured intensity value of miRNA i in sample j. Let $X(:,j)$ be the j-th sample, corresponding to the j-th column in X, and $X(i,:)$ be the i-th miRNA, corresponding to the i-th row in X.

Define $B(X(:,j)) = k$ as the experiment batch id during which sample j was collected.

We note the following distinction between missing values in X:

$$X(i,j) = \begin{cases} nan & miRNA\ i\ was\ not\ sampled\ in\ B(X(:,j)) \\ 0 & miRNA\ i\ was\ sampled\ in\ B(X(:,j))\ but\ not\ detected\ in\ sample\ j \\ \geq 0 & otherwise \end{cases}$$

Let $MFP(i,j) = 1$ if miRNA i is missing from platform $B(X(:,j)) = k$ and $MFP(i,j) = 0$ otherwise (indicates if i is missing in the platform j was measured in).

Adjusted Quantile Normalization ($\boldsymbol{X}$):

1.	$\mathcal{D} \leftarrow X + N(0, \epsilon)$	Jitter X to break rank ties.
2.	Let $P(i,j) =$ the percentile of $\mathcal{D}(i,j)$ within $\mathcal{D}(:,j)$.	ignored nans in percentile computation. Note: $P(i,j) \in [0,100]$
3.	$Q(i,j) = \underset{1 \le t \le m}{median}\{\mathcal{D}(s,t): P(s,t) = p(i,j)\}$	Transforms values to the cross-sample-median of the corresponding per-sample-quantile.
4.	$Q(i,j) = nan$ if $MFP(i,j)$=1	

A description of this process in words is that it replaces present expression values with the corresponding median value of all samples within the same percentile. The underlying assumption is that a measured expression is volatile due to technical differences and measurement noise, however, (sample-based) percentiles are assumed to be stable up to the biological differences between samples.

The overall impact of applying AQN to the distribution of expression values and to quantified batch effects as measured by the silhouette coefficient is further presented in Supplementary Figure S20.

Packages implementing AQN are available online for Python, R and Matlab in https://github.com/YakhiniGroup/PyAQN.

4.2.4 Functional experiments

Functional experiments were performed as previously described (Leivonen *et al.*, 2009, 2014) with the breast cancer cell lines MCF7 and KPL-4. The lysate microarray data measuring apoptosis in the form of cleaved PARP (cPARP), HER2 and phosphorylated ERK (pERK) protein levels after 72 hours were previously published (data taken from Supplementary table 2 of the corresponding publication) (Leivonen *et al.*, 2014). Values ±2 × standard deviation (SD) were considered as significant, which corresponded to a threshold of $|1.96|$. For the cell viability data, MCF7 cells were transfected with the Dharmacon miRIDIAN microRNA mimic library v.10.1 (20 nM) and incubated for 72 hours. The cell viability was measured with CellTiter-Glo assay (Promega Corp, Madison, WI, USA) according to manufacturer's protocol. The experiments were done with two biological replicates. The data were normalized by a Loess method (Boutros *et al.*, 2006) and log2-transformed. Values ±2 × SD, were considered as significant, which corresponded to a threshold of $|0.2|$. In both experiments the average of two different miRNA mimic controls from two replicates was used as negative controls (miRIDIAN microRNA Mimic Negative Control #1 from Dharmacon and pre-miR negative control #2 from Ambion).

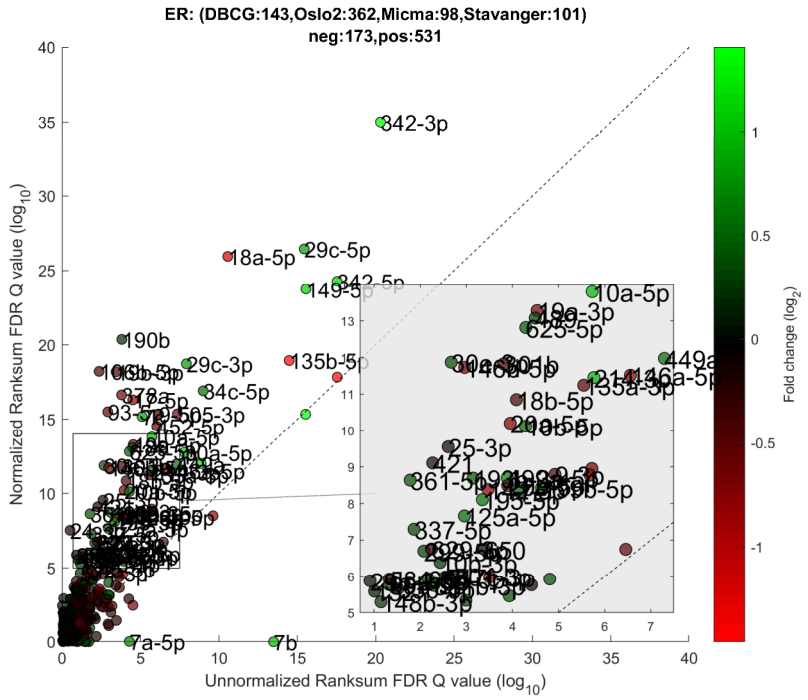
4.3 Results

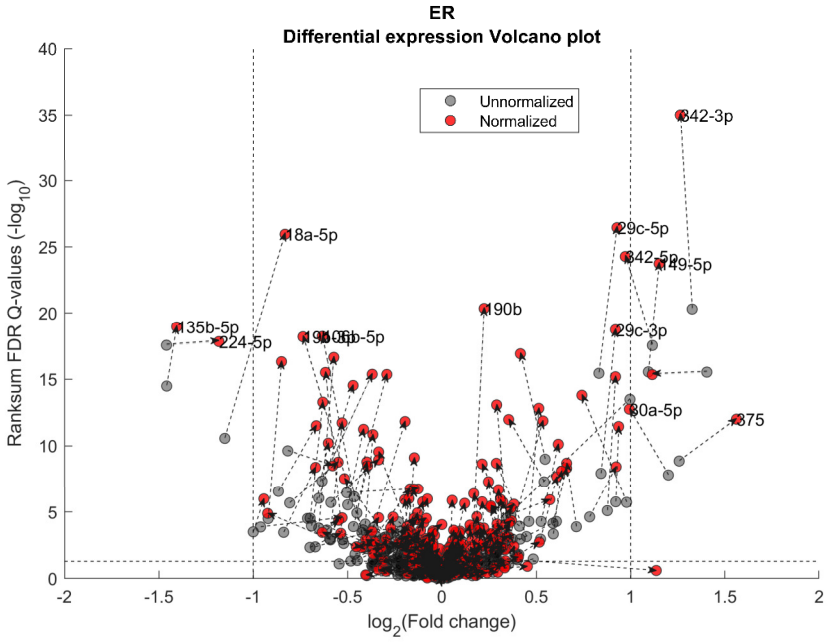
We apply the Adjusted Quantile Normalization (AQN) process to the datasets described in (Enerly *et al.*, 2011; Aure *et al.*, 2017; Myhre *et al.*, 2010; Tramm *et al.*, 2014) and illustrate the benefit and effects of the normalization step as related to data properties and to various downstream analysis steps in the subsections below.

4.3.1 Differential expression reveals novel breast-cancer associated miRNA

We performed a differential expression analysis between clinically relevant subgroups of breast cancer. We measure differential expression of a specific miRNA on a pair of sample subpopulations (e.g. ER positive vs ER negative). Fold-change is defined as the ratio ($\log_2$) between median expression of both sets. We apply Wilcoxon Rank-sum 1-tailed test (where the tail is determined empirically according the sign of the fold-change). Resulting p-values are corrected across miRNAs using false discovery rates (FDR). Figure 21 showcases our differential expression analysis results for ER status. In the top scatter plot, we observe that the normalized dataset presents with more significant results (lower Q-values) for most miRNAs (482/655). The middle volcano plots illustrate that the increase in significance is not necessarily correlated with effect size (i.e. fold change), and that we gain confidence on lower effect sizes as anticipated by a priori power analysis. At the bottom cumulative distribution function (CDF) plot we showcase again the overall trend of increased statistical significance, contrasted by even lower statistical significance that would be obtained from performing the differential expression analysis on each dataset separately (shown as dashed lines). In addition, we present the CDF plots that would be obtained by (individually) applying four commonly used normalization methods (shown as dotted lines). Evaluated normalization methods include:

- Mean ratio: scales each sample by dividing it by its mean intensity.
- Median subtraction: subtracts the median of each sample, then sets the minimum of each sample to the (global) minimum across samples.
- Vanilla quantile: MATLAB's implementation of Quantile Normalization also known as Quantile Standardization (Amaratunga and Cabrera, 2001).
- ComBat (Johnson *et al.*, 2007): empirical Bayes batch effect mitigation employing a design matrix that includes dataset batching along with clinical labels and status of Tumor grade, Subtype, ER, PR, HER2 and TP53.


ER: (DBCG:143,Oslo2:362,Micma:98,Stavanger:101)
neg:173,pos:531
Normalized Ranksum FDR Q value (log10)
Unnormalized Ranksum FDR Q value (log10)
Fold change (log2)
342-3p
18a-5p
29c-5p
190b
29c-3p
135b-5p
84c-5p
10a-5p
19a-3p
625-5p
449a
18b-5p
25-3p
421
361-5p
125a-5p
837-5p
148b-3p
7a-5p
7b


ER
Differential expression Volcano plot
Unnormalized
Normalized
Ranksum FDR Q-values (-log10)
log2(Fold change)
342-3p
18a-5p
29c-5p
342-5p
149-5p
190b
135b-5p
224-5p
29c-3p
30a-5p
375

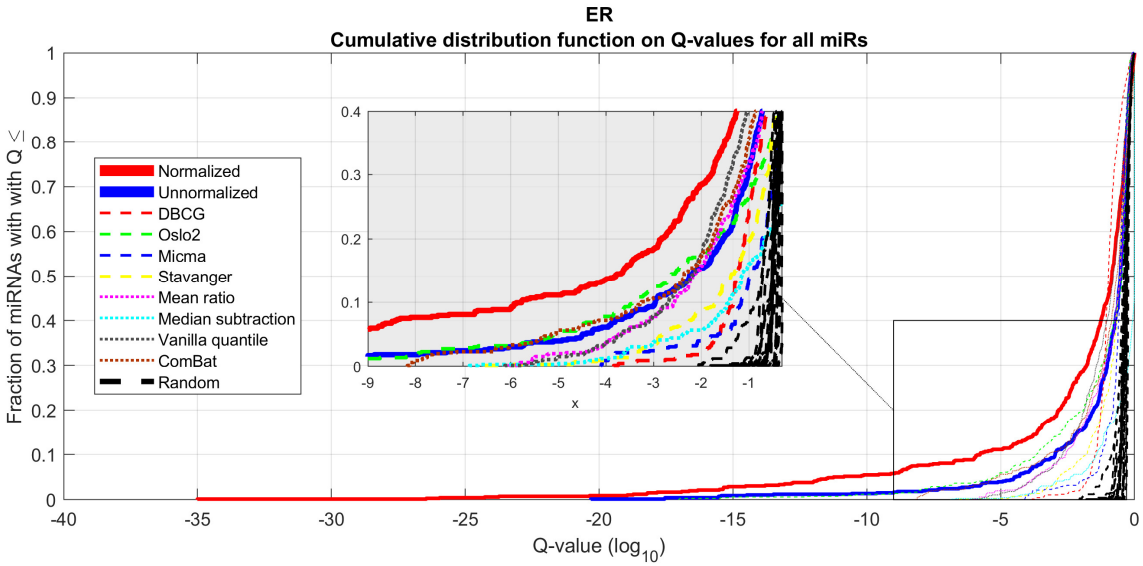


*Figure 21. Differential miRNA expression between ER positive and negative. Title contains sample size details and dataset distribution (Top) A scatter plot of differential expression p-values (-log$_{10}$, Wilcoxon Rank-sum) for the unnormalized (x) vs normalized (y) joint dataset. (Middle) Volcano plot showing the fold change and corresponding Wilcoxon Rank-sum FDR corrected Q value ratio between the normalized and unnormalized datasets. High absolute values in X axis correspond to substantial difference in median expression between ER negative over ER positive samples (for a particular miRNA). High values in Y axis correspond to miRNAs that present substantial difference *after* normalization but not before. Low values in Y axis correspond to miRNAs that present substantial difference *before* normalization but not after. Vertical dashed lines represent a Fold change threshold of 2x (log$_2$(2)=1) and horizontal dashed lines represent a Q-value threshold of 0.05 (-log$_{10}$(0.05)≅1.3) (Bottom) a CDF plot showing many more substantially differentially expressed miRNAs after normalization (red line) than before normalization (blue line), and substantially more than would be expected at random (compared to 20 random permutation of labels, dashed black lines). Also shown are dashed colored lines corresponding to each appropriate single-dataset Q values exemplifying the advantage of a joint-dataset analysis.*

In Figure 22 we demonstrate the impact of normalization on single miRNAs (hsa-miR-190b, hsa-miR-18a-5p) across samples and distinguish between differently labeled samples according to ER status. Previous studies (Cizeron-Clairac *et al.*, 2015) have shown hsa-miR-190b to be linked to ER status and further suggested its use as a potential biomarker. Similarly, hsa-miR-18a-5p is an oncogene and prognostic biomarker (Zhou *et al.*, 2018). As we have shown in the volcano plot in Figure 21, hsa-miR-190b would not have been identified as differentially expressed in ER

positive vs negative samples prior to normalization. Similar plots for the top 40 differentially expressed miRNA (post-normalization) are available in the Supplementary materials.

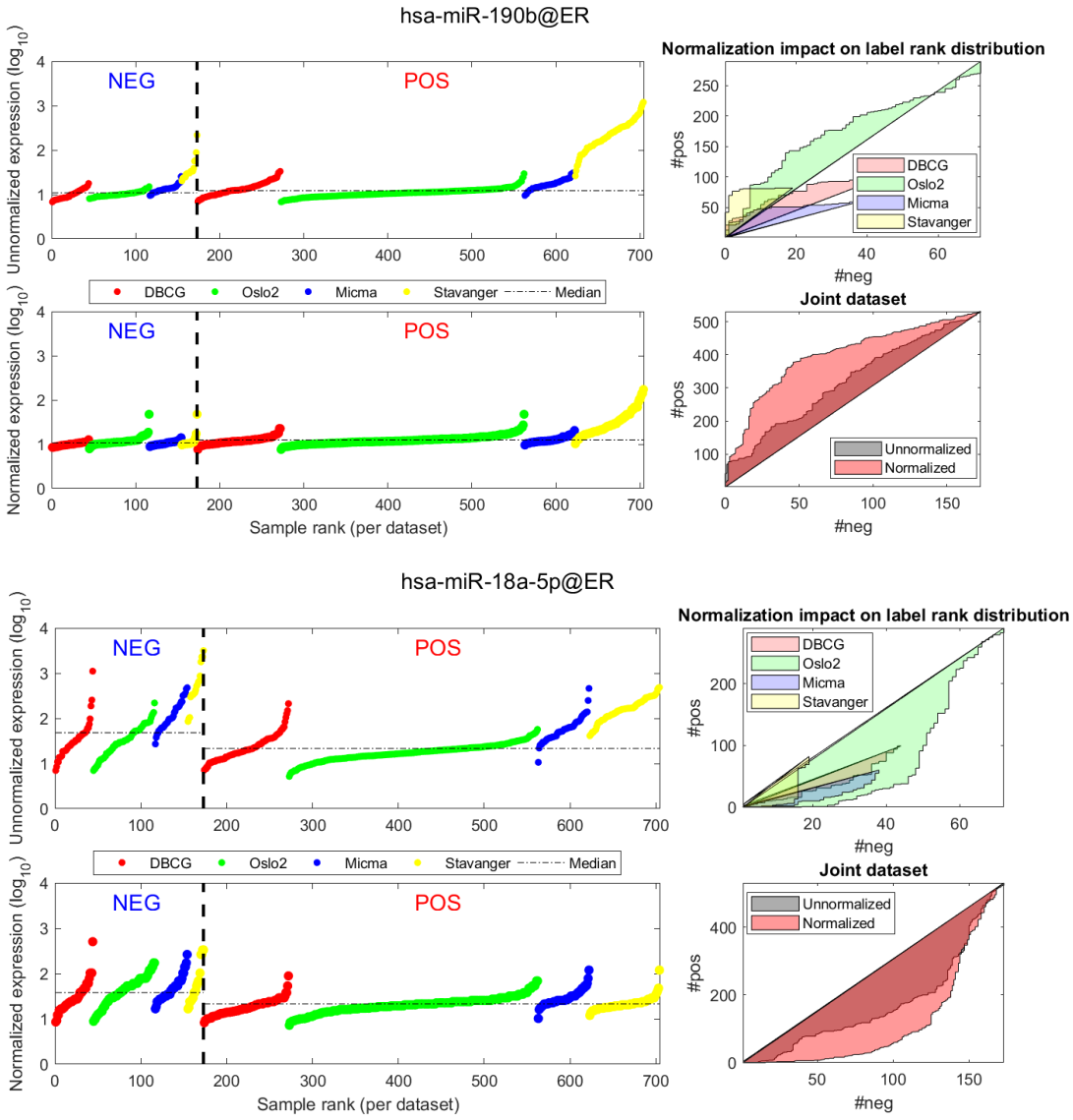


Figure 22. Differential expression behavior of single miRNA. (Top - hsa-miR-190b, bottom – hsa-miR-18a) across datasets and samples for a specific clinical label (estrogen receptor (ER) positive (pos) vs. negative (neg). (Left) Expression values (log_2) of each sample before quantile normalization. Samples are ranked by ER status label, then by dataset and finally by ascending expression value. Top-Unnormalized joint dataset. Bottom-Normalized joint dataset. (Right) Actual vs expected (via a uniform null model) rank distribution of ER negative (neg) vs positive (pos). Diagonal straight lines bounding a polygon represent a null uniform distribution of positive and negative samples (when ranked by expression value). The colored surface represents deviations from a uniform distribution. The boundary of the surface is calculated by the cumulative number of ER negative (x axis) vs ER positive (y axis) samples in the ranked (descending) expression vector. Top-illustrating the rank distribution per-dataset (without normalization). Bottom-comparing the joint-dataset distributions when ranking before or after normalization.

When inspecting the differential expression results of all normalization methods, the unnormalized data and each dataset separately, there are 33 unique miRNAs that are only

shown as significantly ($Q\ value < 0.05$) differentially expressed in ER positive vs ER negative as identified by our normalization method (Supplementary Figure S21). Contrastingly, other approaches yield far fewer significantly differentially expressed miRNAs. Of the 33 miRNAs uniquely detected by our method, we present four in Table 3 that have fold change greater than 0.15 (absolute $\log_2 > 0.15$, i.e. > 10% change between median ER positive and negative expression).

miRNA	Q-value	Fold Change ($\log_2$)
hsa-miR-601	0.048	-0.18
hsa-miR-424-3p	0.0003	-0.17
hsa-miR-936	0.027	-0.15
hsa-miR-193b-5p	0.0002	0.19

Table 3. Top differentially expressed miRNA. We present miRNA detected by applying AQN normalization on the joint dataset and not detected by other approaches.

To study any functional significance of these top differentially expressed miRNAs between ER positive and ER negative tumors, we performed miRNA gain-of-function studies in the ER-positive MCF breast cancer cell line. Here, cell viability was measured as an endpoint after overexpression of the miRNAs. Indeed, one of the miRNAs, hsa-miR-193b-5p, showed a significant reduction in cell viability compared to miRNA negative controls (Figure 23). Furthermore, we looked into data from another functional experiment previously published (Leivonen *et al.*, 2014) in the HER2 positive breast cancer cell line KPL4 and here we found that hsa-miR-193b-5p induced apoptosis (as measured by the levels of cleaved PARP), and downregulated the levels of HER2 and phosphorylated ERK upon overexpression. Altogether, these results suggest that miR-193b-5p may exert a tumor-suppressor function in breast cancer, both in an ER+ and a HER2+ context. Interestingly, the other miRNA originating from the same precursor, hsa-miR-193b-3p has been previously shown to directly target ESR1 mRNA and is thus a direct regulator of ER (Leivonen *et al.*, 2009).

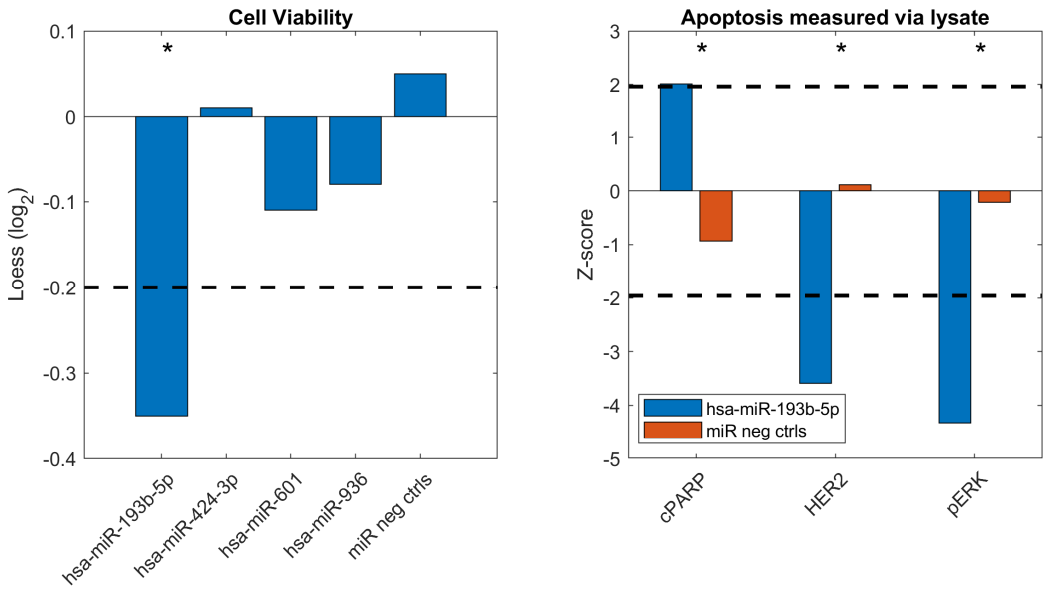


Figure 23. Functional analyses on uniquely identified miRNA. Breast cancer cell lines were transfected with miRNA mimics (20nM) and assayed for functional effects 72 hours after transfection. a) Cell viability measured in MCF7 breast cancer cells. b) Apoptosis measured by levels of cleaved PARP (cPARP), HER2 and phosphorylated ERK (pERK) protein levels measured in KPL4 cells. The dashed lines indicate cut-off points that were considered significant (see Methods). Asterisks denote significant effects. Original data from b) are taken from (Leivonen *et al.*, 2014).

Further investigation of the three other top differentially expressed miRNAs shows prior evidence linking them to cancer. hsa-miR-601 is a known prognostic marker and potential tumor-suppressor in breast cancer (Hu *et al.*, 2016). hsa-miR-936 was identified as a potential tumor-suppressor miRNA in ovarian cancer (Li *et al.*, 2019).

4.3.2 Joint analysis with mRNA data

A similar pipeline to the one described in section 2 (Dataset pre-processing and coverage) was used to parse mRNA data, using Limma.

We want to assess the effect of normalization on the results of enrichment analysis as performed using both mRNA and miRNA data. To this end we first form a ranked list of transcripts as follows. For each miRNA, μ, we rank all mRNAs according to the (ascending) Spearman correlation between the miRNA expression pattern across the entire dataset and the mRNA expression pattern across the entire dataset (paired on matching samples). We denote the resulting ranked gene list, with μ as a pivot, as $\mathcal{G}_\mu$.

4.3.3 Effect on gene target enrichment

For the first analysis we investigated the impact of normalization on correlations between miRNA and the expression levels of their expected mRNA targets. We expect stronger negative correlation after normalization to direct gene targets. To validate this hypothesis, we applied a

non-parametric, rank-based analysis using the MiTEA (Eden *et al.*, 2009b; Steinfeld *et al.*, 2013a) approach. MiTEA is used to evaluate the statistical association between $\mathcal{G}_\mu$ and $\mathcal{C}_\lambda$, where $\mathcal{C}_\lambda$ is a ranked list of genes wherein the ranking is based on the affinity of the gene as a target candidate for the miRNA λ, taken from TargetScan (Agarwal *et al.*, 2015). For each prefix $\Pi_{\mathrm{B}}(\mathcal{C}_\lambda)$ of B most-prominent candidate targets in $\mathcal{C}_\lambda$, MiTEA produces a binary vector, $\mathcal{B}(\mu, \lambda, B)$, such that, g_i, the i-th gene in $\mathcal{G}_\mu$ is "1" if and only if it is in the candidate prefix, i.e. $g_i \in \Pi_{\mathrm{B}}(\mathcal{C}_\lambda)$. MiTEA then computes an approximate minimum hypergeometric (mHG (Eden *et al.*, 2009b, 2007)) P-value to quantify whether the B proposed targets are enriched at the top of the $\mathcal{G}_\mu$ list or not. Finally – MITEA applies an FDR correction (using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995)) across evaluated λs and reports the set of miRNAs associated with the ranked target list $\mathcal{G}_\mu$ and their associated Q-values.

We declare a matching if MiTEA returns a significant (≤ 0.05) Q-value when $\lambda = \mu$. To recapitulate, a matching occurs when the prominent predicted targets of μ are enriched at the top of the list of genes ranked (in ascending order) according to the rank correlation (across samples) between their mRNA levels and the expression levels of μ. When applying this procedure on a non-normalized miRNA expression we find no matchings. When applying the same procedure on normalized data we find 6 matchings as detailed in Table 4. For each matched miRNA we also provide supporting evidence of several studies describing its role in breast cancer.

miRNA	P-value	Q-value	Corroborating studies
hsa-miR-29b	1.28E-08	1.73E-06	(Kwon *et al.*, 2019; Wang *et al.*, 2011; Shinden *et al.*, 2015)
hsa-miR-106b	1.96E-06	1.11E-04	(Ni *et al.*, 2014; Lee *et al.*, 2019; Zheng *et al.*, 2015)
hsa-miR-200b	1.06E-04	5.54E-03	(Ye *et al.*, 2014; Yao *et al.*, 2015; Zheng *et al.*, 2017)
hsa-miR-30d	4.38E-04	1.19E-02	(Zhang; Yang *et al.*, 2017)
hsa-miR-96	9.02E-05	1.53E-02	(Hong *et al.*, 2016; Xie *et al.*, 2018)
hsa-miR-182	4.58E-04	4.43E-02	(Zhang *et al.*, 2017; Chiang *et al.*, 2013)

Table 4. Resulting MiTEA matchings on normalized miRNA expression.P and Q values are color coded by magnitude where from green (more significant results) to red (less significant results). None of these statistically significant associations between pivot miRNAs and their targets is observed when using the raw, un-normalized data. Nor is any other matching miRNA target enrichment observed in the unnormalized data.

We show one such analysis in detail for *hsa-miR-29b* in Figure 24. Here we follow MiTEA's approach to obtain a statistical assessment of target enrichment for $\mu = \lambda =$ *hsa-miR-29b* and $B = \{1, \ldots, |\mathcal{C}_\lambda|\}$ binary vectors $\mathcal{B}(\mu, \lambda, B)$. We present the results on various Bs and the optimal B^* for both unnormalized and normalized miRNA expression.

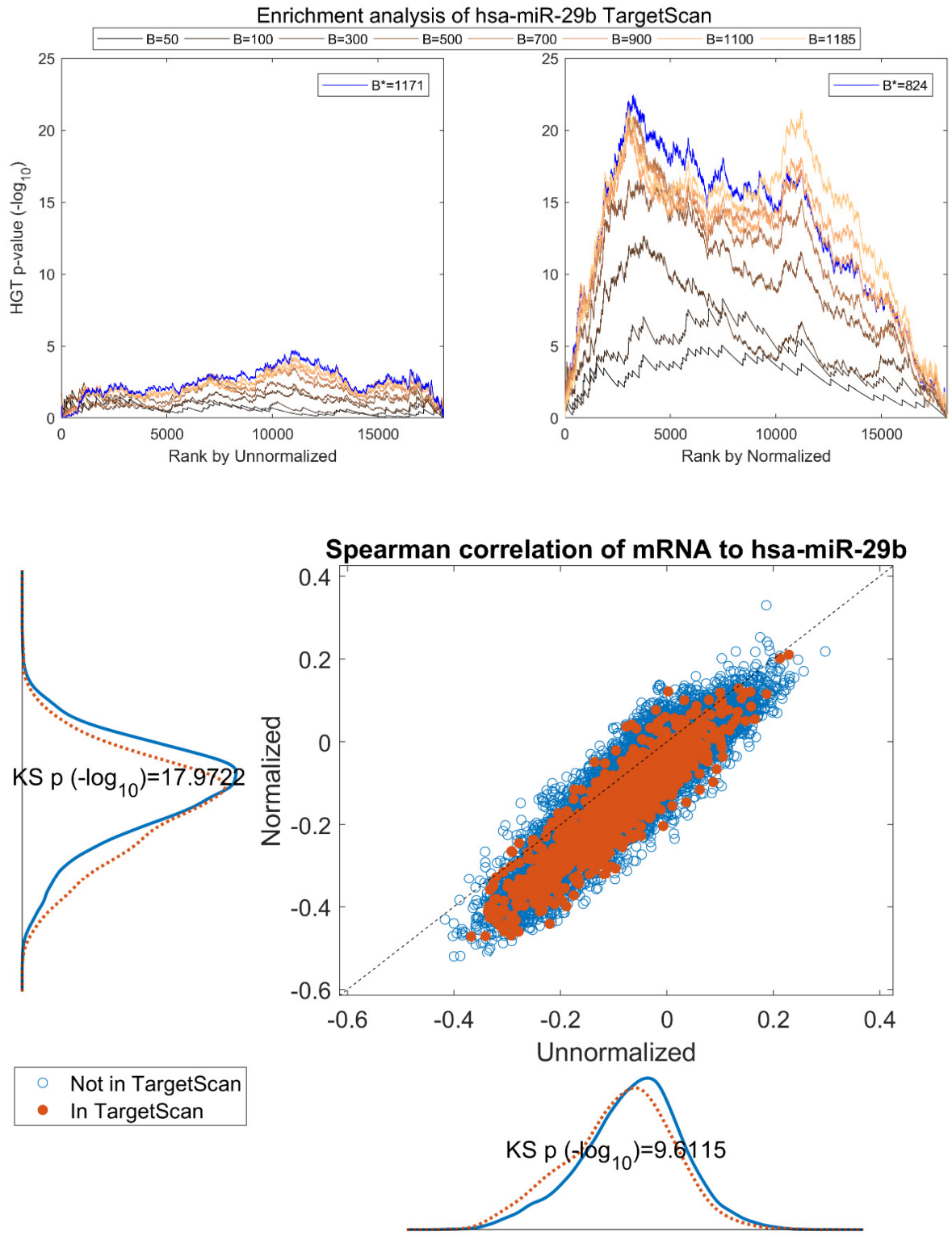
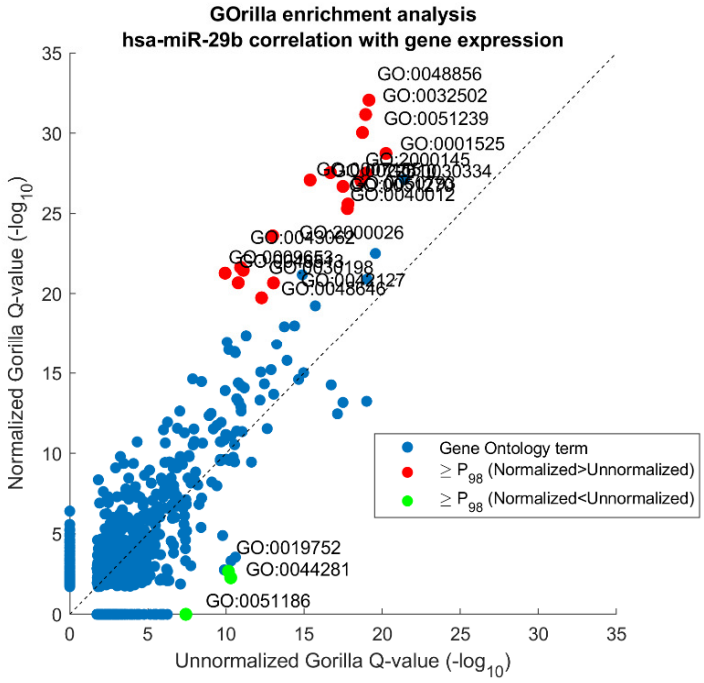


Figure 24. Impact of normalization on the correlation between hsa-miR-29b expression and its in-silico *predicted targets according to TargetScan. Top) Normalized miRNA is more negatively correlated to the prominent hsa-miR-29b targets in TargetScan as evident in stronger enrichment values. Bot) Scatter plot of spearman correlation on normalized miRNA or unnormalized miRNA expression. If the target mRNA appears in TargetScan it is highlighted in orange. The marginal*

distributions and corresponding Kolmogorov-Smirnov test p-values are displayed showing an overall lowered correlation for TargetScan candidates on normalized data.

4.3.4 Effect on Gene Ontology (GO) enrichment

We applied GOrilla (Eden *et al.*, 2009b) to identify gene ontology enrichment in $\mathcal{G}_\mu$ on both unnormalized miRNA expression and on normalized miRNA expression. Given a ranked list $\mathcal{G}_\mu$, GOrilla produces a binary vector $\mathcal{B}(\mathcal{G}_\mu, \omega)$ for each gene ontology term, ω, in which a gene is labeled as binary '1' if it belongs to ω. Next, GOrilla computes mHG p-values, correcting them across GO terms. Figure 25 is a scatterplot comparing between our results on unnormalized and normalized hsa-miR-29b lists. The findings from this analysis are in line with previous studies that have linked the miR-29 family with tumor growth and metastasis (Wang *et al.*, 2011; Luna *et al.*, 2009; Liu *et al.*, 2017).



	Ontology term
GO:0048856	anatomical structure development
GO:0032502	developmental process
GO:0007155	cell adhesion
GO:0009653	anatomical structure morphogenesis
GO:0051239	regulation of multicellular organismal process
GO:0022610	biological adhesion
GO:0043062	extracellular structure organization
GO:2000026	regulation of multicellular organismal development
GO:0030198	extracellular matrix organization
GO:0048513	animal organ development
GO:0050793	regulation of developmental process
GO:0030334	regulation of cell migration
GO:0001525	angiogenesis
GO:2000145	regulation of cell motility
GO:0051270	regulation of cellular component movement
GO:0042127	regulation of cell proliferation
GO:0040012	regulation of locomotion
GO:0048646	anatomical structure formation involved in morphogenesis
GO:0051186	cofactor metabolic process
GO:0044281	small molecule metabolic process
GO:0019752	carboxylic acid metabolic process

Figure 25. GOrilla enrichment analysis comparison of results before and after miRNA normalization. Right) Top 2 percentile of results by Normalized-Unnormalized Q-value (-log$_{10}$)

4.4 Discussion

We have presented an integrative analysis technique and applied it to jointly analyze human breast cancer miRNA expression datasets spanning different studies and utilizing different measurement technologies. Our approach is powerful in its ability to increase statistical power without apparent adverse effects on precision, as exemplified by several downstream tasks. Our normalization method (AQN) is based on a slight adaptation to standard (a.k.a. vanilla) quantile normalization. Vanilla quantile normalization averages values across samples with the same rank, while our method averages values across samples within the same percentiles (computed per sample). This has the effect of lowering the impact of within-quantile noise when computing rank-based statistics. Additionally, our method is defined consistently for normalizing samples with partial miRNA overlaps.

Correctly applying AQN requires a basic understanding of the impact it has on downstream statistics. In this work we focused on applying nonparametric rank-based statistics to downstream analyses. While not deemed a best practice, our normalization approach admits to parametric analyses as well. Further discussing parametric analysis is out of scope for this work.

We distinguish between Sample-wise (a.k.a. column-wise) and miRNA-wise (a.k.a. row-wise) impact. Sample-wise, we apply a monotonic transformation of raw expression values per sample which should not affect rankings of miRNAs within each sample. As we observe in Supplementary Figure S*22*, Left we see these samples almost fully correlated before and after normalization. The minor differences are owed to two effects – jitter and quantization. Jitter can swap miRNA ranks within a sample, especially for miRNA with low expression compared to our jitter scale. We pre-process the data by min-max normalization and select a jitter scale such that ranks are mostly unaffected by jitter. A stronger impact is due to quantization which replaces values within the same percentile with a cross-sample median, creating ties.

miRNA-wise there are no guarantees of monotonicity, as evident in Supplementary Figure S*22*, Right and as shown in improved results for analyses such as differential expression in section 4.1.

4.4.1 Comparison to per-dataset analysis

When comparing downstream analyses of the normalized joint dataset with per-dataset analyses we observe stronger p-values, yielding more statistically significant candidates after applying multiple hypothesis correction procedures. In Figure 21, bottom we illustrate this result through a shift in the cumulative distribution of Wilcoxon Rank-sum FDR corrected Q-values calculated for the differential expression of ER positive and negative samples. In Figure 26 we present a per dataset variation of the analysis as related to Figure 21, Middle. We observe that some miRNA exhibit a tradeoff between higher absolute fold-change and higher rank-sum -$\log_{10}$ Q-values. For example observe hsa-miR-135b that has $> -8 \times$ fold change for

Stavanger, but at a fairly low -$\log_{10}$ Q-value < 4 while after joint analysis it yields only $> -2 \times$ fold change but at -$\log_{10}$ Q-value > 18.

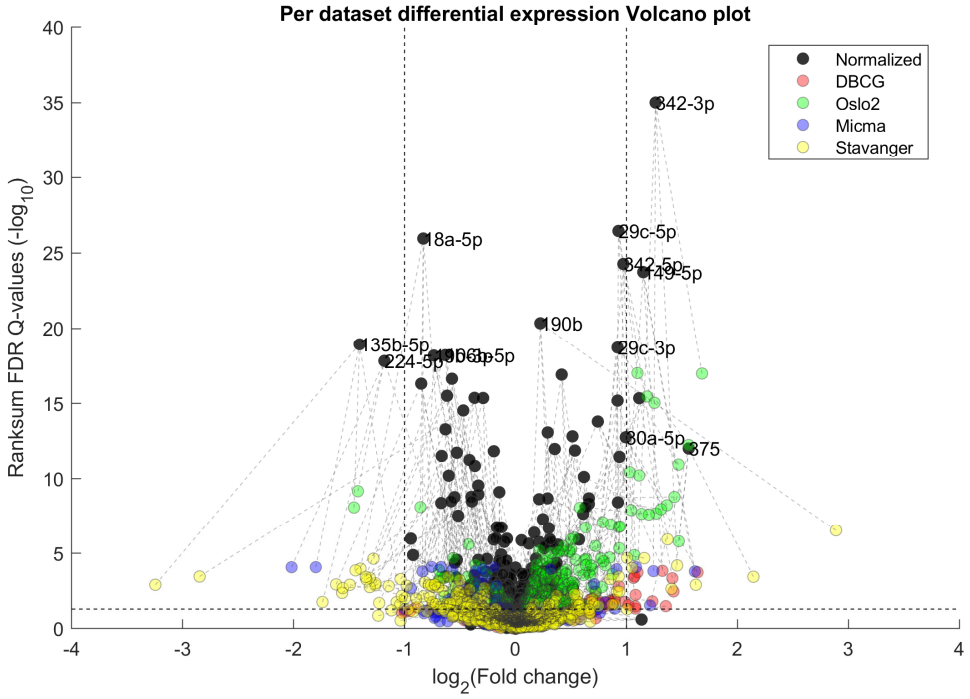


Figure 26. Per dataset Volcano plot of Differential Expression. Showing ER positive vs negative from Figure 21 Compared to joint normalized data.

4.4.2 Statistical power analysis on the impact of increasing sample size

One of the main motivating reasons for jointly analyzing datasets collected in different places, times and possibly using different measurement technologies is the fact that the combined dataset supports higher statistical power.

We present a theoretical statistical a-priori power analysis (Faul *et al.*, 2007) to put in context the advantage of jointly analyzing the datasets investigated in the current work. Remember that power is used in statistics to quantify the recall of a statistical test, i.e. the probability of correctly rejecting the null hypothesis. The test evaluated in this analysis is Wilcoxon rank-sum as applied for our differential expression analysis in section 4.1. Power is only meaningful in the context of an expected effect size, as larger differences and less variance in samples implies a smaller sample size is required to decide there is a difference between two populations. For the purpose of this analysis we assume allocation ratio = 1 (i.e. equal group sizes), while in the ER examples shown in Figure 27 actual ratios of Negative vs Positive ER samples are 0.44, 0.24, 0.63 0.23 and 0.32 for DBCG, Oslo, Micma, Stavanger and Joint, accordingly – further reducing expected power.

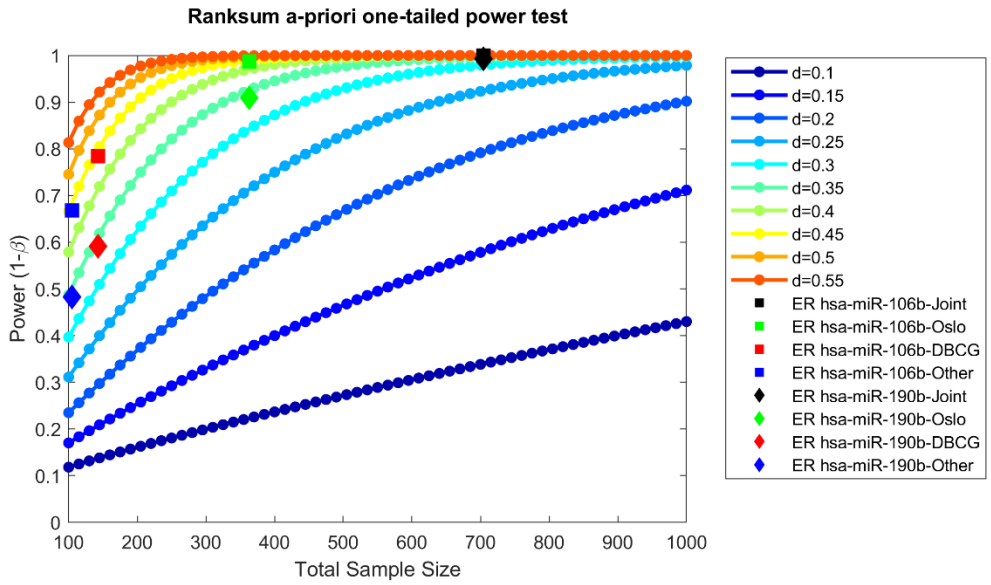


Figure 27. Statistical power as a function of sample size and expected effect size (measured in Cohen's d (Cohen, 1977)*).Overlaid in squares and triangles are effect sizes, d, for the differential expression of hsa-miR-106b and hsa-miR-135b, accordingly, in ER positive vs ER negative samples as estimated empirically over the joint dataset on non-normalized data. Power values are estimated via (linear, 2D) interpolation on different dataset sizes.*

4.4.3 Summary of contribution and next steps

Overall, we provide multiple lines of evidence for the advantageous joint analysis of miRNA expression using nonparametric statistics. Our analysis yields potential novel biomarkers as exemplified by hsa-miR-193b-5p and its potential tumor-suppressor role in breast cancer. While these results require further validation, our approach provides directions to statistically prominent candidates for follow up studies to pursue.

4.5 Chapter Supplementary Materials

Supplementary Methods

Supplementary 1 – Joint one-colored and two-colored analysis.

Stavanger dataset contains a second color with pooled samples deliberately left out of our analysis. Our downstream statistics are rank-based, assuming that, within a margin of error, identical samples measured with different technologies produce similar ranked miRNA vectors. Normalized Stavanger data using a pool reference second channel would cause substantial re-rankings. E.g. housekeeping, or constitutive miRNAs that are highly expressed would effectively "cancel out", and differently expressed miRNAs compared to the background would emerge instead. Therefore, to avoid an apples-to-oranges comparison, we decided to neglect the background expression data available in Stavanger from our analysis.

Supplementary Figures

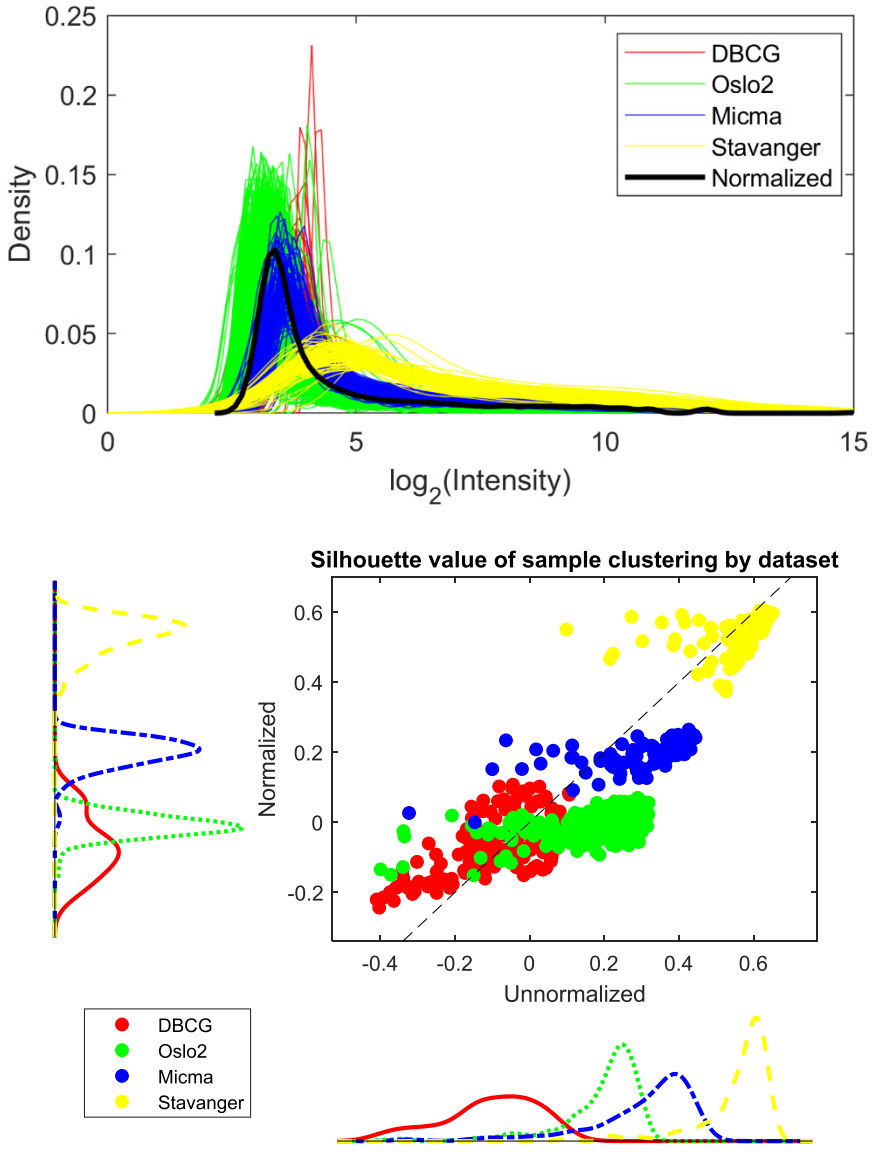


***Figure S20**. Normalization impact on per dataset distributions Top) Kernel density estimates of each sample colored by their corresponding dataset. The resulting normalized distribution is overlaid in black. Bottom) Impact of normalization on per-sample silhouette coefficient measured for clustering by dataset. 602/745 samples have lower silhouette coefficients after normalization in comparison to before normalization, demonstrating an overall alleviation of batch effect per dataset. Marginal distributions are shown to highlight differences between datasets.*

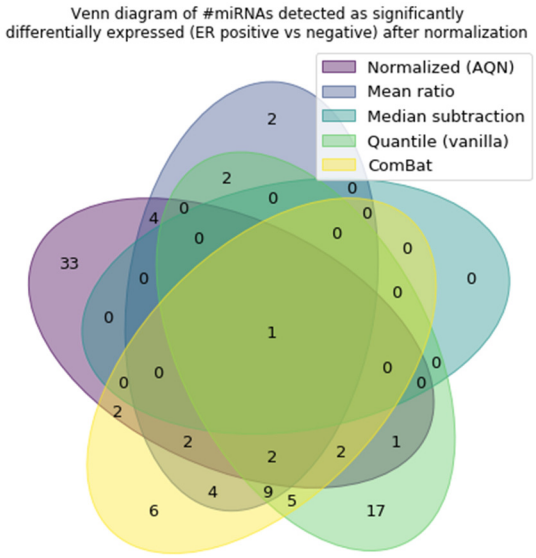


***Figure S21**. Venn diagram of differentially expressed miRNAs surfaced by different normalizations. We observe a larger set of unique miRNAs detected by our normalization approach compared to other approaches.*

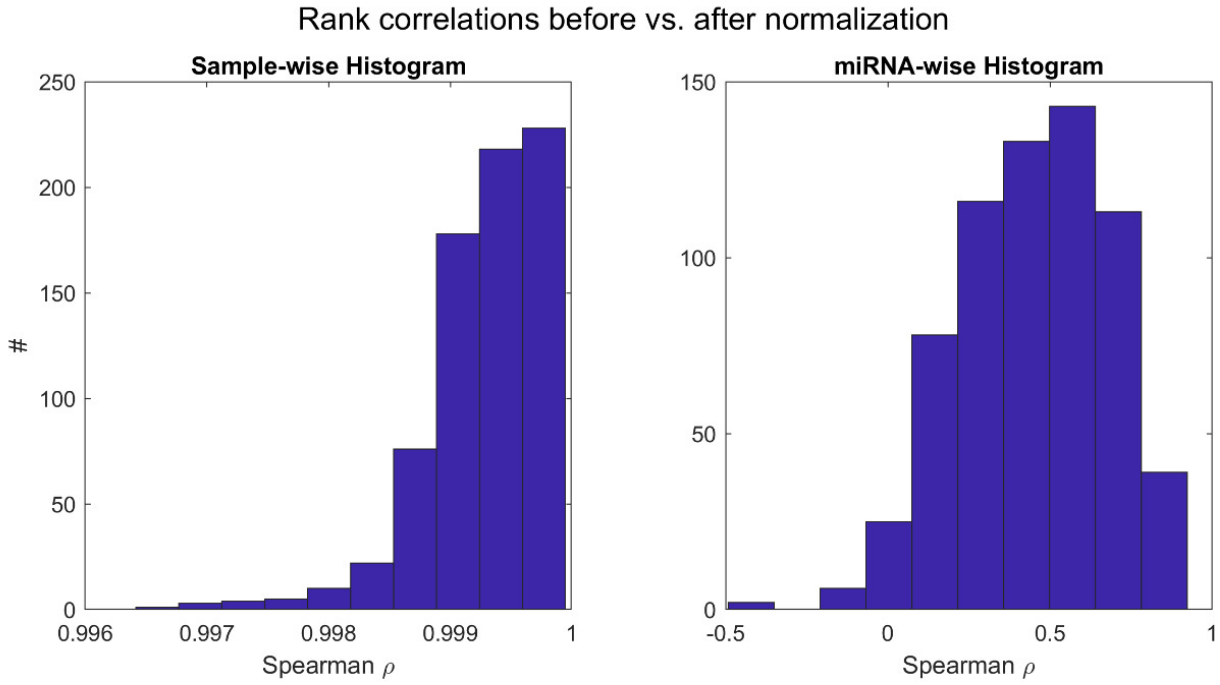


***Figure S22**. Correlations before and after normalization.Histograms of Sample-wise and miRNA-wise Spearman correlation coefficient (ρ) between expression before and after normalization.*

Chapter 5:

Discussion

In this thesis we have developed computational approaches for studying genomic spatial structure and properties. We demonstrated the applicability of our methods to biological data and described our findings, which, pending additional experimental validation, may offer novel biological insights. In the following chapter we summarize the algorithms detailed above offer additional observations and characterize possible extensions of them by outlining future research directions that may continue our work.

In Chapter 2 we presented an algorithmic framework to jointly completing a partial-haplotyping and demultiplexing Hi-C reads from homolog chromosomes in diploid organisms. We applied our approach to available ground-truth biological data to showcase its performance comparing to naïve approaches. Our approach is based on a novel sequence mapping algorithm which softly assigns reads to the correct compartment in the Hi-C diploid chromosomal adjacency block matrix by considering SNPs overlapping a sequencing read. We denoise the resulting Hi-C adjacencies by dimensionality reduction and use a simple, but optimal, decoding schema to assign each homologous pair of blocks a binary identity. The binary identity is assigned to effectively phase the blocks into their homolog chromosomal copies by maximum likelihood.

A natural extension of this work would be to add support for higher-ploidy organisms. One can consider replacing our argmax decoding algorithm with a dynamic programming one such as Viterbi to that end. Another direction worth exploring is in depth analysis of the impact of phasing Hi-C data on 3D modeling of genome conformation, co-localization including validating reproducibility of results observed in Hi-C studies which ignore the phasing problem, etc. A third direction of interest is resolving the need for a fixed binning resolution of Hi-C data which we determined by hyperparameter optimization aimed at yielding optimal phasing due to the tradeoff between sparsity when using small bins and averaging effects when using large bins.

In Chapter 3 we revisited our earlier work on spatial co-localization (Ben-Elazar *et al.*, 2013b) to devise improved methodologies of identifying spatial co-localization using more rigorous definitions and algorithmics. Using this new approach developed herein we are able to better detect whether a given binary property on a set of points exhibits 3D spatial co-localization, manifested as convex compartments with many target elements and few background elements. We quantify co-localization using a non-parametric statistical model, the minimum hypergeometric. By ignoring distances and considering ranks mHG offers an appropriate scale-free interpretation of the embedded conformation. We note that this is an appropriate approach considering our embedding methodology, NMDS (Seber, 1984), optimizes for rank-consistency rather than distance measurements. An additional advantage of focusing on ranks rather than distances is that the search space of possible compartments that one needs to consider becomes finite, and as we show in fact polynomial in the number of input points.

It is worth noting the apparent connection between our definition of the co-localization problem and a well-studied NP-hard problem, maxFS - maximum feasible subsystem (Amaldi and Kann, 1995). In spatial-mHG, we seek the minimal mHG score across tessellation cells induced by linear inequalities corresponding to the bisecting hyperplanes of pairs of differently-labeled input points. In maxFS we are interested in finding a solution to satisfy a maximal subset of a given set of linear inequalities. Spatial-mHG might be formalized as a more refined optimization problem, where we may in fact prefer a solution that forgoes satisfying several constraints that were induced by distant point pairs, in favor of satisfying few constraints that are induced by nearby point-pairs. We have thus far been unable to find a direct reduction from maxFS to spatial-mHG, however, this relaxation appears to make spatial-mHG at least as hard, if not harder than maxFS. It is possible that some approximation schemes that apply to maxFS could apply to spatial-mHG and may be used to initialize a solution, or when time constraints do not permit more extensive search.

During our work we have also explored approaches based on optimization algorithms to more efficiently traverse the bisector tessellation space. Our experiments show that the overhead of relatively optimized data structures has overall underperformed compared to sampling cells uniformly with replacement. However, we propose that a hybrid approach which quickly finds a

local minimum and then applies local search methods on its neighbors (e.g. discrete gradient descend) might empirically outperform our current approach.

Other directions to extend our work include: statistics to support target sets, e.g. in a polyploid organism there are multiple copies of each gene. We may not care which copy of each set is co-localized and want to reflect this in our search. Spatial co-localization for non-binary properties, e.g. we present an analysis on Pombe CGH data that required binarization to admit to our methodology. We could consider extending our work to support this input directly, for example using mmHG (Steinfeld *et al.*, 2013b).

In chapter 4 we present an adaptation of quantile normalization applied to integrative analysis of four miRNA expression breast-cancer datasets. In this work we attempted to overcome several challenges in jointly analyzing four miRNA datasets: partial miRNA target overlap, strong batch effects due to the technological differences between collection platforms and correctly interpreting normalized measurements in downstream statistical analyses. To overcome partial miRNA target overlap we devised an adaptation to quantile normalization that acts on percentile-binned rather than rank-binned miRNA expression. Our analysis provides evidence that our normalization is capable of detecting statistically consistent differences at smaller effect size than several standard methods, however this is by no means an exhaustive list of normalization approaches, nor is it necessarily consistent across datasets. A more rigorous understanding of the effect of different normalization approaches to different data distributions and edge cases (outliers) is necessary to fully characterize and assign normalization-approach-to-dataset and to analysis task. One direction to extend our work is to compare with more normalization techniques and on other integrative datasets. We have also deferred several downstream analysis tasks on available data to follow-up papers, including but not limited to measuring impact on correlation between miRNA expression and copy number, inclusion of more related miRNA expression datasets, such as TCGA, inclusion of more mRNA datasets in the mRNA validation section.

Overall, this thesis embodies computational approaches to analyze properties of genomes ranging from methodology to statistically analyzing their folding in space to improving the

interpretability of measured expression for miRNA of cancerous genomes. With the rise of more methods to measure genomes and increase in data availability, our approaches promise to aid in correctly interpreting and basing conclusions as we have shown in this work.

Acronyms

3C – Chromosome Conformation Capture

2D/3D – two/three dimensional

AQN – Adjusted quantile normalization

TF – transcription factor

ER – Estrogen receptor

GEO – Gene expression omnibus

Hi-C – High-throughput Chromosome Conformation Capture

RNA - ribonucleic acid

mRNA – messenger RNA

miRNA – micro RNA

nan – not a number

MFP – missing from platform

SNPs – Single nucleotide polymorphisms

mHG – minimum hypergeometric

smHG – Spatial mHG

SD – standard deviation

mmHG – minimum-minimum hypergeometric

maxFS – maximum feasible subsystem / subset

CGH – comparative genomic hybridization

TCGA – the cancer genome atlas

Bibliography

Adelfalk,C. *et al.* (2009) Cohesin SMC1beta protects telomeres in meiocytes. *J. Cell Biol.*, **187**, 185–99.

Agarwal,V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**.

Ahrens,H. (2007) Seber, G. A. F.: Multivariate Observations. J. Wiley & Sons, New York 1984. *Biometrical J.*, **28**, 766–767.

Amaldi,E. and Kann,V. (1995) The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theor. Comput. Sci.*, **147**, 181–210.

Amaratunga,D. and Cabrera,J. (2001) Analysis of Data From Viral DNA Microchips. *J. Am. Stat. Assoc.*, **96**, 1161–1170.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Aure,M.R. *et al.* (2015) Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer. *Genome Med.*, **7**, 21.

Aure,M.R. *et al.* (2017) Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. *Breast Cancer Res.*, **19**, 44.

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Ay,F. *et al.* (2014) Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res.*, **24**, 974–88.

Ay,F. and Noble,W.S. (2015) Analysis methods for studying the 3D architecture of the genome. *Genome Biol.*, **16**, 183.

Bansal,V. and Bafna,V. (2008) HapCUT: An efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, **24**, 153–159.

Bartel,D.P. (2009) MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, **136**, 215–233.

Ben-Elazar,Shay, Chor,B., Yakhini,Z., *et al.* (2016) Extending partial haplotypes to full genome haplotypes using Chromosomal Conformation Capture data. ECCB 2016, p. 9.

Ben-Elazar,S. *et al.* (2016) Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data. *Bioinformatics*, **32**.

Ben-Elazar,Shay, Chor,B., and Yakhini,Z. (2016) Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data. *Bioinformatics*, **32**, i559–i566.

Ben-Elazar,S. *et al.* (2013a) Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. *Nucleic Acids Res.*, **41**, 2191–2201.

Ben-Elazar,S. *et al.* (2013b) Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. *Nucleic Acids Res.*, **41**, 2191–2201.

Ben-Elazar,S. *et al.* (2019) The Functional 3D Organization of Unicellular Genomes. *Sci. Rep.*, **9**, 12734.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful

Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

van Berkum,Nynke L *et al.* (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.*, **6**, 1869.

van Berkum,Nynke L. *et al.* (2010a) Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *J. Vis. Exp.*

van Berkum,Nynke L. *et al.* (2010b) Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *J. Vis. Exp.*

Boutros,M. *et al.* (2006) Analysis of cell-based RNAi screens. *Genome Biol.*, **7**, R66.

Bovolenta,L.A. *et al.* (2012) HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, **13**, 405.

Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Capurso,D. *et al.* (2016) Discovering hotspots in functional genomic data superposed on 3D chromatin configuration reconstructions. *Nucleic Acids Res.*, **44**, 2028–2035.

Carlisle,C. *et al.* (2017) Chaperones and the Proteasome System: Regulating the Construction and Demolition of Striated Muscle. *Int. J. Mol. Sci.*, **19**, 32.

Castells-Roca,L. *et al.* (2011) Heat Shock Response in Yeast Involves Changes in Both Transcription Rates and mRNA Stabilities. *PLoS One*, **6**, e17272.

Chiang,C.-H. *et al.* (2013) Up-regulation of miR-182 by β-catenin in breast cancer increases tumorigenicity and invasiveness by targeting the matrix metalloproteinase inhibitor RECK. *Biochim. Biophys. Acta - Gen. Subj.*, **1830**, 3067–3076.

Chung,F. (1994) Spectral Graph Theory.

Cizeron-Clairac,G. *et al.* (2015) MiR-190b, the highest up-regulated miRNA in ERα-positive compared to ERα-negative breast tumors, a new biomarker in breast cancers? *BMC Cancer*, **15**, 499.

Cohen,J. (1977) Statistical power analysis for the behavioral sciences Academic Press.

Cook,P.R. (2010) A Model for all Genomes: The Role of Transcription Factories. *J. Mol. Biol.*, **395**, 1–10.

Dai,Z. and Dai,X. (2012) Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic Acids Res.*, **40**, 27–36.

Davis,R.A. *et al.* (2011) Remarks on Some Nonparametric Estimates of a Density Function. In, *Selected Works of Murray Rosenblatt*. Springer New York, New York, NY, pp. 95–100.

Denker,A. and de Laat,W. (2016) The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.*, **30**, 1357–82.

Diament,A. *et al.* (2014) Three-dimensional eukaryotic genomic organization is strongly correlated with codon usage expression and function. *Nat. Commun.*, **5**, 5876.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Duan,Z. *et al.* (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.

Dulmage,A.L. and Mendelsohn,N.S. (1958) Coverings of bipartite graphs. *Can. J. Math.*, **10**, 517–534.

Dvinge,H. *et al.* (2013) The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature*, **497**, 378–382.

Eden,E. *et al.* (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, 0508–0522.

Eden,E. *et al.* (2009a) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Eden,E. *et al.* (2009b) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Ehrenberg,A.S.C. (1982) A primer in data reduction : an introductory statistics textbook Wiley.

Eisenstein,M. (2015) Startups use short-read data to expand long-read sequencing market. *Nat. Biotechnol.*, **33**, 433–435.

Enerly,E. *et al.* (2011) miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors. *PLoS One*, **6**, e16915.

Faul,F. *et al.* (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods*, **39**, 175–91.

Floudas,C.A. (1995) Nonlinear and mixed-integer optimization : fundamentals and applications Oxford University Press.

Fridman,A. *et al.* (2013) Cell cycle regulation of purine synthesis by phosphoribosyl pyrophosphate and inorganic phosphate. *Biochem. J.*, **454**, 91–99.

Galperin,M.Y. *et al.* (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.

Gibbons,S.M. *et al.* (2018) Correcting for batch effects in case-control microbiome studies. *PLoS Comput. Biol.*, **14**, e1006102.

Git,A. *et al.* (2010) Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA*, **16**, 991–1006.

Glusman,G. *et al.* (2014) Whole-genome haplotyping approaches and genomic medicine. *Genome Med.*, **6**, 73.

Haakensen,V.D. *et al.* (2016) Subtype-specific micro-RNA expression signatures in breast cancer progression. *Int. J. Cancer*, **139**, 1117–1128.

Ham,J. *et al.* (2004) Kernel view of the dimensionality reduction of manifolds.

Hong,E.P. and Park,J.W. (2012) Sample size and statistical power calculation in genetic association studies. *Genomics Inform.*, **10**, 117–22.

Hong,Y. *et al.* (2016) miR-96 promotes cell proliferation, migration and invasion by targeting PTPN9 in breast cancer. *Sci. Rep.*, **6**, 37421.

Hu,J.-Y. *et al.* (2016) miR-601 is a prognostic marker and suppresses cell growth and invasion by targeting PTP4A1 in breast cancer. *Biomed. Pharmacother.*, **79**, 247–53.

Hu,J. *et al.* (2007) Replication-associated purine asymmetry may contribute to strand-biased gene distribution. *Genomics*, **90**, 186–194.

Iborra,A. *et al.* (1996) A Candida albicans gene expressed in Saccharomyces cerevisiae results in a distinct pattern of mRNA processing. *Microbiologia*, **12**, 443–8.

Iborra,F.J. *et al.* (1996) Active RNA polymerases are localized within discrete transcription "factories'' in human nuclei". *J. Cell Sci.*, **109**, 1427–1436.

Imai,J. *et al.* Proteasomes and molecular chaperones: cellular machinery responsible for folding and destruction of unfolded proteins. *Cell Cycle*, **2**, 585–90.

Jain,P. and Kar,P. (2017) Non-convex Optimization for Machine Learning.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Junier,I. *et al.* (2012) CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the β-globin locus. *Nucleic Acids Res.*, **40**, 7718–7727.

Junier,I. *et al.* (2010) Spatial and Topological Organization of DNA Chains Induced by Gene Co-localization. *PLoS Comput. Biol.*, **6**, e1000678.

Kanduri,C. *et al.* (2018) Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics*.

Kauffmann,A. *et al.* (2009) arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**, 415–416.

Kim,J.M. *et al.* (1998) Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res.*, **8**, 464–78.

Klocko,A.D. *et al.* (2016) Normal chromosome conformation depends on subtelomeric facultative heterochromatin in Neurospora crassa. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 15048–15053.

Koonin,E. V *et al.* (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.*, **5**, R7.

Kruskal,J.B. (1964a) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.

Kruskal,J.B. (1964b) Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, **29**, 115–129.

Kuleshov,V. *et al.* (2014) Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.*, **32**, 261–6.

Kwon,J.J. *et al.* (2019) A Systematic Review of miR-29 in Cancer. *Mol. Ther. oncolytics*, **12**, 173–194.

de Laat,W. and Duboule,D. (2013) Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, **502**, 499–506.

Lazar-Stefanita,L. *et al.* (2017) Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *EMBO J.*, **36**, 2684–2697.

Le,T.B.K. *et al.* (2013) High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science (80-.).*, **342**, 731–734.

Lee,J. *et al.* (2019) miR-106b-5p and miR-17-5p could predict recurrence and progression in breast ductal carcinoma in situ based on the transforming growth factor-beta pathway. *Breast Cancer Res. Treat.*, **176**, 119–130.

Leivonen,S.-K. *et al.* (2014) High-throughput screens identify microRNAs essential for HER2 positive breast cancer cell growth. *Mol. Oncol.*, **8**, 93–104.

Leivonen,S.-K. *et al.* (2009) Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines. *Oncogene*, **28**, 3926–36.

Lesurf,R. *et al.* (2016) Molecular Features of Subtype-Specific Progression from Ductal Carcinoma In Situ to Invasive Breast Cancer. *Cell Rep.*, **16**, 1166–1179.

Li,C. *et al.* (2019) MicroRNA-936 targets FGF2 to inhibit epithelial ovarian cancer aggressiveness by deactivating the PI3K/Akt pathway. *Onco. Targets. Ther.*, **12**, 5311–5322.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–93.

Lin,D. *et al.* (2018) Computational methods for analyzing and modeling genome structure and organization. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, e1435.

Liu,J. *et al.* (2018) Unsupervised embedding of single-cell Hi-C data. *Bioinformatics*, **34**, i96–i104.

Liu,Y. *et al.* (2017) Down-regulation of miR-29b in carcinoma associated fibroblasts promotes cell growth and metastasis of breast cancer. *Oncotarget*, **8**, 39559.

Luna,C. *et al.* (2009) Role of miR-29b on the regulation of the extracellular matrix in human trabecular meshwork cells under chronic oxidative stress. *Mol. Vis.*, **15**, 2488–97.

Van Der Maaten,L. *et al.* (2009) Tilburg centre for Creative Computing Dimensionality Reduction: A Comparative Review Dimensionality Reduction: A Comparative Review.

Mahy,N.L. *et al.* (2002) Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. *J. Cell Biol.*, **157**, 579–89.

Marbouty,M. *et al.* (2015) Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol. Cell*, **59**, 588–602.

McCoy,R.C. *et al.* (2014) Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. *PLoS One*, **9**, e106689.

Mead,A. (1992a) Review of the Development of Multidimensional Scaling Methods. *Stat.*, **41**, 27.

Mead,A. (1992b) Review of the Development of Multidimensional Scaling Methods. *Stat.*, **41**, 27.

Meagher,D. (1982) Geometric modeling using octree encoding. *Comput. Graph. Image Process.*, **19**, 129–147.

Mercy,G. *et al.* (2017) 3D organization of synthetic and scrambled chromosomes. *Science (80-.).*, **355**, eaaf4597.

Mestdagh,P. *et al.* (2014) Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nat. Methods*, **11**, 809–815.

Mita,P. and Boeke,J.D. (2016) How retrotransposons shape genome regulation. *Curr. Opin. Genet. Dev.*, **37**, 90–100.

Mizuguchi,T. *et al.* (2014) Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature*, **516**, 432–435.

Myhre,S. *et al.* (2010) In Silico Ascription of Gene Expression Differences to Tumor and Stromal Cells in a Model to Study Impact on Breast Cancer Outcome. *PLoS One*, **5**, e14002.

Ni,X. *et al.* (2014) Downregulation of miR-106b induced breast cancer cell invasion and motility in association with overexpression of matrix metalloproteinase 2. *Cancer Sci.*, **105**, 18–25.

Nora,E.P. *et al.* (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–5.

Nouri,H. *et al.* (2018) Multiple links connect central carbon metabolism to DNA replication initiation and elongation in *Bacillus subtilis*. *DNA Res.*, **25**, 641–653.

Novichkov,P.S. *et al.* (2013) RegPrecise 3.0 – A resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics*, **14**, 745.

Nurick,I. *et al.* (2018) Genomic meta-analysis of the interplay between 3D chromatin organization and gene expression programs under basal and stress conditions. *Epigenetics Chromatin*, **11**, 49.

Nygaard,P. and Saxild,H.H. (2005) The purine efflux pump PbuE in Bacillus subtilis modulates expression of the PurR and G-box (XptR) regulons by adjusting the purine base pool size. *J. Bacteriol.*, **187**, 791–4.

Nygaard,V. *et al.* (2016) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, **17**, 29–39.

Osborne,C.S. *et al.* (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.*, **36**, 1065–1071.

Parzen,E. (1962) On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.*, **33**, 1065–1076.

Patterson,M. *et al.* (2015) W hats H ap : Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.*, **22**, 498–509.

Paulsen,J. *et al.* (2013) Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Res.*, **41**, 5164–74.

Peng,Y. and Croce,C.M. (2016) The role of MicroRNAs in human cancer. *Signal Transduct. Target. Ther.*, **1**, 15004.

Pirola,Y. *et al.* (2016) HapCol: accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics*, **32**, 1610–7.

Rao,Suhas S P *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Rao,Suhas S.P. *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**, 1665–1680.

Rieder,D. *et al.* (2012) Transcription factories. *Front. Genet.*, **3**, 221.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.

Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Sanborn,A.L. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.*, **112**, E6456–E6465.

Sanyal,A. *et al.* (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.

Seber,G.A. (1984) Multivariate Observations Seber,G.A.F. (ed) John Wiley & Sons, Inc., Hoboken, NJ, USA.

Selvaraj,S. *et al.* (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.*, **31**, 1111–8.

Servant,N. *et al.* (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, **16**, 259.

Shinden,Y. *et al.* (2015) miR-29b is an indicator of prognosis in breast cancer patients. *Mol. Clin. Oncol.*, **3**, 919–923.

Sims,A.H. *et al.* (2008) The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis. *BMC Med. Genomics*, **1**, 42.

Snyder,M.W. *et al.* (2015) Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.*, **16**, 344–358.

Snyman,J.A. and Wilke,D.N. (2018) Practical Mathematical Optimization Springer International Publishing, Cham.

Sofueva,S. *et al.* (2013) Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J.*, **32**, 3119–29.

Spielman,D.A. (2007) Spectral Graph Theory and its Applications. In, *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, pp. 29–38.

Steinfeld,I. *et al.* (2013a) miRNA target enrichment analysis reveals directly active miRNAs in health and disease. *Nucleic Acids Res.*, **41**, e45–e45.

Steinfeld,I. *et al.* (2013b) miRNA target enrichment analysis reveals directly active miRNAs in health and disease. *Nucleic Acids Res.*, **41**, e45–e45.

Sutherland,H. and Bickmore,W.A. (2009a) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457–466.

Sutherland,H. and Bickmore,W.A. (2009b) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457–466.

Taddei,A. *et al.* (2010) The budding yeast nucleus. *Cold Spring Harb. Perspect. Biol.*, **2**, a000612.

Tahiri,A. *et al.* (2014) Deregulation of cancer-related miRNAs is a common event in both benign and malignant human breast tumors. *Carcinogenesis*, **35**, 76–85.

Tanizawa,H. *et al.* (2017) Architectural alterations of the fission yeast genome during the cell cycle. *Nat. Struct. Mol. Biol.*, **24**, 965–976.

Tanizawa,H. *et al.* (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.*, **38**, 8164–8177.

Tatebayashi,K. *et al.* (1998) Isolation of a Schizosaccharomyces pombe rad21ts mutant that is aberrant in chromosome segregation, microtubule function, DNA repair and sensitive to hydroxyurea: possible involvement of Rad21 in ubiquitin-mediated proteolysis. *Genetics*, **148**, 49–57.

Teixeira,M.C. *et al.* (2018) YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **46**, D348–D353.

Tewhey,R. *et al.* (2011) The importance of phase information for human genomics. *Nat. Rev. Genet.*, **12**, 215–23.

The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.

Thévenin,A. *et al.* (2014) Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic Acids Res.*, **42**, 9854–9861.

Tramm,T. *et al.* (2014) Development and Validation of a Gene Profile Predicting Benefit of Postmastectomy Radiotherapy in Patients with High-Risk Breast Cancer: A Study of Gene Expression in the DBCG82bc Cohort. *Clin. Cancer Res.*, **20**, 5272–5280.

Trask,B.J. *et al.* (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, **437**, 94–100.

Treut,G. Le *et al.* (2018) A polymer model for the quantitative reconstruction of 3d chromosome architecture from HiC and GAM data.

Varoquaux,N. *et al.* (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, **30**, i26-33.

Varoquaux,N. *et al.* (2015) Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic Acids Res.*, **43**, 5331–5339.

Wang,C. *et al.* (2011) miR-29b regulates migration of human breast cancer cells. *Mol. Cell. Biochem.*, **352**, 197–207.

Wang,M. and Xu,S. (2019) Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity (Edinb).*, **123**, 287–306.

Witten,D.M. and Noble,W.S. (2012) On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res.*, **40**, 3849–3855.

Xia,K. *et al.* (2018) Multiscale Persistent Functions for Biomolecular Structure Characterization. *Bull. Math. Biol.*, **80**, 1–31.

Xie,W. *et al.* (2018) miR-96 promotes breast cancer metastasis by suppressing MTSS1. *Oncol. Lett.*, **15**, 3464–3471.

Xu,T. *et al.* (2018) miRBaseConverter: An R/Bioconductor Package for Converting and Retrieving miRNA Name, Accession, Sequence and Family Information in Different Versions of miRBase. *bioRxiv*, 407148.

Yaglom,A.M. and Yaglom,I.M. (1987) Challenging mathematical problems with elementary solutions Dover.

Yang,S.-J. *et al.* (2017) The miR-30 family: Versatile players in breast cancer. *Tumor Biol.*, **39**, 101042831769220.

Yao,Y. *et al.* (2015) MiR-200b expression in breast cancer: a prognostic marker and act on cell proliferation and apoptosis by targeting Sp1. *J. Cell. Mol. Med.*, **19**, 760–769.

Ye,B.-C. *et al.* (2009) Time-Resolved Transcriptome Analysis of Bacillus subtilis Responding to Valine, Glutamate, and Glutamine. *PLoS One*, **4**, e7073.

Ye,F. *et al.* (2014) miR-200b as a prognostic factor in breast cancer targets multiple members of RAB family. *J. Transl. Med.*, **12**, 17.

Yeeles,J.T.P. *et al.* (2013) Rescuing stalled or damaged replication forks. *Cold Spring Harb. Perspect. Biol.*, **5**, a012815.

Yildirim,A. and Feig,M. (2018) High-resolution 3D models of Caulobacter crescentus chromosome reveal genome structural variability and organization. *Nucleic Acids Res.*, **46**, 3937–3952.

Zhang,L. The role of microRNA, mir-30d, in the initiation and progression of cancer.

Zhang,X. *et al.* (2017) MicroRNA-182 promotes proliferation and metastasis by targeting FOXF2 in triple-negative breast cancer. *Oncol. Lett.*, **14**, 4805–4811.

Zheng,Q. *et al.* (2017) miR-200b inhibits proliferation and metastasis of breast cancer by targeting fucosyltransferase IV and α1,3-fucosylated glycans. *Oncogenesis*, **6**, e358–e358.

Zheng,R. *et al.* (2015) Prognostic value of miR-106b expression in breast cancer patients. *J. Surg. Res.*, **195**, 158–165.

Zhou,L. *et al.* (2018) Identification of miR-18a-5p as an oncogene and prognostic biomarker in RCC. *Am. J. Transl. Res.*, **10**, 1874.

מבנה ותרומת התיזה

חיבור זה מבוסס על שלושת המאמרים הבאים. בזמן הכתיבה, שניים מתוך שלושת המאמרים מטה פורסמו בכתבי עת מדעיים, כאשר אחד מאלה הוצג בנוסף בכנס מוביל והמאמר השלישי נשלח לפרסום וממתין להערכת עמיתים.

1. **Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data**
   Shay Ben-Elazar, Benny Chor, Zohar Yakhini
   Published in *Bioinformatics 2016,* Presented as poster and orally at *ECCB 2016*

2. **The functional 3D organization of unicellular genomes**
   Shay Ben-Elazar, Benny Chor, Zohar Yakhini
   Published in *Nature Scientific Reports 2019*

3. **miRNA normalization enables joint analysis of several datasets to increase sensitivity and to reveal novel miRNAs differentially expressed in breast cancer**
   Shay Ben-Elazar, Miriam Ragle Aure, Kristin Jonsdottir, Suvi-Katri Leivonen, Vessela N. Kristensen, Emiel A.M. Janssen, Kristine Kleivi Sahlberg, Ole Christian Lingjærde and Zohar Yakhini
   Submitted to *PLOS Computational Biology 2019*

תקציר

לכידת תצורה כרומוזומלית (Chromosome conformation capture - 3C) ונגזרותיה (כגון High-throughput 3C, או Hi-C), הם אוסף שיטות ונהלים ניסויים המבוססים על טכנולוגיות ריצוף DNA אשר מייצרות מיפוי דליל של שכיחות זוגות רצפי קריאות על גבי כרומוזומים. שכיחות קריאות אלה פרופורציונים (בקירוב) לקירבה בין זוגות של מיקומים על גבי הכרומוזומים (Nynke L. van Berkum *et al.*, 2010a). גישות רבות ומגוונות משבצות את מספרי הקריאות מניסוי Hi-C למודלים תלת מימדיים מייצגים (באופן איכותי) על מנת ולהחליק רעשי דגימה ולספק הצצה אינטואיטיבית אל המבנה הגנומי אשר בבסיסם. 3C ושיטות הקשורות לו סללו את הדרך באופן אמפירי לשרטוט התכונות המבניות התלת מימדיות של גנומים בתאים חיים ובפירוט אשר לא נגיש לטכניקות מיקרוסקופיה נפחיות.

תגליות מרכזיות שניתן לייחס באופן חלקי ל-3C כוללות: TAD (topologically associated domain) - יחידת ארגון תפקודית הזוהתה כמנגנון אפיגנטי מבני אשר אוכף מגע בין מקדם ומעצם ומאפשר בידוד של שכונות גנומיות. ראיות שיטתיות להיפותזת "מפעלי השיעתוק". כלומר, גורמים שיעתוק משותפים ומנגנון השיעתוק מתגייסים לתתי-תאים בגרעין (אאוקריוטים) / גרעינון (פרוקריוטים) יחד עם המטרות הגנומיות שלהם.

בפרק 1.1 אנו מציגים גישה אשר ממנפת מידע אשר הושג על ידי לכידת תצורה כרומוזומלית על מנת להתייחס לבעית "מייל אחרון" בריצוף גנטי של קביעת הפלוטיפ. קביעת הפלוטיפ הינו התהליך שבו משייכים שונויות וסטיות ברצף חומצות הגרעין לאחת משתי העותקים ההומולוגים של כרומוזום. בעבודתנו (S. Ben-Elazar *et al.*, 2016) אנו מציגים שיטות אשר שימושיות עבור 1) פיצול מפות קרבה מנתוני Hi-C אשר מוצעו "באופן מסורתי" לכדי קבלת מפות Hi-C הכוללות נתוני שכיחות לכל זוג הומולוגים. 2) הפרדה (סידור) של מפות Hi-C הכוללות נתוני שכיחות להומולוגים ושיוך נכון של בלוקים הומולוגים אחד לשני. מפות Hi-C מופרדות שכאלה חשובות לשיפור דיוק וישימות של פירושים נוספים מנתוני Hi-C בעזרת ניתוחי המשך.

בפרק 1.2 אנו חוזרים בשנית להתמודד עם בעיית המשך בניתוח מודלים תלת מימדיים שנגזרו מנתוני Hi-C. בעבודה זו (Ben-Elazar *et al.*, 2019) אנו מפתחים תשתית אלגוריתמית וסטטיסטית לזיהוי תאים כדורים בחלל התלת מימדי אשר בתוכם רכיבים גנומיים בעלי תכונה ביולוגית משותפת מתמקמים יחדיו באופן מובהק סטטיסטית. גישה זו מתגברת על מגבלה של עבודה קודמת שלנו אשר בה תאים כדוריים אשר מועמדים לבדיקה חייבים להיות ממורכזים על גבי הגנום החד-מימדי. אנו מספקים ניתוח ריגורוזי של שיטה זו וממחישים את יתרונה בלזהות תבניות ייחודיות אשר מספקות פירושים ביולוגים מתקבלים על הדעת. אנו מתארים ממצאים במספר אורגניזמים.

מיקרו-חומצות-ריבונוקלאיות (MicroRNAs (miRNAs הן מולקולות RNA אשר לרוב מבצעות פעילות כלשהי למרות שהן לא עוברות תרגום לחלבון. מיקרו-RNA התפתחו אבולוציונית בכדי לשחק תפקיד כמנגנון בקרה של

ביטוי גנים וכמו-כן בפעילות וויסות המערכת החיסונית. מיקרו-RNA התגלו כמעורבים בדיכוי וכמו כן בזירוז התפתחות של גידולים ממאירים בתלות לתנאי סביבה שונים (Peng and Croce, 2016). ישנו עניין מיוחד באפיון מדויק של הקשר שלהם לתתי סוגים של סרטן ובתור ביו-סמנים פוטנציאלים להנעת טיפולים קלינים מותאמים אישית.

בפרק 1.3 אנו מציגים שיטה לנירמול ולאנליזה משולבת של נתוני ביטוי מיקרו-RNA. השיטה שלנו מקלה על תופעות הקשורות לאפקט אצוה (batch). אנו מפעילים שיטה זו בכדי לנתח באופן משולב ארבעה עקבות של נתוני ביטוי מיקרו-RNA בסרטן שד, מציגים ביו-סמנים פוטנציאלים חדשים ומתדיינים ביתרונות הסטטיסטיים של הגישה שלנו. בנוסף, אנו מתארים תצפיות מסויימות אשר לא היו צפות ללא הנירמול.

בחלקים השונים של פרק זה אנו מספקים רקע נוסף וסוקרים לעומק את כל אחת מהשאלות המוזכרות לעיל. בפרקים שלאחר מכן אנו מציגים שיטות חישוביות על מנת לנסות ולענות על שאלות אלה ומנתחים לעומק ומתדיינים על ממצאים פוטנציאלים חדשניים שהשיטות שלנו הציפו. לבסוף, אנו מסכמים את החיבור בדיון על העבודה ובהצעות להמשכה במחקרים עתידיים.

1.1 Hi-C וקביעת הפלוטיפ

קביעת הפלוטיפ הינו תהליך קבלת ההחלטה על שיוכם של שינויים גנומיים שזוהו לאורך כרומוזומים הומולוגים לעותק הכרומוזומלי הפיזי שלהם ביצורים דיפלואידים או פוליפלואידים. קיימות מספר שיטות לקביעת הפלוטיפ אשר נעות בין גישות המבוססות על ניתוח גנטי של אוכלוסיות ובכך זקוקות למידע גנומי של הגנומים של ההורים, ועד גישות מתוחכמות ובעלות עלות זמן גבוהה של בידוד מולקולות להפרדת עותקי הכרומוזומים פיזית לפני ריצוף בתפוקה גבוהה של כל עותק לחוד.

אנחנו פיתחנו צינור עיבוד נתונים אשר משלב נתוני Hi-C עם מידע חלקי לגבי הפלוטיפים על מנת להסיק את ההפלוטיפ המלא וכמו כן מפות Hi-C מופרדות לפי ההפלוטיפ (Shay Ben-Elazar, Chor, Yakhini, *et al.*, 2016). בעבודה זו שיבצנו מיקומים גנומיים על סמך תיקון למפות Hi-C והשתמשנו בייצוג זה בכדי להפעיל שיטה לייצירת חיץ באופן חמדני אשר מובילה לפיענוח השיוך של אללים שונים שהתקבלו מניתוח הפלוטיפ חלקי לעותק ההומולוגי הנכון שלהם. בניסוח סטטיסטי של הבעיה אנחנו מוכיחים שהפתרון שלנו מוביל לקונפיגורציה שממקסמת את הניראות באופן גלובלי. אנחנו מראים ששיבוץ מידע Hi-C מהווה מדד טוב יותר לפיענוח ההפלוטיפ, דבר אשר מרמז לכך ששיבוץ הוא צעד הכרחי בהחלקת נתוני Hi-C אשר יכולים להיות דלילים ורועשים. עובדה זו עלולה להיות נכונה וחשובה בצינורות עיבוד נתוני Hi-C למטרות שונות. בנוסף, אנו מנתחים קריאות רצפים אשר חופפות מיקומים הומוזיגוטיים (מונו-אלליים), מקומות שלא ניתן לשייך לעותק האמהי או אבהי ולא רק קריאות רצפים אשר חופפות אתרים הטרוזיגוטיים (בי-אלליים) באדם בו נערכה המדידה. רצפים

מונו-אלליים שכאלה משוייכים באופן רך תחת הנחה של הסתברות פריורית אחידה על גבי העותקים ההומולוגים הרלוונטים עבורו. שיטות עדכניות רבות נוהגות להתעלם מקריאות שכאלה בניתוח נתוני Hi-C.

בפרק 2 אנחנו חוקרים את הבעיה של שחזור מידע Hi-C מופרד לפי הפלוטיפ ושחזור הפלוטיפ מלא מנתוני הפלוטיפ חלקיים ומידע Hi-C. הפתרון שהצענו מודגם על ידי ניתוח הדיוק שלו על מידע Hi-C מבן אנוש, דיפלואידי, ונתוני אמת אשר הושגו מקביעת הפלוטיפ בשלשה (Auton et al., 2015). התוצאות שלנו מראות שהשיטה שהצענו מובילה להפלוטיפים אשר מסכימים עם נתוני האמת ב-98% (בממוצע על גבי הכרומוזומים). אנו מראים ערך מוסף אפשרי בניתוח נתוני Hi-C דיפלואידים מופרדים בעזרת אנליזה של מיקום משותף אשר מראה תבניות מיקום משותף שבהן גנים מעותקים של כרומוזומים הומולוגים שונים שוכנים בתוך מפעל שיעתוק אפשרי. למען שלמות העבודה, אנו מספקים ישירות לאחר פרק 2 תוספת ובה הגדרות מתמטיות ריגורוזיות יותר לבעיה הגיאומטרית שניסינו לתת לה מענה במאמר.

1.2 Hi-C, העשרה מרחבית ומפעלי שיעתוק

בפרק 3 אנו דנים באלגוריתמים וסטטיסטיקות לאמידה של העשרה מרחבית של תכונה בינארית הניתנת על נתונים מאורגנים מרחבית (קואורדינטות ב-R^2, R^3). פיתחנו שיטה לזיהוי של מיקומים במרחב דו או תלת מימדי אוקלידי אשר סביבן תת קבוצה מסויימת של אלמנטים ממוקמים בצפיפות גבוהה באופן משמעותי סטטיסטית. בחנו את תוקף והיעילות של השיטה הזו על נתונים מלאכותיים ויישמנו אותה על שיבוצים של נתוני Hi-C ממספר אורגניזמים חד-תאיים בתלת מימד ולאורך אנוטציות גנומיות מרובות (Ben-Elazar *et al.*, 2019). אנו משווים את השיטה הזו לבחינה ישירה של מספרי זוגות קריאות רצפים בין איזורים גנומיים ודנים ביתרונותיה.

מחקרים קודמים שנערכו על ידנו ועל ידי אחרים הציעו היוריסטיקות לביצוע ניתוח נתוני Hi-C לזיהוי העשרה מרחבית. בעבודה זו חקרנו הגדרה רשמית ריגורוזית של בעיית ההעשרה המרחבית. אנו מציגים ראיות משכנעות אשר תומכות במתודולוגיה שלנו ביחס לאלו שהשתמשו בהן לפנינו ומפעילים את השיטה שלנו לקבלת תוצאות משמעותיות באופן סטטיסטי אשר מרמזות על תגליות ביולוגיות חדשות.

קיים עניין רב בהפעלת ניתוח העשרה מרחבית על מנת לקבל אפיון יותר מדיוק של ועל מנת לזהות מפעלי שיעתוק. מפעלי שיעתוק הם מנגנון בקרת ביטוי גנים אשר מתבטא באיזורים תחומים במרחב הגרעין שבהם מכונת השיעתוק מגייסת רכיבי בקרה ורצפים גנומיים על מנת לבקר את ביטויה של תוכנית פעילות תאית מסויימת (F. J. Iborra *et al.*, 1996; Sutherland and Bickmore, 2009a). מחקרים קודמים ניסוי לאמוד באופן סטטיסטי את קיומם של מפעלי שיעתוק. הכותבים של (Dai and Dai, 2012) השוו את מספר האינטראקציות בין קבוצות גנים בעלי פעילות משותפת וזיהוי העשרה סטטיסטית תחת מודל בעל השערת אפס היפרגאומטרית לאינטראקציות בין גנים שהם מטרות של פקטורי שיעתוק מסויימים.

מצד שני, מחקר עוקב (Witten and Noble, 2012) טען כי קשתות בגרף המושרה על ידי ניסוי ה-Hi-C הן לא בלתי תלויות סטטיסטית, כפי שהמודל ש Dai and Dai השתמשו בו מניח, וכי לפי כן ארועי העשרה מרחבית יספרו באופן רב מכפי שהנחת מודל תקינה הייתה מספקת. על מנת להציע תיקון לבעיה זו, Witten ו-Noble הפעילו מתודולוגיית דגימה אשר תחתה לא הופיע אות שמצביע על העשרה מרחבית למטרות של פקטורי שיעתוק. הגישה שלנו, אשר הפעלנו ב- (Ben-Elazar *et al.*, 2013b) וב- (Shay Ben-Elazar, Chor, Yakhini, *et al.*, 2016), נמנעת מהשוואה בין אוכלוסיות של דגימות קירבה לגמרי, ובכך נמנעת מבעיות תלות סטטיסטיות אשר צצות בשיטות הקודמות. במקום זאת, אנו מתמקדים במרחקים לנקודת ציר מסויימת – נקודת ייחוס אשר סביבה מודדים העשרה מרחבית באופן סטטיסטי כפי שמתואר בהמשך.

במחקר קודם זיהינו אתרים מועמדים להיות מפעלי שיעתוק על ידי פיתוח מודל סטטיסטי מבוסס על המבחן הסטטיסטי של ההיפרגיאומטרי המינימלי, mHG (Eden *et al.*, 2007, 2009a). בפירוט, קחו בחשבון מיקום גנומי, l. דרגו את כל המיקומים הגנומיים האחרים $l_1, \dots, l_N$ בעזרת פונקציית מרחק ל-l, $d(l_i, l)$. בהינתן פקטור שיעתוק וסט גני המטרות שלו, T, כלומר המיקומים הגנומים שבהם שיעתוק מבוקר על ידי פקטור השיעתוק. נגדיר ווקטור בינארי, λ, באורך N, שבו $\lambda(i) = 1$ אם ורק אם $l_i \in T$. עבור $1 \leq n \leq N$ נגדיר $\Lambda_n =$ $[\lambda(1), \dots, \lambda(n)]$ בתור הרישא באורך n של הווקטור הבינארי λ. יהי $b_n = \Sigma\Lambda_n$, $B = \Sigma\Lambda_N$. התוצאה של mHG מוגדרת על ידי הסף n אשר מביא למקסימום את הזנב הימני של פונקציית ההתפלגות המצטברת ההיפרגיאומטרית כשנצפים b_n ערכים. כלומר,

$$mHG(\lambda) = \min_{1 \leq n \leq N} \sum_{i=b_n}^{\min(n,B)} \frac{\binom{n}{i}\binom{N-n}{B-i}}{\binom{N}{B}}$$

השערת האפס ב-mHG היא שבהינתן מספר האחדות B, הם מפולגים באופן שווה לאורכו של הווקטור הבינארי באורך N. בהקשר שלנו, לדחות את השערת האפס מרמז על כך שהמטרות של פקטור השיעתוק ממוקמות בסמיכות לנקודת הציר שבחרנו בריכוז מפתיע באופן סטטיסטי. אנו חוזרים על ניסוי זה לכל נקודות הציר לאורך הגנום ועבור כל פקטורי השיעתוק, ומתקנים השערות מרובות בעזרת תיקון Bonferroni.

הסטטיסטי של mHG שימש בכדי למדוד את ההסתברות שבה דירוג נצפה של גנים על פי מרחקם מנקודת הציר מציף מספר 'לא סביר' של גנים שהם מטרות ידועות של פקטור השיעתוק לראש רשימת הגנים המדורגת. הערך הנצפה של הסטטיסטי מוארך כנגד מודל רקע של פרמוטציות על גבי רשימת הגנים, אשר עליו אנו מוסיפים בקרות בכדי לשלוט בהשפעות הסדר של הגנים לאורך הגנום החד-מימדי ועל מנת לבודד את האות מהשערה מרחבית ממשית. אילוסטרציה של שיטה זו מוצגת באיור 1 אשר נלקח מ (Ben-Elazar *et al.*, 2013b).

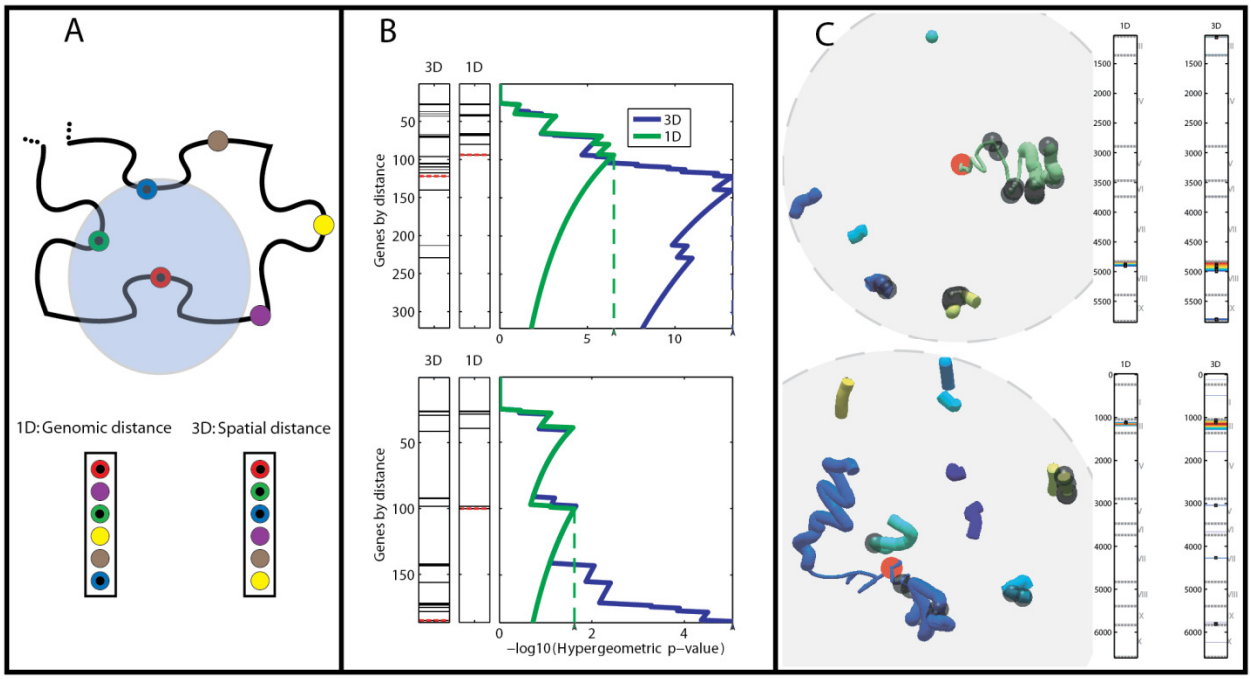


Figure 1. Comparing functional enrichment between the genomic and spatial regions of the genome. (A) Two genomic distances. The schematic shows the gene neighborhood surrounding a particular gene (red). The neighboring genes may be ranked by their genomic proximity (left) or their spatial proximity (right). (B) Detecting areas of enrichment for TF-cohorts. In ranked gene lists, generated by either genomic or spatial proximity, the genes annotated as targets of a particular TF are indicated as black lines. The p-value of the enrichment of the targets for each threshold is indicated on the right. The threshold with the best p-value is indicated by the dashed line (see Methods). This analysis is shown for two genomic loci surrounding genes YCL012C and YHL050C respectively and querying for targets of GLN3. (C) Local structures of the two loci examined in B. Colors indicate distinct yeast chromosomes. The red circles indicate the center gene around which co-localization was tested. The center genes shown are YCL012C (top) and both YHL050C and YHL050W-A (bottom). The content shown in each sphere is the environment which corresponds to the mHG threshold, dictated by the most enriched spatial environment for GLN3 targets. Bars on the right mark the loci along the linear genome which participate in the most enriched environment by both the genomic and spatial rankings. Black dots, both in the bars and the visualized structure, indicate gene targets of GLN3.

בגישה המוצגת לעיל, סרקנו גנים לאורך הגנום החד-מימדי בתור נקודות ציר וזיהינו מיקום והופעת מפעלי שיעתוק אפשריים על ידי מדידת ההעשרה המרחבית בתלת מימד סביב כל נקודת ציר שכזו. עם זאת, מפעלי שיעתוק לא מוכרחים להיות ממורכזים סביב גן או אפילו סביב נקודת ציר שנמצאת על גבי הגנום החד-מימדי, כפי שאנו מראים במאמר. בעבודה זה הרחבנו את הישימות של בדיקת ההעשרה המרחבית שתיארנו לעיל בכדי להקל על הדרישה שנקודות ציר יהיו על גבי הגנום החד מימדי.

לצאת מהמרחב הדיסקרטי של נקודות ציר אפשריות לאורך הגנום החד-מימדי על מנת לכסות את כל נקודות הציר האפשריות בתלת מימד זו בעיה קשה באופן כללי מפני שיש מספר אינסופי של נקודות ציר אפשריות. למרות כן, מכיוון שמבחן ההעשרה הסטטיסטי שלנו מבוסס על סדרי הדירוג של הגנים ולא על המרחקים הממשיים בינהם, אין צורך לבדוק את כל נקודות הציר האפשריות. אנו מראים כי ישנו מספר פולינומי של קבוצות של נקודות ציר שכל קבוצה משרה דירוג אחר על הגנים ובכך ערכי mHG אפשריים שונים. בעבודתנו אנו מאפיינים את המרחב הקומבינטורי אשר תחת בעיה זו באופן מדוייק ומספקים אלגוריתם מקוון בשיטת הסתעף-וחסום על מנת לסרוק נקודות ציר רלוונטיות שרירותיות.

האלגוריתמיקה שלנו מסתמכת על תכונות של ההתפלגות ההיפרגאומטרית על מנת לדלג על איזורים מועמדים באופן יעיל ולחדד באופן רקורסיבי העשרות פוטנציאליות. בהמשך לעבודתנו הקודמת, הפעלנו שיטה זו בכדי לנתח נתוני Hi-C ולגלות נקודות במרחב שמהוות מפעלי שיעתוק אפשריים בהסתמך על תצורות תלת מימדיות. הערכנו את שיטה זו על מספר קבצי נתונים בנוסף לנתונים תלת מימדיים מלאכותיים. חשוב מכך – יישמנו את השיטה על נתוני Hi-C מאורגניזמים חד-תאיים ונדונו בכמה תוצאות ביולוגיות מעניינות: תופעת הגדלה של מספר עותקי הרצפים הפרי-טלומריים מרוכזת במרחב במוטאנט הנוקאאוט של Rad21, דבר אשר מצביע על קשר עמוק בין קומפלקס Cohesin מתפקד לבין שלמותו של הרצף הפרי-טלומרי. תופעה זו מוצגת באיור 2 שנלקחה מהמאמר שלנו. ריכוז משותף במרחב של גנים האחריים לשכפול הגנום מועתקים לשני עותקים אשר בסמיכות גבוהה לאיזורי ה-*ori* וה-*ter*, עובדה המספקת ראייה להתפתחות אבולוציונית של תבנית "גיבוי" היכולה לשמש בכדי להציל שכפול מושהה, וכו'.

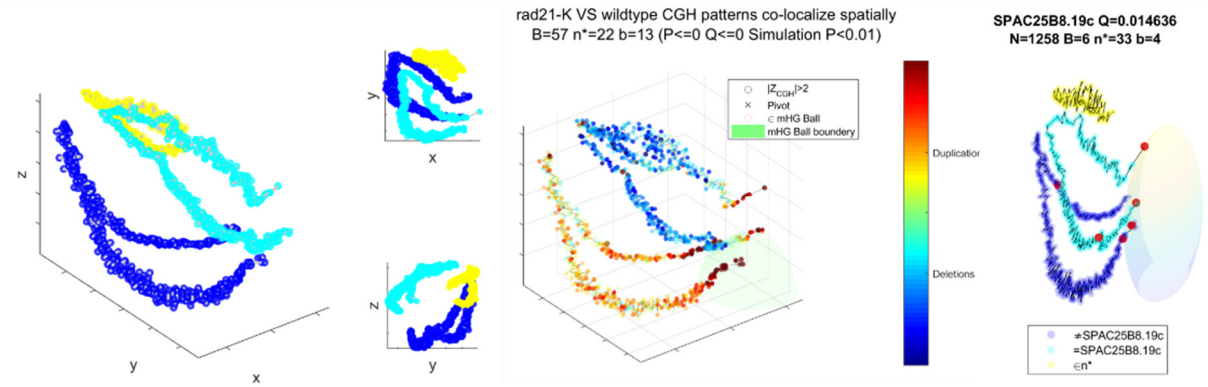


Figure 2. example of spatial co-localization identified by our method. Left: sNMDS embedding for S. pombe *with colour coded chromosomes. Middle (animation available as Supplementary Video 5): Bins are colour coded by average aCGH value, with marked outliers (opaque red for Z>2 and blue for Z<-2). We can observe a weak duplication signal on ChrII, and deletion on ChrI, ChrIII. Strongest duplication is evident at the telomeres. Right (animation available as Supplementary Video 6): Red bins contain Loz1 transcription factor targets. The resulting $smHG$ pivot and corresponding ball are visible containing 4/6 TF targets.*

1.3 miRNA וניתוח משולב של נתוני ביטוי

פרק 4 מתאר גישה מותאמת המבוססת על נורמליזציה בשברונים (AQN) למטרה של ניתוח משולב של נתוני ביטוי miRNA ממספר ניסויים אשר נדגמו בעזרת טכנולוגיות שונות. מיקרו-RNA (miRNAs) הם מולקולות RNA קטנות (~22 נוקלאוטידים) לא מקודדות לחלבון, אשר נקשרות לאתרי מטרה ספציפיים שלרוב נמצאים בצד ה-3' הלא מתורגם (UTRs) של RNA שליח המטרה שלהן. על ידי הקשרות זו, miRNA מבקרים את רמות הביטוי בעזרת דיכוי פעילות תרגום או על ידי הגרעה של mRNA. קביעת פרופיל הביטוי של miRNA הינו כלי חשוב למחקר הביולוגיה וסיווג של גידולים והתגלה כחשוב באבחון וקביעת פרוגנוזה.

ניתן לחלק לשני משפחות את הגישות השונות בניתוח משותף של נתוני ביטוי ממקורות שונים (בעלי הטיות התלויות במקור): ניתוח-על וניתוח משולב. בניתוח-על אנו בוחנים כל קובץ נתונים באופן נפרד ומשלבים את

התוצאות על מנת לקבל מסקנות יותר חסינות סטטיסטית. ניתוח-על נחשב כמרוויח פחות מיתרונת של תוספת לכוח הסטטיסטי שמתקבלות מהגדלת גודל המדגם ביחס לניתוח משולב. מנגד, ניתוח משולב מנסה להתגבר על תופעות אצווה על ידי הסטה של התפלגויות רמות הביטוי מניסויים שונים כך שהם יהיו ברי השוואה בהתחשב בהנחות מסויימות.

אנו מפתחים גרסה מכומתת ומורטטת של נורמליזציה בשברונים (Quantile normalization), בשם AQN, אשר מורידה השפעות של תופעות אצווה הבאות לידי ביטוי באשכול. אנו מראים שכשמצמידים את הניתוח עם מבחנים סטטיסטים מתאימים לאנליזות שניוניות, השיטה שלנו מציפה יותר miRNAs שמבטאים רמות ביטוי דיפרנציאליות בין חולים בעלי קולטני אסטרוגן בתאי הגידול לעומת כאלו שאין להם קולטני אסטרוגן, ובפרט miRNA שלא תואר בספרות, has-miR-193b-5p, שנמצא כמדכא גידולי. הגישה שלנו מספקת רמות ביטוי אשר במתאם גבוהה יותר עם הביטוי של גני המטרות שלהן, והעשרה גבוהה יותר ל-GO (אונטולוגיות של גנים) עבור מונחים שנראים שייכים לפי מחקרי תצפיות. אנו משווים את השיטה שלנו לשיטות אשר משתמשים בהם בתחום לנירמול ומציגים מספרי קווי ראיות בעדה.

תמצית

גנומים מאכסנים ומקודדים אוסף הוראות שמטרתן להורות על אופן הייצור והבקרה של גנים. גנים משפיעים ופועלים הדדית זה על זה ועם סביבתם במטרה להכתיב את הפנוטיפ והפעילות של תאים ביולוגים. הבהרת דרכי הפעולה של גנומים יכולה להוביל לחדשנות ברפואה, חקלאות, אקולוגיה ואף על דרכי קידוד מידע דיגיטלי וחישוב. קידמה טכנולוגית ונהלי ניסוי חדשניים ופורצי דרך מובילים, באופן שוטף, לאתגרים בהעלאת סדרי הגודל ובפירוש נכון של מדידות ותצפיות גנומיות. תיזה זו מתייחסת להיבטים של ניתוח נתונים משני טכניקות מדידה ניסויות מביולוגיה מולקולרית, ובפרט מתרכזת באתגרים מסויימים בהקשר של פירוש תוצאות אלה. אנו מציגים שלושה פרויקטי מחקר נפרדים בעלי מטרה משותפת אחת – פיענוח תכונות גנומיות וחוץ-גנומיות (אפי-גנומיות) מנתוני מדידה. אנו מציגים אלגוריתמים חדשניים, נותנים להם מוטיבציה בעזרת סימולציות, מפעילים אותם על על נתוני אמת ומציגים ראיות סטטיסטיות ופרשנות ביולוגית לממצאים מניתוח התוצאות. הגישות עליהן אנו דנים בתיזה זו מקדמות את העדכני ביותר (state-of-the-art) בתחומם ומספקות תובנות חדשות על מאפיינים גנומיים וחוץ-גנומיים של תאים ופעילותם התפקודית. בפרט, התרומה של עבודה זו כוללת:

- גישה לשימוש בנתוני Hi-C להסקה של הפלוטיפים מגנוטיפים מרובבים.
- גישה סטטיסטית לאפיון הארגון הפונקציונלי של גנומים באורגניזמים חד-תאיים בתלת-מימד בעזרת נתוני Hi-C.
- גישה חדשנית לנירמול נתוני miRNA אשר מאפשרת שילוב של מספרי ערכות נתונים ומובילה להגדלת הכוח הסטטיסטי.

הפקולטה למדעים מדוייקים ע"ש סאקלר
ביה"ס למדעי המחשב ע"ש בלווטניק

שיטות סטטיסטיות-חישוביות למחקר תכונות מרחביות של מבנים גנומיים

חיבור לשם קבלת התואר
"דוקטור לפילוסופיה"
מאת

שי בן-אלעזר

המחקר בתיזה זו בוצע תחת הנחייתם שם

פרופ' בני שור

פרופ' זהר יכיני

הוגש לסנאט של אוניברסיטת תל אביב
נובמבר 2019