Computational methods and reproducible analysis in regulatory genomics

Pablo Monteagudo Mesas

München 2023

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig–Maximilians–Universität München

Computational methods and reproducible analysis in regulatory genomics

Pablo Monteagudo Mesas

aus

Barcelona, Spain

2023

Erklärung:
Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Dr. Stefan Canzar betreut.

Eidesstattliche Versicherung:
Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 16.08.2023

Pablo Monteagudo Mesas

Dissertation eingereicht am 27.06.2023
1. Gutachter: Dr. Stefan Canzar
2. Gutachter: Prof. Dr. Johanna Klughammer
Mündliche Prüfung am 01.08.2023

Contents

Acknowledgments **ix**

Summary **xi**

1 Introduction **1**
1.1 Genome regulation . . . 2
1.1.1 Gene Expression . . . 3
1.1.2 Sequencing technologies . . . 4
1.2 Transcriptional and post-transcriptional regulation . . . 5
1.2.1 Transcriptional regulation . . . 5
1.2.2 Post-transcriptional regulation . . . 6
1.3 Gene expression dynamics . . . 7
1.3.1 Single-cell RNA-seq and trajectory inference (TI) methods . . . 7
1.3.2 Alignment of complex single-cell trajectories . . . 8
1.4 List of peer-reviewed articles . . . 10
1.5 List of pre-print articles . . . 10

2 Ccr4–Not complex reduces transcription efficiency in heterochromatin **11**
2.1 Introduction . . . 12
2.2 Methods . . . 13
2.2.1 t-distributed stochastic neighbor embedding (t-SNE) . . . 13
2.2.2 IP-Input subtraction in ChIP-seq data . . . 14
2.2.3 Transcription Efficiency, RNA Stability and Pol II Occupancy . . . 14
2.2.4 Calculation of pathways' contributions to silencing . . . 15
2.3 Results . . . 16
2.3.1 Effect of heterochromatin on silencing pathways . . . 18
2.3.2 Contribution of heterochromatin factors to distinct silencing pathways 19
2.3.3 Contribution of individual proteins to the heterochromatic silencing pathways . . . 23
2.4 Discussion . . . 26

3 Dynamic pseudo-time warping of complex trajectories **29**
3.1 Introduction . . . 30
3.2 Preliminaries . . . 32
3.2.1 Dynamic time warping . . . 32
3.2.2 Arboreal matchings . . . 33

3.2.3 Computing arboreal matchings with Trajan 35
3.3 Methods 38
3.3.1 Overview of the method 38
3.3.2 TrajanR: integration with dynverse 39
3.3.3 Experiments using simulated data with Dyngen 41
3.4 Results 44
3.4.1 From paths to trees: DTW vs arboreal matchings 44
3.4.2 Metric conformity 49
3.4.3 Experiments: Real Data 53
3.5 Conclusion 64

4 Conclusion and outlook **65**
4.1 Conclusion 65
4.2 Outlook 66
4.2.1 Ccr4–Not complex reduces transcription efficiency in heterochromatin 66
4.2.2 Dynamic pseudo-time warping of complex trajectories 67

Appendix A Supplementary Material **69**
A.1 Ccr4-Not complex reduces transcription efficiency in heterochromatin 69
A.2 Dynamic pseudo-time warping of complex trajectories 83

Acknowledgments

First and foremost, I offer my sincerest gratitude to my supervisor Dr. Stefan Canzar. I appreciate all the help, guidance, and encouragement I have gotten from you over the years, as well as the chance to work on such interesting projects.

I want to thank Dr. Mario Halic, Dr. Cornelia Brönner and Dr. Maria Spletter for all their work, input, and discussions throughout our respective collaborations. I would like to express my appreciation to all members in my thesis defense committee for their valuable feedback and insightful suggestions. In particular, I would like to thank Prof. Dr. Johanna Klughammer for all of her assistance, caring about me and my colleagues, and integrating us into her group.

I would like to express my big thanks to past and current members of the Canzar Lab for their support. Especially to Francisca, Parastou, Israa, Shounak, Hoan, Shuang, and Alice, I can not stress enough how important you have been to keep me going through these last years. I would also like to thank everyone in the Klughammer Lab for the fun scientific discussions, social interactions, and always being friendly and welcoming. In addition, I would like to thank the Graduate School of Quantitative Biosciences Munich (QBM) and the Gene Center Munich for their financial support and for creating such an excellent scientific education environment that allowed connecting with other researchers.

To all my friends, whether in Germany, Spain, or anywhere else in the world, thank you so much for all the good times and the ones to come. Thanks to my family, especially my parents, Manuel and Angela, for their love, encouragement, and support in shaping who I am today. And to my "no longer so little" cousins, Unai and Ainara, for showing me the things that truly matter in life. Finally, a special thanks goes to Anna for her patience and support and for sharing this Munich adventure with me. I am sure we will forever treasure our time here.

Summary

While the modern field of genomics is exploding with new experimental techniques that push the limits of what is possible, computational methods designed to process and extract useful information from these data try to keep up in order to reliably improve our understanding of gene regulation. In this thesis, we developed reproducible computational pipelines and algorithms to study gene regulation at different levels or stages. We begin by reviewing core ideas required for understanding cells and their regulatory mechanisms and the main experimental sequencing-based assays used to characterize biological systems at the molecular level. We then introduce the two separate projects that comprise this thesis.

In our first project, we dissect the contribution of three competing pathways involved in heterochromatic silencing, namely, Pol II occupancy (PO), transcription efficiency (TE), and RNA stability (RS), by comparing heterochromatic and euchromatic regions in Schizosaccharomyces pombe (S.pombe). We characterize each of these regulatory pathways as ratios between expression levels of corresponding high-throughput sequencing assays, PO (^{mu}ChIP-seq/^{wt}ChIP-seq), TE (RIP-seq/ChIP-seq) and RS (pA-RNA/RIP-seq), and quantify the relative effects that mutants lacking core components associated with each pathway (i.e., chromatin modifiers, RNAi, and RNA degradation) have on heterochromatic silencing.

In our second project, we study how dynamic biological processes, such as development, are regulated and can be characterized at the molecular level by complex (non-linear) single-cell RNA sequencing (scRNA-seq) trajectories, focusing on how such processes can be compared using our novel tool Trajan. We introduce TrajanR, our accompanying R package, that facilitates the standardization and pre-processing of Trajan input data, trajectory inference, and alignment computation under different parameter schemes and provides various visualization options, enabling the analysis of scRNA-seq trajectories in complex settings. We demonstrate the accuracy of Trajan's alignments through extensive experimentation on simulated data. We also showcase how our TrajanR package facilitates the study of scRNA-seq data based on the analysis of two independent real-world datasets.

Finally, we conclude with a discussion of both projects presented in this thesis and an outlook for the future.

Chapter 1

Introduction

The *cell* is the fundamental unit of all living organisms; it provides structure to the organism, absorbs and converts nutrients into energy, and executes other specialized functions. Since its discovery by Robert Hooke (Hyllner et al., 2015) in 1665, made possible by the earlier invention of the microscope, there have been substantial efforts to categorize the many human and non-human cell types and link them to their distinct functions. Initially, efforts were limited to morphological features accessible via light microscopy, such as size and shape, which are still significant predictors of cellular function. However, over time, other experimental approaches, such as stainings and immunochemical assays, were developed to further characterize tissues and cell types.

In the latter half of the 20th century, following the discovery of the genome and its structure and, ultimately, the construction of the first human reference genome in 2003 (Lander et al., 2001; Venter et al., 2001), a plethora of sequencing techniques were created to advance our understanding of the inner workings of cells and the genetic code. In the last decades, the development of high-throughput, next-generation sequencing technologies, such as RNA sequencing (RNA-seq) (Mortazavi et al., 2008) and single-cell RNA sequencing (scRNA-seq) (Tang et al., 2009), has enabled the unbiased characterization of tissues and cellular programs at the molecular level, while related sequencing-based biochemical assays, such as chromatin immunoprecipitation sequencing (ChIP-seq) and RNA immunoprecipitation sequencing (RIP-seq), have aided in the understanding of genomic regulatory mechanisms by targeting specific, DNA-protein and RNA-protein interactions, respectively.

We now know that two cells in an organism can differ in many ways. Nevertheless, they both possess identical *genetic material* (i.e., DNA sequence) carrying the full set of instructions to synthesize all the RNA molecules and proteins the organism will ever need. The differences between these cells arise at multiple levels, in a cell-type and developmental-stage-specific manner, where not only the code but also its tight regulation is critical. All these complex regulatory mechanisms can be summarized under the term: *genomic regulation*, which is diagrammatically depicted in Figure 1.1 and extends the principles of possible information flow as first stated by the central dogma of molecular biology (Crick, 1958; Hewitt, 2020), as discussed below.

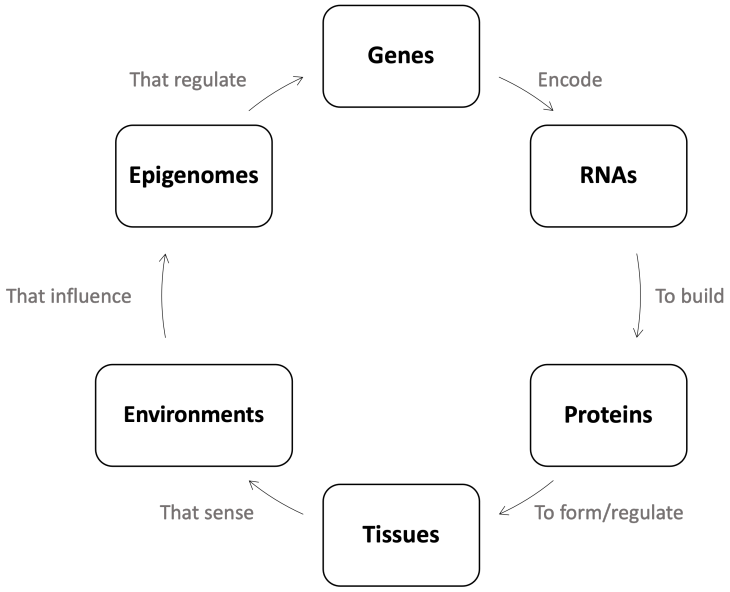

Figure 1.1: Flow of biological information. Adapted from (Mukherjee, 2016)

In this thesis, we developed computational methods to study different aspects of gene regulation based on various high-throughput sequencing technologies. In the following Sections, we introduce some of the basic mechanisms used by cells to regulate the expression of genes. More biological sections in this chapter are based on the classic reference in cell biology (Alberts et al., 2022); see the original publication for a more detailed and in-depth discussion. Finally, before delving deeper into some of these regulatory mechanisms, we review and introduce a set of sequencing-based experimental techniques used throughout this thesis to investigate the molecular composition of tissues and cells.

1.1 Genome regulation

According to a simplified view of the so-called "central dogma of molecular biology" (Watson, 1965; Crick, 1958) (Figure 1.2), genetic information in the form of genomic sequences, encoded into functional units called *genes*, flows only in one direction: from Deoxyribonucleic acid (DNA) to Ribonucleic acid (RNA), during *transcription*, and from RNA to protein, during *translation*, from where it can no longer escape. Thus, genes serve as templates for producing *proteins* that perform a wide range of tasks, including constructing bodily structures, regulating chemical reactions, and transmitting information between and within cells. It is now evident that not only the actual genomic sequences but also their interactions are essential for genome regulation, which is a significantly more intricate process than previously believed, with information flowing in multiple directions and including multiple feedback mechanisms, such as those illustrated in Figure 1.1.

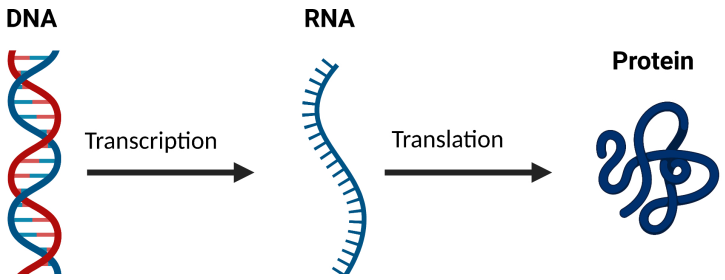

Figure 1.2: Simplified central dogma of molecular biology: DNA makes RNA makes proteins

1.1.1 Gene Expression

Chromosomes, threadlike structures that store genetic material, comprise a single, extraordinarily long DNA molecule containing a linear array of numerous genes and many proteins. However, only a small fraction of all this DNA encodes for proteins or RNA molecules with known functional properties. The first level of genomic control involves DNA's composition and spatial organization into a higher-order structure called *chromatin* and its different states: euchromatin and heterochromatin. In Section 1.2, we review how these distinct states facilitate or restrict access to the transcriptional machinery and regulate the subsequent transcriptional process, called *transcriptional regulation*.

During transcription, the DNA sequence of a gene is transformed into an RNA molecule known as precursor messenger RNA (pre-mRNA), initiating a series of activities known collectively as gene expression that result in the transformation of a gene into a functional product. A gene is said to be "on" if it is being transcribed into RNA, and only a fraction of the genes in a cell are active at any given moment. For example, a typical human cell expresses only 30-60% of its approximately 25.000 genes at a relevant level (Alberts et al., 2022). Thus, distinct cell types can be characterized by the set of specific genes they express. Moreover, by examining changes in gene expression patterns over time or across conditions, one can capture the dynamic nature of biological processes or responses to perturbations and environmental changes, such as signaling from neighboring cells. In Section 1.3, we review how such dynamic biological processes can be characterized at the gene expression level from scRNA-seq snapshot data using trajectory inference (TI) methods.

In addition to transcription initiation and transcription regulation, the expression of most genes is controlled at multiple levels, co-transcriptionally or once transcription has been completed. These *post-transcriptional regulatory* mechanisms control gene expression at the RNA level and regulate the distribution and stability of transcripts within the cell. In Section 1.2, we mention all known post-transcriptional regulatory mechanisms, focusing on mRNA degradation in particular. The final phase of gene expression is *translation*, during which mRNAs are converted into the amino acid sequences that constitute proteins. Although translational and post-translational mechanisms are essential for proper cellular regulation, a comprehensive review of these mechanisms is beyond the scope of this thesis.

Experimental sequencing-based methods, combined with appropriate downstream computational methods, are routinely used and have become the standard technique for profiling gene expression and, thus, the most popular approach for investigating genome regulation. In the following Section, we provide a brief overview of how these technologies have evolved over the last few decades.

1.1.2 Sequencing technologies

The process of sequencing DNA involves identifying the specific nucleic acid sequence, or arrangement of nucleotides in DNA, which has become an essential technique in basic and applied biological research. In the early 1970s, the first DNA sequences were obtained using laborious methods to isolate oligonucleotides and their subsequent separation based on two-dimensional chromatography (Jou et al., 1972). In 1977, an alternative method called the Sanger chain termination method (Sanger et al., 1977) was used to sequence the first ever complete genome for the bacteriophage ΦX174, for which a Nobel prize was awarded in 1980. Instruments based on Sanger sequencing, called "first-generation", with important improvements in safety and automation in the late 1980s, eased the experimental procedure and increased sequencing throughput. Such instruments dominated the sequencing world for the next decades and were extensively employed in the Human Genome Project (Lander et al., 2001; Venter et al., 2001), but their non-quantitative nature was their main not yet foreseeable limitation.

In 2005, Solexa (acquired by Illumina in early 2007) introduced the first short-read sequencer, with significantly higher throughput, based on a fundamentally different principle called sequencing by synthesis (SBS) (McPherson, 2009). Similar short-read instruments, built on massive parallel sequencing, followed in a wave of "next-generation" (NGS) or "second-generation" sequencing methods. The throughput of NGS methods increased rapidly, outpacing Moore's law and marking the beginning of high throughput sequencing (HTS), with one of the most well-known applications being whole-transcriptome sequencing, also known as RNA-seq. This, linked to an exponential decrease in sequencing costs and the more quantitative nature of the data available, completely revolutionized the field of genomics. Previous to the invention of RNA-seq, hybridization-based methods were designed to target the expression of a specific gene or set of genes (Wang et al., 2009). Hybridization-based techniques were limited by the necessity of prior knowledge of the genomic sequences being targeted, making it difficult to distinguish transcript isoforms of the same gene, and less precise expression levels due to high-background levels originating from cross-hybridization, as well as a more limited dynamic range due to the continuous nature of the signal. Furthermore, comparing the expression levels of different samples is not simple and involves complicated normalizing techniques. One of the key advantages of RNA-seq, concerning previous methods, is that it allows simultaneously discovering and quantifying transcriptomes.

In the last decade, RNA-seq and related HTS technologies have become one of the most precise methods for studying functional elements in the genome and other molecular components of cells and tissues under different conditions at an unprecedented resolution (Wold and Myers, 2007). Thanks to these next-generation sequencing platforms, researchers may now easily examine any organism's genome, transcriptome, and epigenome. Note that although the majority of sequencing technologies used in this thesis are referred to as RNA sequencing, they are sequencing cDNA that has been reverse-transcribed from the original input RNA and are, in fact, DNA sequencing technologies. The only exception is ChIP-seq, where the input for sequencing is actual DNA. We should emphasize that this is a common abuse of the term in the field and that methods differ mostly in how DNA or RNA is extracted from samples and matching sequencing libraries are generated. However, the sequencing remains fundamentally the same and is typically done using Illumina instruments. We use RNA sequencing technologies that aim to identify and quantify the expression of

genomic sequences in a given biological sample at the tissue-level (bulk RNA-seq) or the individual cell-level (scRNA-seq), and additional techniques based on DNA/RNA-seq coupled with other biochemical assays to identify and quantify corresponding genomic interactions (ChIP-seq and RIP-seq). We have focused on the analysis of Illumina short-read data, but most steps remain the same for other alternative sequencing technologies. For additional information on the various next-generation sequencing methods, see Goodwin et al. (2016). Third-generation sequencing technologies, which allow for the sequencing of much longer reads but present their own challenges, are reviewed in Stark et al. (2019) but are beyond the scope of this thesis.

1.2 Transcriptional and post-transcriptional regulation

1.2.1 Transcriptional regulation

Prior to initiating the production of a certain protein, the appropriate mRNA molecule must be created through transcription. The enzymes that carry out the transcription of DNA into RNA are called RNA polymerases, and in order for transcription to begin, an RNA polymerase needs to gain access to exposed DNA. In eukaryotes, DNA is tightly bonded and wrapped around an octameric core of histone proteins, forming repetitive arrays of DNA-protein particles known as *nucleosomes*. Around 200 nucleotide pairs separate nucleosomes that interact with adjacent nucleosomes to form a denser *chromatin fiber*. Such chromatin structure must be highly dynamic for DNA to be accessible. Unwrapping and rewrapping occur spontaneously, and chromatin-remodeling complexes are the general strategy for reversibly altering local chromatin structure in a cell. These remodeling complexes permit nucleosome cores to be relocated, rebuilt with various histones, or removed entirely to reveal the underlying DNA. Furthermore, once the RNA polymerase has gained access to a DNA template, a group of other proteins known as general transcription factors are required to *initiate* transcription, and additional proteins such as transcription activator proteins, chromatin remodeling complexes, and histone-modifying enzymes, are required to *activate* transcription. These regulatory complexes are abundant in cells and are localized to certain chromatin regions at the appropriate moments.

Different chromatin configurations are possible based on a broad range of reversible covalent modifications of the four distinct histones in the nucleosome core, including methylation, acetylation, phosphorylation and ubiquitination. These chromatin modifications are heritable and result in different regulatory properties essential for proper genome stability, yet their source is not the actual genetic code sequence. Therefore, they are an example of epigenetic ("on top of" genetics) regulatory mechanisms. The specific combination of modifications that nucleosomes are marked with governs their interactions with other proteins. When protein modules from a larger protein complex attach to modified nucleosomes in a chromatin region, these marks are read and attract other proteins with different functions.

Broadly speaking, the two major chromatin forms are euchromatin and heterochromatin, which were first discovered due to their distinct staining properties (Passarge, 1979). Denser regions of chromatin that stain strongly are typically located at the nucleus's periphery and are known as *heterochromatin*. In contrast, less stainable regions located largely in the nucleus's center are known as *euchromatin*. Euchromatin is rich in genes and correlates with high transcription levels, whereas a key role of heterochromatin's dense DNA packing is to silence the few genes it contains. Within the chromosome, heterochromatin is typically lo-

cated near centromeres and telomeres, although it is also present in many other chromosomal locations. There are two major types of heterochromatin, each with its own set of covalent histone modifications spread by different reader-writer enzyme complexes. *Constitutive* heterochromatin is established on repetitive DNA sequences and identified by the H3K9me3 mark. Constitutive heterochromatin is mostly present around centromeres and telomeres and inhibits both gene expression and genetic recombination. *Facultative* heterochromatin spreads over the genome, its presence can be adjusted, and it plays an important role in gene regulation. Facultative heterochromatin is identified by H3K27me3 and, in some cases, H3K9me3 marks.

For a comprehensive review of the different biochemical pathways involved in the regulation of heterochromatin and the associated silencing of these regions, see Brönner (2017).

1.2.2 Post-transcriptional regulation

During the *elongation* phase of transcription, the nascent RNA undergoes three crucial forms of processing:

- 5'-end capping
- Alternative RNA splicing
- 3'-end polyadenylation

Only correctly processed mRNAs can pass through nuclear pore complexes to be later translated into protein in the cytoplasm. Other post-transcriptional regulatory mechanisms exist, i.e., all processes that occur between the *transcription* and *translation* stages, and thus control the generation of distinct gene expression patterns observed in cells at the RNA level, including:

- RNA editing
- Export from the nucleus to the cytosol
- Localization of mRNAs in the cytoplasm
- Translation initiation
- mRNA degradation (or mRNA stability)

mRNA stability, is one particularly important post-transcriptional regulatory mechanism. Due to the constant turnover of mRNA, steady-state mRNA levels are defined by RNA synthesis (transcription) rates and corresponding decay (degradation). If a newly synthesized transcript is rapidly degraded, it will not have time to be translated or to participate in any regulatory pathway; consequently, it will appear as though the corresponding gene is silent. The 5'-end capping and 3'-end polyadenylation, or poly(A) tail, increase the stability of transcribed mRNAs. Thus, the shortening of the poly(A) tail is one of the first steps in RNA degradation, mediated by the Ccr4-Not complex, the predominant deadenylase complex. Deadenylation of the transcript is followed by either decapping of the 5' end and subsequent 5'-3' digestion by the Xrn1/2 exonucleases or by 3'-5' degradation via the exosome.

In Chapter 2, we dissect the contributions of 3 different pathways to the silencing observed in heterochromatic regions of *Schizosaccharomyces pombe* (S.pombe) through a combination of both transcriptional silencing and RNA degradation. The first mechanism under study suggests reduced accessibility of RNA Polymerase II (RNA Pol II) to heterochromatic regions. We quantify changes in Pol II occupancy between WT and mutant strains devoid of the necessary machinery to form heterochromatin by measuring per-gene occupancy levels of RNA Pol II with Chromatin immunoprecipitation sequencing (ChIP-seq). The next mechanism at the level of transcriptional silencing involves the possibility that although RNA Pol II can, in principle, access DNA, other mechanisms might be hindering the actual transcription of RNAs. To investigate this, we quantify per-gene nascent RNA levels with Pol II bound nascent RNA sequencing (RIP-seq), and define corresponding transcription efficiencies in each mutant as the ratio of nascent RNA levels to previously computed RNA Pol II occupancy levels. The last mode of silencing considered in the study relates to the recruitment of RNA degradation machinery by heterochromatin, hindering the subsequent expression of certain types of RNAs. We quantify per-gene steady-state RNA levels with RNA-seq, both for total RNA and poly A enriched RNA (pA-RNA), and define corresponding RNA stabilities as the ratio of pA-RNA (or total RNA) levels to previously computed nascent RNA levels.

1.3 Gene expression dynamics

In the previous Section, we have focused on studying certain regulatory mechanisms from a bulk and static perspective. We quantified the effect perturbations have on a large population of cells at the molecular level by comparing them to a control group or homeostatic state. When using bulk RNA-seq, for instance, measurements are limited to the average gene expression levels across a vast population of cells, which may be insufficient to characterize highly diverse systems or complex tissues. In addition, many cellular processes of interest, such as cell differentiation or reprogramming, are not static but rather dynamic biological processes.

1.3.1 Single-cell RNA-seq and trajectory inference (TI) methods

To address these limitations, single-cell RNA sequencing (scRNA-seq) was developed to measure the transcriptomic patterns of individual cells. Various protocols and technologies for high-throughput scRNA-seq are now available, and new ones are continuously being developed. In comparison to its bulk counterpart, a typical scRNA-seq experiment begins with the dissociation of cells from a tissue and the isolation of single-cells using specialized techniques, such as the deposition of individual cells in micro-well plates or capturing individual cells employing microfluidic droplet-based platforms. Despite its own challenges, such as the relatively small amounts of mRNA collected from single cells and the typical use of unique molecular identifiers (UMI) to correct for amplification biases, the subsequent sequencing steps are similar to the analysis of pooled bulk RNA-seq libraries, except that each sample corresponds to a single cell. Moreover, a single snapshot of scRNA-seq data collected during a dynamic process will comprise cells at various stages along this process. Therefore, cell-level information can be used to infer lineage relationships across cell types and states computationally using trajectory inference (TI) methods (Cannoodt et al., 2016a).

In scenarios with larger time scales, where some cell populations arise before or after a single sampling instance, time-series scRNA-seq data can be collected by obtaining multiple scRNA-seq snapshots at different time points and analyzed using the same methods (Tran and Bader, 2020). TI approaches aim to reconstruct a trajectory in which individual cells are ordered based on their transcriptomic similarity, using pseudo-time as a parametrization of these continuous transitions between observed cell states. In this sense, the pseudo-time associated with a given cell is the distance in transcriptomic space between that cell and the origin of the trajectory. Although pseudo-time can often be interpreted as an increasing function of chronological time, their specific functional relationship is arbitrary and trajectory-specific. Monocle (Trapnell et al., 2014) was the first TI method to be developed, and like other early proponents, was limited to the analysis of biological processes that simple linear trajectories could characterize. Newer versions of Monocle (Qiu et al., 2017; Cao et al., 2019), and other state-of-the-art TI methods allow the reconstruction of more complex topologies, which are characteristic of more complex biological processes such as development, in which trajectories typically contain multiple branching patterns leading to the various possible cell fates. A recent extensive benchmark of 45 TI methods by Saelens et al. (2019) concluded that while there are TI methods that perform significantly better than others, like Slingshot (Street et al., 2018), Monocle2 (Qiu et al., 2017), PAGA (Wolf et al., 2019), and Scorpius (Cannoodt et al., 2016b), there is no "one-size-fits-all" method that works well on every dataset and highlights that performance was strongly dependent on the type of trajectory topology present in the underlying data.

1.3.2 Alignment of complex single-cell trajectories

When a pair of trajectories from a related but potentially distinct process is available, it is natural to be interested in identifying differences and similarities between the two. By comparing the molecular programs executed by cells following biological processes under different conditions, we can improve our understanding of cellular regulatory mechanisms, as well as of health and disease. For example, comparing two single-cell trajectories, one characterized by normal muscle cell differentiation and the other by fibroblast reprogramming into muscle cells, revealed the critical molecular determinants for effective reprogramming (Cacchiarelli et al., 2018). Alternatively, comparing trajectories derived from two species can illuminate on the evolutionary differences between both organisms (Alpert et al., 2018). The main issue preventing direct comparison between these pseudo-time trajectories is that pseudo-time is not a global attribute of a dataset but is defined in a trajectory-specific manner, which means that pseudo-times defined in two separate trajectories are in different systems of reference. In order to make two trajectories comparable, gene expression patterns of cells ordered along different pseudo-times can be used to match or align cells along a common pseudo-time axis. Trajectories in the examples above were aligned using dynamic time warping (DTW) (Vintsyuk, 1968), in which two time-series that evolve at different times or speeds can be stretched and compressed to find an optimal "warping path" such that the two signals are mapped into a common system of reference. Due to the inherent one-dimensional nature of time-series, DTW is limited to the analysis of simple linear trajectories. Typically, when dealing with complex trajectories, analysis often relies on identifying and selecting a single path from each trajectory, followed by alignment using DTW. This means that prior biological knowledge, which may not always be available, or additional computational methods are required to identify a "core" path of interest that contains biologically relevant information

previous to alignment. Lastly, information from other cell lineages represented by alternate paths distinct from the core path, which could in theory help guide the alignment, is completely ignored by such approaches.

In Chapter 3, we introduce Trajan, a novel method that generalizes the comparison and alignment of simple linear scRNA-seq trajectories to arbitrary trajectory types that can be represented as rooted trees, such as bifurcating or binary graph topologies. Trajan aligns all paths in a pair of complex single-cell trajectories automatically and consistently, identifying the correspondence between biological processes and improving on the multiple pair-wise alignment between individual lineages. We made an effort to integrate our software with the pre-existing dynverse framework (Saelens et al., 2019), enabling and facilitating the inference of single-cell trajectories with any of the 50+ methods available. The result is our novel R package, TrajanR, which standardizes the pre-processing of Trajan input data, allows alignment computation under different parameter schemes, and provides various visualization options.

1.4 List of peer-reviewed articles

- Van Hoan Do*, Mislav Blažević*, **Pablo Monteagudo-Mesas**, Luka Borozan, Khaled Elbassioni, Soeren Laue, Francisca Rojas Ringeling, Domagoj Matijevic and Stefan Canzar. *Dynamic pseudo-time warping of complex single-cell trajectories.* RECOMB 2019, Lecture Notes in Computer Science (2019).

- **Pablo Monteagudo-Mesas***, Cornelia Brönner*, Parastou Kohvaei, Haris Amedi, Stefan Canzar, Mario Halic. *Ccr4–Not complex reduces transcription efficiency in heterochromatin.* Nucleic Acids Research (2022).

- In preparation: **Pablo Monteagudo-Mesas***, Van Hoan Do*, Mislav Blažević*, Luka Borozan, Khaled Elbassioni, Soeren Laue, Francisca Rojas Ringeling, Domagoj Matijevic and Stefan Canzar. *TrajanR enables the accurate alignment and comparison of complex scRNA-seq trajectories.* (2023).

- Accepted - Bioinformatics: Luka Borozan, Francisca Rojas Ringeling, Shao-Yen Kao, Elena Nikonova, **Pablo Monteagudo-Mesas**, Domagoj Matijevic, Maria L. Spletter, Stefan Canzar. *Counting pseudoalignments to novel splicing events.* Bioinformatics (2023).

- In preparation: Shao-Yen Kao, Vanessa Todorow, Peter Meinke, Benedikt Schoser, Keshika Ravichandran, Oliver Weinert, Paul Walper, Linus Hoelzel, Michael Menden, **Pablo Monteagudo-Mesas**, Israa Alqassem, Stefan Canzar, Mara Barone, Andrea David Re Cecconi, Rosanna Piccrillo, Tobias Straub, Rippei Hayashi, Maria L. Spletter. *Splicing mediated by U2-associated Scaf6/CHERP is necessary for myogenesis in Drosophila and vertebrates.* (2023).

1.5 List of pre-print articles

- Pre-print: Van Hoan Do*, Mislav Blažević*, **Pablo Monteagudo-Mesas**, Luka Borozan, Khaled Elbassioni, Soeren Laue, Francisca Rojas Ringeling, Domagoj Matijevic and Stefan Canzar. *Dynamic pseudo-time warping of complex single-cell trajectories.* bioRxiv (2019).

- Pre-print: Jianfeng Sun, Martin Philpott, Danson Loi1, Shuang Li, **Pablo Monteagudo-Mesas**, Gabriela Hoffman, Jonathan Robson, Neelam Mehta, Vicki Gamble, Tom Brown Jr, Tom Brown Sr, Stefan Canzar, Udo Oppermann1, Adam P Cribbs. *Correcting PCR amplification errors in unique molecular identifiers to generate absolute numbers of sequencing molecules.* bioRxiv (2023).

*indicates equal contribution.

Chapter 2

Ccr4–Not complex reduces transcription efficiency in heterochromatin

This Chapter is adapted with minimal modifications from:

- **Pablo Monteagudo-Mesas***, Cornelia Brönner*, Parastou Kohvaei, Haris Amedi, Stefan Canzar, Mario Halic. *Ccr4–Not complex reduces transcription efficiency in heterochromatin*. Nucleic Acids Research, 2022.

2.1 Introduction

Heterochromatin is essential to maintain genome stability and transcriptional regulation. Defects in heterochromatin formation lead to aberrant centromere and telomere function, aneuploidy and cancer. In fission yeast, constitutive heterochromatin is established at centromeres, subtelomeres and the silent mating type (mat) locus (Allshire and Ekwall, 2015). At centromeric repeats, RNAi is essential for heterochromatin formation as Argonaute, guided by small RNAs, recruits the H3K9 methyltransferase complex CLRC to chromatin (Halic and Moazed, 2010; Holoch and Moazed, 2015; Marasovic et al., 2013; Martienssen and Moazed, 2015; Ugolini and Halic, 2018; Verdel et al., 2004; Volpe et al., 2002). This leads to deposition of the repressive H3K9 methylation (H3K9me) mark by Clr4, recruitment of HP1 proteins Chp2 and Swi6, and heterochromatin formation (Allshire and Ekwall, 2015; Holoch and Moazed, 2015; Martienssen and Moazed, 2015).

Current data suggest that heterochromatic silencing occurs through a combination of transcriptional silencing and RNA degradation. At the level of transcriptional silencing, HP1 proteins bind H3K9me nucleosomes and recruit downstream-acting complexes (Castel and Martienssen, 2013; Motamedi et al., 2008; Zocco et al., 2016). HP1 protein Chp2 recruits the complex SHREC to deacetylate chromatin and to remodel nucleosomes in heterochromatin, which is required for silencing (Motamedi et al., 2008; Creamer et al., 2014; Sugiyama et al., 2007). These activities were suggested to reduce RNA Pol II access (Chen and Widom, 2005; Schuettengruber et al., 2007).

Heterochromatin is also thought to promote recruitment of the RNA degradation machinery to degrade nascent transcripts (Marasovic et al., 2013) (Brönner et al., 2017; Bühler et al., 2008; Cotobal et al., 2015; Pisacane and Halic, 2017; Reyes-Turcu et al., 2010; Reyes-Turcu and Grewal, 2012; Sugiyama et al., 2016). In fission yeast, RNA Pol II-transcribed heterochromatic transcripts are polyadenylated (pA) products and undergo degradation by the RNAi pathway, by the Ccr4–Not complex and by the exosome (Brönner et al., 2017). The first step of mRNA degradation is generally shortening of the 3′ pA tail by the Ccr4–Not complex and Pan nucleases (Wahle and Winkler, 2013). This induces removal of the 5′ cap which enables 5′-3′ exonucleases Xrn1/2 (Exo2 and Dhp1 in S. pombe) and 3′-5′ degradation by the exosome (Rrp6) (Houseley et al., 2006).

How transcriptional silencing and RNA degradation pathways collaborate and their relative contributions to silencing are not known. In this study, we quantified the contribution of these pathways to heterochromatic silencing by analyzing RNA Pol II occupancy, nascent RNA and steady-state RNA in different fission yeast mutants. We also defined the heterochromatic factors that contribute to transcriptional silencing and/or RNA degradation. We found that transcriptional silencing occurs through reduced RNA Pol II accessibility, as previously proposed, but, unexpectedly, also through reduced transcriptional efficiency, a mechanism not previously implicated in silencing. Our data revealed that RNA Pol II transcriptional output is lower at heterochromatic loci compared to euchromatic loci relative to levels of RNA Pol II occupancy. We determined that the Ccr4–Not complex and H3K9 methylation are essential for the reduced transcriptional efficiency at heterochromatin and quantified the contributions of heterochromatin factors to the reductions in RNA Pol II occupancy, transcriptional efficiency and RNA stability.

2.2 Methods

Supplementary Material A.1 contains information on strain construction and the preparation of sequencing libraries.

Analysis of sequencing data

Sequencing reads obtained in the poly(A) RNA sequencing (pA RNA), Pol II ChIP, nascent RNA sequencing (Pol II RIP) and total RNA sequencing experiments were mapped to the S. pombe reference genome (PomBase, release 2018) using splice-aware alignment tool STAR version 2.7.3a (Dobin et al., 2013). Alignment of RIP-seq and RNA-seq data was performed with STAR default parameters. Unspliced alignments of ChIP-seq data was enforced through parameters '–alignIntronMax 1' and '–alignEndsType EndToEnd'. Reads mapping to ribosomal RNA have been removed from further analysis.

Genomic read counts were obtained using a custom script that extended basic htseq-count (Anders et al., 2015) functionality with '–mode intersection-strict' option. Additionally, for RNA assays pA RNA, Pol II RIP and total RNA, we used the '–stranded yes' option to identify the strand the read originated from. Only reads that mapped uniquely and reads that mapped to less than 16 locations within heterochromatic regions were counted. We chose 16 as multi-mapping threshold for heterochromatic genes to eliminate low-complexity reads without discarding reads originating from heterochromatic regions (dg/dh have 12–13 copies). We did not try to resolve the origin of multi-mapping reads, but instead counted all reads that mapped to one representative copy of dg/dh. We normalized gene counts by gene length and sequencing depth using Transcripts Per Million (TPM). Finally, average TPM values were computed for protein-coding and heterochromatic genes across biological replicates with a Spearman correlation coefficient of log transformed gene expression values of at least 0.8.

For the analysis in Supplemental S1B, we obtained relative intronic read counts by first counting the number of reads overlaping each intronic region, normalizing this count by the length of the intron, and finally dividing this normalized count by the total number of reads mapping to the corresponding gene, normalized again by the gene length.

Read coverages (as shown for example in Figure 2.1A) were obtained using a custom python script that uses STAR read alignments as input and returns corresponding coverages in wiggle format. The script generates a coverage profile x_i, by counting the number of read alignments x overlaping each genomic location i. Multi-mapping reads in this case fractionally contribute 1/NH to corresponding locations in the coverage profile, where NH is the number locations the read maps to. For visualization we used a custom R script that normalizes each profile by sequencing depth, making them comparable across datasets. Here, sequencing depth is computed as milions of reads that map to protein-coding genes or heterochromatic regions.

2.2.1 t-distributed stochastic neighbor embedding (t-SNE)

For each mutant, we created a high-dimensional vector containing Pol II occupancy and transcription efficiency log transformed TPM values of all heterochromatic genes. In order to visualize this high-dimensional data and their relationships, we used scikit-learn's (Pedregosa

et al., 2011) implementation of t-SNE with 'perplexity = 5' parameter, to reduce the data to two dimensions as shown in Figure 2.5A.

2.2.2 IP-Input subtraction in ChIP-seq data

For normalization of ChIP-seq data based on Input data, we used the three core centromeric regions to define the background component of the IP data. These are the longest regions with very low transcription and we thus scale Input data to match the IP data along these regions (Diaz et al., 2012). More specifically, for each of the three regions we computed for each ChIP-seq sample a coverage profile x_i and a coverage profile y_i for the corresponding IP-Input sample, and computed a scaling factor λ using the following equation:

$$\sum_{i=1}^{n} x_i - \lambda \sum_{i=1}^{n} y_i = 0 \tag{2.1}$$

where n is the gene length. We take the median of the three λ values as our final scaling factor. In Supplemental S1A we show that this normalization notably reduces the read coverage within these regions in wild-type cells.

2.2.3 Transcription Efficiency, RNA Stability and Pol II Occupancy

For each mutant, let θ_{ChIP}, θ_{RIP} and θ_{pA-RNA} denote the average TPM value of a given gene obtained from Pol II ChIP-seq, Pol II RIP-seq and pA RNA-seq experiments, respectively, as described above. We compute transcription efficiency as the amount of newly synthesised RNA (Pol II RIP) relative to the level of Pol II Occupancy (Pol II ChIP):

$$\text{Transcription Efficiency} = \frac{\theta_{RIP}}{\theta_{ChIP}} \tag{2.2}$$

RNA levels being the result of RNA synthesis and degradation, we define RNA stability as the ratio of steady state RNA levels (pA RNA-seq) over the amount of newly synthesised RNA (Pol II RIP):

$$\text{RNA Stability} = \frac{\theta_{pA-RNA}}{\theta_{RIP}} \tag{2.3}$$

Changes in Pol II occupancy were quantified by the log fold change of Pol II occupancy levels (Pol II ChIP) in each mutant (mu) relative to wild-type (wt) cells:

$$\Delta\text{Pol II} = \log_2 \left(\frac{\theta^{mu}_{ChIP}}{\theta^{wt}_{ChIP}} \right) \tag{2.4}$$

As described above, we compute ratios of averaged TPM values across replicates, similar to Thoreen et al. (2012), to obtain quantitative measures of transcription efficiency, RNA stability, and changes in Pol II occupancy. Since different experiments are unpaired, we visualize variability in these ratios by combining all possible pairs of replicates between the two assays involved in each of the three quantities and provide the standard error of the mean (SEM) in Supplemental S2.

2.2.4 Calculation of pathways' contributions to silencing

Since the transcriptional output is proportional to Pol II occupancy, transcriptional efficiency, and RNA stability, we quantify the RNA output by:

$$\text{RNA output} = [\Delta\text{Pol II}] \cdot [\text{Transcription Efficiency}] \cdot [\text{RNA Stability}] \tag{2.5}$$

$$\text{RNA output} = \theta_{ChIP} \cdot \left(\frac{\theta_{RIP}}{\theta_{ChIP}}\right) \cdot \left(\frac{\theta_{pA-RNA}}{\theta_{RIP}}\right) \tag{2.6}$$

2.3 Results

Transcriptional silencing and degradation of heterochromatic RNA

To dissect the contributions of transcriptional silencing and RNA degradation to heterochromatic silencing, we analyzed RNA Pol II occupancy, nascent RNA, and steady-state RNA at heterochromatic and euchromatic regions in fission yeast (Figure 2.1A shows data for specific genomic regions from wild-type cells). For Pol II occupancy, we performed ChIP-seq analyses of serine 2 phosphorylated (S2P)-RNA Pol II-bound DNA. To correct for noise, we subtracted scaled input data from all RNA Pol II ChIP-seq datasets. For normalization, the background component was defined based on the centromeric central core (Cenp-A containing chromatin), a region with low occupancy in RNA Pol II ChIP-seq; we thus subtracted input so that this region would show near zero signal (Supplemental S1A, see also Methods). For nascent RNA, we performed RNA-seq of S2P-RNA Pol II-bound RNA (Pol II RIP) and assessed the quality of nascent RNA by examining retention of intronic sequences, which are strongly enriched in nascent RNA (Supplemental S1B, C). For steady-state RNA, we sequenced polyadenylated (pA) RNA and total RNA; both datasets showed comparable levels of heterochromatic transcripts, indicating that these are mostly polyadenylated (Supplemental S1D and E).

To determine how heterochromatin changes RNA Pol II accessibility, we compared RNA Pol II occupancy in wild-type cells and in clr4Δ cells, which lack H3K9 methyltransferase Clr4 and thus do not have H3K9me or heterochromatin (Marasovic et al., 2013) (Figure 2.1B). Error bars for all the data are shown in Supplemental S2A. We did not detect substantial differences in RNA Pol II occupancy at protein-coding genes between wild-type and clr4Δ cells. At heterochromatic regions, the effects varied according to the specific locus: at centromeric dg/dh and subtelomeric tlh repeats (tlh and SPAC212.10), RNA Pol II occupancy was ∼4-fold higher in clr4Δ cells compared to wild-type cells (Figure 2.1B and Supplemental S3A, B). At other subtelomeric regions and at the mat locus, we observed smaller change in RNA Pol II occupancy in the absence of heterochromatin (Figure 2.1B and Supplemental Table S2).

Our data also showed that RNA Pol II complexes can be present in heterochromatic regions, but they do not always actively transcribe or produce nascent RNA. For example, in wild-type cells, heterochromatic tlh1 and subtelomeric ftm1 loci show similar RNA Pol II occupancy, but ftm1 produces substantially more nascent RNA (Figure 2.1A). To analyze this relationship further, we plotted the nascent RNA levels over chromatin-bound RNA Pol II for individual loci in wild-type cells (Figure 2.1C); we also calculated the ratio between those measurements, which informs on how much nascent RNA is synthesized relative to RNA Pol II occupancy at any given locus (Figure 2.1C). We define this parameter as transcription efficiency. Error bars for all the data are shown in Supplemental S2B.

We found a linear relationship between RNA Pol II occupancy and nascent RNA when examining ∼5000 euchromatic, protein-coding loci in wild-type cells, indicating that transcription efficiency was constant across the examined euchromatic loci. In contrast, heterochromatic loci showed lower transcriptional efficiency (8-fold on average) compared to protein-coding loci (Figure 2.1C), and this observation applied to all heterochromatic regions in fission yeast. Thus, our data indicate that, in wild-type cells, heterochromatic loci are less efficiently transcribed than protein-coding loci that have the same amount of RNA Pol II, suggesting that RNA Pol II does not productively transcribe heterochromatic regions.

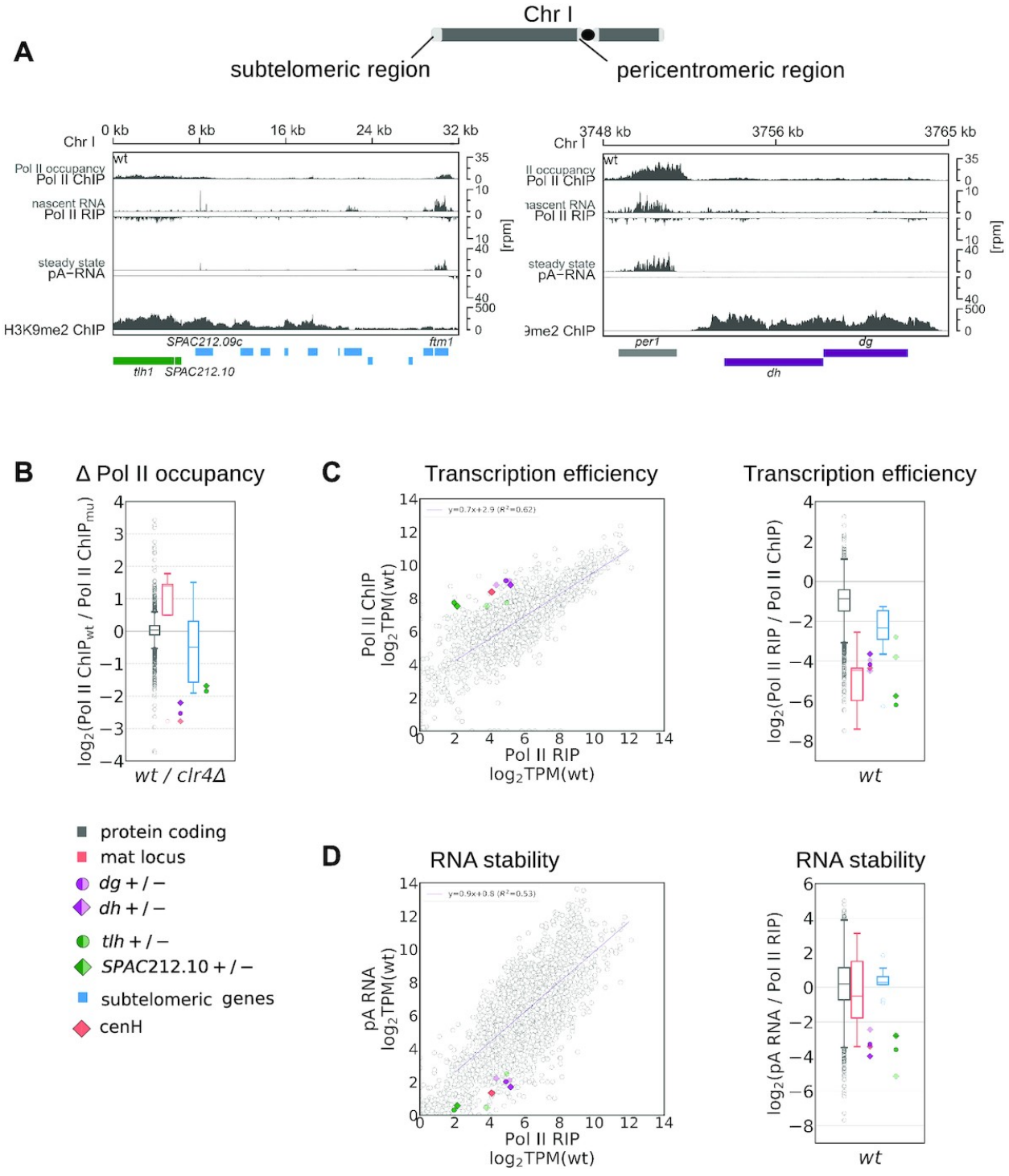
A
Chr I
subtelomeric region
pericentromeric region
Chr I
0 kb
8 kb
16 kb
24 kb
32 kb
Pol II occupancy
Pol II ChIP
nascent RNA
Pol II RIP
steady state
pA−RNA
H3K9me2 ChIP
wt
[rpm]
SPAC212.09c
ftm1
tlh1
SPAC212.10
3748 kb
3756 kb
3765 kb
per1
dg
dh
B
Δ Pol II occupancy
log2(Pol II ChIPwt / Pol II ChIPmu)
wt / clr4Δ
protein coding
mat locus
dg + / −
dh + / −
tlh + / −
SPAC212.10 + / −
subtelomeric genes
cenH
C
Transcription efficiency
Pol II ChIP
log2TPM(wt)
Pol II RIP
log2TPM(wt)
log2(Pol II RIP / Pol II ChIP)
wt
D
RNA stability
pA RNA
log2TPM(wt)
Pol II RIP
log2TPM(wt)
log2(pA RNA / Pol II RIP)
wt

Figure 2.1: Heterochromatic repeats have reduced RNA Pol II occupancy, Transcription Efficiency and RNA stability. (A) Analysis of the next-generation sequencing data showing occupancy of S2P-RNA Pol II (ChIP-seq), nascent RNA (S2P-Pol II RIP-seq), steady-state RNA levels (pA RNA-seq) and H3K9me2 levels (ChIP-seq) at subtelomeric and pericentromeric regions in S. pombe wild-type cells. Gene locations are indicated as boxes below the coverage and color-coded: gray, protein-coding genes; purple, centromeric dg, dh; green, subtelomeric loci tlh and SPAC212.10; blue, other subtelomeric genes. (B) Box plot showing RNA Pol II occupancy (S2P-Pol II ChIP-seq) in wild-type cells relative to clr4Δ cells for protein-coding genes (gray), mat locus (orange), and subtelomeric genes (blue). The sequencing data for clr4Δ are shown in S4B. For protein-coding genes, individual transcript are shown as circles; bottom and top of the box correspond to lower and upper quartiles of the data, bar is the median and whiskers are median ±1.5 times interquartile range. Colored symbols on the right show centromeric dg and dh (dark purple), subtelomeric tlh and SPAC212.10 (dark green), and cenH (orange). Solid/transparent color show + and – strand respectively. Each data point is the average of at least two independent samples. (C) Transcription efficiency in wild-type cells. Left, S2P-Pol II ChIP-seq (Pol II occupancy) data plotted over S2P-Pol II RIP-seq data (nascent RNA). TPM, transcripts per million. Gray circles are individual protein-coding genes; regression line is also shown in purple. Also plotted are centromeric dg and dh (dark purple for + strand, bright purple for - strand) and tlh and SPAC212.10 (dark green for + strand, bright green for – strand) and cenH (orange). Each data point is the average of at least two independent samples. Right, box plot showing transcription efficiency distributions by gene categories, data are plotted and color-coded as in panel B. (D) RNA stability in wild-type cells. Left, pA RNA-seq (steady-state RNA) data plotted over S2P-Pol II RIP seq data (nascent RNA). TPM, transcripts per million. Data are plotted as defined for panel (C). Right, box plot showing RNA stability (pA RNA/Pol II RIP) distribution by gene categories, data are plotted and color-coded as in panel (B). Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

Next, we calculated the ratio of steady-state RNA to nascent RNA, which informs on the stability of a specific transcript. Error bars for all the data are shown in Supplemental S2C. Notably, we found that a subset of heterochromatic transcripts was less stable than the average transcript from a protein-coding loci (Figure 2.1D). This subset of unstable heterochromatic RNAs includes transcripts from centromeric dg/dh and subtelomeric tlh regions, which are known to be degraded by RNAi and by the Ccr4–Not complex, respectively (Halic and Moazed, 2010; Marasovic et al., 2013; Brönner et al., 2017). For transcripts derived from the other subtelomeric genes or from the mat locus, RNA stability was comparable to transcripts from protein-coding genes (Figure 2.1D), indicating that those loci primarily undergo transcriptional silencing, with no major role for RNA degradation in their silencing.

2.3.1 Effect of heterochromatin on silencing pathways

Our data indicate that three different pathways can contribute to silencing of heterochromatin: RNA Pol II occupancy, transcriptional efficiency and RNA degradation. To further examine the impact of heterochromatin structure on those pathways, we compared the contributions from each of them at heterochromatic loci in wild-type and clr4Δ cells.

We found that RNA Pol II transcribes repetitive regions more efficiently in the clr4Δ cells compared to wild-type cells. RNA Pol II occupancy at centromeric dg/dh and subtelomeric

tlh repeats was ∼four-fold higher in clr4Δ cells relative to wild-type cells, but the increase in nascent RNA was ∼twenty-fold (Figure 2.2A and B). The data also demonstrate that heterochromatin structure contributes to reduced transcriptional efficiency: in clr4Δ cells, all heterochromatic regions (centromeres, subtelomeres and the mat locus) reach transcriptional efficiency that is comparable to euchromatic regions (Figure 2.2C). Moreover, transcription efficiency is the single mode of heterochromatic silencing that acts in all heterochromatic regions.

We next assessed RNA stability in the absence of heterochromatin structure. We observed that subtelomeric tlh RNA has similarly low stability in clr4Δ compared to wild-type cells, indicating that multiple pathways degrade RNA from subtelomeric tlh repeats, independently of heterochromatin structure, in agreement with previous observations (Figure 2.2D and E) (Brönner et al., 2017). In contrast, centromeric dg/dh transcripts showed increased stability in clr4Δ compared to wild-type cells, indicating that heterochromatin structure is required for degradation of those transcripts.

These data from clr4Δ cells also confirm our observations with wild-type cells: at centromeric dg/dh, silencing results from a combination of reduced RNA Pol II occupancy (Figure 2.2A), reduced transcriptional efficiency (Figure 2.2C) and increased RNA degradation (Figure 2.2E), whereas for subtelomeric tlh repeats, silencing occurs primarily at the transcriptional level.

We quantified the relative contribution of each pathway to heterochromatic silencing, by comparing data from wild-type and clr4Δ cells. In wild-type cells, RNA degradation and transcriptional silencing contribute similarly to silencing of centromeric repeats (∼38% and ∼62%, respectively) (Figure 2.2F) wherein transcriptional silencing can be further parsed out into RNA Pol II occupancy (∼36%) and transcriptional efficiency (∼26%). At subtelomeric tlh repeats, heterochromatic silencing occurs primarily by transcriptional silencing (∼92%) (Figure 2.2G), which is predominately mediated through reduced transcriptional efficiency (∼66%). A similar pattern is observed at the remaining subtelomeric regions and mat locus: at mat locus silencing is mostly transcriptional, with transcriptional efficiency (∼70%) being the major silencing pathway (Supplemental S1F, G).

2.3.2 Contribution of heterochromatin factors to distinct silencing pathways

Next, we analyzed RNA Pol II occupancy, transcriptional efficiency and RNA degradation in strains defective in the RNAi machinery (ago1Δ), lacking chromatin modifiers (clr3Δ, mit1Δ, chp2Δ and swi6Δ) or RNA degradation components (rrp6Δ, exo2Δ, caf1Δ, ccr4Δ and mot2Δ) (Supplemental S3-S7); error bars between replicates shown in Supplemental S2. Chp2 and Swi6 are HP1 family proteins; wherein Chp2 recruits the SHREC complex to chromatin. SHREC subunits, Mit1 and Clr3, are a chromatin remodeler and a histone deacetylase, respectively (Motamedi et al., 2008; Sugiyama et al., 2007). Rrp6 is a component of the exosome complex; Exo2 is a 5′ to 3′ exonuclease. Caf1, Ccr4 and Mot2 are components of the Ccr4–Not deadenylase complex.

Our data show that reduced RNA Pol II occupancy at centromeric repeats requires RNAi (Figure 2.3A and Supplemental S3A, S4A–C), HP1 proteins, components of SHREC (Figure 2.3A and Supplemental S4D, E, S5A, B) and Exo2 and Rrp6 of the RNA degradation machinery (Figure 2.3A and Supplemental S5C-D). At subtelomeric tlh repeats, only chromatin modifiers are required to limit RNA Pol II occupancy (Figure 2.3A and Supplemental S3B, S4D, E, S5A, B), whereas at the remaining subtelomeric genes and at the cenH element of

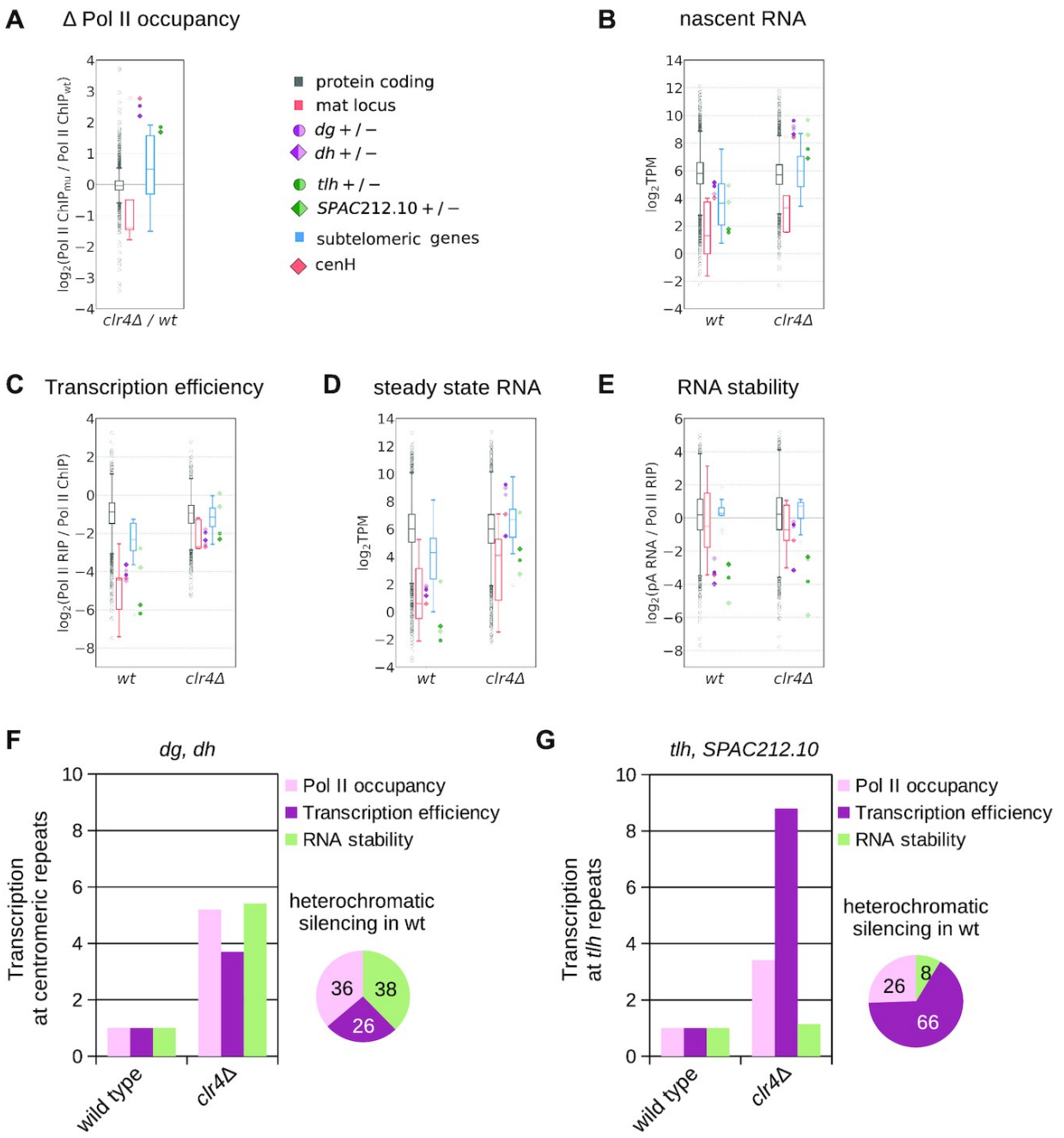

Figure 2.2: Contribution of distinct pathways to heterochromatic silencing. (A) RNA Pol II occupancy (Pol II ChIP-seq) in clr4Δ cells relative to wild-type cells for indicated gene categories; data are the same as in Figure 2.1B, with inverted ratios. (B) Nascent RNA (Pol II RIP) in wild type and clr4Δ cells. Data are plotted as defined for Figure 2.1B. (C) Transcription efficiency (Pol II RIP / Pol II ChIP) in wild-type and clr4Δ cells. Data are plotted as defined for Figure 2.1B. (D) Steady state RNA (pA RNA-seq) shown for wild-type and clr4Δ cells. Data are plotted as defined for Figure 2.1B. (E) RNA stability (pA RNA / Pol II RIP) in wild type and clr4Δ cells. Data are plotted as defined for Figure 2.1B. (F, G) Bar chart showing fold change in quantitative measures (ratios of average TPM, see Methods) of the three pathways (Pol II occupancy, transcription efficiency and RNA stability) at centromeric dg and dh (D) and at subtelomeric tlh (E) in clr4Δ cells, relative to wild type. Pie charts show relative contribution of each pathway to heterochromatic silencing at repeats in wild-type cells. Average of at least two independent samples is shown for all figures. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

the mat locus, RNA Pol II occupancy was increased in absence of several chromatin modifiers and rrp6 (Figure 2.3A and Supplemental S3C, D, S4E, S5A). Finally, the Ccr4–Not complex components showed little effect on RNA Pol II occupancy at all heterochromatic regions (Figure 2.3B and Supplemental S6A–C).

The reduced transcriptional efficiency at heterochromatic loci depends on most chromatin modifiers (Clr4, HP1 proteins and SHREC complex), but not on RNA degradation factors Rrp6 and Exo2 (Figure 2.3C and Supplemental S4–S6). Those chromatin modifiers seem to reduce transcriptional efficiency more than RNA Pol II occupancy (Supplemental S3), an effect similar to what we had observed in clr4Δ cells. The strongest effect on transcriptional efficiency in heterochromatic regions was observed for components of the Ccr4–Not complex (Figure 2.3D and Supplemental S3). The Ccr4–Not deadenylase complex could affect transcriptional efficiency indirectly, through changes in RNA levels of other factors involved in heterochromatin formation. However, analysis of nascent RNA, total RNA and pA RNA data show that RNA levels of factors involved in heterochromatin formation do not change substantially in caf1Δ cells compared to wild-type cells (Supplemental S7E). This suggests a direct effect of Ccr4–Not on transcriptional efficiency.

Although RNA Pol II occupancy at tlh repeats is comparable in caf1Δ and wild-type cells, caf1Δ cells produce ∼10 fold more nascent RNA from that locus than wild-type cells (Figure 2.3E). Thus transcriptional efficiency is increased in all heterochromatic regions in caf1Δ cells, to a level comparable to protein-coding genes (Figure 2.3D, F). To determine if Ccr4–Not mediated reduction in transcription efficiency occurs post-heterochromatin formation, we analyzed H3K9me levels in wild-type and caf1Δ cells. Notably, H3K9me levels are only slightly affected at tlh and centromeric dg/dh repeats in caf1Δ cells (Supplemental S7F), indicating that increased transcriptional efficiency does not interfere with H3K9me deposition. Thus, Caf1 reduces transcriptional efficiency at heterochromatin loci after H3K9me is deposited and heterochromatin is established; in fact, Caf1 requires H3K9me and heterochromatin for its activity. In absence of heterochromatin, transcriptional efficiency is strongly increased, as seen by our data on clr4Δ (Figure 2.2C). Moreover, in caf1Δago1Δ cells, which lack both H3K9me and small RNAs (Brönner et al., 2017), transcriptional efficiency is not further increased (S8), suggesting that Ccr4–Not is the primary cause of reduced transcriptional efficiency in heterochromatin. In sum, our data show that the Ccr4–Not complex is required to reduce transcriptional efficiency in heterochromatic regions and that it modulates RNA Pol II activity in a heterochromatin-dependent way.

To test if the deadenylase activity of the Ccr4–Not complex is required for reduction of transcriptional efficiency in heterochromatin, we introduced point mutations into the active site of the two deadenylases in the complex, Caf1 and Ccr4 (Brönner et al., 2017). The mutations in the active sites led to an increase in transcriptional efficiency at heterochromatin loci that was comparable to the effect seen with gene deletions (Figure 2.3D), suggesting that the nuclease activity is required for reduced transcriptional efficiency. The deadenylase activity of the Ccr4–Not complex requires an accessible 3′ RNA end, which is not the case with nascent RNA, where the 3′ end is engaged with RNA Pol II. It is possible that our Pol II-RIP assay also detects chromatin-bound RNA that are targeted by the Ccr4–Not complex (Brönner et al., 2017), thus contributing to reducing transcriptional efficiency. Alternatively, Ccr4–Not might bind backtracked nascent RNAs (Dutta et al., 2015; Kruk et al., 2011)) which would have accessible 3′ ends, and this could potentially stall RNA Pol II.

In support of a direct effect on transcription, we observed changes in RNA Pol II dis-

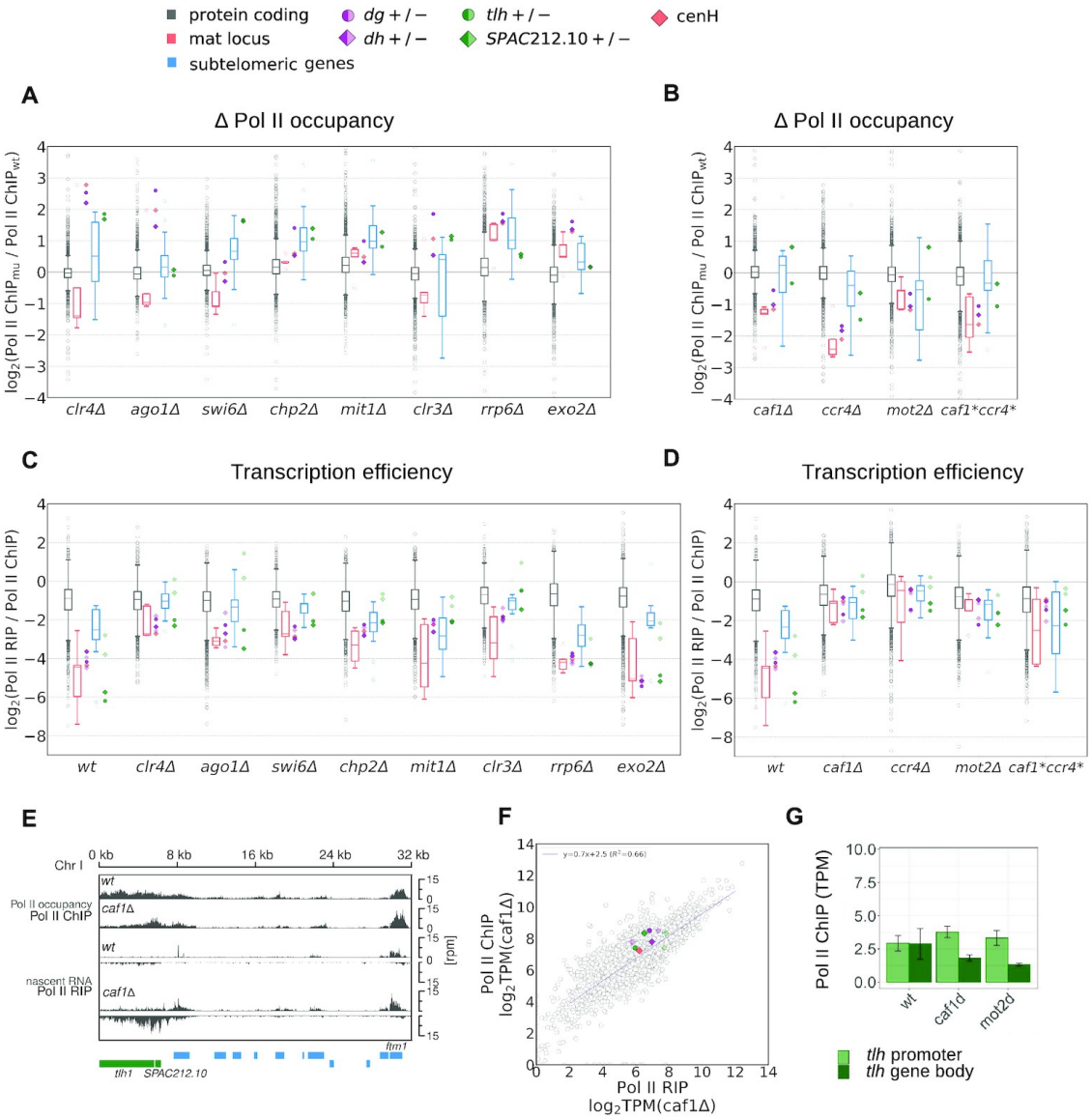

Figure 2.3: RNA Pol II occupancy and transcription efficiency in different mutants. (A, B) Box plots showing ratio of RNA Pol II occupancy (S2P-Pol II ChIP-seq) in mutants compared to wild type over indicated genes. Shown are mutants in factors involved in heterochromatin formation and RNA degradation (A) or Ccr4–Not complex components (B). Data are plotted as defined for Figure 2.1B. (C, D) Box plot showing transcription efficiency (Pol II RIP / Pol II ChIP) over indicated genes in wild type and mutants in factors involved in heterochromatin formation and RNA degradation (C) or in Ccr4–Not complex components (D). Data are plotted as defined for Figure 2.1B. (E) Analysis of the next-generation sequencing data showing occupancy of S2P RNA Pol II (ChIP-seq) and nascent RNA (S2P-Pol II RIP-seq) at subtelomeric regions in S. pombe wild-type and caf1Δ cells. Gene locations are indicated as boxes below the coverage and color-coded: green, subtelomeric loci tlh and SPAC212.10; blue, other subtelomeric genes. (F) Transcription efficiency in caf1Δ cells. S2P-Pol II ChIP-seq (Pol II occupancy) data plotted over S2P-Pol II RIP-seq data (nascent RNA). TPM, transcripts per million. Gray circles are individual protein-coding genes; regression line is also shown in purple. Also plotted are centromeric dg and dh (dark purple for + strand, bright purple for – strand) and tlh and SPAC212.10 (dark green for + strand, bright green for - strand). Each data point is the average of at least two independent samples. (G) Quantification of RNA Pol II occupancy (S2P-Pol II ChIP-seq) at tlh promoter region and tlh gene body in indicated wild type and mutant strains. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

tribution upon deletion of the Ccr4–Not complex at the tlh locus (Figure 2.3E, G and Supplemental S9A). Whereas RNA Pol II occupancy at the promoter and 5′ end of the tlh gene in caf1Δ and wild-type cells is comparable (Figure 2.3E), we observe strong reduction of RNA Pol II occupancy in the gene body in the mutant, suggesting higher RNA Pol II processivity (Figure 2.3E, G). Similar changes are observed in deletions of all components of the Ccr4–Not complex (Supplemental S9A), especially in mot2Δ cells, but not in deletions of chromatin complexes such as swi6 or clr3 (Supplemental S9B). Notably, deletion of Ccr4–Not components did not change RNA Pol II profiles in euchromatic regions (Supplemental S9C). These data indicate that the Ccr4–Not complex directly affects the RNA Pol II distribution in heterochromatic regions.

Our data reveal that low RNA stability contributes primarily to the silencing of centromeric dg and dh transcripts (Figure 2.4A and Supplemental S3A), with exception of dh+ transcripts which are degraded in an RNAi- and heterochromatin-independent way. This effect requires RNAi (Ago1) and H3K9me (Clr4), but not other heterochromatic factors or individual RNA degradation factors examined (Figure 2.4A). Among the latter, only deletion of exo2 showed a small increase in RNA stability at the centromeric region, suggesting that multiple RNA degradation pathways act redundantly to degrade heterochromatic transcripts (Figure 2.4A and Supplemental S3A). In contrast, deficiency in the Ccr4–Not complex components increased degradation of heterochromatic transcripts compared to wild-type cells. This effect is consistent with increased transcriptional efficiency in this mutant, leading to higher amounts of nascent RNA (Figure 2.4B and Supplemental S3, S9D), which would recruit the RNAi machinery to these regions, thus leading to increased degradation (Brönner et al., 2017).

2.3.3 Contribution of individual proteins to the heterochromatic silencing pathways

For all the mutant strains examined here, we quantified the level of silencing imposed by each of the three pathways (Pol II occupancy, transcriptional efficiency, RNA stability) at different heterochromatic loci (Figure 2.4C, D and Supplemental S9E, F). It should be also noted that the three pathways do not operate independently in cells. The quantification of each pathway's contribution to heterochromatic silencing was calculated relative to clr4Δ, which was defined as a complete loss of heterochromatic silencing. We observed that Ago1 is essential for all three silencing pathways at the centromeric region; in fact, heterochromatic silencing is completely lost in strains lacking RNAi. In the strains bearing deletions of chromatin modifiers, silencing overall was strongly reduced, but each pathway was still active. In contrast, deletions of RNA degradation factors led to more limited loss of silencing, with varying effects between centromeric and subtelomeric regions.

We had previously shown that the Ccr4–Not complex degrades subtelomeric RNA redundantly with RNAi (Brönner et al., 2017). Our new data show that the Ccr4–Not complex is also required for silencing at the transcriptional level, regulating transcriptional efficiency at all heterochromatic loci (Figure 2.4C, D and Supplemental S3). Although transcriptional efficiency is increased in caf1Δ, ccr4Δ and mot2Δ cells, the overall loss of silencing at centromeric repeats is small and increased transcriptional efficiency is compensated by reduced RNA Pol II occupancy and increased RNA degradation (Figures 3B, D, 4B and Supplemental S3). The loss of silencing in caf1Δ, ccr4Δ and mot2Δ cells was more pronounced at subtelomeric tlh repeats (compared to centromeric loci), since transcriptional efficiency is the dominant silencing pathway in those regions, and increased degradation by RNAi (Brönner

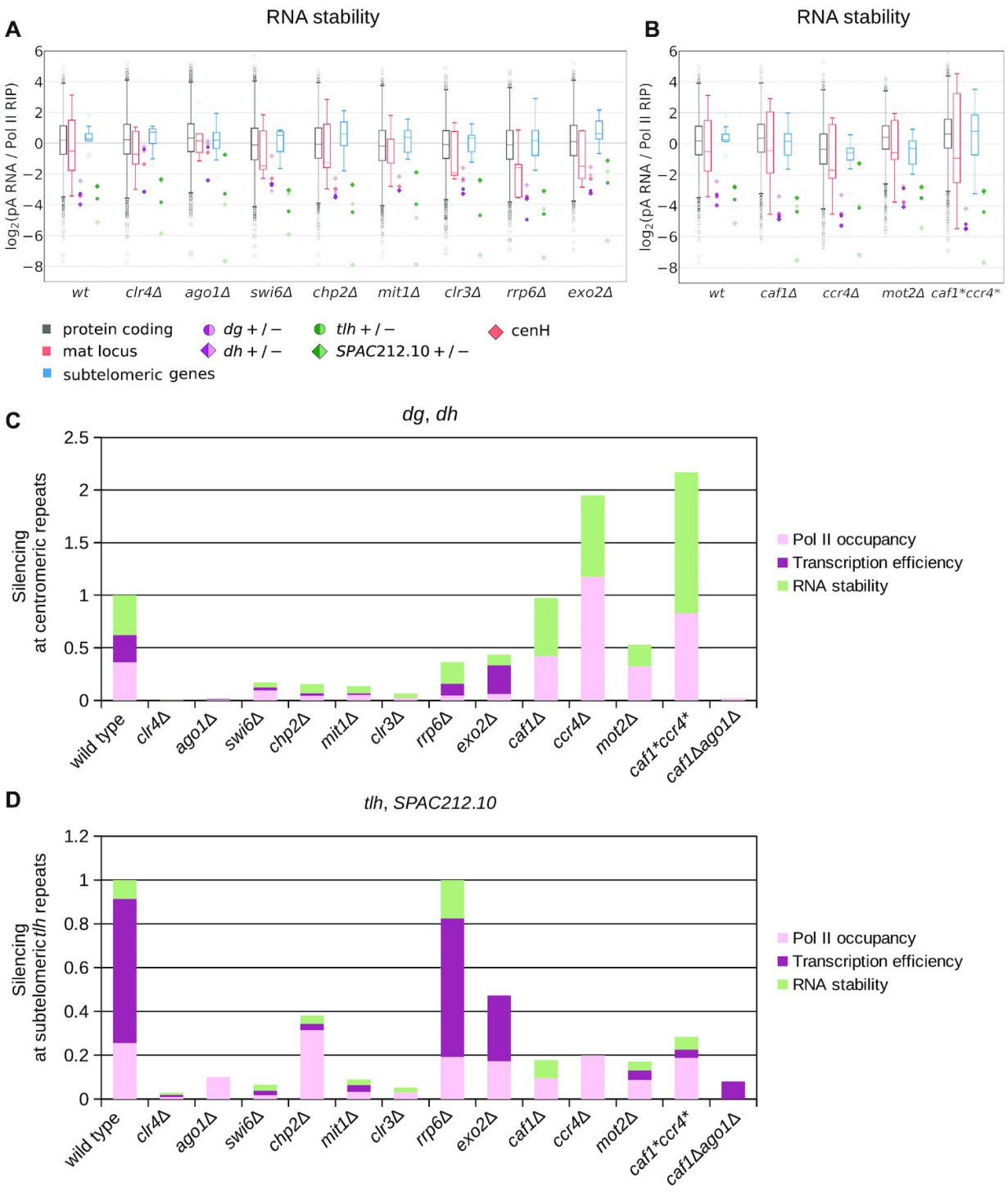

Figure 2.4: Contribution of individual factors to each pathway of heterochromatic silencing. (A, B) Box plot showing RNA stability (pA RNA/Pol II RIP) over indicated genes in wild type and mutants in factors involved in heterochromatin formation and RNA degradation (A) or Ccr4–Not complex components (B). Data are plotted as defined for Figure 2.1B. (C, D) Bar charts displaying contribution of each of the pathways that are still active in the mutants to silencing at centromeric dg/dh regions (C) and subtelomeric tlh regions (D). The height of each bar corresponds to the fold change in RNA output relative to wild-type. The relative contribution of each pathway was computed as fold change in quantitative measures (ratios of average TPM, see Materials and Methods) relative to clr4Δ. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

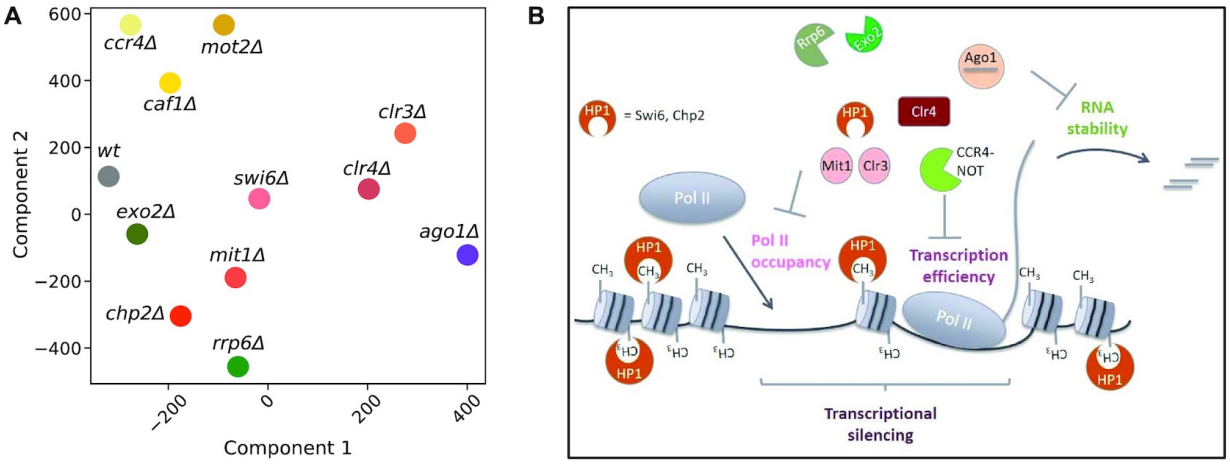

Figure 2.5: Heterochromatic silencing is a combination of reduced RNA Pol II occupancy, Transcription Efficiency and RNA stability. (A) The t-distributed stochastic neighbor embedding (t-SNE) plot showing two dimensional embedding of Pol II occupancy and transcription efficiency. Close proximity of mutants visualizes similarities in transcriptional silencing. (B) Schematic presentation of how different proteins involved in heterochromatin formation or RNA degradation contribute to heterochromatic silencing. Three pathways are important for heterochromatic silencing: transcriptional silencing (consisting of Pol II occupancy and transcription efficiency) and RNA degradation. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

et al., 2017) is not sufficient to compensate for its loss (Figure 2.4C, D). Notably, at subtelomeric regions other than tlh and at the mat locus, the transcriptional efficiency pathway remains active in almost all mutant strains examined; the exceptions were mutants for the components of the Ccr4–Not complex, in which transcriptional efficiency is abolished but other pathways were fully functional (Supplemental S9E, F). These data show that Ccr4–Not complex specifically affects transcriptional efficiency.

We projected log transformed TPM values of Pol II occupancy and transcription efficiency across all heterochromatic genes into two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE). In agreement with our previous calculations (Figures 3 and 4), the t-SNE plot shows that the components of the Ccr4–Not complex caf1, ccr4 and mot2 co-localize, indicating their specialized role in transcriptional silencing (Figure 2.5A).

2.4 Discussion

Our data show that heterochromatic silencing occurs through different mechanisms. First, RNA Pol II accessibility is reduced at heterochromatic regions, which also reduces overall transcription at those loci (Figure 2.5B). Second, we identified transcriptional efficiency as a new mode of heterochromatic silencing; although RNA Pol II is present at heterochromatic loci, transcriptional efficiency is reduced by heterochromatin and the Ccr4–Not complex (Figure 2.5B). This mode of silencing operates/occurs at all heterochromatic regions in fission yeast. Third, the final component of heterochromatic silencing is RNA degradation by RNAi and several RNA degradation factors, including the Ccr4–Not complex.

Reduced RNA Pol II access was initially proposed as the major mode of heterochromatic silencing and has been observed in many organisms (Feng and Michaels, 2015; Grewal and Elgin, 2007). In fact, we observed that RNA Pol II occupancy at centromeres strongly increased in cells bearing mutations that greatly affect H3K9me levels, such as clr4Δ or ago1Δ (Okita et al., 2019). Components of the heterochromatin pathway that act downstream of H3K9me, such as HP1 proteins, are required to reduce RNA Pol II occupancy, but to a lesser extent than Clr4. Notably, RNA degradation factors Exo2 and Rrp6 contribute to reduced RNA Pol II occupancy at the centromeric region, in agreement with our previous finding that RNA degradation is required for formation of heterochromatic domains (Brönner et al., 2017).

In addition to reduced RNA Pol II occupancy, we identified reduced transcriptional efficiency as another mode of transcriptional silencing. Although RNA Pol II is present at heterochromatic regions, the RNA it produces (nascent RNA levels) is proportionally lower than in euchromatic regions. This silencing mode occurs at all heterochromatic regions in fission yeast cells and might be conserved in other organisms as well. In fact, this mode of silencing might be analogous to SIR-mediated silencing in S. cerevisiae, wherein transcription is initiated but elongation is blocked by the SIR complex, which maintains RNA Pol II in a stalled conformation (Johnson et al., 2013). We found that heterochromatin is essential to reduce transcriptional efficiency: in the absence of H3K9me, transcription efficiency at heterochromatic loci is increased to the level comparable to euchromatic genes. These results also show that reduced transcriptional efficiency is not encoded in the DNA sequence itself, but it is controlled by the heterochromatin. We show that reduced transcriptional efficiency is mediated by the Ccr4–Not complex, as transcriptional efficiency increased to the level of protein-coding genes in caf1Δ cells. Notably, H3K9me levels at subtelomeric tlh and centromeric dg/dh repeats were only modestly affected in caf1Δ cells (Brönner et al., 2017; Cotobal et al., 2015; Sugiyama et al., 2016), indicating that heterochromatin formation is functional in those cells. Thus, the Ccr4–Not complex regulates transcriptional efficiency post-heterochromatin formation, but H3K9me and heterochromatin are required for this regulation by the Ccr4–Not complex. The observed reduction in transcription efficiency is likely a combination of reduced transcription by RNA Pol II and degradation of chromatin bound RNA. RNAi, Ccr4–Not and the Rix1 complex were suggested to co-transcriptionally degrade heterochromatic transcripts in fission yeast (Holoch and Moazed, 2015; Verdel et al., 2004; Brönner et al., 2017; Holla et al., 2020; Shipkovenska et al., 2020). In wild-type fission yeast, centromeric dg/dh transcripts are targeted by RNAi, whereas subtelomeric tlh transcripts are not. Our data show that transcriptional efficiency is reduced at both centromeric dg/dh transcripts and subtelomeric tlh transcripts, suggesting that the reduction does not occur via

co-transcriptional degradation by RNAi. Moreover, in caf1Δ cells transcriptional efficiency in heterochromatin is comparable to protein coding genes, however, both tlh and centromeric dg/dh transcripts remain to be degraded by RNAi (Brönner et al., 2017). This indicates that RNAi degradation has only a minor contribution to the observed reduction in transcriptional efficiency. Notably, our data show that mutations in the active site of the Ccr4–Not deadenylases lead to an increase in transcriptional efficiency comparable to that seen with gene deletions, suggesting that nuclease activity is required for reduced transcriptional efficiency. This also suggests that the Ccr4–Not complex might degrade chromatin-bound RNA, which might co-purify with nascent RNA. Alternatively, the deadenylase activity of Ccr4–Not might directly regulate RNA Pol II. In support of this, we observe changes in the RNA Pol II distribution, which was reduced in tlh gene body upon deletion of Ccr4–Not subunits. This suggests higher processivity of elongating RNA Pol II in those mutants compared to wild-type cells. Ccr4–Not was initially described as a chromatin-associated complex involved in transcription (Miller and Reese, 2012), and later shown to act as a transcription elongation factor that would reactivate arrested RNA Pol II (Dutta et al., 2015; Kruk et al., 2011). It is plausible that, in the presence of heterochromatic marks, Ccr4–Not exhibits the opposite effect and stalls RNA Pol II, perhaps by targeting backtracked nascent RNAs that have accessible 3′ ends. Recently, Ccr4–Not was shown to be required for DNA-damage dependent ubiquitination and degradation of RNA Pol II (Jiang et al., 2019), and stalling of RNA Pol II by Ccr4–Not could require RNA Pol II ubiquitination. Furthermore, the Ccr4–Not complex is recruited to RNA Pol II by the histone chaperone Spt6 (Dronamraju et al., 2018), which was also implicated in heterochromatin formation in fission yeast (Kiely et al., 2011). Altogether, these various observations support the concept that Ccr4–Not is recruited to chromatin and regulates RNA Pol II transcription.

Our data show that RNA degradation contributes to heterochromatic silencing at centromeric repeats and tlh, but not at other subtelomeric genes or at the mat locus. RNA degradation at the centromeric region is dependent on RNAi and heterochromatin, but those are not required for RNA degradation at subtelomeric tlh repeats. This observation is in agreement with the previous report that RNA is degraded at subtelomeric tlh repeats by parallel mechanisms that are heterochromatin-dependent and -independent (Brönner et al., 2017).

In conclusion, we identified a new mode of heterochromatic silencing termed transcriptional efficiency. This mode of silencing depends on H3K9me and the Ccr4–Not complex and acts as a dominant silencing pathway at most heterochromatic loci in fission yeast.

Chapter 3

Dynamic pseudo-time warping of complex trajectories

This Chapter builds on top of our previous work:

- Van Hoan Do*, Mislav Blažević*, **Pablo Monteagudo-Mesas**, Luka Borozan, Khaled Elbassioni, Soeren Laue, Francisca Rojas Ringeling, Domagoj Matijevic and Stefan Canzar. *Dynamic pseudo-time warping of complex single-cell trajectories.* bioRxiv, 2019.

The main contributions presented in this Chapter are being summarized into a manuscript:

- In preparation: **Pablo Monteagudo-Mesas***, Van Hoan Do*, Mislav Blažević*, Luka Borozan, Khaled Elbassioni, Soeren Laue, Francisca Rojas Ringeling, Domagoj Matijevic and Stefan Canzar. *TrajanR enables the accurate alignment and comparison of complex scRNA-seq trajectories.* (2023).

Section 3.1 (Introduction), is based on our prior work (Do et al., 2019) and has been revised in order to contextualize our method in light of recent published work. In Section 3.2 (Preliminaries), we motivate and introduce the problem of aligning complex single-cell RNA sequencing (scRNA-seq) trajectories, as well as the limitations of current approaches based on DTW to align linear scRNA-seq trajectories. This content has also been adapted from our prior work (Do et al., 2019).

After that, our main contributions are presented:

- Implementation of TrajanR package
- Extensive experimentation on simulated datasets
- Metric conformity analysis
- Analysis of two real datasets

Additional experiments and theoretical work, which were included in our original publication, are described in (Do, 2021) and (Borozan, 2021), and are referenced for completeness but omitted from this Chapter.

3.1 Introduction

Single-cell RNA sequencing (scRNA-seq) has allowed the detailed dissection of biological processes such as differentiation, development, and cell reprogramming. Using so-called trajectory inference (TI) methods (Saelens et al., 2019), like Monocle2 (Qiu et al., 2017), PAGA (Wolf et al., 2019) or Slingshot (Street et al., 2018), cells in a scRNA-seq dataset are ordered based on their transcriptomic similarities into trajectories describing continuous transitions between cells states. Progression in these scRNA-seq trajectories is parametrized by pseudotime, defined for each cell as the distance along the trajectory from the trajectory's origin or root. The analysis of such trajectories enables the characterization of changes in gene expression driving these dynamic processes. An alternative approach to studying cells undergoing dynamic processes is RNA velocity, where splice and unspliced counts are used to indirectly infer cells' future gene expression states based on their current state. Although it has limitations, RNA velocity provides valuable and complementary information. For example, it can help in finding a trajectory's root or, more recently, in guiding the whole trajectory inference process (e.g., CellPath (Zhang et al., 2021) and CellRank (Lange et al., 2022)). As described throughout this Chapter, much can be learned from the comparative analysis of scRNA-seq trajectories. Comparing gene expression dynamics along trajectories from two conditions can aid in elucidating the key differences between them and the regulatory programs underpinning the biological process under study. For example, comparing the trajectories underlying a given differentiation process in two species would shed light on the evolutionary differences between these organisms. Comparing the trajectory describing a normal developmental process to that affected by a particular mutation would yield insights into disease mechanisms.

The main issue preventing the comparison of gene expression dynamics along a given pair of trajectories is that pseudotime is defined in a trajectory-specific manner, rendering the actual pseudotime values between trajectories not directly comparable. Dynamic time warping (DTW) is a class of algorithms for comparing two time-series that advance at different speeds (Vintsyuk, 1968). It was originally developed in the context of automatic speech recognition but has gained increasing popularity in the comparison of single-cell trajectories (Alpert et al., 2018; Cacchiarelli et al., 2018; Ellwanger et al., 2018). Similar to a pairwise sequence alignment that allows for insertions and deletions, DTW finds a mapping (warping) between similar elements in the two sequences to overcome locally stretched and compressed sections. In single-cell trajectories, where cells are ordered along pseudotime, gene expression values are used to establish a common pseudotime axis along which expression kinetics become comparable between different conditions by matching similar cells in both trajectories. An alternative approach, in which one combines both scRNA-seq datasets and attempts to learn a joint trajectory, is likely to fail due to batch effects confounding the trajectory inference process or differences between biological processes that cannot be easily reconciled into a single trajectory (Cacchiarelli et al., 2018). Moreover, in Sugihara et al. (2022), the authors demonstrated the superior accuracy and robustness of DTW alignments when compared against state-of-the-art data integration methods, where the datasets were integrated and subsequently, a common trajectory was inferred on the merged dataset.

The main limitation of DTW is that it can only compare two time-series at a time. Thus current methods for comparing single-cell trajectories are restricted to linear trajectories. Complex single-cell trajectories (e.g., branching, tree-like, etc.) are needed to characterize many biological processes in development and differentiation and in response to perturbations (Qiu et al., 2017). When dealing with complex trajectories, one must manually select individual paths or lineages from both trajectories before applying DTW. In these cases, prior information, such as a set of specific marker genes, would be necessary to pick the correct or most relevant path to align, but this information is often unavailable. Another potential caveat of DTW is that it ignores cells on alternative paths and could amplify the signal used to infer the mapping between trajectories. Lastly, very recent work (Laidlaw et al., 2022; Sumanaweera et al., 2023)has highlighted further limitations of DTW when aligning trajectories containing unmatched regions due to its constraint to match every time-point in one trajectory to at least one time-point in the other.

In our previous work (Do et al., 2019), we introduced Trajan, a novel method to compare and align complex scRNA-seq trajectories with multiple branch points diverting cells into alternative fates (Figure 3.3). Trajan automatically identifies the correspondence between biological processes in two trajectories and aligns all of them simultaneously, taking into account their overlap. Since cells diverted into different fates share a common ancestry, they cannot be treated as independent. Their independent pairwise alignment (using DTW) could introduce inconsistencies concerning the mapping of common progenitor cells. Akin to the extension of pairwise alignments to multiple sequence alignment, we seek the best alignment between all corresponding pairs of paths that agree on common progenitor cells. To this end, Trajan adopts arboreal matchings (Böcker et al., 2013) to capture globally consistent similarities between trajectories.

In the following Sections, we demonstrate the accuracy of Trajan alignments through extensive experimentation on simulated data. We also introduce TrajanR, an R package designed to improve the standardization and pre-processing of Trajan's input data and showcase its utility on two separate real datasets.

3.2 Preliminaries

In the following sections, we review the concept of DTW, the current approach to the alignment of *linear* trajectories, and how to generalize to the alignment of *complex* trajectories using arboreal matchings. Arboreal matchings were introduced in the context of phylogenetic trees (Böcker et al., 2013) and set the theoretical basis for alignment of scRNA-seq trajectories with Trajan. The review on DTW and generalization to arboreal matchings has been adapted from our original publication (Do et al., 2019). We have summarized the different implementations available in Trajan and refer the reader to our previous work for additional information and mathematical proofs.

3.2.1 Dynamic time warping

Dynamic time warping (DTW) is the algorithmic workhorse underlying current alignment methods used to compare *linear* single-cell RNA-seq trajectories (Alpert et al., 2018; Cacchiarelli et al., 2018). As in classical sequence alignment, DTW matches similar elements in two sequences while preserving their order. However, since the main idea is to account for the different speeds at which the two sequences advance, each element in one sequence can be mapped to one or more elements in the other sequence, see Figure 3.1 (left).

More formally, given two *time-series*:

$$(x_i)_{i=1}^{n}, (y_j)_{j=1}^{m} \tag{3.1}$$

and a *distance* or *similarity measure* between time-points:

$$d(x_i, y_j) \geq 0, \quad \forall x_i \in (x_i)_{i=1}^{n}, \forall y_j \in (y_j)_{j=1}^{m} \tag{3.2}$$

We define a *warping*:

$$p = (p_1, \ldots, p_L) \tag{3.3}$$

as a sequence of index-pairs: $p_\ell = (n_\ell, m_\ell) \in [1:n] \times [1:m]$ for $\ell \in [1:L]$, that satisfies the following three conditions:

i *Boundary*: $p_1 = (1,1)$ and $p_L = (n,m)$.

ii *Monotonicity*: $n_1 \leq n_2 \leq \cdots \leq n_L$ and $m_1 \leq m_2 \leq \cdots \leq m_L$.

iii *Step size*: $p_{\ell+1} - p_\ell \in \{(1,0),(0,1),(1,1)\}$ for $\ell \in [1:L-1]$.

Note that, as exemplified in Figure 3.1, an alignment that preserves the order between two sequences, represented as matched edges between time-points, may not contain a pair of crossing edges.

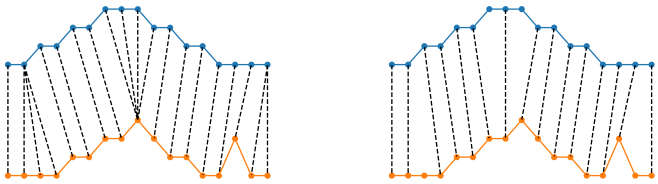

Figure 3.1: An example of a warping **(left)** and an arboreal matching **(right)** between two time-series. Figure reproduced from original publication (Do et al., 2019).

The goal of DTW is to find a *warping* p such that the total distance between mapped elements is minimized:

$$c_p(x, y) := \sum_{\ell=1}^{L} d(x_{n_\ell}, y_{m_\ell}).$$

In classic DTW, the optimal warping can be computed by a dynamic program (DP) that solves:

$$D(i, j) = d(x_i, y_j) + \min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\}. \tag{3.4}$$

There are various extensions of the classic DTW described above that can be mainly classified as:

- Restricting the range of the mapping to a certain window (Laidlaw et al., 2022)
- Assigning different weights to different types of steps
- Using different step patterns (e.g. $p_{\ell+1} - p_\ell \in \{(1, 1), (1, 2), (2, 1)\}$)

If not stated otherwise, throughout this Chapter we consider the classic DTW as the default scheme (Zhao and Itti, 2018), similar to previous publications (Alpert et al., 2018; Cacchiarelli et al., 2018; Cannoodt et al., 2021). Classic DTW provides enough flexibility for most single-cell alignment tasks, however, we note that DTW might not be robust under the choice of these parameters. In Section 3.4.1, we used DTW with different step patterns to compute alignments for the same dataset, and show that results vary substantially between different schemes. Similarly, as discussed later, the choice of a penalty scheme for individual cells during Trajan alignments will influence the overall results.

3.2.2 Arboreal matchings

We propose a generalization of DTW from time-series (i.e. linear trajectories or individual paths) to rooted trees (i.e. non-linear or complex trajectories). Each lineage or path in tree T_1 should be aligned to at most one lineage/path in T_2, and vice versa, while preserving the order of nodes along the paths. In addition, we require all alignments to be consistent, that is, every node must be matched to the same node in all pairwise alignments it is part of. Such notion of consistent path-by-path alignment of trees, was first introduced in Böcker et al. (2013) and termed *arboreal matching*.

More formally, given two *rooted trees*:

$$T_1 = (V_1, E_1), \quad T_2 = (V_2, E_2) \tag{3.5}$$

and a *distance* or *similarity measure* between nodes in the trees:

$$d(u_i, v_j) \geq 0, \quad \forall u_i \in V_1, \forall v_j \in V_2, \tag{3.6}$$

We define an *arboreal matching*:

$$M = (m_1, \ldots, m_L) \tag{3.7}$$

as a one-to-one correspondence between nodes in trees T_1 and T_2 such that:

$$u_2 \text{ is a descendant of } u_1 \iff v_2 \text{ is a descendant of } v_1, \quad \forall (u_1, v_1), (u_2, v_2) \in M \tag{3.8}$$

The task at hand is an *optimization* problem in which we want to find an arboreal matching M that minimizes the cost in Equation 3.9, so that the total distance between mapped elements is minimized (1st term), while flexibly penalizing those elements that remain unmatched in M (2nd and 3rd terms):

$$c(M) := \sum_{(u,v)\in M} d(u, v) + \sum_{\substack{u \in V_1 \\ u \text{ unmatched}}} d(u, -) + \sum_{\substack{v \in V_2 \\ v \text{ unmatched}}} d(-, v), \tag{3.9}$$

where the *cost* or *penalty scheme* of leaving a node unmatched is:

- $d(u, -) > 0$ for node $u \in V_1$
- $d(-, v) > 0$ for node $v \in V_2$

In contrast to DTW, an arboreal matching M matches each node (time-point/cell) to at most one similar node in the other tree (trajectory), allowing for nodes in each pair of trees to be unmatched. In light of recent work (Laidlaw et al., 2022; Sumanaweera et al., 2023), the flexibility to leave regions of the trajectory unmatched should be regarded as a strength of our method, as it addresses a limitation of DTW, where mismatched regions between trajectories are not properly handled.

An example arboreal matching between two simple paths is shown in Figure 3.1 (right). In this case, non-crossing edges match similar regions between the two time-series, while compressed/stretched or distinct regions are represented by unmatched nodes. Note that, when comparing time-series or simple paths (i.e. linear trajectories), arboreal matchings are as flexible as DTW. In the methods and experiments of our original publication (Do et al., 2019), we show that given an appropriate penalty for unmatched nodes, the optimal DTW and the optimal arboreal matching produce comparable distance/similarity measures. In Section 3.4.1, we further demonstrate that when aligning linear trajectories DTW and Trajan produce solutions of comparable accuracy based on a novel metric proposed in Cannoodt et al. (2021), the ABWAP score.

In the following Section, we summarize the different strategies available in Trajan for computing arboreal matchings, including both exact/optimal and heuristic/sub-optimal solutions.

3.2.3 Computing arboreal matchings with Trajan

Trajan provides 2 main implementations for computing arboreal matchings:

- An efficient **branch-and-cut** algorithm
- An **FPT** algorithm for small number of cell fates

Although not explicitly implemented in Trajan, in the following Section we briefly describe a **naive ILP** formulation for solving the *minimum weight* arboreal matching problem.

For a more in depth discussion of each implementation see the original publication (Do et al., 2019). For further details on how to run Trajan, and an example workflow using trajectories generated with Monocle2 (Qiu et al., 2017), we refer the reader to Trajan's GitHub repository (`https://github.com/canzarlab/Trajan`).

Naive ILP formulation

As described in (Böcker et al., 2013), we aim to find a *minimum weight* arboreal matching between two rooted trees $T_1 = (V_1, E_1)$, $T_2 = (V_2, E_2)$, i.e. an arboreal matching that minimizes the cost in Equation 3.9. In order to solve this minimization problem, we rephrase it as a maximization problem (Equation P), such that the cost of the optimal alignment differs as much as possible from the cost of the worst possible alignment, where all cells are unmatched. We formulate an ILP with constraints (3.12) that explicitly forbid the two possible types of *ancestry violations* ($\mathcal{I}$), see Figure 3.2.

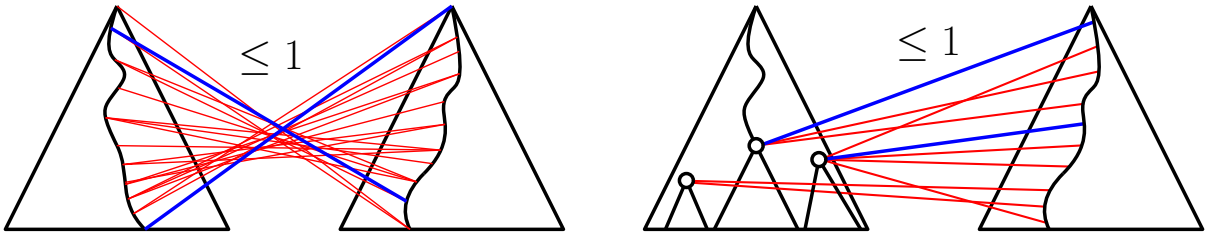

Figure 3.2: Pair of crossing edges (blue) extended to a clique of crossing edges **(left)** and pair of semi-independent edges (blue) extended to a clique of semi-independent edges **(right)**. Figure reproduced from original publication (Do et al., 2019).

Feasible arboreal matchings can **not** include:

i A pair of crossing edges, ensuring that the order of nodes in both trees is preserved equivalently to DTW, Figure 3.2 (left).

ii A pair of nodes on the same root-to-leaf path that match a pair of nodes on different root-to-leaf paths, ensuring the consistenncy of the alignment, Figure 3.2 (right).

The first type of constraint is analogous to enforcing monoticity in DTW, while the second type of constraint stems from the simultaneous comparison of multiple paths and prevents arbitrary jumps between branched biological processes.

$$\max \sum_{i=1}^{|V_1|} \sum_{j=1}^{|V_2|} w(i,j) x_{i,j} \tag{P}$$

$$\text{s.t.} \sum_{j=1}^{|V_2|} x_{i,j} \leq 1 \qquad \forall i = 1 \dots |V_1|, \tag{3.10}$$

$$\sum_{i=1}^{|V_1|} x_{i,j} \leq 1 \qquad \forall j = 1 \dots |V_2|, \tag{3.11}$$

$$x_{i,j} + x_{k,l} \leq 1 \qquad \forall \{(i,j),(k,l)\} \in \mathcal{I}, \tag{3.12}$$

$$x_{i,j} \in \{0,1\}, \tag{3.13}$$

where:

- $x_{i,j}$: indicator variables that denote the presence or absence of an edge (i,j)
- $w(i,j) = d(i,-) + d(-,j) - d(i,j)$: weights associated with an edge (i,j)

Note that, additional constraints (3.10) and (3.11) ensure that nodes in each respective tree are selected at most once.

As demonstrated in our previous experiments (Do et al., 2019), such a *naive* ILP formulation does not allow for the practical alignment of trajectories consisting of as few as 100 cells. In Do et al. (2019), we presented a meticulously designed *branch-and-cut algorithm* that enables the comparison of realistic complex single-cell trajectories. In the following Section, we outline the main ingredients of such implementation.

An efficient branch-and-cut algorithm

A branch-and-cut algorithm is a combinatorial optimization technique used to solve ILPs. It typically entails executing a branch-and-bound algorithm and utilizing cutting planes to refine LP-relaxations. Our *branch-and-cut algorithm* main ingredients are:

- Cuts that trim the LP-relaxation closer to the convex hull of feasible arboreal matchings
- Polynomial-time algorithms that can find these cuts on demand
- Several strategies to obtain integral solutions: optimal or sub-optimal
  - An *exact* branch-and-bound (bnb) scheme
  - Multiple sub-optimal *heuristic* strategies

While the first step provides a tighter relaxation of the naive ILP formulation, the corresponding solutions are still fractional in general. As outlined above, in order to obtain *optimal* integral solutions Trajan implements an exact branch-and-bound scheme, whose worst case exponential run-time can become computationally expensive for certain instances. For that reason, Trajan also implements several *heuristic* strategies to find integral solutions. These faster but approximate approaches address the need for tailored trade-offs between

accuracy and speed imposed by different single-cell sequencing technologies, which assay a variable number of genes in hundreds to millions of cells. Improvements provided by our branch-and-cut scheme can be seen in Do et al. (2019) run-time experiments.

An FPT algorithm for small number of cell fates

Finally, since the typical number of alternative fates in trajectories inferred by current TI methods is not excessively high, we have also implemented a fixed-parameter tractable (FPT) algorithm parametrized by k: the total number of cell fates or lineages in the trajectories. Typically, FPT algorithms are exponential in the size of a fixed parameter (e.g. number of cell fates) but polynomial in the size of the input (e.g. number of cells). This means that the problem can be solved efficiently for small values of the fixed parameter, and in our case, that there is an efficient algorithm for solving arboreal matchings for trajectories with small number of cell lineages. In brief, the FPT algorithm guesses the correspondence between paths in the two trajectories and applies a dynamic program similar to Zhang and Shasha (2006) to align them *optimally*. The total run-time is $\mathcal{O}(n^2m^2k!)$, where k is the smaller number of leaves among the two trees comprising n and m nodes, respectively.

3.3 Methods

In the following Section, we provide an overview of our method for the alignment of complex scRNA-seq trajectories, Trajan. Next, we describe TrajanR, our accompanying R software package to Trajan, and one of the main contributions to this Chapter. Finally, we describe how the scRNA-seq data used in our experiments was simulated using the state-of-the-art simulator, *dyngen* (Cannoodt et al., 2021), as well as the definition of various metrics used to evaluate accuracy.

3.3.1 Overview of the method

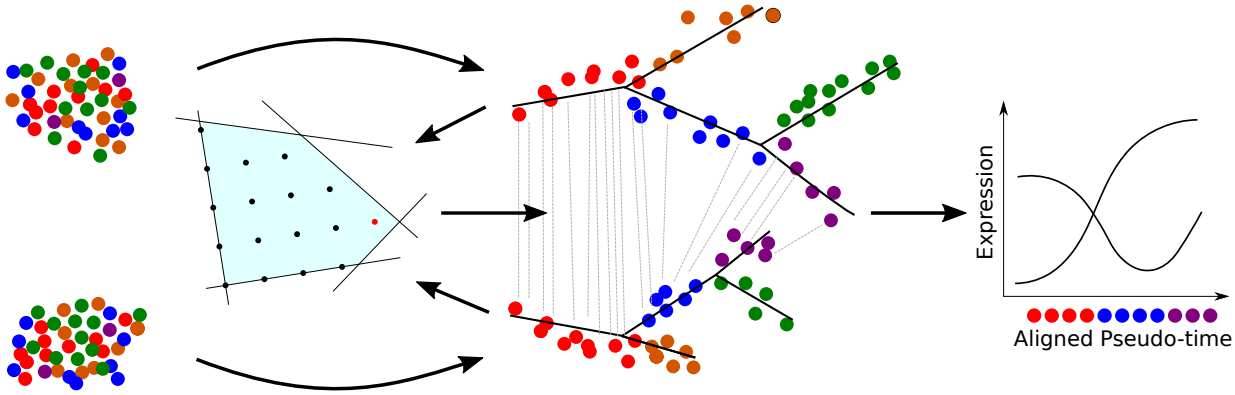

Figure 3.3: Trajan workflow. From complex scRNA-seq trajectory inference to alignment of pseudo-time scales into a common system of reference. Figure reproduced from original publication (Do et al., 2019).

The main workflow of our novel trajectory alignment tool Trajan is outlined in Figure 3.3. First, a pair of complex trajectories needs to be reconstructed from scRNA-seq data using any of the existing TI methods (Saelens et al., 2019). After adequate pre-processing of the inferred trajectories, which typically involves the selection and smoothing of highly-variable gene expression profiles, we need to transform the data into a format suitable for Trajan. The following files need to be provided for each trajectory (input/output formats are described in more detail in Trajan's GitHub repository):

- **Edge set**: t{i}.tree
- **Map**: t{i}.map

Additionally, a *distance matrix* between all pairs of cells needs to be provided, with an extra row and column that assigns a *penalty* to the corresponding cell in each trajectory.

- **Distance matrix**: distance_matrix.csv

For further details on the installation of Trajan, documentation and an example workflow based on Monocle2 generated trajectories, we refer the reader to Trajan's GitHub repository (`https://github.com/canzarlab/Trajan`).

Once all the necessary files are provided, Trajan solves a minimum weight arboreal matching problem (see Section 3.2.2) by computing an *optimal* or *approximate* alignment between trajectories, based on the user needs, by running any of the available implementations. For simplicity, in Figure 3.3, only the alignment between a pair of paths or lineages is shown, but note that Trajan finds a consistent alignment between all paths in both trajectories. Trajan's output consists of the actual *alignment* between cells in both trajectories, and the *cost* associated to the alignment solution as defined in Equation 3.9. Using matched cells as anchors, individual pseudotimes scales are set into a common system of reference along which expression kinetics are directly comparable. Moreover, the cost associated with the alignment defines a notion of distance/similarity between trajectories (see Section 3.4.2). As described in Section 3.2.3, Trajan implements multiple strategies to compute arboreal matchings between any pair of scRNA-seq trajectories. An *exact* FPT algorithm that returns solutions at a significantly reduced computational cost for trajectories with small number of cell fates. An efficient branch-and-cut algorithm with an *optimal* branch-and-bound scheme or alternatively several *heuristic* strategies that return sub-optimal but faster solutions.

In the next Section, we introduce our TrajanR package, whose integration with dynverse enables the inference of trajectories using any of the available TI methods and facilitates the pre-processing, standardization, alignment and visualization of data.

3.3.2 TrajanR: integration with dynverse

In recent work, the authors of Saelens et al. (2019) introduced the *dynverse* framework, where more than 50 trajectory inferece (TI) methods were wrapped into a common *abstraction model*. TrajanR builds upon this work by integrating Trajan with dynverse, providing all functions essential to process output data obtained from any TI method available in dynverse into input suitable for Trajan. Moreover, we provide an object-oriented framework that simplifies the computation of trajectory alignments with Trajan. This becomes specially useful when computing multiple alignments under slightly different parameter schemes, or using multiple TI methods, enabling and facilitating more complex experimentation. TrajanR also provides multiple visualization options for trajectories, trajectory alignments, and gene expression dynamics. For further details on the installation of TrajanR and use-case workflows based on test data, we refer the reader to TrajanR's GitHub repository (`https://github.com/pablommesas/trajanR`).

In brief, TrajanR builds on top of the trajectory abstraction model proposed in dynverse (Saelens et al., 2019), see Figure 3.4, which offers a common multi-layered framework for describing scRNA-seq trajectories:

- **Topology**: overall structure of the trajectory described by a network of milestones.
- **Branch assignment**: association between cells and edges in the milestone network.
- **Cell positions**: describe specific order of cells within each associated branch in the milestone network.

The abstraction model is sufficiently general that the output of most TI methods can be easily represented in this format, which, after suitable pre-processing, will serve as Trajan's

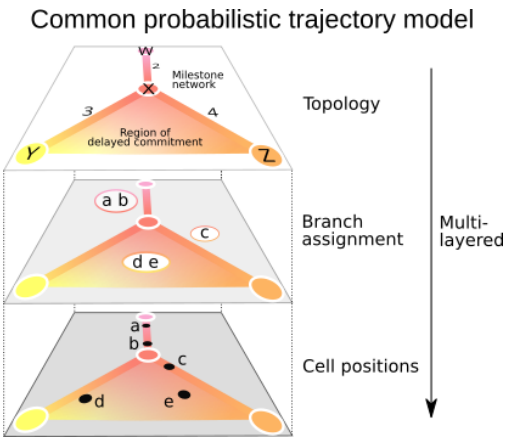

Figure 3.4: Trajectory Model Abstraction. Figure reproduced with permission from Springer Nature (Saelens et al., 2019).

input. In fact, Trajan is somewhat more restrictive than the original model in terms of the types of structures that are permitted. Thus, TrajanR is limited to the analysis of tree-like topologies (i.e. topologies with cycles are not allowed) and cells must be uniquely associated with one branch (i.e. delayed commitment regions are not allowed).

Initially, preparing input data for Trajan required the manual pre-processing of each TI method-specific output. In our original publication (Do et al., 2019), inspired by previous work (Alpert et al., 2018; Cacchiarelli et al., 2018), we created example scripts to adapt the output of Monocle 2 (Qiu et al., 2017) into a format suitable for Trajan. With the introduction of TrajanR, the pre-processing of data from any TI method has been standardized through the use of dynverse's common abstraction model. Starting from a single or multiple dyno objects containing pre-computed scRNA-seq trajectories, a Trajan object, the main data structure in TrajanR, can be easily initialized. Then, various built-in methods defined for the Trajan object ease the application of individual subsequent computational steps: export data, run alignment, visualization, etc. Moreover, this modular framework facilitated the implementation of multiple parameter schemes in TrajanR, enabling the study of the robustness of Trajan alignments. The most important parameter, the *cell representation*, has a huge influence on the results and refers to what "cells" actually represent in the abstraction model. The motivation behind the cell representation parameter is similar to the "trajectory alignment" use-case in Cannoodt et al. (2021), where the authors evaluated the impact of interpolating data when aligning linear trajectories, by comparing classic DTW (Vintsyuk, 1968) to cellAlign (Alpert et al., 2018). It can take two possible values:

- *Raw cells (r)* represent cells as present in the original scRNA-seq count/expression matrices.
- *Smoothed cells (s)* represent a, typically, smaller set of interpolated cells constructed by smoothing the gene expression profiles of raw cells.

Smoothing helps de-noising single cell measurements (Alpert et al., 2018) and, potentially, scaling computation in downstream analysis for very large number of cells. As demonstrated by our metric conformity analysis (see Section 3.4.2), these should be considered as complementary views of the data, as each mode captures different aspects of the biological process

under investigation.
Other parameters that need to be defined in the alignments, relate to more technical aspects (i.e. *distance metric*) or directly to the optimization process (i.e. *penalty scheme*), and have important but milder effects on the results. In practice, any *distance metric* can be used to compute cell-to-cell dissimilarity, which is summarized into a distance matrix for Trajan to use. In our metric conformity analysis, we explored both the Euclidean metric (*eucl*) and the Pearson correlation (*pear*). Similarly, any *penalty scheme* can be used that defines the cost of leaving cells unmatched in the alignment. In our metric conformity analysis, we explored two such penalty schemes: the average (*avg*) and the maximun (*max*) scheme, where the cost of leaving a particular cell unmatched is defined by the average and maximun distance, respectively, from that cell to every other cell.

In Section 3.4.3, we showcase how Trajan's integration with dynverse can ease and help with the analysis and visualization of complex scRNA-seq trajectories, by examining two separate *real datasets*.

3.3.3 Experiments using simulated data with Dyngen

In order to simulate data for our experiments, we used dyngen (Cannoodt et al., 2021), a recent simulation engine capable of generating scRNA-seq data based on pre-defined network topologies or backbones. Dyngen generates scRNA-seq gene expression matrices, based on a given backbone type (e.g. linear, bifurcating, etc.), meant to represent different dynamic biological processes at the single-cell level. In Section 3.4, we perform 3 separate experiments based on simulated data with dyngen. We have simulated two independent datasets, containing linear and complex trajectories, respectively, and a third dataset obtained by perturbing the dataset with complex trajectories. Thus, each dataset contains trajectories of different characteristics, that will be used to quantify different aspects of Trajan's alignments.

Similar to the "trajectory alignment" use-case presented in Cannoodt et al. (2021), our first dataset includes a total of 40 pairs of *linear* trajectories simulated using 4 slightly different linear backbones (10 pairs for each backbone type). Each trajectory pair is characterized by sharing a common gene regulatory network (GRN), but generated using different kinetics resulting in two "sub-datasets" describing a similar but non-identical dynamical process.

Our second simulated dataset includes a total of 40 pairs of *complex* trajectories, based on a binary tree backbone model with increasing levels of complexity (10 pairs for each complexity level). Complexity levels are determined by the number of branching events present in the trajectory (num_modifications), with bifurcating trajectories being the simplest (num_modifications=1) and binary trees with five alternative outcomes being the most complex ones (num_modifications=4), see Figure 3.5. The simulation process is identical to that of linear trajectories outlined above, resulting in two instances of a similar but non-identical dynamical process.

In both datasets, we will use the the Area Between Worst And Prediction (ABWAP) metric to assess the accuracy of our alignments, a novel metric described in Cannoodt et al. (2021). First, before computing the ABWAP scores, the respective pseudotimes in each trajectory need to be re-scaled between 0 and 1. As previously described, a trajectory alignment induces a matching between trajectories $(x_i)_{i=1}^n$, $(y_j)_{j=1}^m$, represented as a sequence of index pairs $\mathcal{M} = (p_1, ..., p_L)$ with $p_l = (n_l, m_l) \in [1 : n] \times [1 : m]$ for $\ell \in [1 : L]$,

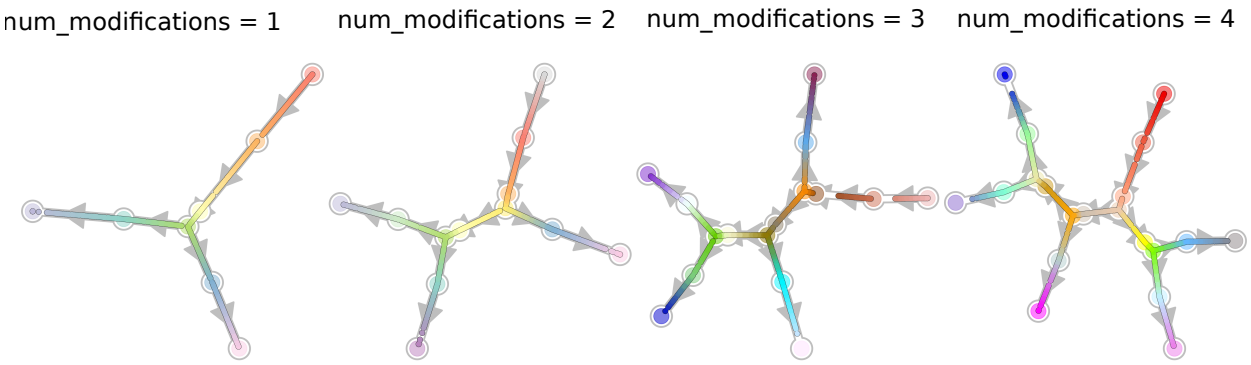

Figure 3.5: Example representations for trajectories included in the complex trajectories dataset. Depicted are the different binary tree backbones with increasing levels of complexity (num_modifications).

see Figure 3.6 (left) for a visual representation of two example matchings. Note that, the pseudotimes pt_1, pt_2 associated with each index pair are by construction ordered in ascending order. The ABWAP metric is defined as:

$$\text{ABWAP}\,(M, pt_1, pt_2) = 1 - \text{Area\_under\_curve}\,(pt_1[1, ..., L], pt_2[1, ..., L]) \tag{3.14}$$

In Figure 3.6 (right), we illustrate the intuition behind this metric.

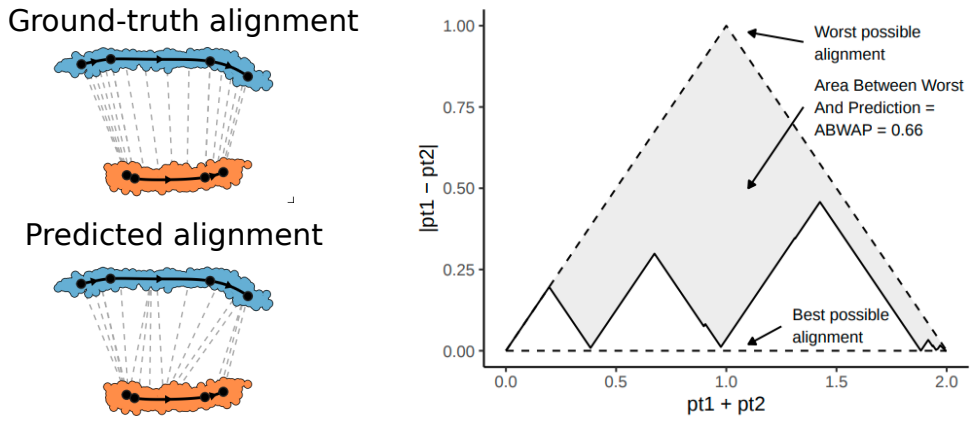

Figure 3.6: Area Between Worst And Prediction (ABWAP) metric for linear scRNA-seq trajectories. **(left)** Ground-truth alignment according to the trajectories respective pseudotimes and predicted alignment using DTW. **(right)** Evaluation of the ABWAP score for the predicted alignment, which is equal to the area between the path generated by our prediction and the path corresponding to the worst possible matching of pseudotimes. The largest the difference, the closer we get to the best possible prediction with ABWAP=1, where matched pseudotimes are identical for every pair of aligned cells. Figure adapted from (Cannoodt et al., 2021) licensed under CC BY 4.0.

Our third simulated dataset, used in the sub-sampling experiment in Section 3.4.1, is actually a perturbed version of the second dataset containing complex trajectories. It also includes a total of 40 pairs of tree-like trajectories with different levels of complexity. To

generate it, we picked the first trajectory in each trajectory pair and randomly sub-sampled it twice to get a new pair of trajectories, discarding both trajectories in the original pair.

Given the natural correspondence between cells in a pair of sub-sampled trajectories, we can evaluate alignments based on the number of true positives (TP), false positives (FP) and false negatives (FN) matchings, and their associated precision and recall. In brief, the problem we are interested in can be reduced to evaluating the consistency between two sets: the *ground-truth* matchings ($\mathcal{P}$) and the *predicted* ($\mathcal{M}$) matchings. The set $\mathcal{P}$ is defined by matching identical cells present in both trajectories and leaving cells unique to each trajectory unmatched, whereas the set $\mathcal{M}$ is defined by a Trajan alignment solution. Given $\mathcal{M}$ and $\mathcal{P}$, we define the following sub-sets:

$$\begin{aligned} TP &= \mathcal{P} \cap \mathcal{M} \\ FP &= \mathcal{M} \setminus \mathcal{P} \\ FN &= \mathcal{P} \setminus \mathcal{M} \end{aligned} \tag{3.15}$$

Note that, the TNs set is not relevant in our context since it describes all possible cell matchings not present in the ground-truth, which is huge and very little informative. Moreover, given the absolute numbers of TPs, FPs and FNs we define:

$$\begin{aligned} \text{Precision: } & \frac{TP}{TP + FP} \\ \text{Recall: } & \frac{TP}{TP + FN} \end{aligned} \tag{3.16}$$

Precision and recall, describe rates rather than absolute numbers, which helps to eliminate the confounding effect of having different set sizes.

3.4 Results

3.4.1 From paths to trees: DTW vs arboreal matchings

In our original publication (Do et al., 2019), we illustrated through experimentation that, for *linear* trajectories, optimal solutions computed using DTW are identical to those obtained by Trajan given an adequate penalty scheme. In the following Section, we complement our previous results by showing that alignments of simulated linear trajectories using classic DTW, cellAlign or Trajan, yield comparable scores based on the novel metric introduced in Cannoodt et al. (2021): the Area Between Worst And Prediction (ABWAP). Moreover, we demonstrate that the choice of a given step pattern has a big influence on the final DTW alignment solution. Next, we aim to demonstrate that for *complex* trajectories, Trajan alignments based on arboreal matchings, that try to simultaneously and consistently align all paths in both trajectories, provide and improvement in ABWAP scores with respect to multiple pair-wise alignments between individual lineages, where information from alternative paths is ignored, as in DTW. Finally, analogously to Do et al. (2019), where we demonstrated Trajan's accuracy through a series of sub-sampling experiments on real data, we complement our previous perturbation experiment with a similar experiment based on simulated data with dyngen. The goal here is twofold: to further assess Trajan's accuracy using precision and recall measures, and to see if we can detect an improvement in accuracy by using full-graph information during alignment, to multiple pair-wise alignments between individual lineages.

Linear trajectories comparison to DTW

We begin by analyzing a simulated dataset containing a total of 40 pairs of *linear* trajectories, see Section 3.3.3 for details on the data generation and metric definition. We align each pair of trajectories using classic DTW, cellAlign and Trajan, and evaluate the corresponding alignments based on the ABWAP metric. Here, classic DTW and cellAlign, were run with default parameters using Euclidean distance and 'symmetric2' step pattern, whereas for Trajan alignments we used the raw cells representation, average penalty scheme and Euclidean distance parameter combination.

In Figure 3.7, we recapitulate the results obtained for classic DTW and cellAlign in the original publication (Cannoodt et al., 2021), and demonstrate that Trajan results are comparable to those obtained by cellAlign. We should point out that, in general, special care should be taken when comparing ABWAP scores between *warpings* (the output of DTW and cellAlign) and *matchings* (the output of Trajan). In contrast to warpings, where every cell must be matched to at least one other cell, in warpings we can leave cells unmatched, which could result in sparse alignments with very few cells matched but very high ABWAP scores in extreme cases. In practice, such behavior is not observed in any of our datasets.

Next, we assessed the influence of different DTW step patterns and Trajan alternative distance and penalty parameter combinations in the resulting alignments, see Figure 3.8. Results demonstrate a big influence on the parameters used, which is likely due to the simulated data characteristics. Both trajectories represent very similar dynamical processes, with the same number of cells, and based on the definition of ABWAP, trivial matchings where each cell is matched to the corresponding cell in increasing order tend to have very high ABWAP scores. Thus, DTW step patterns that favor oblique steps (e.g. symmetric1),

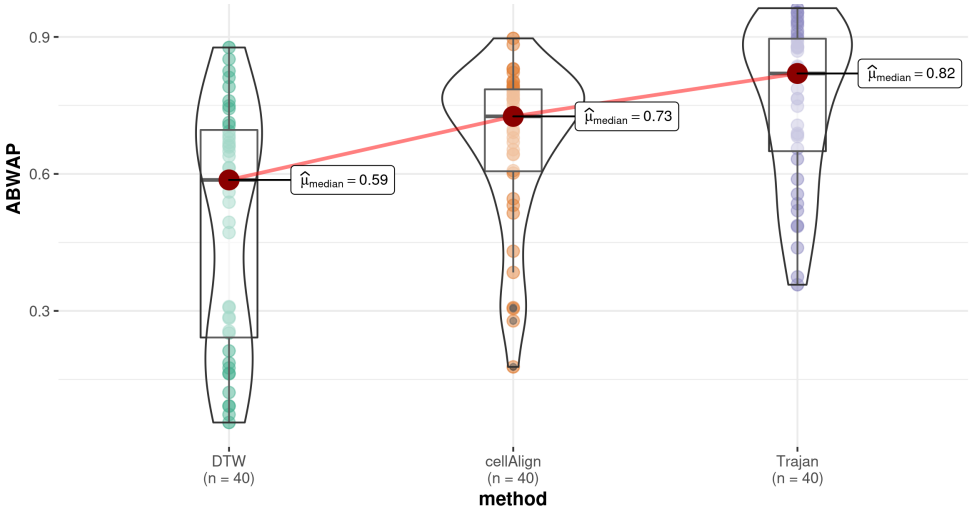

Figure 3.7: Comparison of trajectory alignment methods classic DTW, cellAlign and Trajan using ABWAP scores for dataset of 40 linear trajectories generated with dynverse.

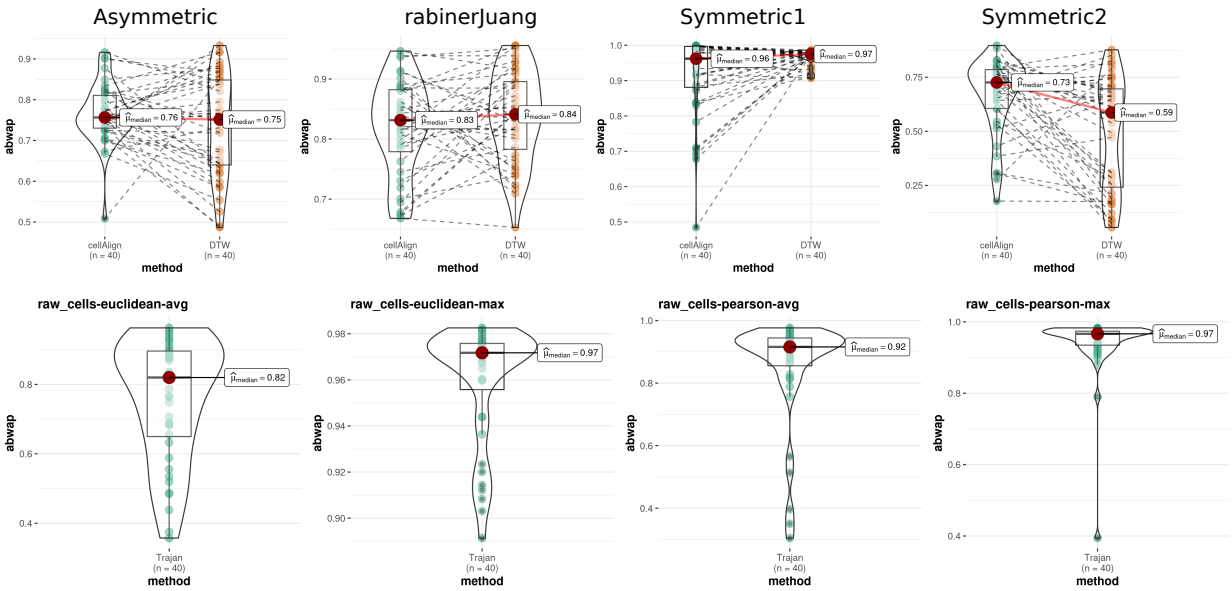

Figure 3.8: Influence of step pattern, metric and penalty scheme on trajectory alignment. Comparison of trajectory alignment methods using ABWAP scores for dataset containing 40 pairs of linear trajectories generated with dynverse. **(top)** DTW and cellAlign ABWAP scores for different step patterns: asymmetric, rabinerJuang, symmetric1 and symmetric2. **(bottom)** Trajan ABWAP scores for different metrics: euclidean and pearson correlation; and penalty schemes: max and avg. ABWAP scores in our dataset are not robust to the choice of step pattern, metric or penalty scheme.

and similarly, Trajan penalty schemes for which leaving nodes unmatched becomes quite costly (e.g. max), tend to favor such trivial matchings and lead to ABWAP scores close to 1.

Binary tree trajectories: Graph vs Path

In this Section, we extend the previous "trajectory alignment" use-case to the study of *complex* trajectories. We simulated a total of 40 pairs of complex trajectories with different levels of complexity (num_modifications=1,2,3,4), determined by the number of branching events present in the trajectory, see Section 3.3.3 for details on the data generation. As depicted in Figure 3.9, by focusing on Trajan alignments alone, we compare the quality of alignments obtained using full-graphs as input (i.e. simultaneously and consistently aligning all paths in both trajectories) to pair-wise alignments between individual paths determined by lineages in each trajectory (i.e. completely ignoring information from alternative lineages). In order to make both schemes comparable based on ABWAP scores, we "linearize" full-graph alignments into individual lineage alignments. The goal is to demonstrate the potential benefit of pooling information from different lineages during the alignment of full-graphs. Note that, in the second case, the correct matching between lineages is provided by construction, whereas in the first case it is inferred from the data alone, which is one of the key strengths of our method.

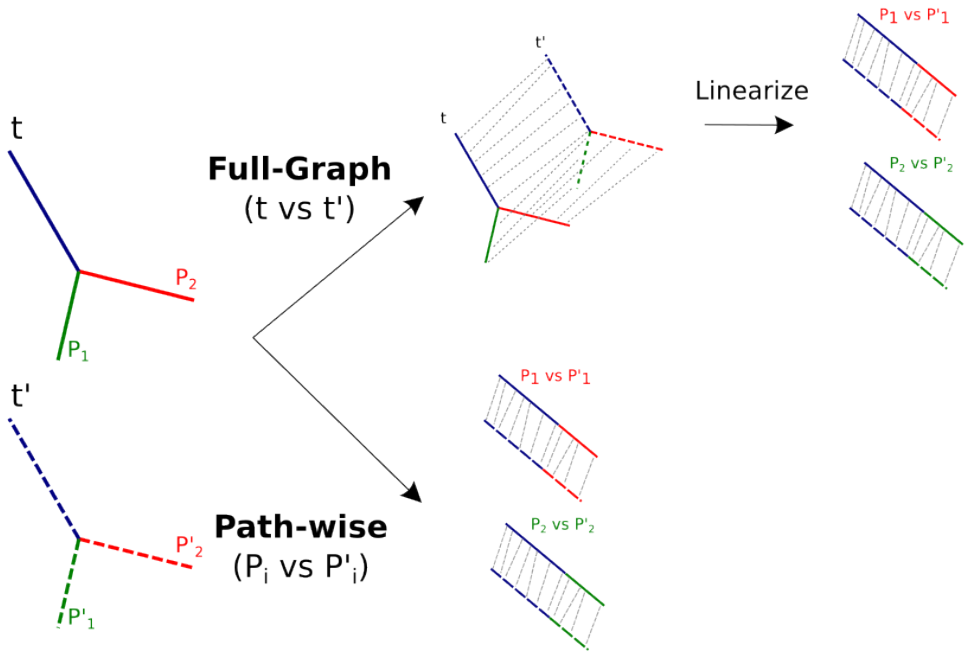

Figure 3.9: Schematic representation of the two different alignment schemes under investigation: full-graph alignments (t vs t') and pair-wise alignments between individual paths (p_i vs p'_i).

In Figure 3.10, we show ABWAP scores associated with Trajan alignments using the raw cells representation, average penalty scheme and Euclidean distance parameter combination. We have stratified trajectories by complexity-level and compared corresponding alignments at the lineage-level between both schemes: **full-graph** (red) and **path-wise** (green and blue) alignments. Note that, for path-wise alignments we considered two different approaches: own and fix penalty. First, we implemented a naive approach (*own penalty*) where linear trajectories associated to each lineage are extracted from corresponding trees prior to alignment, and given directly to Trajan as input. In this scenario, cell penalties used in the alignment will be calculated based only on cells present in that specific lineage,

meaning that the same cell might have different penalties in different lineages/alignments. Thus, results from such an approach will be confounded by cell penalties, since cell penalties will be lineage/alignment-specific, hindering the comparison we are interested in, meaning that we will not know if the observed differences originate from Trajan successfully pooling information across lineages or simply due to a different penalty scheme. For that reason, we implemented a second strategy (*fix penalty*) where we fix the penalty of each cell to that used in the full-graph alignment. In this scenario, cell penalties are calculated based on all cells in the trajectory, ensuring that differences in results originate from topology alone. Comparing both the own (green) and fix (blue) penalty schemes, we note that indeed using different cell penalties has a measurable effect on the alignments, although is there is no superior approach. On the other hand, comparing **full-graph** (red) alignments to alignments using either of the **path-wise** (green and blue) approaches, demonstrate the superior accuracy of the full-graph approach in every complexity level and almost every lineage. Thus, our results suggest that pooling of information across alternative paths, indeed increases the accuracy of the alignments.

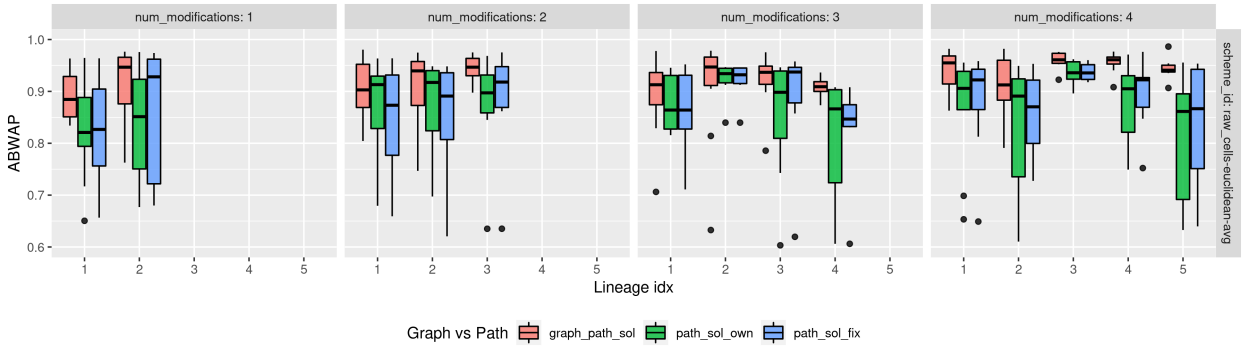

Figure 3.10: Comparison of full-graph and multiple path-wise Trajan's alignments based on ABWAP scores for dataset with 40 complex trajectories generated with dynverse. Trajan alignments were based on: raw cells representation, avg penalty scheme and Euclidean distance parameter combination. Trajectories are stratified by complexity-level (num_modifications=1,2,3,4) and each trajectory is further split into individual lineages (1,2,3,4,5) for comparison purposes. Shown are ABWAP scores associated to each linear alignment computed under 3 different schemes: full-graph (red) and path-wise (green and blue) alignments.

In Supplemental Figure S10, we show equivalent ABWAP scores for other Trajan schemes.

Sub-sampling Experiment: precision and recall

In our third experiment using dyngen simulated data, we perform a perturbation experiment re-using the data from our previous Section, see Section 3.3.3 for details on the data generation. We picked the first trajectory in each trajectory pair and randomly sub-sample it twice to get a new pair of trajectories, discarding both trajectories in the original pair. We perform this procedure at different sub-sampling levels (i.e. 90%, 80%, 70% and 20% of cells). As in our previous experiments, we use Trajan to compute alignments for these new dataset of perturbed trajectories and evaluate their accuracy. However, since in our sub-sampled dataset there is a natural correspondence between cells in both trajectories, instead of using ABWAP scores, we can evaluate the alignments based on the more intuitive

number of true positives (TP), false positives (FP) and false negatives (FN) matchings, and their associated precision and recall. Moreover, since we are still interested in assessing the potential influence of pooling information across alternative paths, we compute alignments using full-graphs (red), with subsequent linearization, and compare them to corresponding multiple pair-wise alignments between individual lineages, using either own penalty (green) or fix penalty (blue) approaches as described above.

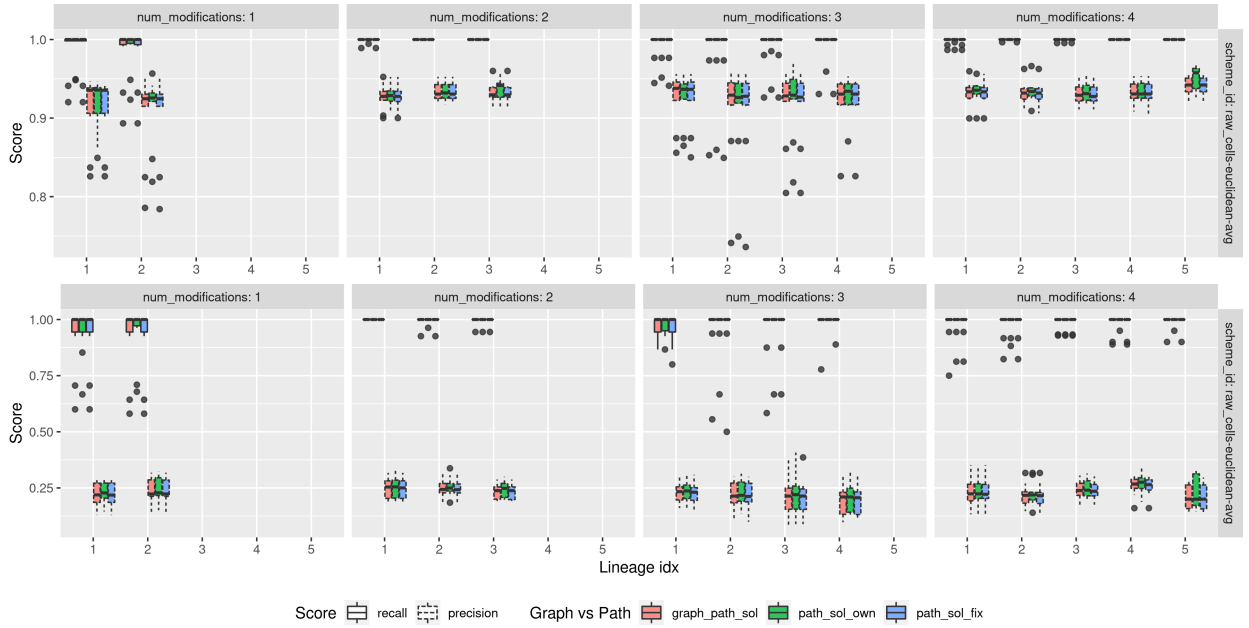

Figure 3.11: Sub-sampling experiment evaluating Trajan alignments by precision and recall scores in 40 complex trajectories generated with dynverse and perturbed at different levels. Sub-sampling of 80% of cells **(top)** and Sub-sampling of 20% of cells **(bottom)**. Trajan alignments were based on: raw cells representation, avg penalty scheme and Euclidean distance parameter combination. Trajectories are stratified by complexity-level (num_modifications=1,2,3,4) and each trajectory is further split into individual lineages (1,2,3,4,5) for comparison purposes. Shown are precision and recall associated to each linear alignment computed under 3 different schemes: full-graph (red) and path-wise (green and blue) alignments.

In Figure 3.11 (top), we show the precision and recall scores obtained for the 80% sub-sampling experiment. Median precision values are very high, above 0.9 in all complexity levels, lineages and schemes, indicating low amount of FP's (Supplementary Figure S11), and demonstrating no benefit in the use of full-graph alignments. Similarly, recall values are even higher, very close or identical to one in all complexity levels, lineages and schemes, meaning that there are almost no FN's (Supplementary Figure S11). Sub-sampling experiments of 70% and 90% of cells were also considered (not shown), and lead to similar results. In order to investigate the alignments behaviour in a more extreme case and assess the effect that the number of cells has in these metrics, we considered the sub-sampling of only 20% of cells and show results in Figure 3.11 (bottom). Since a smaller percentage of cells were sampled, the precision falls significantly due to the smaller overlap between cells in both trajectories, leading to an increase in the number of FP's predictions. This is to be expected, since matching between neighboring cells that are similar but not identical should not be

significantly penalized. Notably, the persistent extremely high recall means that there are still very low number of FNs, which implies that if a cell is present in both trajectories Trajan will almost never make the wrong match. All these results together, demonstrate the overall high accuracy of Trajan alignments. On the other hand, they also suggests that the task at hand might be too simple and show no benefit in the full-graph approach. Matching identical cells will contribute no cost at all, whereas matching close but non-identical cells will increase the cost slightly, with very little room for full-graph alignments to improve the solutions.

3.4.2 Metric conformity

In order to complement our previous results on the accuracy of Trajan, in the following Section we carry out a full metric conformity analysis equivalent to the one presented in Saelens et al. (2019), and demonstrate that Trajan's objective value can be used as an accurate metric to compare any pair of scRNA-seq trajectories. In other words, we show that Trajan can be used as an integrative measure of similarity/dissimilarity between pairs of scRNA-seq trajectories, successfully capturing their most relevant biological aspects, and opening the door to many different types of analysis.

As previously described, see Section 3.3.1, Trajan's output does not only include the actual matching/alignment between a given trajectory pair, but also provides a cost associated with the alignment solution. As first introduced in Böcker et al. (2013), the cost associated with an optimal alignment between any pair of trajectories is a quantitative measure that can be used to define a notion of similarity/dissimilarity between them, the higher the cost the more dissimilar the trajectories. Motivated by the need to evaluate the accuracy of trajectories inferred with different TI methods, a *metric conformity* analysis was developed in the landmark TI benchmark (Saelens et al., 2019) to assess the suitability of different metrics, based on a set of rules that should be fulfilled by a good or "faithfull" (Agrawal et al., 2021) similarity/distance measure. The main idea is to perturb a trajectory at multiple levels, in order to evaluate how these changes are reflected by the distance measure when comparing perturbed and unperturbed versions of the same trajectory, and whether these changes conform to our expectations based on a set of pre-defined rules for each type of perturbation. A robust metric should detect certain changes and be unaffected by others, consequently, conformity will be defined differently in each rule. For example, shuffling the position of cells with-in each branch of a trajectory for an increasing percentage of cells, and comparing these to the unperturbed trajectory, intuitively, should lead to increasing distances between trajectories (or decreasing similarity scores, defined between 0 and 1, with 1 representing identical trajectories). In our metric conformity analysis, we used a subset of the rules in the original study (Saelens et al., 2019) because some of them could not be accommodated into our framework. One additional rule was included to measure the impact of the trajectory's root, which was not considered in the original publication, since all previous metrics ignored this information. Rules marked with '*' denote perturbations where neither the ordering of cells nor their gene expressions have been altered, for which we have deviated from the authors original definitions and modified the meaning of "conformity". Based on the assumption that there is no absolute measure of pseudotime, which motivates the whole idea of pseudotime alignment, cell sequences should be allowed to be locally stretched and compressed to match one another. Thus, as long as cell orderings are

not modified, we suggest that such perturbations should leave the metric unaffected. Lastly, in our metric conformity analysis, datasets with very small number ($n = 10, 20$) of cells were discarded because after being perturbed resulted in artifacts.
Based on the framework provided in Saelens et al. (2019), we simulated all necessary trajectories using the *dyntoy* package.
The final set of rules included in our analysis are the following:

1. **Same score on identity**: The score should be approximately the same when comparing the trajectory to itself.

2. **Local cell shuffling**: Shuffling the positions of cells within each edge should lower the score. This is equivalent to changing the cellular position locally.

3. **Edge shuffling**: Shuffling the edges in the milestone network should lower the score. This is equivalent to changing the cellular positions only globally.

4. **Local and global cell shuffling**: Shuffling the positions of cells should lower the score. This is equivalent to changing the cellular position both locally and globally.

5. **Changing positions locally and/or globally**: Changing the cellular position locally AND globally should lower the score more than any of the two individually.

6. **Cell filtering**: Removing cells from the trajectory should lower the score.

7. **Move cells to start milestone**: Moving the cells closer to their start milestone should lower the score. Note that, cells were moved closer to the start milestone using $percentage_{new} = percentage^{warp\ magnitude}$.

8. **Move cells to closest milestone***: Moving the cells closer to their nearest milestone without altering cell order should leave the score unaffected.

9. **Length shuffling***: Shuffling the lengths of the edges of the milestone network without altering cell order should leave the score unaffected.

10. **Cells into small subedges**: Moving some cells into short subedges should lower the score.

11. **Bifurcation merging**: Merging the two branches after a bifurcation point should lower the score.

12. **Bifurcation merging and changing cell positions**: Merging the two branches of a bifurcation and changing the cells positions should lower the score more than any of the two individually.

13. **Bifurcation concatentation**: Concatenating one branch of a bifurcation to the other bifurcation branch should lower the score.

14. **Linear splitting**: Splitting a linear trajectory into a bifurcation should lower the score.

15. **Change of topology**: Changing the topology of the trajectory should lower the score.

16. **Cells on milestones vs edges**: A score should behave similarly both when cells are located on the milestones (as is the case in real datasets) or on the edges between milestones (as is the case in synthetic datasets).

17. **Change root node**: Changing the root node defines a different trajectory and should lower the score.

For more details on formal conformity definitions and intuitive visualizations of the types of perturbations associated to each rule, we refer the reader to the Supplementary Material in Saelens et al. (2019).

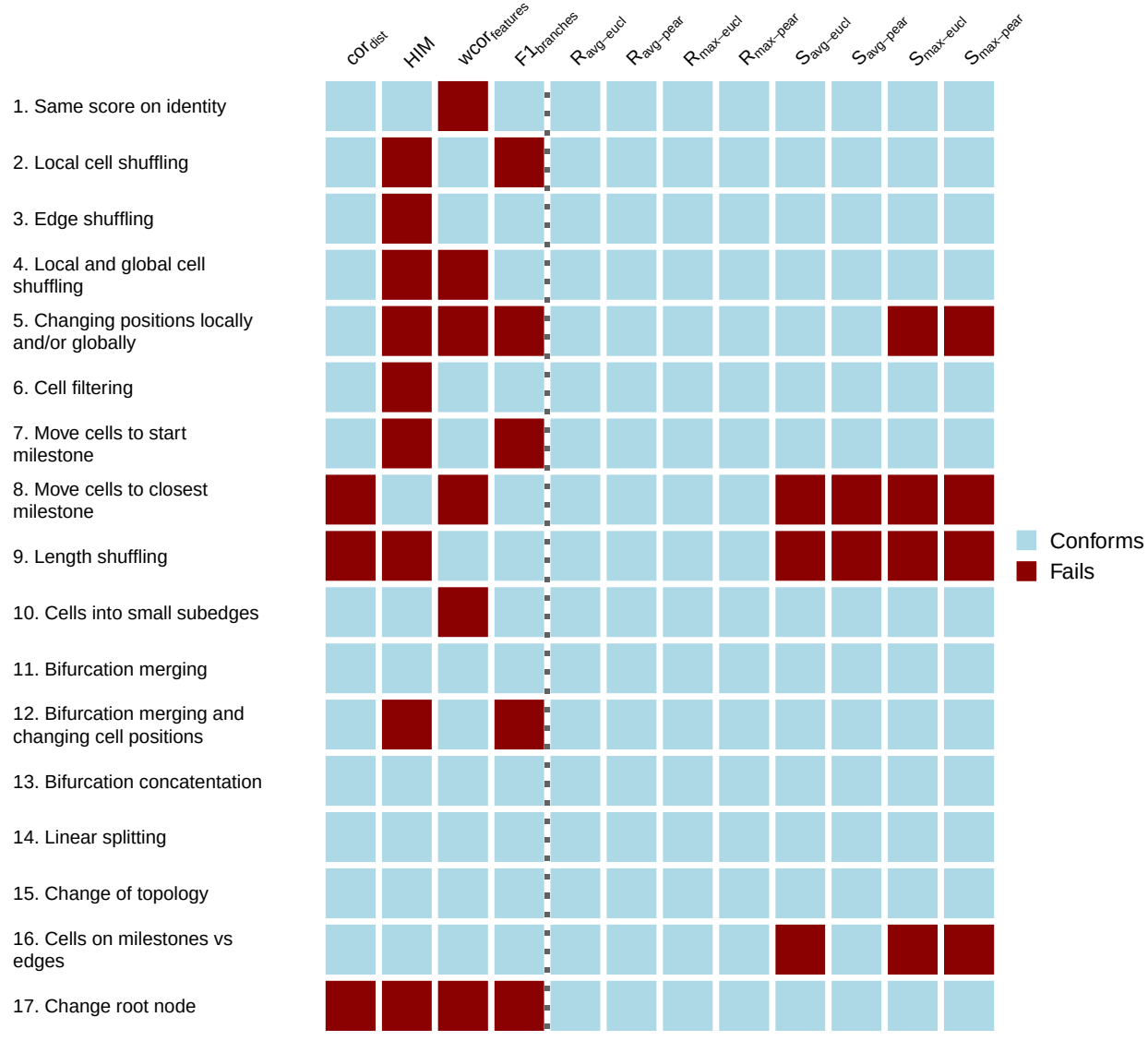

Figure 3.12: Overview Metric Conformity

Our results are summarized in Figure 3.12. In total 17 rules (rows) were evaluated using 12 different metrics (columns). The color scheme depicts whether the metric captures (*blue*) the expected behavior of a given rule, or fails to capture it (*red*). The first 4 columns represent the metrics used in Saelens et al. (2019), which look at individual aspects of the trajectory separately, the rest of columns represent multiple runs of Trajan using different parameter scheme combinations (see Section 3.3.2). One of the main advantadges of Trajan with respect to the other metrics is that it integrates multiple aspects, topology, cell ordering and gene expression, into a single measure.

As previously described (see Section 3.3.2), the *cell representation* parameter defines the two possible representations of what is meant by "cells" in the abstraction model. As shown in 3.12, the choice of cell representation has a big influence on the results. It can take one of two different values: *raw cells (R)*, where cells in the original scRNA-seq count/expression matrices are used directly, or *smooth cells (S)*, where a set of interpolated cells are constructed by smoothing the gene expression profiles of raw cells. The other parameters, *distance metric* used to compute the cell-to-cell dissimilarity and the *penalty scheme* describing the cost of leaving cells unmatched during the optimization process, are still important but have milder effect on the results. In our metric conformity analysis we explored the following parameter combinations (n=2^3=8):

- Cell representation:
  - Raw cells (**R**)
  - Smooth cells (**S**)
- Distance metric:
  - Euclidean metric (**eucl**)
  - Pearson correlation (**pear**)
- Penalty scheme:
  - Average (**avg**)
  - Maximun (**max**)

Our results recapitulate those obtained by Saelens et al. (2019) for metrics: $corr_{dist}$, *HIM*, $wcor_{features}$ and $F1_{branches}$, where each separate metric aims to capture different aspects of the trajectory (e.g. topology, gene expression, etc.). The correlation between geodesic distances (corr_{dist}), builds on the idea that if the position of a cell is the same in both trajectories, its relative distances to all other cells in the trajectory should also be the same, and computes a correlation between geodesic distances from the two trajectories. The $\text{wcor}_{features}$, ranks genes according to their importance in predicting the positions of cells in the trajectory and compares the two rankings by calculating their pearson correlation. The HIM metric (Hamming-Ipsen-Mikhailov distance) assesses the similarity in the topology between two trajectories, regardless of where the cells were positioned. $\text{F1}_{branches}$ assesses whether cells are clustered similarly in both trajectories, first by mapping each cell to its closest branch and evaluate the consistency between clusters based on the Jaccard similarity. For Trajan, our results demonstrate that we are capable of integrating all those different aspects into a single measure and conform to all those rules that we would expect to. Note that, for Trajan

schemes based on *raw cells*, first four columns after dashed line in Figure 3.12, our distance metrics successfully conform to Rules 8 and 9, where the perturbation has not altered cell ordering. On the other hand, Trajan schemes based on *smooth cells* will be indirectly influenced by these perturbations during the interpolation process, giving rise to differences in smoothed gene expression profiles that are successfully detected. This emphasizes the complementary view of both cell representations, depending on the context, we may want to highlight or just ignore these changes. Note that, because the smoothing process is very sensitive to global changes, additional local changes (see Rule 5: "Changing positions locally and/or globally) do not have much influence, specially when coupled with max penalty scheme. Potential issues related to smoothing, and how much influence should cells have in their local neighborhood, are flexibly controlled through a kernel size parameter. Also noteworthy is that, for most *smooth cells* schemes, grouping cells into milestones will have an impact (see Rule 16: "Cells on milestones vs edges"). This challenges the grouping of cells into milestones, as is done in reference trajectories for real datasets in Saelens et al. (2019), which is an obvious limitation when no ground-truth is available. Interestingly, this did not affect raw cells because the way cells were collapsed into the milestones did not affect cell ordering. Finally, Rule 17: "Change root node", which measures the influence of the rooting and directionality of the trajectory and is of crucial biological interest, is not captured by any of the previously defined metrics but is captured by every Trajan scheme.

In summary, these results demonstrate that Trajan can be used as a measure or notion of similarity/dissimilarity between trajectories, providing a framework for comparing complex trajectories and opening the door to many different types of analysis. For example, given a ground-truth trajectory, Trajan could be used to compare trajectories inferred by different TI methods against the reference to assess individual method performance. Alternatively, when no ground-truth is present, one could use Trajan to compare trajectories inferred with different TI methods to each other, to assess similarity and consistency between them. In Section 3.4.3, we investigate two separate real datasets and showcase how our TrajanR package facilitates both of these types of analyses.

3.4.3 Experiments: Real Data

In our original publication (Do et al., 2019), we demonstrated Trajan's utility by analyzing several real-world datasets. In particular, we re-analyzed two public single-cell datasets (Cacchiarelli et al., 2018) characterizing two related dynamic biological processes, where complex trajectories had been previously described: a human skeletal muscle myoblast (HSMM) *differentiation* dataset and a dataset containing human fibroblasts undergoing MYOD-mediated myogenic *reprogramming* (hFib-MyoD) We showed that Trajan was capable of recovering the key biological findings in Cacchiarelli et al. (2018), without the need to manually select core-paths from each trajectory prior to alignment, as required by DTW. Since our initial publication, another method called CAPITAL (Sugihara et al., 2022) has been published that finds correspondence between paths in both trajectories using cluster-level information, before computing an alignment between a selected pair of paths at the cell-level using DTW. In this Section, we aim to showcase the benefits of TrajanR's integration with dynverse in the analysis of complex scRNA-seq trajectories using two additional real-world datasets. In the first dataset (Treutlein et al., 2016), we re-analyze the direct reprogramming of mouse embryonic fibroblast (MEF) to induced neuronal cells (iN). Our aim is to show how, given a

single dataset, multiple TI methods can be used to infer trajectories characterizing the same biological process, and how Trajan can be used to compare these trajectories quantitatively to improve our understanding of the underlying process as characterized by the individual TI methods. In the second dataset (Klaus et al., 2019), we compare the differentiation from neural progenitor cells (NPC) to mature neurons in control individuals and patients with neuronal heterotopia, derived and sequenced using two different experimental protocols, 3D organoids (Smart-seq2) and 2D cultures (10x Genomics). Our goal is to show how, given a pair of datasets from related biological processes obtained under different conditions, or even sequenced using different protocols, Trajan alignments can act as an alternative to data integration methods, allowing direct comparison of trajectories inferred separately to better understand similarities/differences between these inferred processes. Moreover, by leveraging our TrajanR package, we demonstrate the importance of complementary analysis using different TI methods and different Trajan parameter combinations.

Direct reprogramming from fibroblast to neurons

The first dataset under consideration, Treutlein et al. (2016), is one of the real datasets included in the *dynbenchmark* package (Saelens et al., 2019), where it was used to evaluate different trajectory inference (TI) methods against a reference or ground-truth trajectory provided by the package's authors. The issue with such reference trajectories derived from real datasets is that, unlike simulated data where all necessary information is known, knowledge about the process under investigation is limited. If we knew how to identify exact trajectories from scRNA-seq data alone, we would not need TI methods in the first place, and, in general, additional sources of information are needed to guide the reference construction but may not always be available. For that reason, the authors did not aim to create exact references, but rather construct approximate trajectories that are biologically relevant, and complemented their analysis with extensive evaluation on simulated data. In dynbenchmark, real datasets were labeled as '*gold standard*' if their reference trajectory was obtained independently of gene expression, such as using cell sorting information, and as '*silver standard*' otherwise, where the reference trajectory was typically obtained by clustering gene expression values. For the Treutlein et al. (2016) dataset, the reference trajectory was constructed based on the cluster annotations provided by the authors of the original publication and it was classified as silver standard. Independently of whether silver or gold standard, cells in reference trajectories from both types of real datasets are grouped into discrete milestone, see Figure 3.14, with no attempt to further infer the relative order of cells or their pseudotimes. Consequently, reference trajectories provided by dynverse define very "coarse" representations and do not have the same level of resolution as trajectories obtained by TI methods. However, a few interesting points can be made by leveraging dynverse to compute multiple trajectories with different TI methods for single dataset and using TrajanR to compute corresponding alignments, not only against the reference but also between trajectories obtained for each method.

It has previously been pointed out that there is no 'one-size-fits-all' method that works well on every dataset (Saelens et al., 2019; Todorov et al., 2020). As a result, employing multiple TI methods and comparing their results may help in developing a better understanding of the underlying biological process being studied. As demonstrated in our metric conformity analysis, see Section 3.4.2, *Trajan* provides a quantitative framework for comparing any pair of complex trajectories, whereas *TrajanR* facilitates the inference of multiple trajectories and efficiently computes corresponding alignments between every pair of trajectories, while providing several visualization options. We start by computing scRNA-seq trajectories with any of the 50+ TI methods that have been wrapped into the dynverse framework (Saelens et al., 2019). Due to visualization limitations, we restrict our analysis to 4 state-of-the-art TI methods: *Slingshot* (Street et al., 2018), *PAGA* (Wolf et al., 2019), *Monocle* (Qiu et al., 2017) and *SCORPIUS* (Cannoodt et al., 2016b).

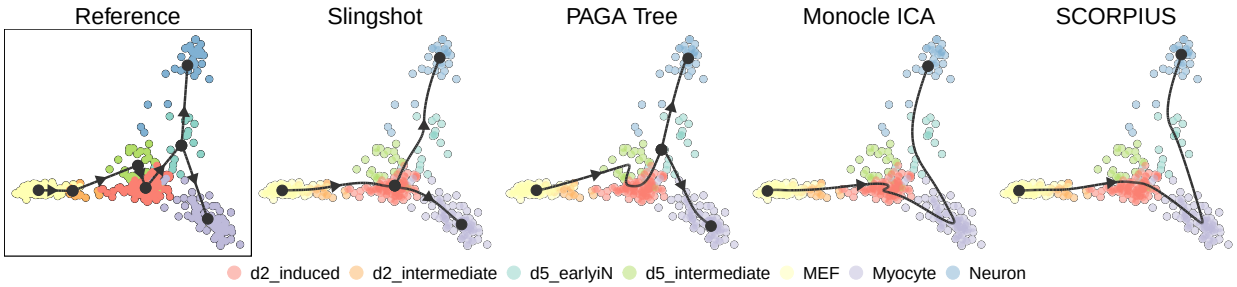

Figure 3.13: Overview of low-dimensional cell embeddings for *dynverse's* (Treutlein et al., 2016) dataset, overlayed with corresponding scRNA-seq trajectories inferred using 4 state-of-the-art TI methods. Trajectories describe the direct reprogramming of mouse embryonic fibroblast (MEF) to induced neuronal cells (iN). Dynverse provides a reference or ground-truth trajectory of the biological process under consideration. Cells are colored using the reference cell state annotation

Figure 3.13 depicts a low-dimensional representation of cells in our dataset describing the direct reprogramming from mouse embryonic fibroblast (MEF) to induced neuronal cells (iN), as well as a representation of the different trajectories inferred using the aforementioned TI methods. Cells were sequenced at different time points (0 days, 2 days, 5 days, and 20 days) after induced over-expression of the Ascl1 transcription factor, which shifts gene expression towards transcription factors specific to the neural program. However, as the process progresses, a myogenic program interferes with the neural program, resulting in the formation of unwanted myocyte-like cells and decreasing the effectiveness of the direct reprogramming process (Parra et al., 2019). In Figure 3.13 (left), the reference trajectory characterizes the dynamical process followed by cells: from MEF (yellow), through two intermediate states (day 2: orange and day 5: green) and an induced state (red), before cells in an early induced neuron stage (turquoise) branch into either of the two alternative cell fates: neurons (blue) or myocytes (purple). Note that, the use of different TI methods results in a wide variety of scRNA-seq trajectories with different topologies. For example, Slingshot and PAGA correctly inferred a complex branching topology, whereas Monocle and SCORPIUS inferred a simple linear one.

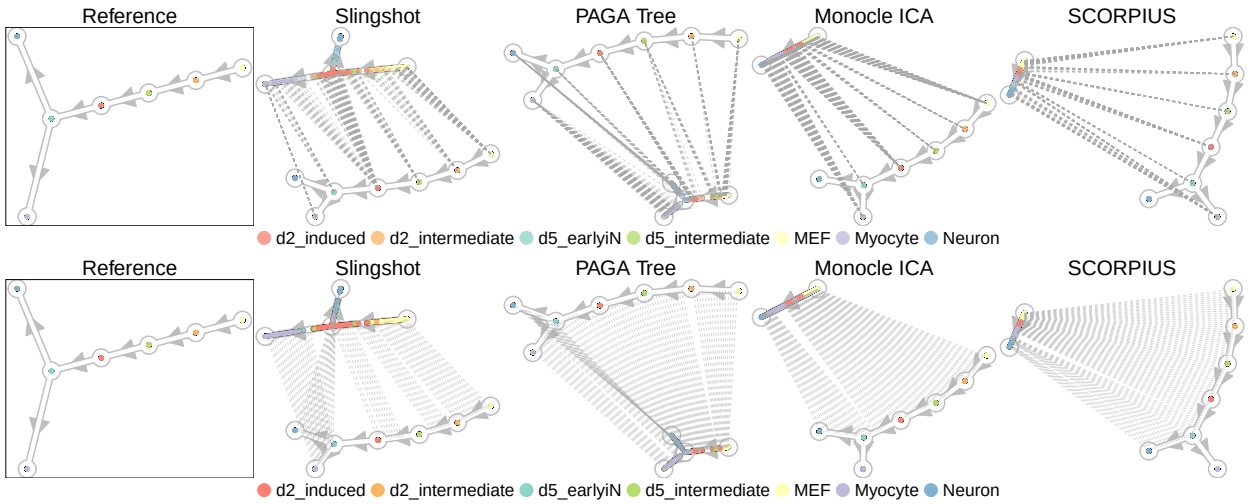

Figure 3.14: Overview of Trajan alignments for inferred scRNA-seq trajectories from *dynverse's* (Treutlein et al., 2016) dataset using 4 state-of-the-art TI methods. The reference or ground-truth trajectory is fixed and we show the alignments against all inferred trajectories. Both Trajan cell representations are shown: raw cells scheme **(top)** and smoothed cells scheme **(bottom)** representations. Cells are colored using the reference cell state annotation.

In Figure 3.14 (top), we show a representation of the Trajan alignments obtained by comparing reference and inferred trajectories using the raw cells representation, Euclidean distance and avg penalty scheme parameter combination. Note that, as previously stated, cells in the reference trajectory are grouped into discrete clusters associated with milestone in the trajectory, and thus the relative ordering between cells within these cluster, i.e. the cells local pseudotimes, is just being ignored. When aligned against the reference, it becomes very clear that PAGA correctly inferred the overall trajectory, while other methods fail to correctly identify the early iN to mature neuron branch (Monocle and SCORPIUS), or suggest an earlier branching event (Slingshot) that hinders the alignment of the neuronal fate branch. In this particular case, such effect can be explained by the unbalanced distribution of cells between the iN ($n = 32$) and the myocyte ($n = 89$) cell fate branches, which creates an almost linear trajectory that favors linear methods over more complex ones that struggle to find the correct location of the branching event. Noteworthy, PAGA is capable of finding the correct branching point and therefore has the highest score (Slingshot: 39.289; **PAGA: 32.785**; Monocle: 39.044; SCORPIUS: 38.206). One way to remedy this issue, at least partially, is to instead use the smoothed cell representation of the trajectory, which will even out or balance the number cells in each cell fate branch as can be seen in Figure 3.14) (bottom). Note that, in the the smoothed cell representation of Slingshot's trajectory, Trajan correctly aligns both cell fate branches: myocytes and neurons.

Finally, in 3.15 we showcase the visualization of aligned gene expression profiles along pseudotime, representing the gene expression dynamics of MEF cells undergoing neuronal reprogramming in both cell representations: (top) raw and (bottom) smoothed cells representations. The color scheme indicates the assignment of cells to branches in each trajectory. We only show the top four highly variable genes, which are method-specific in the smoothed cell representation. We note that, due to scRNA-seq nature and inherent variability between cells, smoothed cells provide a better representation of individual gene behavior.

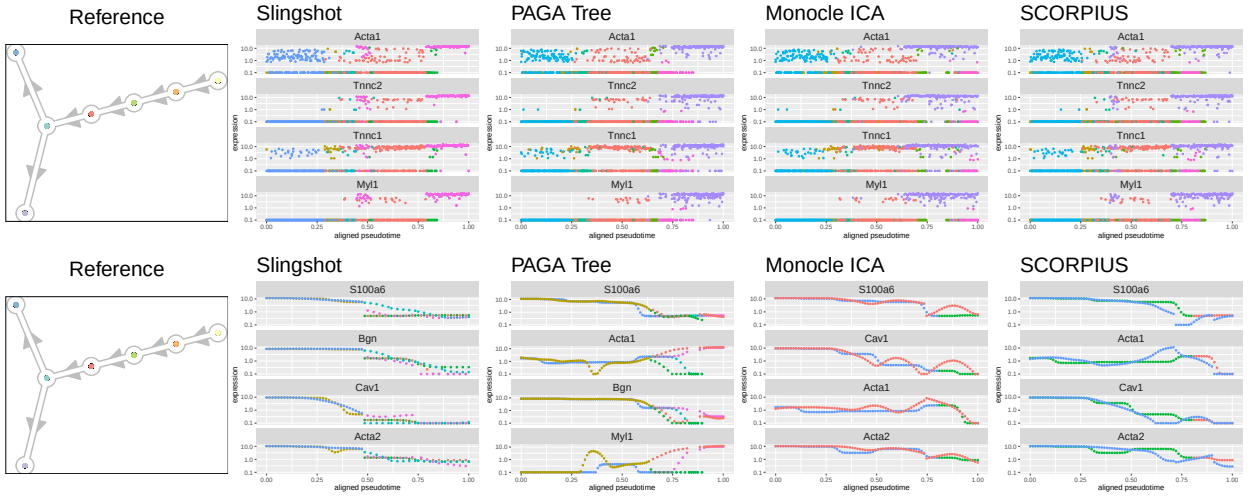

Figure 3.15: Overview of aligned gene expression profiles between reference and inferred scRNA-seq trajectories from *dynverse's* (Treutlein et al., 2016) dataset. Both Trajan cell representations are shown: raw cells scheme **(top)** and smoothed cells scheme **(bottom)** representations. Only a subset of highly-variable genes are shown.

Altered neuronal migratory trajectories

The second dataset under consideration, Klaus et al. (2019), aimed at characterizing an altered neuronal state that arises during neuronal differentiation in patients with periventricular heterotopia (PH). The differentiation process from neural progenitor cells (NPCs) to mature neurons, was studied by comparing samples derived from control individuals to patient samples with mutations in either of the cadherin receptor-ligand pair DCHS1/FAT4, which result in the altered phenotype. Initially, control and mutant states were characterized using samples derived from three-dimensional (3D) organoids, and these results were subsequently validated using samples derived from two-dimensional (2D) cell cultures. Additionally, these two separate datasets were sequenced using different scRNA-seq protocols, 3D organoids were sequenced using the Smart-seq2 protocol, whereas 2D cell cultures were sequenced using 10x Genomics. In the following analysis, we focus on the comparison of trajectories derived from these two different protocols rather than explicitly comparing healthy and diseased states. We use this dataset to motivate the usefulness of our alignment approach, by providing an alternative to data integration methods (Sugihara et al., 2022). Note that, in this case there is no 'ground-truth' or reference trajectory, so we can't use Trajan to compare inferred trajectories against a reference. Moreover, rather than comparing trajectories inferred from a single dataset using different TI approaches, we will demonstrate how to compare trajectories derived from two different datasets under different conditions.

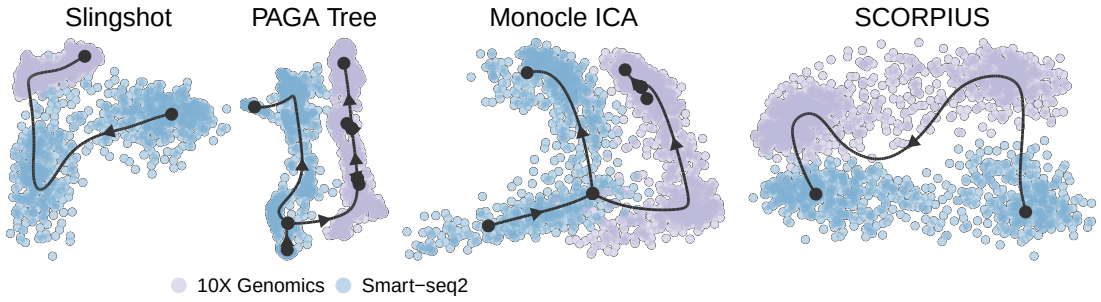

Figure 3.16: Overview of low-dimensional cell embeddings for (Klaus et al., 2019) datasets, overlayed with corresponding scRNA-seq trajectories inferred using 4 state-of-the-art TI methods. Trajectories describe the differentiation of neural progenitor cells (NPCs) to mature neurons where we have merged both 3D organoids (Smart-seq2) and 2D culture (10x Genomics) datasets into a single dataset. Cells are colored by sequencing-protocol.

Similarly to the previous Section, we computed trajectories using four state-of-the-art TI methods. First, we analyzed a combined or merged version of the neural heterotopia datasets in order to reconstruct a joint trajectory, shown in Figure 3.16, where we have colored cells by sequencing protocol. Interestingly, while all approaches are capable of reconstructing trajectories for such dataset, inferred trajectories are heavily influenced by the origin of the data, meaning that cells sequenced using different protocols do not mix correctly. This is a common phenomenon in high-throughput data, known as *batch-effect*, observed even when the same technologies are used to sequence the data but samples were processed separately or in different batches, leading to non-interesting or non-biological factors to significantly influence variation in the resulting data (Leek et al., 2010). The recent emergence of multimodal-

omics data at single-cell resolution, as well as the desire to combine measurements collected from multiple technological sources of a common biological process, has prompted the development of *data integration* methods that attempt to answer an even more general problem. How do we link or connect data together, derived from different sources, by removing non-interesting technical variation while keeping relevant biological one? (Luecken et al., 2022) Data integration is a very challenging problem, as well as an ongoing open and active field of research, in which individual solutions are based on distinct principles and assumptions, and hence optimize for different goals (Argelaguet et al., 2021). The authors of Sugihara et al. (2022), highlighted the limitations of current data integration approaches by inferring a common trajectory after integrating separate scRNA-seq datasets, which could lead to unexpected results, such as mixing cells that are not closely related. Here, we showcase how Trajan can offer an alternative to such methods when analyzing scRNA-seq trajectories, by inferring trajectories for each dataset separately and using the alignment of trajectories as the "integration" step, effectively putting datasets into a common system of reference.

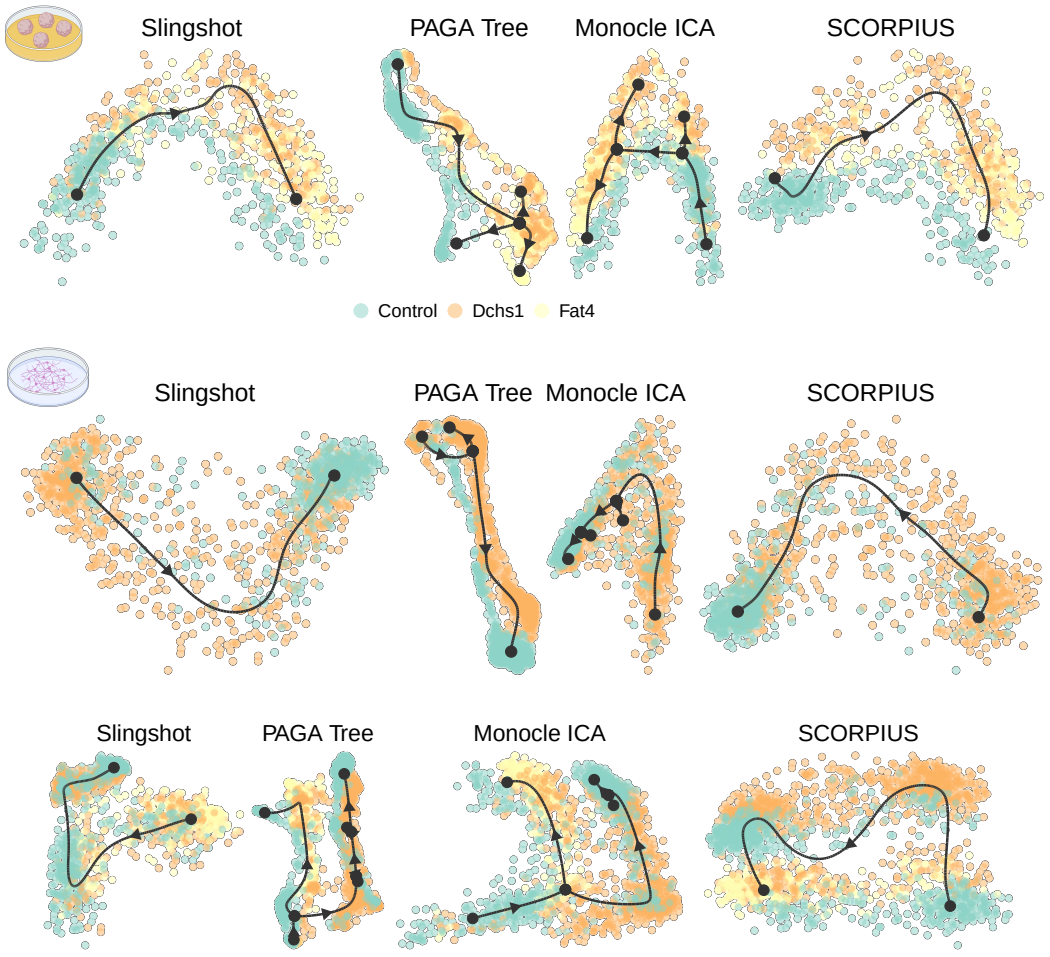

Figure 3.17: Overview of low-dimensional cell embeddings for (Klaus et al., 2019) datasets, overlayed with corresponding scRNA-seq trajectories inferred using 4 state-of-the-art TI methods. Trajectories describe the differentiation of neural progenitor cells (NPCs) to mature neurons in 3D organoids **(top)**, 2D culture **(middle)** and merged datasets **(bottom)**, respectively. Cells are colored using sample condition.

We used the same 4 TI methods to compute trajectories for both neuronal heterotopia datasets separately, see Figure 3.17: (top) organoids (middle) 2D culture, and kept the results for the merged dataset (bottom) for completeness. In this case, we have colored cells by sample condition. Again, the use of different TI methods results in a variety of scRNA-seq trajectory topologies. Here, Monocle and PAGA inferred trajectories with more complex topologies, whereas Slingshot and SCORPIUS inferred linear ones. Note that, all trajectories seem to consistently mix DCHS1 and FAT4 mutants, suggesting that those share common transcriptional features different from the control, as was previously noted in Klaus et al. (2019). Interestingly, Slingshot and SCORPIUS fail to identify the altered neuronal subpopulation present in both datasets, described in Klaus et al. (2019).

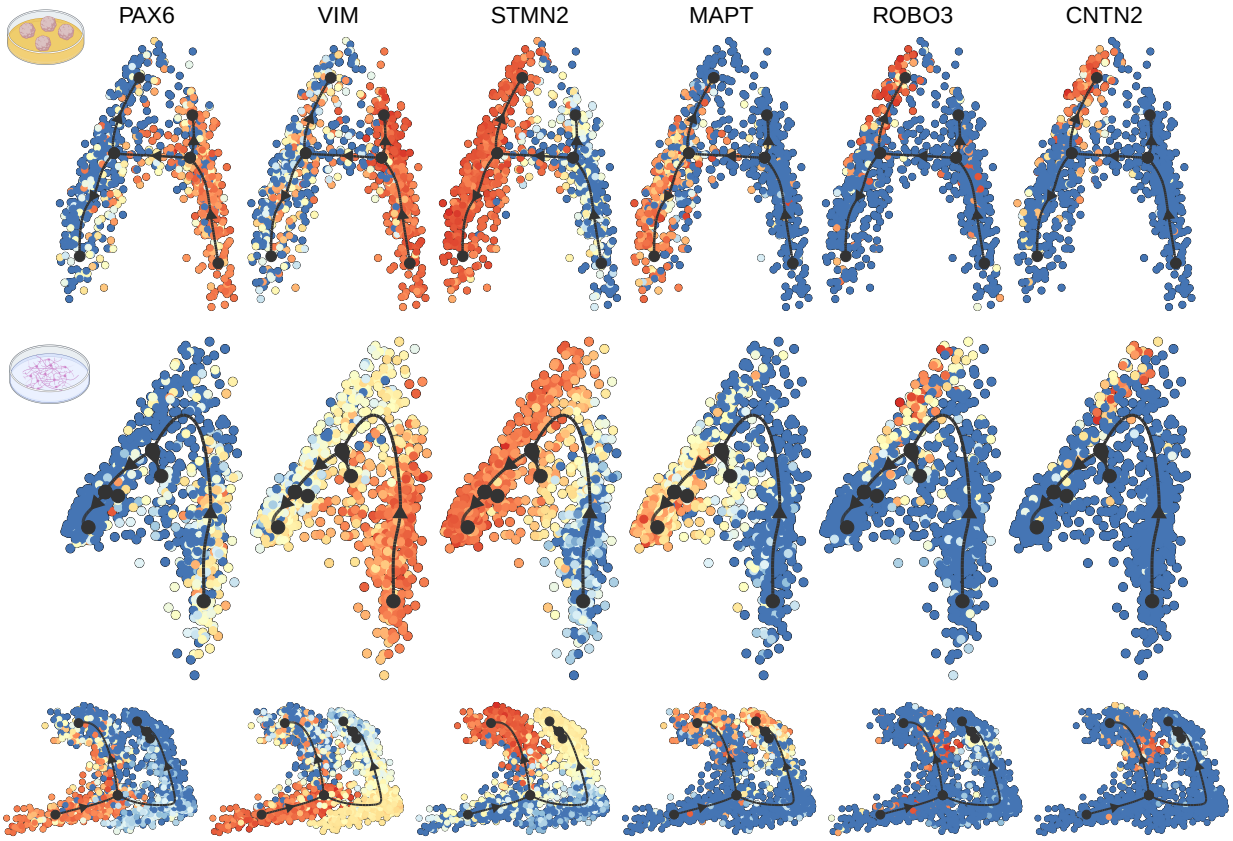

Figure 3.18: Low-dimensional cell embeddings for (Klaus et al., 2019) datasets, overlayed with corresponding scRNA-seq trajectories inferred with Monocle. Trajectories describe the differentiation of neural progenitor cells (NPCs) to mature neurons in 3D organoids **(top)**, 2D culture **(middle)** and merged datasets **(bottom)**, respectively. Cells are colored by expression of selected marker genes: NPCs (PAX6, VIM), neurons (STMN2, MAPT) and altered neuronal state (ROBO3, CNTN2). Blue: low expression and red: high expression.

For simplicity and based on the results of Klaus et al. (2019), we focus on the trajectories obtained using *Monocle*, the same method used in the original publication (see their Figure 4). Our results recapitulate those reported in the original publication for both datasets, and we extend the analysis by visualizing the expression of specific marker genes using dynverse and by computing alignments and corresponding gene expression dynamics using TrajanR.

In Monocle's trajectory for the organoid dataset, see Figure 3.17, we detect a specific subpopulation of neurons that diverges from the main NPC to neuron differentiation trajectory, and is unique to the mutant conditions. This altered neuronal population is characterized by the expression of neuronal marker genes such as STMN2 and MAPT, and the expression of genes specific to the altered neuronal state such as ROBO3 and CNTN2, see Figure 3.18 (top). It is worth noting, that there appears to be an earlier smaller branching event, with a very high percentage of mutant cells (96%), which we believe characterizes NPCs cells that fail to reprogram, since they lack expression of neuronal marker genes STMN2 and MAPT but express NPC marker genes PAX6 and VIM (see Figure 3.18). Equivalently, in Monocle's trajectory for the 2D culture dataset, see Figure 3.17 (middle), we detect two subpopulations of cells that bifurcate from the main NPC to neuron differentiation trajectory. In this case, however, subpopulations include cells from both the control and mutant conditions. The earlier branching event includes a high percentage of cells from the mutant condition (81%), whereas the second branching event includes a moderate percentage of cells from the mutant condition (66%). Simple visual inspection of the lower dimensional embedding is insufficient to fully understand the biological process under investigation. Moreover, due to the dataset's unbalanced nature (63% mutant and 37% control), one question that arises naturally is whether these branches are the corresponding counterparts to those present in the 3D organoid dataset.

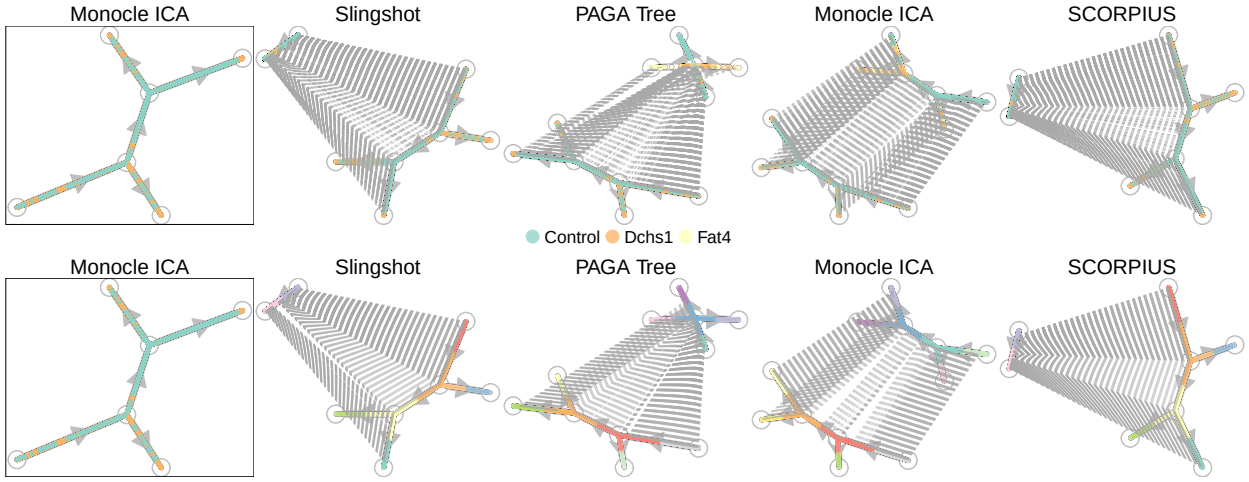

Figure 3.19: Overview of Trajan alignments between inferred scRNA-seq trajectories from Klaus et al. (2019) datasets. Use Monocle's trajectory for the 2D culture dataset as "reference" and show alignments against all trajectories inferred for the 3D organoid dataset. Both Trajan cell representations are shown: raw cells scheme **(top)** and smoothed cells scheme **(bottom)** representations. Cells are colored using sample condition, for raw cells, and by assignment to closest milestone, for smoothed cells.

Similar to the previous section, we leverage TrajanR to compute all pair-wise alignments across trajectories obtained for each of these datasets. To showcase the visualization of such alignments, in Figure 3.19 we fix the Monocle trajectory from the 2D culture dataset and plot the corresponding alignments to all other trajectories in the 3D organoid dataset. We used both cell representations, (top) raw cells and (bottom) smoothed cells, and fixed Euclidean distance and avg penalty scheme parameters. Except for PAGA's alignment using

the raw cell representation, which returned a trajectory in apparent contradiction to the "reference" obtained with Monocle, these alignments demonstrate that Trajan is capable of successfully aligning the main NPC to neuron differentiation path between distinct trajectories. Note that, in PAGA's alignment using the smoothed cell representation we are indeed capable of recovering similar results, highlighting once again the importance of these two complementary approaches. Next, for simplicity and to make our results more comparable to Klaus et al. (2019), we again focus the rest of the analysis on a single alignment using both Monocle's trajectories, the same trajectories as in the original publication.

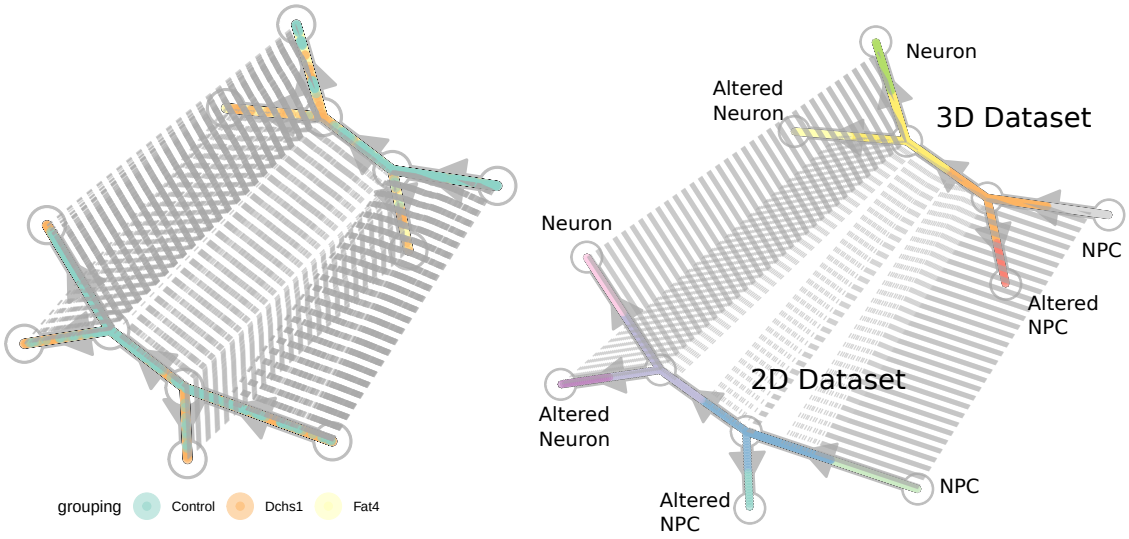

Figure 3.20: Trajan alignment of both Monocle's scRNA-seq trajectories obtained for the 2D culture and 3D organoid datasets in Klaus et al. (2019). Both Trajan cell representations are shown: raw cells scheme **(left)** and smoothed cells scheme **(right)** representations. Cells are colored using sample condition, for raw cells, and by assignment to closest milestone, for smoothed cells.

In Figure 3.20 we visualize the alignment of both Monocle's trajectories, on the left using the raw cells representation and on the right using the smoothed cells representation. Note that, in the raw cells representation all end state branches are matched to their counterpart between the two datasets, whereas on the smoothed cell representation we see that the altered NPC branch is left unmatched. This inconsistency between schemes, and the sparsity of alignments on the central region, demands for further investigation. In Figure 3.21, we visualize the corresponding gene expression dynamics. The actual aligned gene expression profiles are shown on the left, while the middle and right gene expression profiles correspond to the individual, 2D culture and 3D organoid trajectories, respectively, prior to alignment. The color scheme indicates the assignment of cells to branches in each trajectory, but are not consistent across aligned and individual profiles. In order to map branches in the aligned gene expression profiles (left) to corresponding branches in the individual gene expression profiles (middle and right) one needs to look at the actual expression values. By looking at the aligned gene expression profiles (left), note that the two first marker genes VIM (NPC) and STMN2 (neuron) are nicely aligned between the two datasets, whereas for the last marker gene ROBO3 (altered neuron) the aligned profiles do not match very well, specially right after the central region. By closer inspection of the individual gene expression profiles, we note that for the 2D culture dataset (middle) there is a peak of expression for the ROBO3

marker gene around the central region before any branching point, and no expression at all of ROBO3 on the branch supposed to contain the altered neuronal state. This results suggest that the trajectory inferred by Monocle for the 2D culture dataset is not consistent with our biological assumptions, meaning that, although there is a subpopulation of altered neurons it has been placed in the middle of the trajectory rather than as a separate branch. On the other hand, for the 3D organoid dataset (middle), as expected there is no expression at all of the ROBO3 marker gene previous to the branching point, followed by a sudden increase in expression for the branch containing the altered neuronal state.

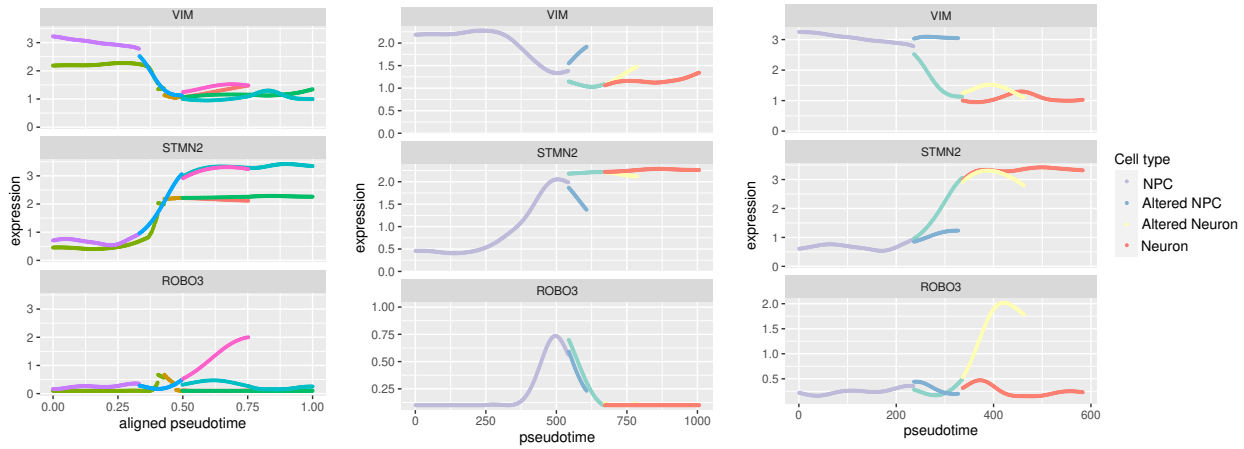

Figure 3.21: Aligned gene expression profiles **(left)** between 2D culture and 3D organoids inferred scRNA-seq trajectories from Klaus et al. (2019) datasets. Gene expression profiles for 2D culture **(middle)** and 3D organoid **(middle)** inferred scRNA-seq trajectories, previous to alignment. Only a subset of marker genes are shown: NPCs (VIM), neurons (STMN2) and altered neuronal state (ROBO3).

Multiple lessons can be learned from our analysis. First, every trajectory inference method will almost certainly return a trajectory given an input dataset, but it is the user's responsibility to assess the trajectory's meaningfulness and whether it accurately describes the biological process under consideration. A tool like Trajan, which can compare different trajectories and assess their similarity and consistency, is a good starting point. Special care must be taken during the feature selection step, as each set of features will tell a different story, and multiple stories may be played simultaneously at different strength levels. In our analysis, we used the same 286 genes that were found to be relevant for this particular biological process in the original publication (Klaus et al., 2019), while the differences between healthy and altered neurons were pin-pointed to just a few marker genes, highlighting that differences between different but related biological processes can be very subtle. In this case, the ROBO3 was strong enough to distinguish between the two neuronal states in the 3D organoid dataset, but it was not for the 2D culture. These issues highlight the importance of complementary analysis, not only by employing different TI methods, but also by being able to compare them under various Trajan schemes and by inspecting corresponding gene expression profiles. Our package TrajanR facilitates the computation and visualization of these multiple schemes simultaneously.

3.5 Conclusion

We described our novel tool Trajan, first introduced in Do et al. (2019), that allows for the first time the alignment of complex (non-linear) scRNA-seq trajectories at the cell-level. In Trajan, we adopt arboreal matchings (Böcker et al., 2013) to perform alignments between any pair of single-cell trajectories both, enabling the meaningful comparison of gene expression dynamics along a common pseudotime scale, and defining a notion of similarity/distance between trajectories.

Trajan does not make any assumptions concerning the algorithm used to reconstruct the trajectory and can in principle be coupled with any available reconstruction method. We have introduced our accompanying R package, TrajanR, which integrates with previous work (Saelens et al., 2019), making Trajan seamlessly compatible with any of the 50+ TI methods implemented in the *dynverse* package. TrajanR also facilitates the standardization and pre-processing of Trajan input data, alignment computation using multiple parameter combinations, and a variety of visualization options, enabling the analysis of scRNA-seq trajectories in complex settings.

Through extensive experimentation on synthetic data generated with the recent simulation engine, dyngen (Cannoodt et al., 2021), we demonstrate the accuracy of Trajan's alignments. In a set of linear trajectories, we compare DTW and Trajan alignments based on ABWAP scores, a novel metric introduced in Cannoodt et al. (2021), to demonstrate the equivalence of both approaches in the case of simple/linear trajectories. In a set of more complex, tree-like, trajectories we focus on Trajan alignments to showcase the improvement achieved by the simultaneous and consistent alignment between all paths using the full-graph information, as opposed to multiple individual path-wise alignments where information about alternative lineages is being ignored, as in DTW. In a sub-sampling experiment, we demonstrate the high accuracy of Trajan alignments based on precision and recall, as an alternative metric to ABWAP scores. Through a full metric conformity analysis, we demonstrate empirically that Trajan's notion of similarity/dissimilarity between single-cell trajectories conforms to our biological expectations, providing a framework for comparing complex trajectories and opening the door to many different types of analysis. Finally, we showcase how the analysis of real data is facilitated using our TrajanR package to examine two independent real-world datasets. In the first dataset, where a ground-truth trajectory is available, Trajan can be used to compare trajectories inferred by different TI methods against the reference to assess method performance. In the second dataset, which contained two datasets derived and sequenced using different experimental protocols, we showcase how Trajan alignments can act as an alternative to data integration methods, and how assessing consistency between different trajectories helps to better understand the underlying biological processes driving these trajectories. Importantly, throughout the whole Chapter we demonstrate the importance of complementary analysis using different Trajan parameter combinations, and based on our real data experiments we highlight the need to evaluate multiple TI methods, all which is enormously facilitate by our TrajanR package.

Chapter 4

Conclusion and outlook

4.1 Conclusion

In this thesis, we developed reproducible computational pipelines an algorithms that enable the study of genomic regulation in biological systems.

In Chapter 2, we aimed to quantify the relative contribution of transcriptional silencing and RNA degradation to heterochromatic silencing. In order to do that, we analyzed RNA Pol II occupancy (ChIP-seq), levels of nascent RNA (RIP-seq) and levels steady-state RNA (RNA-seq) in different mutants of Schizosaccharomyces pombe (S. pombe). We found that transcriptional silencing consists of two components, reduced RNA Pol II accessibility and, unexpectedly, reduced transcriptional efficiency. Heterochromatic loci showed lower transcriptional output compared to euchromatic loci, even when comparable amounts of RNA Pol II were present in both types of regions. We determined that the Ccr4–Not complex and H3K9 methylation are required for reduced transcriptional efficiency in heterochromatin and that a subset of heterochromatic RNA is degraded more rapidly than euchromatic RNA. Finally, we quantified the contribution of different chromatin modifiers, RNAi and RNA degradation to each silencing pathway. Our results show that several pathways contribute to heterochromatic silencing in a locus-specific manner and reveal transcriptional efficiency as a new mechanism of silencing.

In Chapter 3, we describe our novel tool Trajan, previously introduced in Do et al. (2019), the first method for the alignment of complex (non-linear) scRNA-seq trajectories at the cell-level. In Trajan, we adopt arboreal matchings (Böcker et al., 2013) to automatically identify the correspondence between biological processes, characterized by a pair of single-cell trajectories, by aligning all their lineages simultaneously and consistently, enabling the direct comparison of gene expression dynamics along a common pseudotime axis and providing a notion of similarity/distance between trajectories. Using simulated data with different characteristics, we performed multiple experiments to evaluate different aspects of Trajan's alignments based on ABWAP scores, a novel metric introduced in Cannoodt et al. (2021). In a dataset of linear trajectories, we show that Trajan alignments are comparable in accuracy to those obtained by linear trajectories restricted, DTW-based methods, providing an alternative that generalizes to the alignment of complex trajectories. Since our initial publication, a method called CAPITAL (Sugihara et al., 2022) was published that is capable of finding the correspondence between paths in a pair of complex trajectories using cluster-level information. CAPITAL uses those matched lineages to provide an actual alignment at the cell-level, but as in previous methods it is limited to independent pair-wise DTW-based

alignments, completely ignoring information from alternative lineages that could help improving the alignment, as demonstrated in our experiments with complex trajectories. In a sub-sampling experiment, we further show the overall high accuracy of Trajan alignments based on alternative precision and recall metrics. Moreover, through a full metric conformity analysis, we demonstrate empirically that Trajan's notion of similarity/dissimilarity between single-cell trajectories conforms to our biological expectations, providing a framework for comparing complex trajectories and opening the door to many different types of analysis. Lastly, we introduced TrajanR, our accompanying R package that integrates with the dynverse framework (Saelens et al., 2019), allowing the computation of single-cell trajectories with any of the 50+ TI methods implemented in the package and used for subsequent analysis with Trajan. TrajanR facilitates the standardization and pre-processing of Trajan input data, alignment computation using multiple parameter combinations, and provides a variety of visualization options, enabling the analysis of scRNA-seq trajectories in complex settings. Throughout the Chapter, we demonstrate the importance of complementary analysis, either by evaluating multiple TI methods or trying multiple parameter combinations, all which is enormously facilitated by our TrajanR package. Finally, we showcase how our TrajanR package enables and facilitates the study of scRNA-seq data based on the analysis of two independent real-world datasets.

4.2 Outlook

4.2.1 Ccr4–Not complex reduces transcription efficiency in heterochromatin

In our study, we created an end-to-end Snakemake workflow for the quantification of RNA-seq, ChIP-seq and RIP-seq data, together with several Jupyter Notebooks for downstream analysis and visualization. We established a framework to characterize specific regulatory pathways by a rate or ratio between expression levels of corresponding high-throughput sequencing assays, e.g. Pol II occupancy (^{mu}ChIP-seq/^{wt}ChIP-seq), transcription efficiency (RIP-seq/ChIP-seq) and RNA stability (pA-RNA/RIP-seq), and compare those rates under different conditions. Based on the analysis of over a hundred samples from different S. pombe mutant strains and sequencing-protocols, we identified the impact that certain mutations have, in a loci and strand-specific manner, on the distinct regulatory pathways. Our modular, reproducible and scalable implementation, permits the extension of the study to many more samples and conditions, and should be easily adaptable to the study of other regions of interest and even different species were a reference genome and transcriptome are available. Finally, a similar framework, could be used to study other genomic regulatory pathways by defining alternative ratios and obtaining the corresponding sequencing data under the adequate conditions.

4.2.2 Dynamic pseudo-time warping of complex trajectories

We have extensively benchmarked and showcased the utility of Trajan and TrajanR in both real and simulated data. We hope that our framework will be well received by users and method developers in the single-cell community, as it facilitates and sheds light on the study and interpretation of scRNA-seq trajectories, an increasingly common approach in scRNA-seq analysis. We anticipate several possible extensions to our work. While our analysis was restricted to arboreal matchings between rooted trees, in the Appendix of Do et al. (2019) we discuss the generalization of arboreal matchings to directed acyclic graphs (DAGs), which describe biologically relevant processes such as the convergence of previously divergent cell states. In order to keep up with the increasing size of scRNA-seq datasets, TrajanR's pre-processing will need to be significantly improved and sped up. At present, the smoothed cell representation can be used to infer a smaller set of interpolated cells, which aids in scaling subsequent computations. An alternative strategy would be to couple TrajanR with a geometrical sketching method (Do et al., 2020) that down-samples the total number of cells while preserving the full transcriptomic diversity of the data. More importantly, the importance of TrajanR's integration with dynverse in conjunction with our framework for comparing complex single-cell trajectories cannot be emphasized enough. For example, using different TI methods to infer multiple trajectories for the same dataset and leveraging Trajan to compute pair-wise alignments between each pair of trajectories, one could provide a consensus prediction of which method performs best for that dataset based on the similarities and distances between methods as well as their consistencies and inconsistencies. In addition, using a similar supervised consensus approach and a sufficiently diverse and comprehensive set of simulated trajectories, it should be possible to train a classifier that learns under what circumstances a method outperforms others.

Appendix A

Supplementary Material

A.1 Ccr4-Not complex reduces transcription efficiency in heterochromatin

Strain construction

All S. pombe strains used in this study are listed in Supplemental Table S1. Strains were generated like described in Brönner et al. (2017). The strains were constructed by electroporation (Biorad MicroPulser program ShS) with a PCR-based gene targeting product leading to deletion of specific genes. For genomic integration, a PCR with long overhang primers according to Bähler et al. (1998) was performed and the product transformed. Positive transformants were selected on YES plates containing 100–200 μg/ml antibiotics and were confirmed by PCR and sequencing.

Total RNA isolation

Total RNA was isolated of a 2 ml yeast culture with OD600 of 1.0 applying the hot phenol method. The pellet was resuspended in 500 μl lysis buffer (300 mM NaOAc pH 5.2, 10 mM EDTA, 1% SDS) and 500 μl phenol-chloroform-isoamylalcohol (25:24:1, Roth) and incubated at 65°C for 10 min with constant mixing. The organic and aqueous fractions were separated by centrifugation at 20 000 × g for 10 min. Nucleic acids in the aqueous fraction were precipitated with ethanol and then treated with DNAse I (Thermo Scientific) for 1 h at 37°C. DNAse was removed by a second phenol-chloroform-isoamylalcohol extraction and ethanol precipitation.

poly(A) RNA sequencing

The poly(A) RNA library was obtained using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB) including the NEBNext Poly(A) mRNA Magnetic Isolation Module. Libraries were sequenced on Illumina HiSeq platform.

Chromatin immunoprecipitation sequencing (ChIP-seq)

50 ml yeast cultures with an OD600 of 1.2 were cross-linked with 1% formaldehyde (Roth) for 15 min at room temperature. The reaction was quenched with 125 mM glycine for 5 min. The frozen pellet was resuspended in 500 μl lysis buffer (250 mM KCl, 1× Triton-X, 0.1% SDS, 0.1% Na-desoxycholate, 50 mM HEPES pH 7.5, 2 mM EDTA, 2 mM EGTA, 5

mM MgCl2, 0.1% Nonidet P-40, 20% glycerol) with 1 mM PMSF and Complete EDTA free Protease Inhibitor Cocktail (Roche). Lysis was performed with 0.25–0.5 mm glass beads (Roth) and the BioSpec FastPrep-24 bead beater (MP-Biomedicals), 8 cycles at 6.5 m/s for 30s and 3 min on ice. DNA was sheared by sonication (Bioruptor, Diagenode) 35 times for 30 s with a 30 s break. Cell debris was removed by centrifugation at 13 000 × g for 15 min. The crude lysate was normalized based on the RNA and Protein concentration (Nanodrop, Thermo Scientific) and incubated with 1.2 μg immobilized (Dynabeads Protein A, Thermo Scientific) antibody against Anti-RNA polymerase II CTD repeat YSPTSPS (phospho S2) antibody - ChIP Grade (ab5095, abcam) for at least 2 h at 4°C. The resin with immunoprecipitates was washed five times with each 1 ml of lysis buffer and eluted with 150 μl of elution buffer (50 mM Tris–HCl pH 8.0, 10 mM EDTA, 1% SDS) at 65°C for 15 min. Cross-linking was reversed at 95°C for 15 min and subsequent RNase A (Thermo Scientific) digest for 30 min followed by Proteinase K (Roche) digest for at least 2 h at 37°C. DNA was recovered by phenol–chloroform–isoamylalcohol (25:24:1, Roth) extraction with subsequent ethanol precipitation. For sequencing, a ChIP-seq library was made using the NEBNext Ultra II DNA Library Prep Kit for Illumina kit (NEB). Libraries were sequenced on Illumina HiSeq platform.

Pol II bound nascent RNA sequencing (RIP-seq)

RNA IP was performed like ChIP but without RNase A digest, using Anti-RNA polymerase II CTD repeat YSPTSPS (phospho S2) antibody—ChIP Grade (ab5095, abcam). After phenol-chloroform-isoamylalcohol extraction, DNA was digested with DNAse I (Thermo Scientific) for 2 h at 37°C. RNA was recovered with a second phenol-chloroform-isoamylalcohol purification and ethanol precipitation. Sequencing libraries were produced using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB). Libraries were sequenced on Illumina HiSeq platform.

Figures

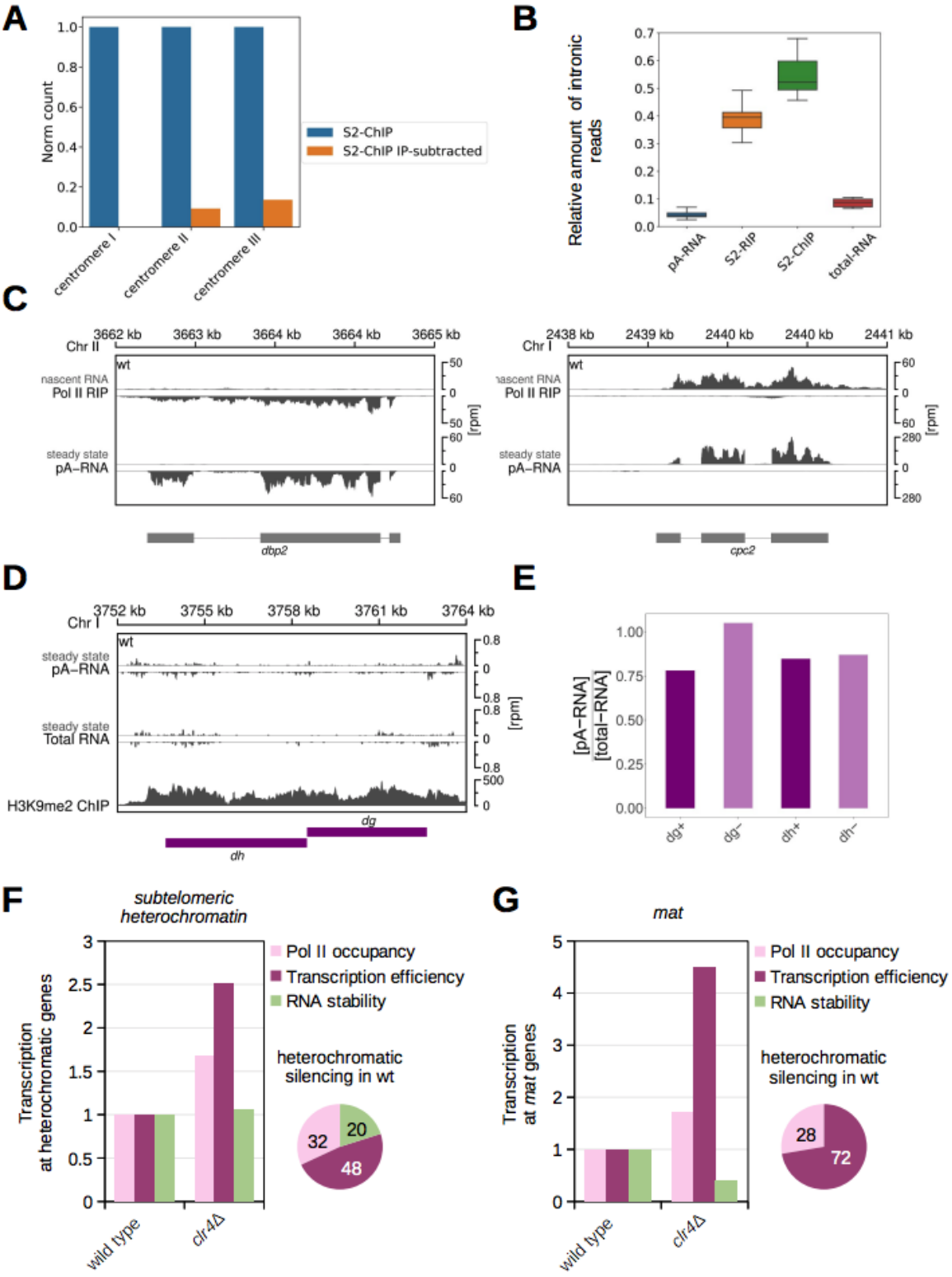
A
Norm count
S2-ChIP
S2-ChIP IP-subtracted
centromere I
centromere II
centromere III
B
Relative amount of intronic reads
pA-RNA
S2-RIP
S2-ChIP
total-RNA
C
Chr II
3662 kb
3663 kb
3664 kb
3664 kb
3665 kb
wt
nascent RNA Pol II RIP
steady state pA–RNA
[rpm]
dbp2
Chr I
2438 kb
2439 kb
2440 kb
2440 kb
2441 kb
cpc2
D
Chr I
3752 kb
3755 kb
3758 kb
3761 kb
3764 kb
steady state pA–RNA
steady state Total RNA
H3K9me2 ChIP
dg
dh
E
[pA−RNA] [total−RNA]
dg+
dg−
dh+
dh−
F
subtelomeric heterochromatin
Transcription at heterochromatic genes
Pol II occupancy
Transcription efficiency
RNA stability
heterochromatic silencing in wt
32
20
48
wild type
clr4Δ
G
mat
Transcription at mat genes
Pol II occupancy
Transcription efficiency
RNA stability
heterochromatic silencing in wt
28
72
wild type
clr4Δ

Figure S1: Silencing in wild-type cells. (A) Analysis of S2P-Pol II ChIPseq read counts from wild type cells at centromeric central core (CenpA chromatin) before and after normalized input subtraction. Note that after subtraction most reads from this region, defined as noise, are removed. (B) Quantification of intronic reads in pA RNA, total RNA, Pol II RIP and Pol II ChIP data. The data reveal high retention of intronic reads in Pol II RIP data. Bottom and top of the box correspond to lower and upper quartiles of the data, bar is the median and whiskers are median +/-1.5 times interquartile range. (C) Analysis of the next-generation sequencing data showing comparison between nascent RNA (RIP) and total RNA (pA) sequencing. Note the presence of intronic reads in Pol II RIP data, indicating nascent RNAs. (D) Analysis of the next-generation sequencing data showing steady state RNA levels (total RNA-seq and pA RNA seq) and H3K9me2 levels (ChIP-seq) at pericentromeric regions in S. pombe wild-type cells. Locations of genes are indicated as boxes below the coverage according the color code: purple = dg, dh. (E) Quantification of total and pA RNA reads over dg and dh transcripts showing that dg+/dg- and dh+/dh- transcripts are similarly polyadenylated. (F, G) Bar chart showing fold change in quantitative measures (ratios of average TPM, see Methods) of the three pathways (Pol II occupancy, transcription efficiency and RNA stability) at (F) other subtelomeric genes and (G) mat locus. Pie charts show relative contribution of each pathway to heterochromatic silencing at repeats in wild-type cells. Average of at least two independent samples is shown for all figures. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

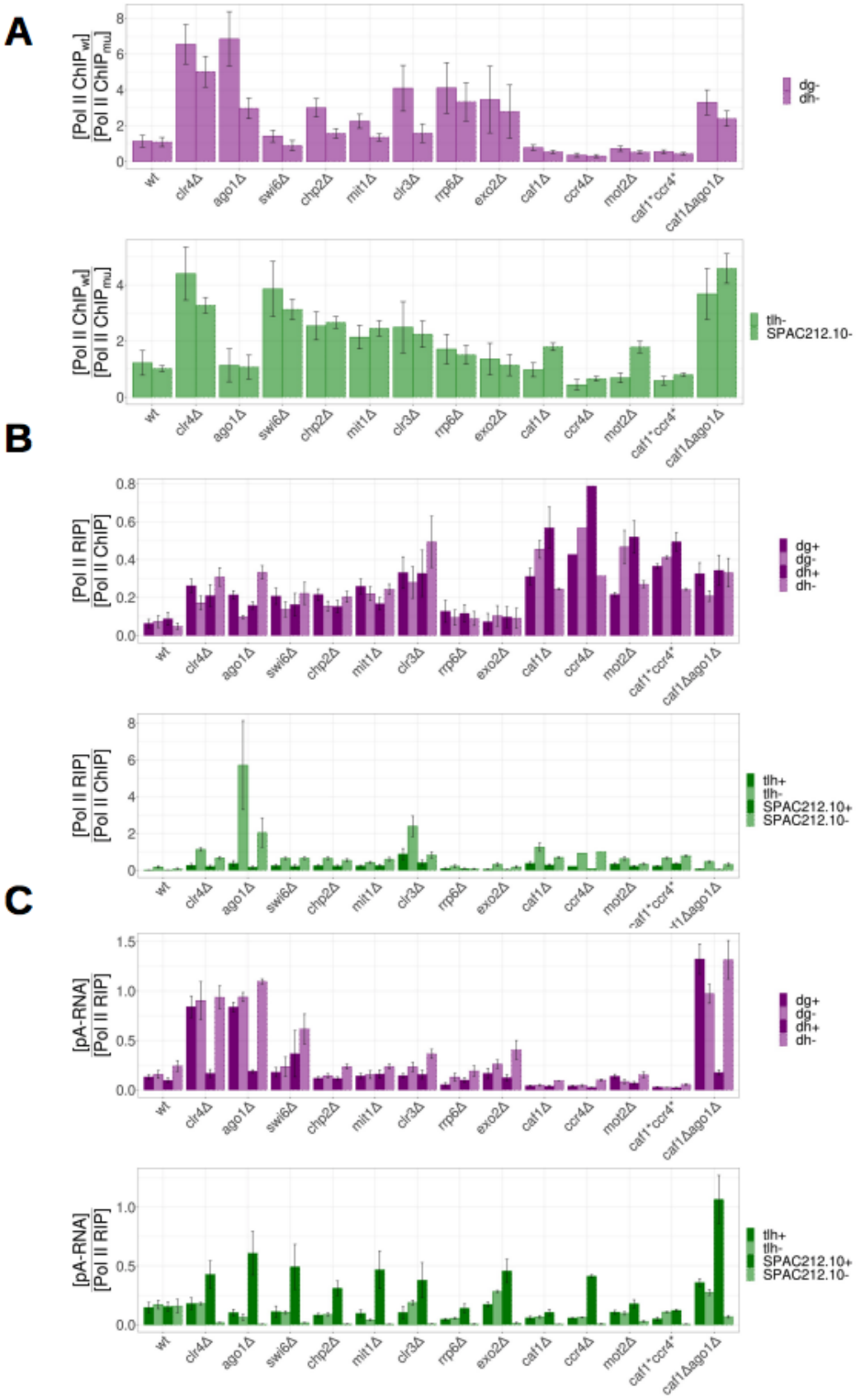

Figure S2: Confidence intervals. Standard error between replicates for RNA Pol II occupancy, transcriptional efficiency and RNA degradation at centromeric dg/dh and subtelomeric tlh repeats. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

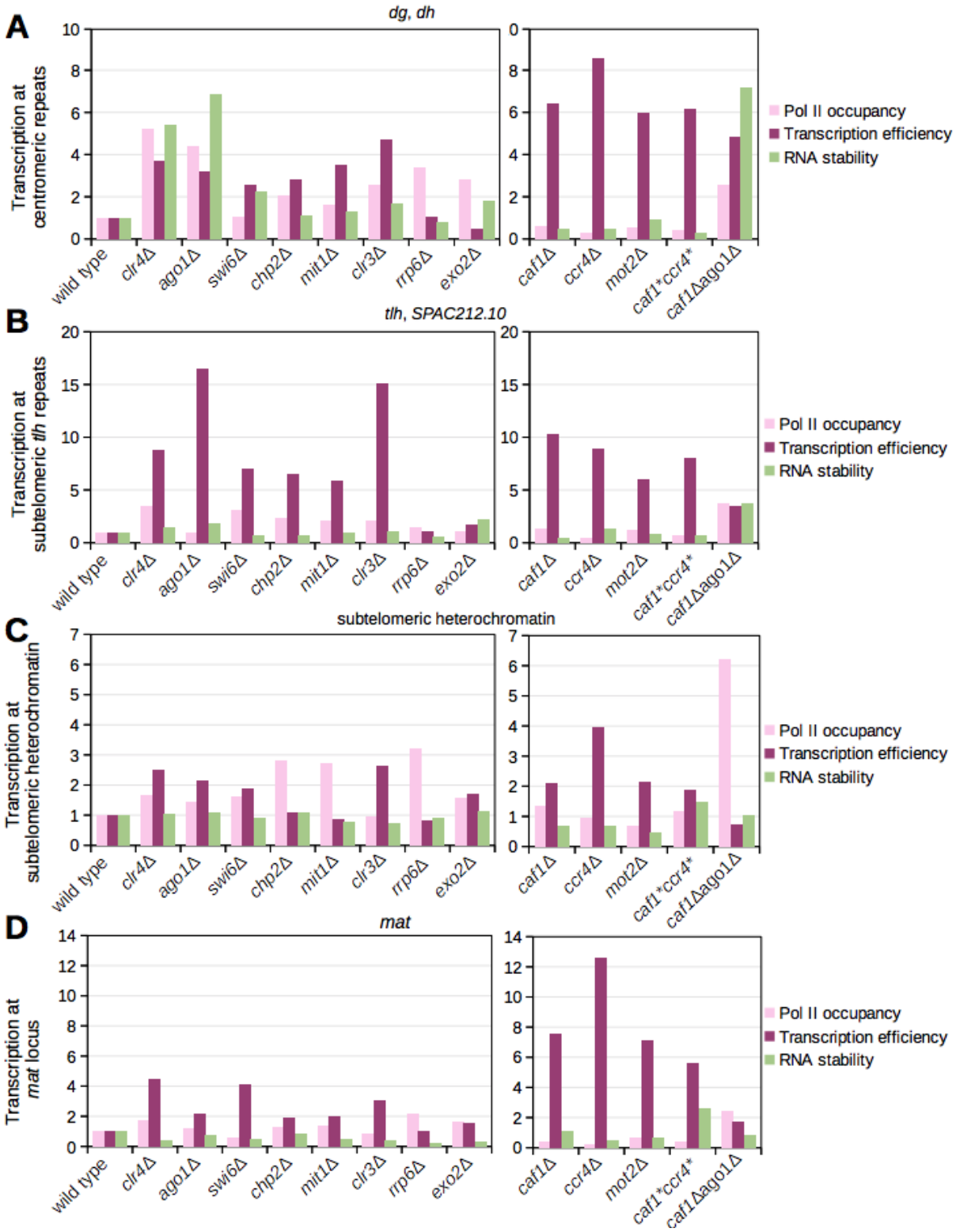

Figure S3: Silencing in mutant cells. Bar chart showing fold change in quantitative measures (ratios of average TPM, see Methods) of the three pathways (Pol II occupancy, transcription efficiency and RNA stability) at (A) centromeric repeats, (B) subtelomeric tlh repeats, (C) other subtelomeric genes and (D) mat locus. Pie charts show relative contribution of each pathway to heterochromatic silencing at repeats in wild-type cells. Average of at least two independent samples is shown for all figures. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

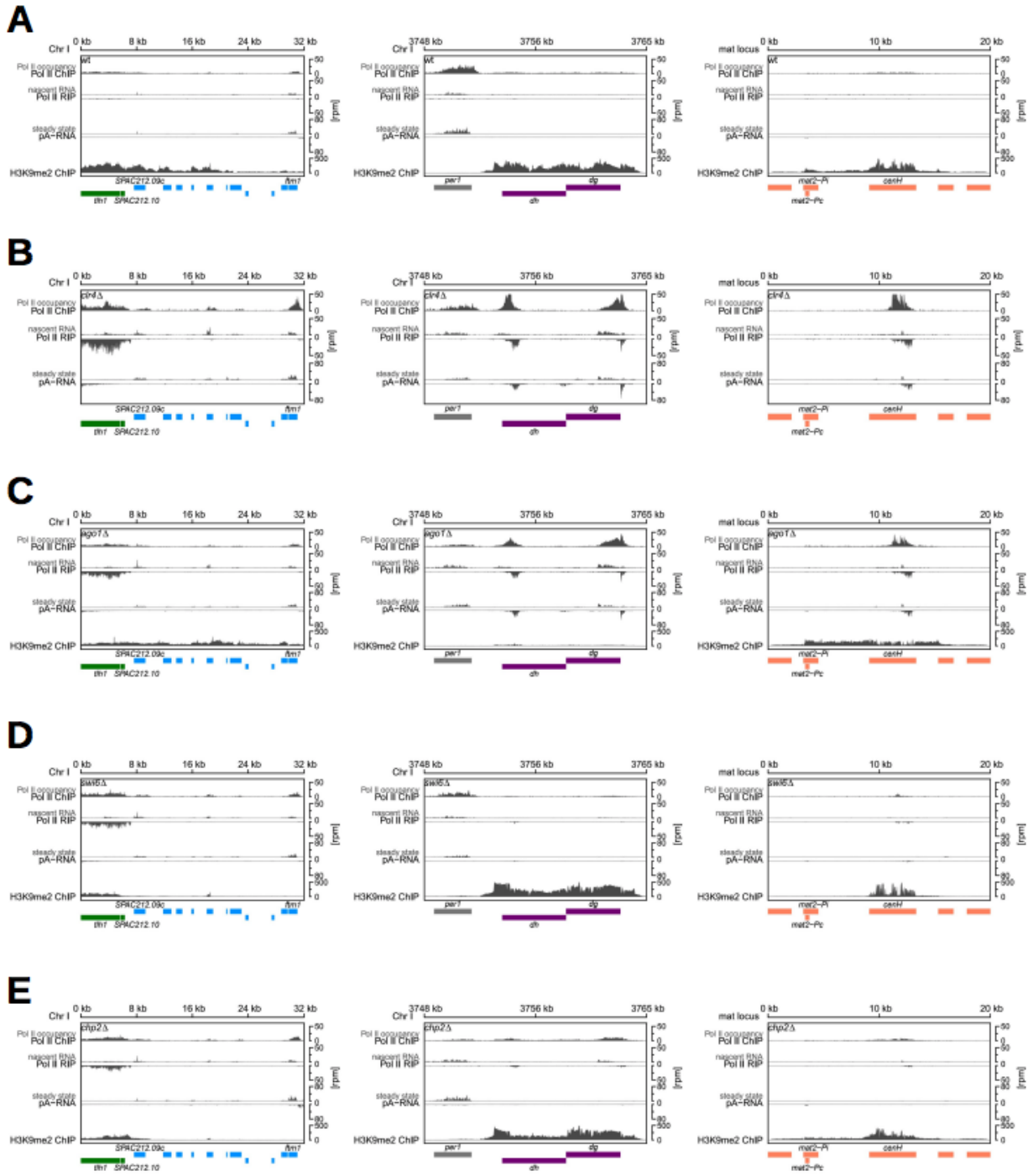

Figure S4: Next-Generation-Sequencing analysis of S2P-Pol II ChIP-seq (Pol II occupancy), S2P-Pol II RIP-seq (nascent RNA) and pA RNA-seq (steady state RNA) at subtelomeric and centromeric repeats, and at the mat locus. Locations of genes are indicated as boxes below the coverage according the color code: gray = protein coding; purple = dg, dh; green = tlh, SPAC212.10, blue = other heterochromatic genes, orange = mat locus. (A) wild type (B) clr4Δ (C) ago1Δ (D) swi6Δ (E) chp2Δ. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

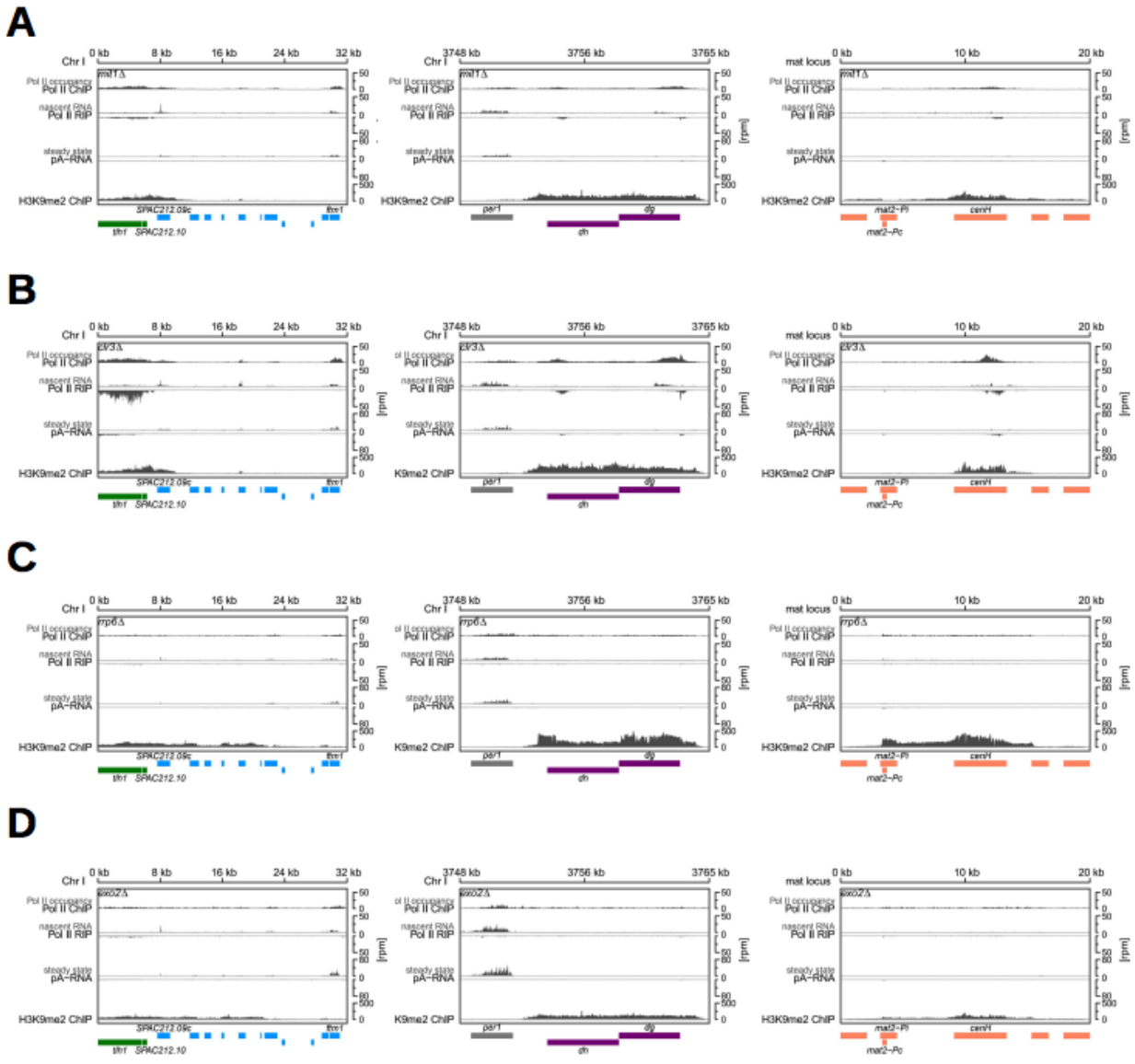

Figure S5: Next-Generation-Sequencing analysis of S2P-Pol II ChIP-seq (Pol II occupancy), S2P-Pol II RIP-seq (nascent RNA) and pA RNA-seq (steady state RNA) at subtelomeric and centromeric repeats, and at the mat locus. Locations of genes are indicated as boxes below the coverage according the color code: gray = protein coding; purple = dg, dh; green = tlh, SPAC212.10, blue = other heterochromatic genes; orange = mat locus. (A) mit1Δ (B) clr3Δ (C) rrp6Δ (D) exo2Δ. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

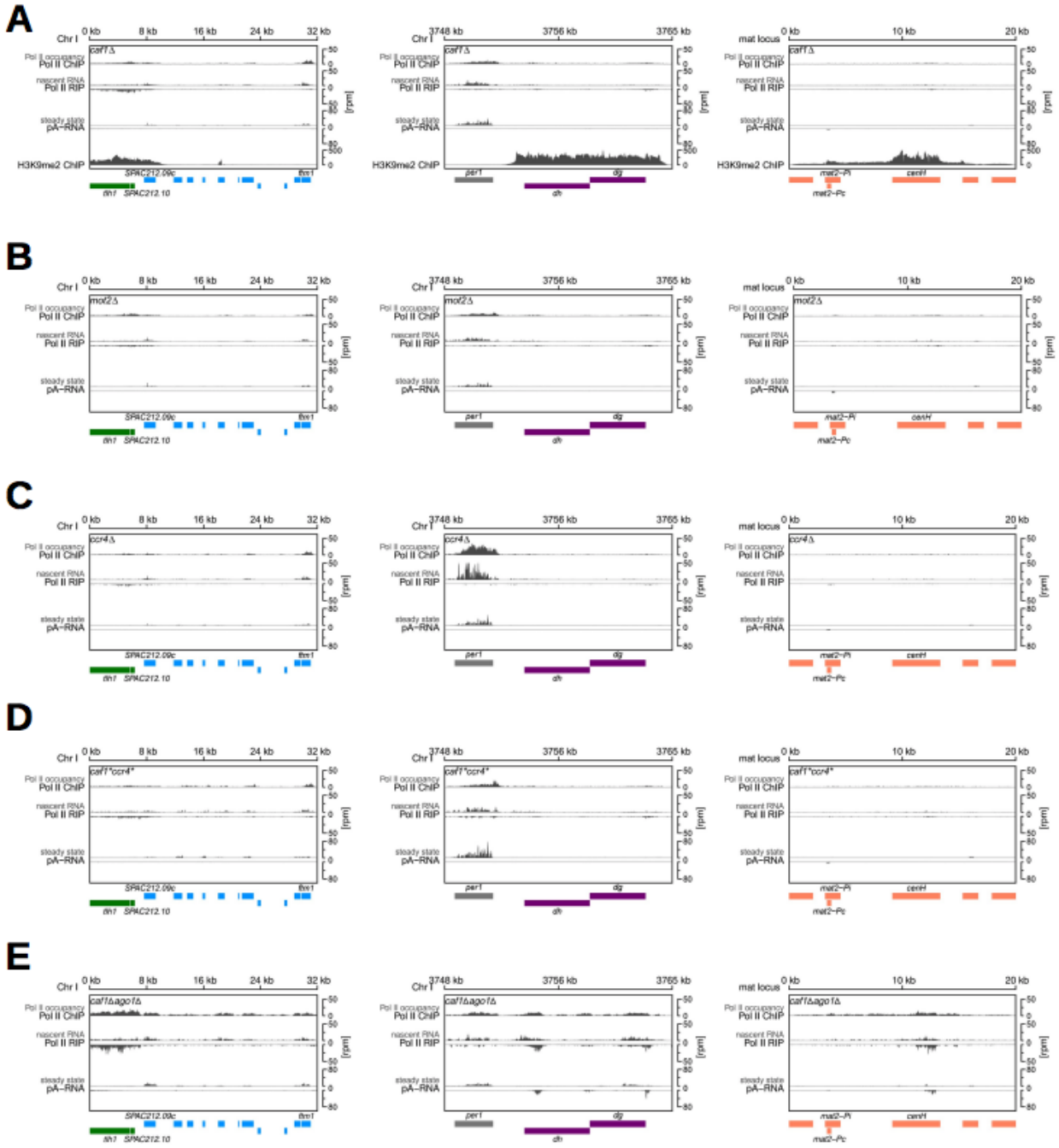

Figure S6: Next-Generation-Sequencing analysis of S2P-Pol II ChIP-seq (Pol II occupancy), S2P-Pol II RIP-seq (nascent RNA) and pA RNA-seq (steady state RNA) at subtelomeric and centromeric repeats, and at the mat locus. Locations of genes are indicated as boxes below the coverage according the color code: gray = protein coding; purple = dg, dh; green = tlh, SPAC212.10, blue = other heterochromatic genes; orange = mat locus. (A) caf1Δ (B) mot2Δ (C) ccr4Δ (D) caf1*ccr4* (E) caf1Δago1Δ. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

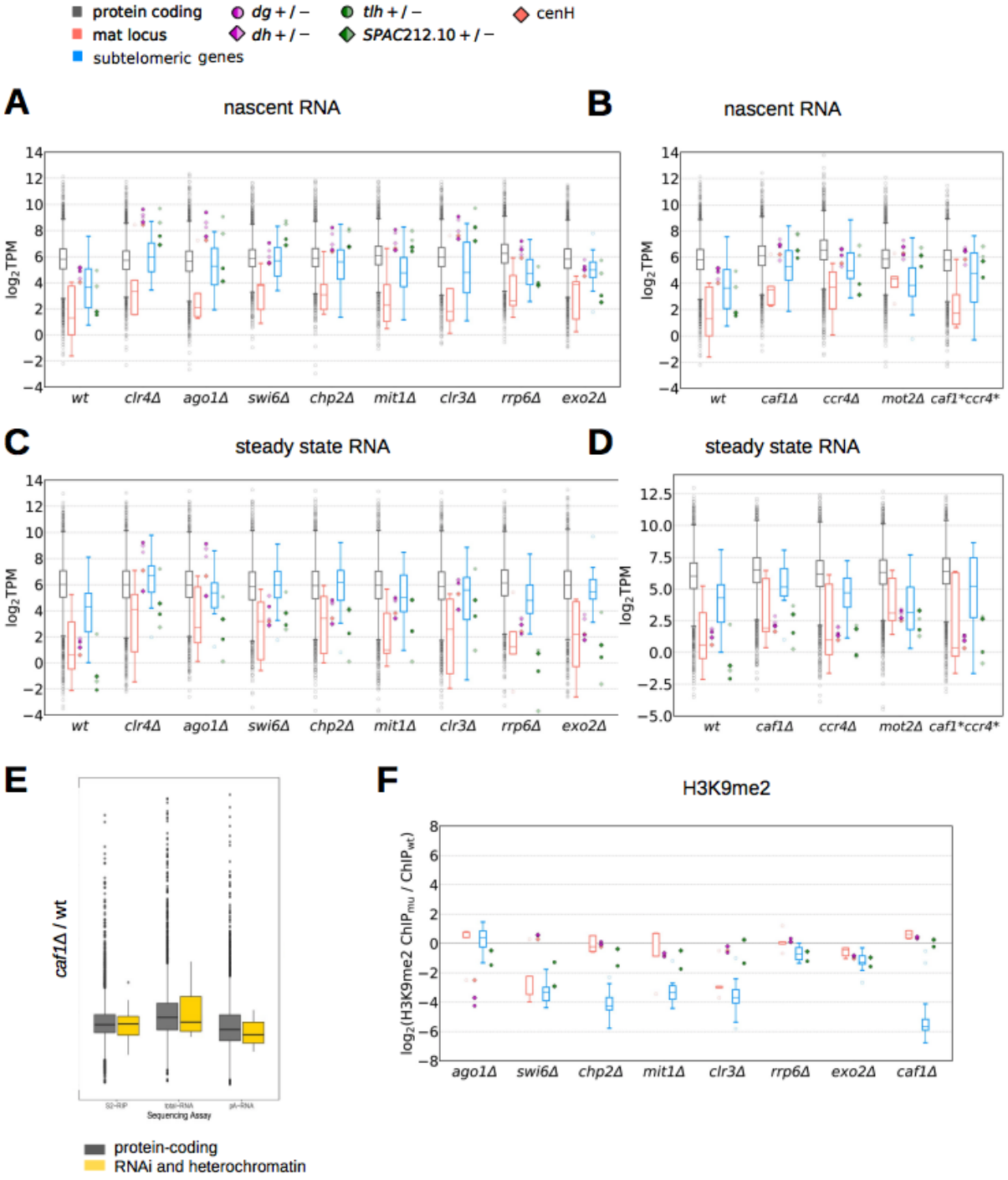

Figure S7: (A, B) Box plot showing S2P-Pol II RIP-seq data in wild-type and mutant cells. Nascent RNA analysis is shown for individual mutants affecting (A) heterochromatin formation and RNA degradation or (B) the Ccr4-Not complex. Data are plotted as defined for Figure 1B. (C, D) pA RNA-seq results (steady state RNA) shown as box plot for individual wild-type and mutants affecting (C) heterochromatin formation and RNA degradation or (D) the Ccr4-Not complex. Data are plotted as defined for Figure 1B. (E) Box plot showing ratio of RNA levels in S2P-Pol II RIP-seq, total RNA and pA RNA data. Protein coding genes are shown in grey and genes involved in RNAi and heterochromatin formation are shown in yellow. (F) Box plot showing H3K9me2 ChIP-seq data. H3K9me2 analysis is shown for individual mutants affecting heterochromatin formation or RNA degradation. Data are plotted as defined for Figure 1B. Average of at least two independent samples is shown. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

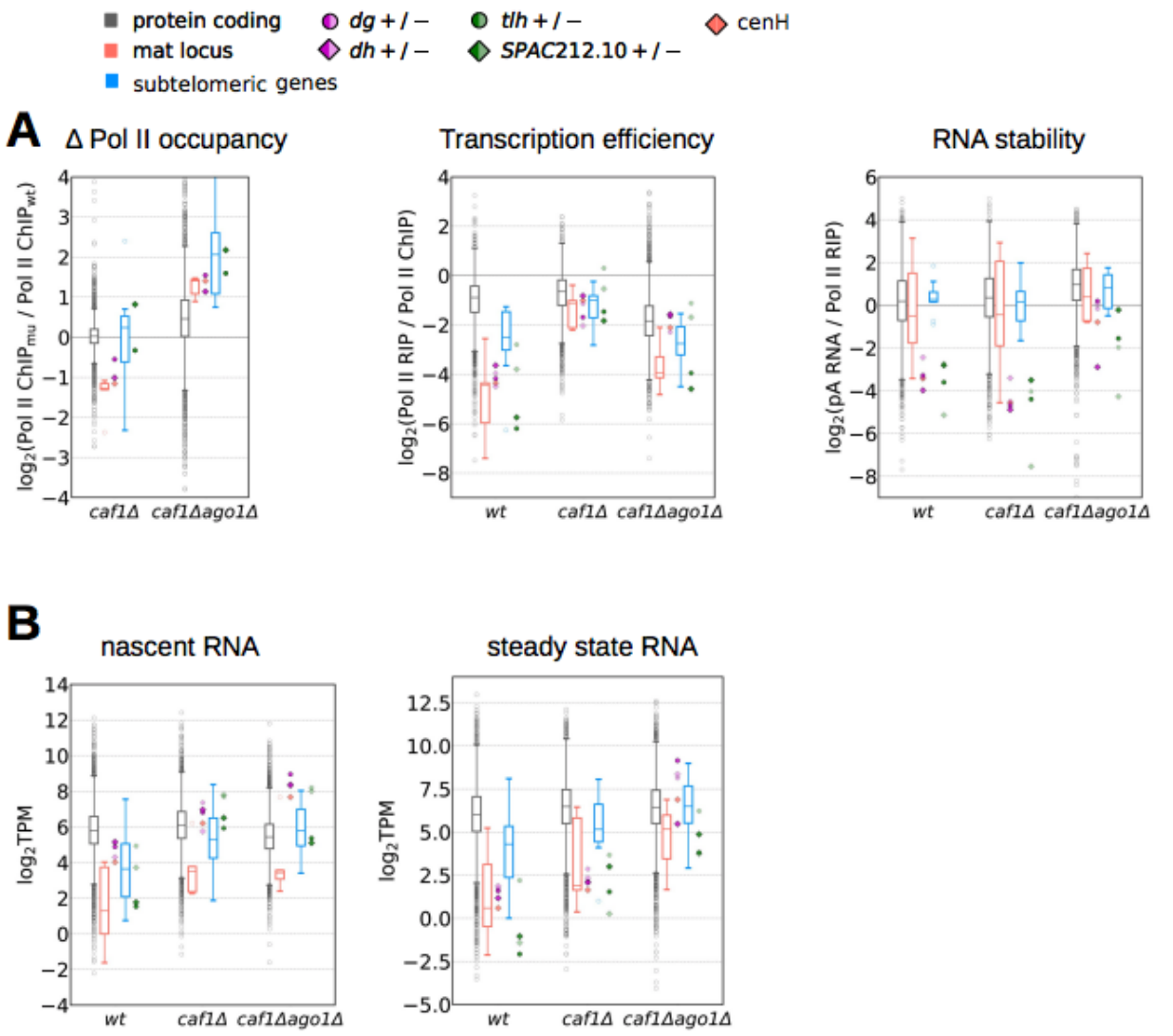

Figure S8: (A) Box plots showing RNA Pol II occupancy, transcription efficiency and RNA stability over indicated genes in caf1Δago1Δ cells. Data are plotted as defined for Figure 1B. (B) Box plots showing nascent and steady state RNA over indicated genes in caf1Δago1Δ cells. Data are plotted as defined for Figure 1B. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

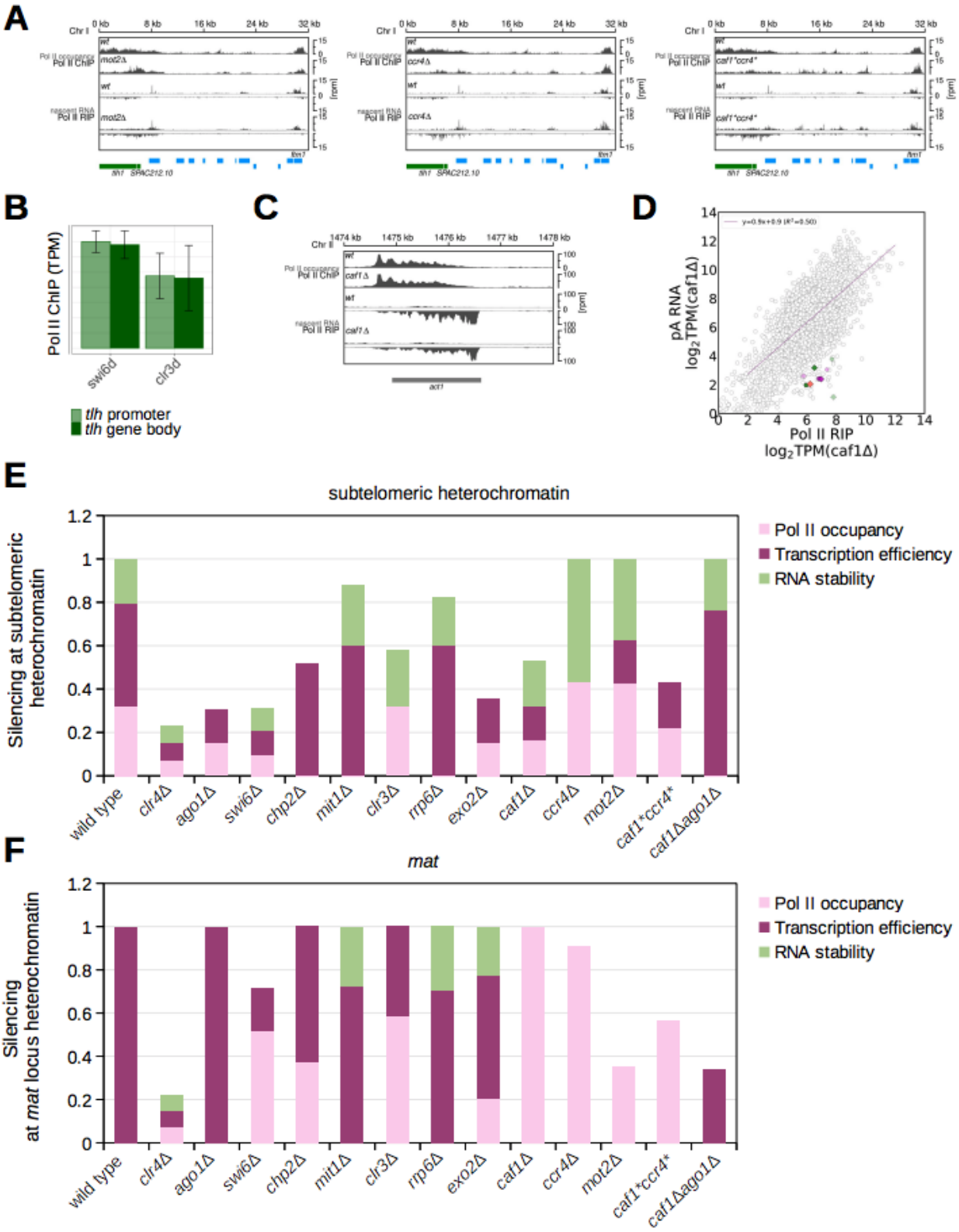
A
B
C
D
E
F
Pol II ChIP (TPM)
swi6d
clr3d
tlh promoter
tlh gene body
pA RNA log2TPM(caf1Δ)
Pol II RIP log2TPM(caf1Δ)
subtelomeric heterochromatin
Silencing at subtelomeric heterochromatin
Pol II occupancy
Transcription efficiency
RNA stability
wild type
clr4Δ
ago1Δ
swi6Δ
chp2Δ
mit1Δ
clr3Δ
rrp6Δ
exo2Δ
caf1Δ
ccr4Δ
mot2Δ
caf1*ccr4*
caf1Δago1Δ
mat
Silencing at mat locus heterochromatin

Figure S9: (A) Analysis of the next-generation sequencing data showing occupancy of S2P RNA Pol II (ChIP-seq) and nascent RNA (S2P-Pol II RIP-seq) at subtelomeric regions in S. pombe mot2Δ, ccr4Δ and caf1*ccr4* cells. Gene locations are indicated as boxes below the coverage and color-coded: green, subtelomeric loci tlh and SPAC212.10; blue, other subtelomeric genes. (B) Quantification of RNA Pol II occupancy (S2P-Pol II ChIP-seq) at tlh promoter region and tlh gene body in indicated wild type and mutant strains. (C) Analysis of the next-generation sequencing data showing occupancy of S2P RNA Pol II (ChIP-seq) and nascent RNA (S2P-Pol II RIP-seq) at euchromatic gene actin in S. pombe caf1Δ cells. Gene locations are indicated as boxes below the coverage. (D) RNA stability in caf1Δ cells. pA RNA-seq (steady state RNA) data plotted over S2P-Pol II RIP-seq data (nascent RNA). TPM, transcripts per million. Gray circles are individual protein-coding genes; regression line is also shown in purple. Also plotted are centromeric dg and dh (dark purple for + strand, bright purple for - strand) and tlh and SPAC212.10 (dark green for + strand, bright green for – strand) and cenH (orange). Each data point is the average of at least two independent samples. (E) Bar chart displaying contribution of each pathway that is still active in the mutants to the silencing of other subtelomeric genes. The height of each bar corresponds to the fold change in RNA output relative to wild-type. The relative contribution of each pathway was computed as fold change in quantitative measures (ratios of average TPM, see Methods) relative to clr4Δ. (F) Bar chart displaying contribution of each pathway that is still active in the mutants to the silencing at the mat locus silencing. The height of each bar corresponds to the fold change in RNA output relative to wild-type. The relative contribution of each pathway was computed as fold change in quantitative measures (ratios of average TPM, see Methods) relative to clr4Δ. Figure reproduced from Monteagudo-Mesas et al. (2022) licensed under Creative Commons CC BY.

Tables

Table S1: Supplemental Table S1: Strains used in this study

Strain id	Description
65	h90 otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 natMX6::3xFLAG-ago1
63	h+ otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210
80	h+ otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 clr4Δ::kanMX6
638	h+ otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 ago1Δ::kanMX6
301	h90 mat3::ura4+ ura4-DS/E leu1-32 ade6-M210 swi6Δ::natMX6
324	h90 mat3::ura4+ ura4-DS/E leu1-32 ade6-M210 chp2Δ::kanMX6
491	h+ leu1-32 ura4-D18 imr1R(NCol)::ura4+ oriI ade6-216 mit1Δ::kanMX6
302	h+ otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 clr3Δ::TAP-kanMX6
504	h+ otr1R::ura4, ura4-DS/E, ade6-M216; leu1-32, his7-366 natMX6::3xFLAG-ago1 rrp6Δ::kanMX6
530	h+ otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 natMX6::3xFLAG-ago1 exo2Δ::kanMX6
510	h90 otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 natMX6::3xFLAG-ago1 caf1Δ::kanMX6
591	h90, ade6-D1, his3-D1, leu1-3, ura4-D18, otr1R(SphI)::ade6$^+$, TAS-his3$^+$-tel1(L), TASura4$^+$-tel2(L), caf1Δ::kanMX6
544	h90 otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 natMX6::3xFLAG-ago1 ccr4Δ::hphMX6
1168	h90, ade6-D1, his3-D1, leu1-3, ura4-D18, otr1R(SphI)::ade6$^+$, TAS-his3$^+$-tel1(L), TASura4$^+$-tel2(L), ccr4H664A-ccr4Terminator::hphMX6, nat::caf1promotercaf1D53AD243AD174A
1022	h90 otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 natMX6::3xFLAG-ago1 mot2Δ::kanMX6
1023	h90 otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 natMX6::3xFLAG-ago1 mot2Δ::kanMX6
523	h90 otr1R(SphI)::ura4+ ura4-DS/E leu1-32 ade6-M210 caf1Δ::kanMX6 ago1Δ::hph

Table S2: Supplemental Table S2: List of heterochromatic genes

Region	Gene id
Subtelomeric	'SPAC212.09c', 'SPNCRNA.70', 'SPAC212.08c', 'SPAC212.07c', 'SPAC212.12', 'SPAC212.06c', 'SPAC212.04c', 'SPAC212.03', 'SPAC212.02', 'SPAC212.01c', 'SPAC977.01', 'SPAC977.18', 'SPAC977.02', 'SPAC977.03', 'SPAC977.04', 'SPAC212.05c'
mat locus	'SPMTR.01', 'FP565355_region_1..2120', 'FP565355_region_9170..13408', 'FP565355_region_15609..16735', 'FP565355_region_18009..20128'

A.2 Dynamic pseudo-time warping of complex trajectories

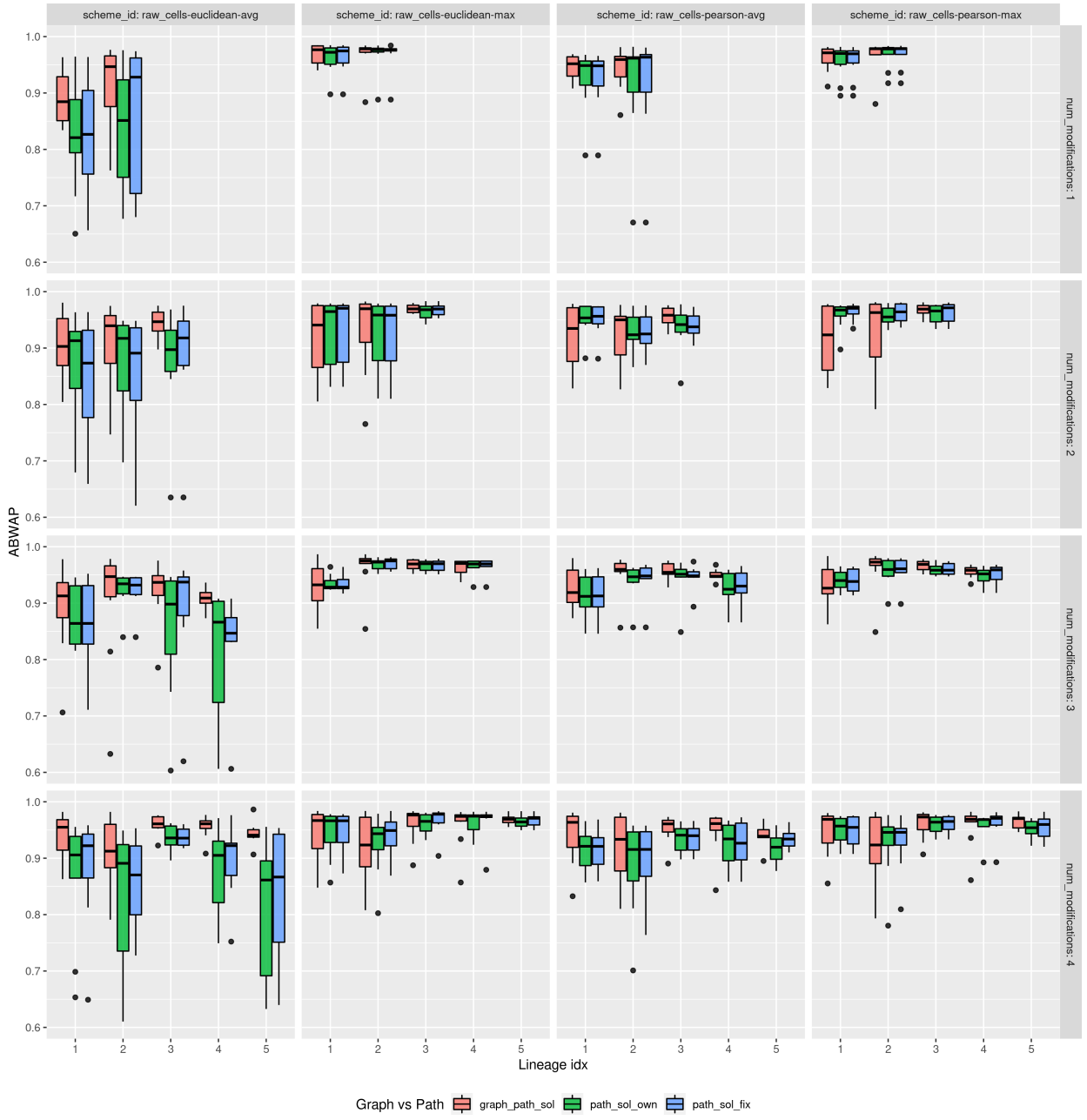

Figure S10: Comparison of full-graph and multiple path-wise Trajan's alignments based on ABWAP scores for dataset with 40 complex trajectories generated with dynverse. All possible parameter combinations for Trajan alignments are shown: raw and smoothed cells representation, avg and max penalty scheme and Euclidean distance and Pearson correlation. Trajectories are stratified by complexity-level (num_modifications=1,2,3,4) and each trajectory is further split into individual lineages (1,2,3,4,5) for comparison purposes. Shown are ABWAP scores associated to each linear alignment computed under 3 different schemes: full-graph (red) and path-wise (green and blue) alignments.

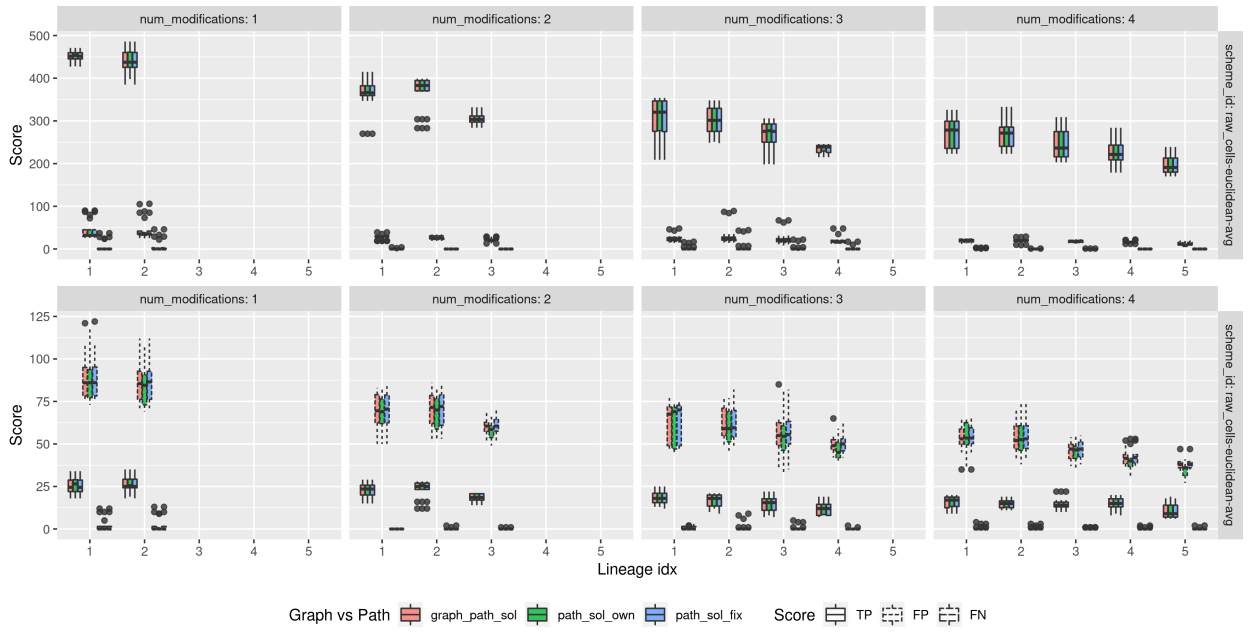

Figure S11: Sub-sampling experiment evaluating Trajan alignments by number of true positives (TP), false positives (FP) and false negatives (FN) in 40 complex trajectories generated with dynverse and perturbed at different levels. **(top)** Sub-sampling of 80% of cells **(bottom)** Sub-sampling of 20% of cells Trajan alignments were based on: raw cells representation, avg penalty scheme and Euclidean distance parameter combination. Trajectories are stratified by complexity-level (num_modifications=1,2,3,4) and each trajectory is further split into individual lineages (1,2,3,4,5) for comparison purposes. Shown are TPs, FPs and FNs associated to each linear alignment computed under 3 different schemes: full-graph (red) and path-wise (green and blue) alignments.

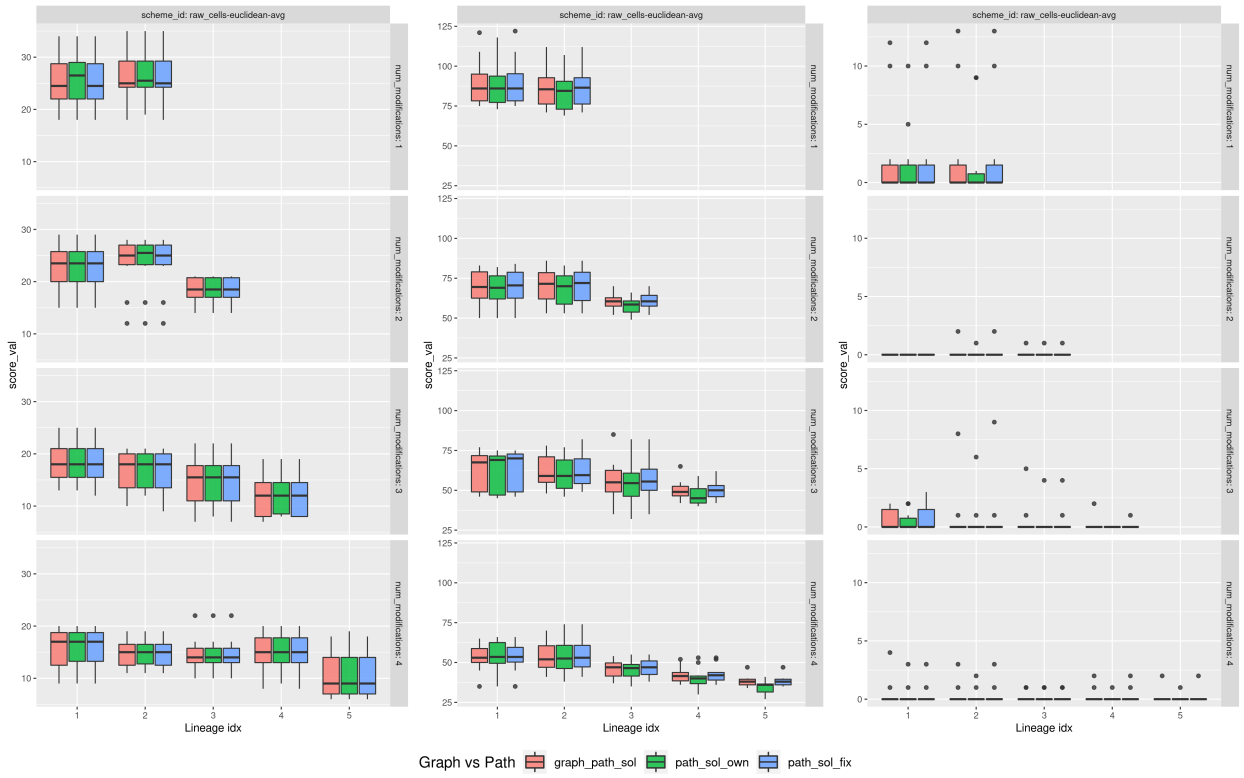

Figure S12: Sub-sampling experiment of 40 complex trajectories generated with dynverse and perturbed using 20% of cells. Evaluate Trajan alignments using number of **(left)** true positives (TP), **(middle)** false positives (FP) and **(right)** false negatives (FN). Trajan alignments were based on: raw cells representation, avg penalty scheme and Euclidean distance parameter combination. Trajectories are stratified by complexity-level (num_modifications=1,2,3,4) and each trajectory is further split into individual lineages (1,2,3,4,5) for comparison purposes. Shown are TPs, FPs and FNs associated to each linear alignment computed under 3 different schemes: full-graph (red) and path-wise (green and blue) alignments.

Bibliography

Akshay Agrawal, Alnur Ali, and Stephen Boyd. Minimum-distortion embedding. *bioRxiv*, 3 2021. URL http://arxiv.org/abs/2103.02559.

Bruce Alberts, Rebeca Heald, Alexander Johnson, David Morgan, and Martin Raff. *Molecular biology of the cell*. Norton and Company, 7th edition, 7 2022. ISBN 0393884856.

Robin C. Allshire and Karl Ekwall. Epigenetic regulation of chromatin states in schizosaccharomyces pombe. *Cold Spring Harbor perspectives in biology*, 7:1–25, 7 2015. ISSN 1943-0264. doi: 10.1101/CSHPERSPECT.A018770. URL https://pubmed.ncbi.nlm.nih.gov/26134317/.

Ayelet Alpert, Lindsay S. Moore, Tania Dubovik, and Shai S. Shen-Orr. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nature Methods 2018 15:4*, 15: 267–270, 3 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4628. URL https://www.nature.com/articles/nmeth.4628.

Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq-a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31:166–169, 1 2015. ISSN 14602059. doi: 10.1093/BIOINFORMATICS/BTU638.

Ricard Argelaguet, Anna S.E. Cuomo, Oliver Stegle, and John C. Marioni. Computational principles and challenges in single-cell data integration, 10 2021. ISSN 15461696.

Luka Borozan. Combinatorial optimization algorithms for (pseudo)alignment in bioinformatics, 7 2021. URL https://urn.nsk.hr/urn:nbn:hr:217:769679.

Cornelia Brönner, Luca Salvi, Manuel Zocco, Ilaria Ugolini, and Mario Halic. Accumulation of rna on chromatin disrupts heterochromatic silencing. *Genome Research*, 27:1174–1183, 7 2017. ISSN 15495469. doi: 10.1101/GR.216986.116.

Cornelia Michaela Brönner. The role of rna degradation in heterochromatin formation, 5 2017.

Jürg Bähler, Jian-Qiu Wu, Mark S. Longtine, Nirav G. Shah, Amos Mckenzie III, Alexander B. Steever, Achim Wach, Peter Philippsen, and John R. Pringle. Heterologous modules for efficient and versatile pcr-based gene targeting in schizosaccharomyces pombe. *Yeast*, 14:943–951, 1998. doi: 10.1002/(SICI)1097-0061(199807)14:10⟨943::AID-YEA292⟩3.0.CO;2-Y. URL https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0061(199807)14:10%3C943::AID-YEA292%3E3.0.CO;2-Y.

Sebastian Böcker, Stefan Canzar, and Gunnar W. Klau. The generalized robinson-foulds metric. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8126 LNBI:156–169, 2013. ISSN 03029743. doi: 10.1007/978-3-642-40453-5_13/COVER. URL https://link.springer.com/chapter/10.1007/978-3-642-40453-5_13.

Marc Bühler, Noah Spies, David P. Bartel, and Danesh Moazed. Tramp-mediated rna surveillance prevents spurious entry of rnas into the schizosaccharomyces pombe sirna pathway. *Nature Structural and Molecular Biology*, 15:1015–1023, 10 2008. ISSN 15459993. doi: 10.1038/NSMB.1481.

Davide Cacchiarelli, Xiaojie Qiu, Sanjay Srivatsan, Anna Manfredi, Michael Ziller, Eliah Overbey, Antonio Grimaldi, Jonna Grimsby, Prapti Pokharel, Kenneth J. Livak, Shuqiang Li, Alexander Meissner, Tarjei S. Mikkelsen, John L. Rinn, and Cole Trapnell. Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. *Cell Systems*, 7:258–268.e3, 9 2018. ISSN 2405-4712. doi: 10.1016/J.CELS.2018.07.006.

Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. Computational methods for trajectory inference from single-cell transcriptomics. *European Journal of Immunology*, 46:2496–2506, 11 2016a. ISSN 00142980. doi: 10.1002/eji.201646347. URL http://doi.wiley.com/10.1002/eji.201646347.

Robrecht Cannoodt, Wouter Saelens, Dorine Sichien, Simon Tavernier, Sophie Janssens, Martin Guilliams, Bart N Lambrecht, Katleen De Preter, and Yvan Saeys. Scorpius improves trajectory inference and identifies novel modules in dendritic cell development. *bioRxiv*, page 079509, 10 2016b. doi: 10.1101/079509. URL https://www.biorxiv.org/content/10.1101/079509v2https://www.biorxiv.org/content/10.1101/079509v2.abstract.

Robrecht Cannoodt, Wouter Saelens, Louise Deconinck, and Yvan Saeys. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nature Communications 2021 12:1*, 12:1–9, 6 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24152-2. URL https://www.nature.com/articles/s41467-021-24152-2.

Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, and Jay Shendure. The single-cell transcriptional landscape of mammalian organogenesis. *Nature 2019 566:7745*, 566:496–502, 2 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0969-x. URL https://www.nature.com/articles/s41586-019-0969-x.

Stephane E. Castel and Robert A. Martienssen. Rna interference in the nucleus: Roles for small rnas in transcription, epigenetics and beyond. *Nature Reviews Genetics*, 14:100–112, 2 2013. ISSN 14710056. doi: 10.1038/NRG3355.

Lingyi Chen and Jonathan Widom. Mechanism of transcriptional silencing in yeast. *Cell*, 120:37–48, 1 2005. ISSN 00928674. doi: 10.1016/J.CELL.2004.11.030.

Cristina Cotobal, María Rodríguez-López, Caia Duncan, Ayesha Hasan, Akira Yamashita, Masayuki Yamamoto, Jürg Bähler, and Juan Mata. Role of ccr4-not complex in heterochromatin formation at meiotic genes and subtelomeres in fission yeast. *Epigenetics and Chromatin*, 8, 8 2015. ISSN 17568935. doi: 10.1186/S13072-015-0018-4.

Kevin M. Creamer, Godwin Job, Sreenath Shanker, Geoffrey A. Neale, Yuan chi Lin, Blaine Bartholomew, and Janet F. Partridge. The mi-2 homolog mit1 actively positions nucleosomes within heterochromatin to suppress transcription. *Molecular and Cellular Biology*, 34:2046–2061, 6 2014. ISSN 0270-7306. doi: 10.1128/MCB.01609-13.

Francis Harry Compton Crick. On protein synthesis. *Cambridge University Press*, 12:138–163, 1958.

Aaron Diaz, Abhinav Nellore, and Jun S. Song. Chance: comprehensive software for quality control and validation of chip-seq data. *Genome biology*, 13:R98, 2012. ISSN 14656914. doi: 10.1186/GB-2012-13-10-R98.

Van Hoan Do. Computational methods for large-scale single-cell rna-seq and multimodal data, 11 2021.

Van Hoan Do, Mislav Blažević, Pablo Monteagudo, Luka Borozan, Khaled Elbassioni, Sören Laue, Francisca Rojas-Ringeling, Domagoj Matijević, and Stefan Canzar. Dynamic pseudo-time warping of complex single-cell trajectories. *bioRxiv*, 1 2019. doi: 10.1101/522672. URL `https://www.biorxiv.org/content/10.1101/522672v1https://www.biorxiv.org/content/10.1101/522672v1.abstract`.

Van Hoan Do, Khaled Elbassioni, and Stefan Canzar. Sphetcher: Spherical thresholding improves sketching of single-cell transcriptomic heterogeneity. *iScience*, 23, 6 2020. ISSN 25890042. doi: 10.1016/j.isci.2020.101126.

Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: Ultrafast universal rna-seq aligner. *Bioinformatics*, 29:15–21, 1 2013. ISSN 13674803. doi: 10.1093/BIOINFORMATICS/BTS635.

Raghuvar Dronamraju, Austin J. Hepperla, Yoichiro Shibata, Alexander T. Adams, Terry Magnuson, Ian J. Davis, and Brian D. Strahl. Spt6 association with rna polymerase ii directs mrna turnover during transcription. *Molecular Cell*, 70:1054–1066.e4, 6 2018. ISSN 10974164. doi: 10.1016/j.molcel.2018.05.020. URL `http://www.cell.com/article/S1097276518303940/fulltexthttp://www.cell.com/article/S1097276518303940/abstracthttps://www.cell.com/molecular-cell/abstract/S1097-2765(18)30394-0`.

Arnob Dutta, Vinod Babbarwal, Jianhua Fu, Deborah Brunke-Reese, Diane M. Libert, Jonathan Willis, and Joseph C. Reese. Ccr4-not and tfiis function cooperatively to rescue arrested rna polymerase ii. *Molecular and Cellular Biology*, 35:1915–1925, 6 2015. ISSN 0270-7306. doi: 10.1128/MCB.00044-15.

Daniel C. Ellwanger, Mirko Scheibinger, Rachel A. Dumont, Peter G. Barr-Gillespie, and Stefan Heller. Transcriptional dynamics of hair-bundle morphogenesis revealed with celltrails. *Cell Reports*, 23:2901–2914.e13, 6 2018. ISSN 2211-1247. doi: 10.1016/J.CELREP.2018.05.002.

Wei Feng and Scott D. Michaels. Accessing the inaccessible: The organization, transcription, replication, and repair of heterochromatin in plants. *Annual Review of Genetics*, 49:439–459, 12 2015. ISSN 15452948. doi: 10.1146/ANNUREV-GENET-112414-055048. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-genet-112414-055048.

Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 2016 17:6*, 17:333–351, 5 2016. ISSN 1471-0064. doi: 10.1038/nrg.2016.49. URL https://www.nature.com/articles/nrg.2016.49.

Shiv I.S. Grewal and Sarah C.R. Elgin. Transcription and rna interference in the formation of heterochromatin. *Nature*, 447:399–406, 5 2007. ISSN 14764687. doi: 10.1038/NATURE05914.

Mario Halic and Danesh Moazed. Dicer-independent primal rnas trigger rnai and heterochromatin formation. *Cell*, 140:504–516, 2 2010. ISSN 00928674. doi: 10.1016/J.CELL.2010.01.019.

Stephen M. Hewitt. Negative consequences of the central dogma, 11 2020. ISSN 15515044.

Sahana Holla, Jothy Dhakshnamoorthy, H. Diego Folco, Vanivilasini Balachandran, Hua Xiao, Ling ling Sun, David Wheeler, Martin Zofall, and Shiv I.S. Grewal. Positioning heterochromatin at the nuclear periphery suppresses histone turnover to promote epigenetic inheritance. *Cell*, 180:150–164.e15, 1 2020. ISSN 10974172. doi: 10.1016/J.CELL.2019.12.004.

Daniel Holoch and Danesh Moazed. Rna-mediated epigenetic regulation of gene expression. *Nature Reviews Genetics*, 16:71–84, 1 2015. ISSN 14710064. doi: 10.1038/NRG3863.

Jonathan Houseley, John LaCava, and David Tollervey. Rna-quality control by the exosome. *Nature Reviews Molecular Cell Biology*, 7:529–539, 7 2006. ISSN 14710072. doi: 10.1038/NRM1964.

Johan Hyllner, Chris Mason, and Ian Wilmut. Cells: from robert hooke to cell therapy—a 350 year journey. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370, 10 2015. ISSN 14712970. doi: 10.1098/RSTB.2015.0320. URL /pmc/articles/PMC4634005/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4634005/.

Haoyang Jiang, Marley Wolgast, Laura M. Beebe, and Joseph C. Reese. Ccr4–not maintains genomic integrity by controlling the ubiquitylation and degradation of arrested rnapii. *Genes and Development*, 33:705–717, 6 2019. ISSN 15495477. doi: 10.1101/GAD.322453.118.

Aaron Johnson, Ronghu Wu, Matthew Peetz, Steven P. Gygi, and Danesh Moazed. Heterochromatic gene silencing by activator interference and a transcription elongation barrier. *Journal of Biological Chemistry*, 288:28771–28782, 10 2013. ISSN 00219258. doi: 10.1074/JBC.M113.460071.

W. Min Jou, G. Haegeman, M. Ysebaert, and W. Fiers. Nucleotide sequence of the gene coding for the bacteriophage ms2 coat protein. *Nature 1972 237:5350*, 237:82–88, 1972. ISSN 1476-4687. doi: 10.1038/237082a0. URL `https://www.nature.com/articles/237082a0`.

Christine M. Kiely, Samuel Marguerat, Jennifer F. Garcia, Hiten D. Madhani, Jürg Bähler, and Fred Winston. Spt6 is required for heterochromatic silencing in the fission yeast schizosaccharomyces pombe. *Molecular and Cellular Biology*, 31:4193–4204, 10 2011. ISSN 0270-7306. doi: 10.1128/MCB.05568-11.

Johannes Klaus, Sabina Kanton, Christina Kyrousi, Ane Cristina Ayo-Martin, Rossella Di Giaimo, Stephan Riesenberg, Adam C. O'Neill, J. Gray Camp, Chiara Tocco, Malgorzata Santel, Ejona Rusha, Micha Drukker, Mariana Schroeder, Magdalena Götz, Stephen P. Robertson, Barbara Treutlein, and Silvia Cappello. Altered neuronal migratory trajectories in human cerebral organoids derived from individuals with neuronal heterotopia. *Nature Medicine 2019 25:4*, 25:561–568, 3 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0371-0. URL `https://www.nature.com/articles/s41591-019-0371-0`.

Jennifer A. Kruk, Arnob Dutta, Jianhua Fu, David S. Gilmour, and Joseph C. Reese. The multifunctional ccr4-not complex directly promotes transcription elongation. *Genes and Development*, 25:581–593, 3 2011. ISSN 08909369. doi: 10.1101/GAD.2020911.

Ross F. Laidlaw, Emma M. Briggs, Keith R. Matthews, Richard McCulloch, and Thomas D. Otto. Tragedy: Trajectory alignment of gene expression dynamics. *bioRxiv*, page 2022.12.21.521424, 12 2022. doi: 10.1101/2022.12.21.521424. URL `https://www.biorxiv.org/content/10.1101/2022.12.21.521424v1https://www.biorxiv.org/content/10.1101/2022.12.21.521424v1.abstract`.

Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie Levine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, Ladeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J.

Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Mei Lee Hong, Joann Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa De La Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G.R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F.A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw Pyng Yang, Ru Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature 2001 409:6822*, 409:860–921, 2 2001. ISSN 1476-4687. doi: 10.1038/35057062. URL `https://www.nature.com/articles/35057062`.

Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B. Schiller, Dana Pe'er, and Fabian J. Theis. Cellrank for directed single-cell fate mapping. *Nature Methods 2022 19:2*, 19:159–170, 1 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01346-6. URL `https://www.nature.com/articles/s41592-021-01346-6`.

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data, 10 2010. ISSN 14710056.

Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19:41–50, 1 2022. ISSN 15487105. doi: 10.1038/s41592-021-01336-8.

Mirela Marasovic, Manuel Zocco, and Mario Halic. Argonaute and triman generate dicer-independent prirnas and mature sirnas to initiate heterochromatin formation. *Molecular Cell*, 52:173–183, 10 2013. ISSN 10972765. doi: 10.1016/J.MOLCEL.2013.08.046.

Robert Martienssen and Danesh Moazed. Rnai and heterochromatin assembly. *Cold Spring Harbor Perspectives in Biology*, 7, 8 2015. ISSN 19430264. doi: 10.1101/CSHPERSPECT.A019323.

John D. McPherson. Next-generation gap. *Nature Methods 2009 6:11*, 6:S2–S5, 10 2009. ISSN 1548-7105. doi: 10.1038/nmeth.f.268. URL `https://www.nature.com/articles/nmeth.f.268`.

Jason E. Miller and Joseph C. Reese. Ccr4-not complex: The control freak of eukaryotic cells. *Critical Reviews in Biochemistry and Molecular Biology*, 47:315–333, 7 2012. ISSN 10409238. doi: 10.3109/10409238.2012.667214.

Pablo Monteagudo-Mesas, Cornelia Brönner, Parastou Kohvaei, Haris Amedi, Stefan Canzar, and Mario Halic. Ccr4–not complex reduces transcription efficiency in heterochromatin. *Nucleic Acids Research*, 50:5565–5576, 6 2022. ISSN 0305-1048. doi: 10.1093/NAR/GKAC403. URL `https://academic.oup.com/nar/article/50/10/5565/6595293`.

Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5:621–628, 7 2008. ISSN 15487091. doi: 10.1038/nmeth.1226. URL `http://www.nature.com/naturemethods`.

Mohammad R. Motamedi, Eun Jin Erica Hong, Xue Li, Scott Gerber, Carilee Denison, Steven Gygi, and Danesh Moazed. Hp1 proteins form distinct complexes and mediate heterochromatic gene silencing by nonoverlapping mechanisms. *Molecular Cell*, 32:778–790, 12 2008. ISSN 10972765. doi: 10.1016/J.MOLCEL.2008.10.026.

Siddhartha Mukherjee. *The gene: an intimate history*. Scribner, 5 2016. ISBN 1476733503.

Akiko K. Okita, Faria Zafar, Jie Su, Dayalini Weerasekara, Takuya Kajitani, Tatsuro S. Takahashi, Hiroshi Kimura, Yota Murakami, Hisao Masukata, and Takuro Nakagawa. Heterochromatin suppresses gross chromosomal rearrangements at centromeres by repressing tfs1/tfiis-dependent transcription. *Communications Biology*, 2, 12 2019. ISSN 23993642. doi: 10.1038/S42003-018-0251-Z.

R. Gonzalo Parra, Nikolaos Papadopoulos, Laura Ahumada-Arranz, Jakob El Kholtei, Noah Mottelson, Yehor Horokhovsky, Barbara Treutlein, and Johannes Soeding. Reconstructing complex lineage trees from scrna-seq data using merlot. *Nucleic Acids Research*, 47:8961–8974, 9 2019. ISSN 13624962. doi: 10.1093/nar/gkz706. URL `https://github.com/soedinglab/csgraph`.

E. Passarge. Emil heitz and the concept of heterochromatin: longitudinal chromosome differentiation was recognized fifty years ago. *American Journal of Human Genetics*, 31: 106, 1979. ISSN 00029297. URL `/pmc/articles/PMC1685768/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1685768/`.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. ISSN 1533-7928. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.

Paola Pisacane and Mario Halic. Tailing and degradation of argonaute-bound small rnas protect the genome from uncontrolled rnai. *Nature Communications*, 8, 5 2017. ISSN 20411723. doi: 10.1038/NCOMMS15332.

Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A. Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods 2017 14:10*, 14:979–982, 8 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4402. URL `https://www.nature.com/articles/nmeth.4402`.

Francisca E. Reyes-Turcu and Shiv I.S. Grewal. Different means, same end-heterochromatin formation by rnai and rnai-independent rna processing factors in fission yeast. *Current Opinion in Genetics and Development*, 22:156–163, 4 2012. ISSN 0959437X. doi: 10.1016/J.GDE.2011.12.004.

Francisca E. Reyes-Turcu, Ke Zhang, Martin Zofall, Eesin Chen, and Shiv I.S. Grewal. Defects in rna quality control factors reveal rnai-independent nucleation of heterochromatin. *Nature Structural and Molecular Biology*, 18:1132–1138, 10 2010. ISSN 15459993. doi: 10.1038/NSMB.2122.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature Biotechnology 2019*, page 1, 4 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0071-9. URL `https://www.nature.com/articles/s41587-019-0071-9#Fig7`.

F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phix174 dna. *Nature 1977 265:5596*, 265:687–695, 1977. ISSN 1476-4687. doi: 10.1038/265687a0. URL `https://www.nature.com/articles/265687a0`.

Bernd Schuettengruber, Daniel Chourrout, Michel Vervoort, Benjamin Leblanc, and Giacomo Cavalli. Genome regulation by polycomb and trithorax proteins. *Cell*, 128:735–745, 2 2007. ISSN 00928674. doi: 10.1016/J.CELL.2007.02.009.

Gergana Shipkovenska, Alexander Durango, Marian Kalocsay, Steven P. Gygi, and Danesh Moazed. A conserved rna degradation complex required for spreading and epigenetic inheritance of heterochromatin. *eLife*, 9:1–25, 6 2020. ISSN 2050084X. doi: 10.7554/ELIFE.54341.

Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics 2019 20:11*, 20:631–656, 7 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0150-2. URL `https://www.nature.com/articles/s41576-019-0150-2`.

Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19:477, 12 2018. ISSN 1471-2164. doi: 10.1186/s12864-018-4772-0. URL `https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4772-0`.

Reiichi Sugihara, Yuki Kato, Tomoya Mori, and Yukio Kawahara. Alignment of single-cell trajectory trees with capital. *Nature Communications*, 13, 12 2022. ISSN 20411723. doi: 10.1038/s41467-022-33681-3.

Tomoyasu Sugiyama, Hugh P. Cam, Rie Sugiyama, Ken ichi Noma, Martin Zofall, Ryuji Kobayashi, and Shiv I.S. Grewal. Shrec, an effector complex for heterochromatic transcriptional silencing. *Cell*, 128:491–504, 2 2007. ISSN 00928674. doi: 10.1016/J.CELL.2006.12.035.

Tomoyasu Sugiyama, Gobi Thillainadesan, Venkata R. Chalamcharla, Zhaojing Meng, Vanivilasini Balachandran, Jothy Dhakshnamoorthy, Ming Zhou, and Shiv I.S. Grewal. Enhancer of rudimentary cooperates with conserved rna-processing factors to promote meiotic mrna decay and facultative heterochromatin assembly. *Molecular Cell*, 61:747–759, 3 2016. ISSN 10974164. doi: 10.1016/J.MOLCEL.2016.01.029.

Dinithi Sumanaweera, Chenqu Suo, Daniele Muraro, Emma Dann, Krzysztof Polanski, Alexander S. Steemers, Jong-Eun Park, Bianca Dumitrascu, and Sarah A. Teichmann. Gene-level alignment of single cell trajectories informs the progression of in vitro t cell differentiation. *bioRxiv*, page 2023.03.08.531713, 3 2023. doi: 10.1101/2023.03.08.531713. URL `https://www.biorxiv.org/content/10.1101/2023.03.08.531713v1https://www.biorxiv.org/content/10.1101/2023.03.08.531713v1.abstract`.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, and M. Azim Surani. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods 2009 6:5*, 6:377–382, 4 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1315. URL `https://www.nature.com/articles/nmeth.1315`.

Carson C. Thoreen, Lynne Chantranupong, Heather R. Keys, Tim Wang, Nathanael S. Gray, and David M. Sabatini. A unifying model for mtorc1-mediated regulation of mrna translation. *Nature*, 485:109–113, 5 2012. ISSN 00280836. doi: 10.1038/NATURE11083.

Helena Todorov, Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. Tinga: fast and flexible trajectory inference with growing neural gas. *Bioinformatics*, 36:i66–i74, 2020. doi: 10.1093/bioinformatics/btaa463. URL `https://github.com/Helena-todd/TinGa.`

Thinh N. Tran and Gary D. Bader. Tempora: Cell trajectory inference using time-series single-cell rna sequencing data. *PLOS Computational Biology*, 16:e1008205, 9 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008205. URL `https://dx.plos.org/10.1371/journal.pcbi.1008205`.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J. Lennon, Kenneth J. Livak, Tarjei S. Mikkelsen, and John L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of

single cells. *Nature Biotechnology 2014 32:4*, 32:381–386, 3 2014. ISSN 1546-1696. doi: 10.1038/nbt.2859. URL `https://www.nature.com/articles/nbt.2859`.

Barbara Treutlein, Qian Yi Lee, J. Gray Camp, Moritz Mall, Winston Koh, Seyed Ali Mohammad Shariati, Sopheak Sim, Norma F. Neff, Jan M. Skotheim, Marius Wernig, and Stephen R. Quake. Dissecting direct reprogramming from fibroblast to neuron using single-cell rna-seq. *Nature 2016 534:7607*, 534:391–395, 6 2016. ISSN 1476-4687. doi: 10.1038/nature18323. URL `https://www.nature.com/articles/nature18323`.

Ilaria Ugolini and Mario Halic. Fidelity in rna-based recognition of transposable elements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373, 12 2018. ISSN 14712970. doi: 10.1098/RSTB.2018.0168.

J. Craig Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Yuan Wang, A. Wang, X. Wang, J. Wang, M. H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Lai Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Yu H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. Ni Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck,

M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291:1304–1351, 2 2001. ISSN 00368075. doi: 10.1126/SCIENCE.1058040/SUPPL_FILE/C18_SCIENCE.PDF. URL https://www.science.org/doi/10.1126/science.1058040.

André Verdel, Songtao Jia, Scott Gerber, Tomoyasu Sugiyama, Steven Gygi, Shiv I.S. Grewal, and Danesh Moazed. Rnai-mediated targeting of heterochromatin by the rits complex. *Science*, 303:672–676, 1 2004. ISSN 00368075. doi: 10.1126/SCIENCE.1093686.

T. K. Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics*, 4:52–57, 1 1968. ISSN 15738337. doi: 10.1007/BF01074755/METRICS. URL https://link.springer.com/article/10.1007/BF01074755.

Thomas A. Volpe, Catherine Kidner, Ira M. Hall, Grace Teng, Shiv I.S. Grewal, and Robert A. Martienssen. Regulation of heterochromatic silencing and histone h3 lysine-9 methylation by rnai. *Science*, 297:1833–1837, 9 2002. ISSN 00368075. doi: 10.1126/SCIENCE.1074973.

Elmar Wahle and G. Sebastiaan Winkler. Rna decay machines: Deadenylation by the ccr4-not and pan2-pan3 complexes. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1829:561–570, 6 2013. ISSN 18749399. doi: 10.1016/J.BBAGRM.2013.01.003.

Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 2008 10:1*, 10:57–63, 1 2009. ISSN 1471-0064. doi: 10.1038/nrg2484. URL https://www.nature.com/articles/nrg2484.

James Dewey Watson. *The Molecular Biology of the Gene*. W.A. Benjamin, 1st edition, 1965.

Barbara Wold and Richard M. Myers. Sequence census methods for functional genomics. *Nature Methods 2008 5:1*, 5:19–21, 12 2007. ISSN 1548-7105. doi: 10.1038/nmeth1157. URL https://www.nature.com/articles/nmeth1157.

F. Alexander Wolf, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20:1–9, 3 2019. ISSN 1474760X. doi: 10.1186/s13059-019-1663-x. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1663-x.

Authors Ziqi Zhang, Xiuwei Zhang Correspondence, Ziqi Zhang, and Xiuwei Zhang. Inference of high-resolution trajectories in single-cell rna-seq data by using rna velocity. *Cell Reports Methods*, 1:100095, 2021. doi: 10.1016/j.crmeth.2021.100095. URL https://doi.org/10.1016/j.crmeth.2021.100095.

Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. *https://doi.org/10.1137/0218082*, 18:1245–1262, 7 2006. ISSN 00975397. doi: 10.1137/0218082. URL https://epubs.siam.org/doi/10.1137/0218082.

Jiaping Zhao and Laurent Itti. shapedtw: Shape dynamic time warping. *Pattern Recognition*, 74:171–184, 2 2018. ISSN 0031-3203. doi: 10.1016/J.PATCOG.2017.09.020.

Manuel Zocco, Mirela Marasovic, Paola Pisacane, Silvija Bilokapic, and Mario Halic. The chp1 chromodomain binds the h3k9me tail and the nucleosome core to assemble heterochromatin. *Cell Discovery*, 2, 12 2016. ISSN 20565968. doi: 10.1038/CELLDISC.2016.4.