Computational approaches for the analysis of epigenome and transcriptome characterisation in *Paramecium tetraurelia*

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften
im Fach Bioinformatik

vorgelegt beim Fachbereich Mathematik und Informatik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Sivarajan Karunanithi
aus Kumbakonam, Indien

Frankfurt (2021)
(D30)

vom Fachbereich Informatik und Mathematik
der Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr.-Ing. Lars Hedrich

Gutacher: Prof. Dr. Marcel H. Schulz und Dr. Kathi Zarnack

Datum der Disputation: 25-Oct-2021

“*As you dream, so shall you become.*”

Dr. APJ Abdul Kalam

Abstract

In the last two decades, our understanding of human gene regulation has improved tremendously. There are plentiful computational methods which focus on integrative data analysis of humans, and model organisms, like mouse and drosophila. However, these tools are not directly employable by researchers working on non-model organisms to answer fundamental biological, and evolutionary questions. We aimed to develop new tools, and adapt existing software for the analysis of transcriptomic and epigenomic data of one such non-model organism, *Paramecium tetraurelia*, an unicellular eukaryote. *Paramecium* contains two diploid (2n) germline micronuclei (MIC) and a polyploid (800n) somatic macronuclei (MAC). The transcriptomic and epigenomic regulatory landscape of the MAC genome, which has 80% protein-coding genes and short intergenic regions, is poorly understood.

We developed a generic automated eukaryotic short interfering RNA (siRNA) analysis tool, called RAPID. Our tool captures diverse siRNA characteristics from small RNA sequencing data and provides easily navigable visualisations. We also introduced a normalisation technique to facilitate comparison of multiple siRNA-based gene knockdown studies.

Further, we developed a pipeline to characterise novel genome-wide endogenous short interfering RNAs (endo-siRNAs). In contrary to many organisms, we found that the endo-siRNAs are not acting in *cis*, to silence their parent mRNA. We also predicted phasing of siRNAs, which are regulated by the RNA interference (RNAi) pathway. Further, using RAPID, we investigated the aberrations of endo-siRNAs, and their respective transcriptomic alterations caused by an RNAi pathway triggered by feeding small RNAs against a target gene. We find that the small RNA transcriptome is altered, even if a gene unrelated to RNAi pathway is targeted. This is important in the context of investigations of genetically modified organisms (GMOs). We suggest that future studies need to distinguish transcriptomic changes caused by RNAi inducing techniques and actual regulatory changes.

Subsequently, we adapted existing epigenomics analysis tools to conduct the first comprehensive epigenomic characterisation of nucleosome positioning and histone modifications of the *Paramecium* MAC. We identified well positioned nucleosomes shifted downstream of the transcription start site. GC content seems to dictate, in *cis*, the positioning of nucleosomes, histone marks (H3K4me3, H3K9ac, and H3K27me3), and Pol II in the AT-rich *Paramecium* genome. We employed a chromatin state segmentation approach, on nucleosomes and histone marks, which revealed genes with active, repressive, and bivalent chromatin states. Further, we constructed a regulatory association network of all the aforementioned data, using the sparse partial correlation network technique. Our analysis revealed subsets of genes, whose expression is positively associated with H3K27me3, different to the otherwise reported negative association with gene expression in many other organisms.

Further, we developed a Random Forests classifier to predict gene expression using genic (gene length, intron frequency, etc.) and epigenetic features. Our model has a test performance (PR-AUC) of 0.83. Upon evaluating different feature sets, we found that genic features are as predictive, of gene expression, as the epigenetic features. We used Shapley local feature explanation values, to suggest that high H3K4me3, high intron frequency, low gene length, high sRNA, and high GC content are the most important elements for determining gene expression status.

In this thesis, we developed novel tools, and employed several bioinformatics and machine learning methods to characterise the regulatory landscape of the *Paramecium*'s (epi)genome.

Kurzfassung

In den letzten zwei Jahrzehnten hat sich unser Verständnis der menschlichen Genregulation enorm verbessert. Es gibt eine Fülle von computergestützten Methoden, die sich auf die integrative Datenanalyse von Menschen und Modellorganismen wie Maus und Drosophila konzentrieren. Diese Werkzeuge können jedoch nicht direkt von Forschern eingesetzt werden, die an Nicht-Modellorganismen arbeiten, um grundlegende biologische und evolutionäre Fragen zu beantworten. Unser Ziel war es, neue Werkzeuge zu entwickeln und bestehende Software für die Analyse von transkriptomischen und epigenomischen Daten eines solchen Nicht-Modellorganismus, *Paramecium tetraurelia*, einem einzelligen Eukaryoten, anzupassen. *Paramecium* enthält zwei diploide (2n) Keimbahn-Mikrokerne (MIC) und einen polyploiden (800n) somatischen Makronukleus (MAC). Die transkriptomische und epigenomische Regulationsmechanismen des MAC-Genoms, das 80% proteinkodierende Gene und kurze intergene Regionen aufweist, ist bisher nur wenig verstanden.

Wir haben ein generisches, automatisiertes Analyse-Tool für kurze interferierende RNAs (siRNAs) in Eukaryoten entwickelt, genannt RAPID. Unser Tool erfasst diverse siRNA-Charakteristika aus kleinen RNA (sRNA) Sequenzierungsdaten und bietet leicht navigierbare Visualisierungen. Wir haben auch eine Normalisierungstechnik eingeführt, um den Vergleich von mehreren siRNA-basierten Gen-Knockdown-Studien zu erleichtern.

Darüber hinaus haben wir eine Pipeline zur Charakterisierung neuartiger genomweiter endogener kurzer interferierender RNAs (endo-siRNAs) entwickelt. Im Gegensatz zu vielen anderen Organismen fanden wir heraus, dass die endo-siRNAs in *Paramecium* nicht in *cis* wirken, um ihre Eltern-mRNA zu inhibieren. Wir haben auch die Phasenlage der siRNAs vorhergesagt, die durch den RNA-Interferenz (RNAi) Signalweg reguliert wird. Weiterhin untersuchten wir mit Hilfe von RAPID die Aberrationen von endo-siRNAs und ihre jeweiligen transkriptomischen Veränderungen, die durch einen RNAi-Signalweg verursacht werden, der durch die Zuführung kleiner RNAs gegen ein Zielgen ausgelöst wird. Wir fanden heraus, dass das Transkriptom der kleinen RNAs verändert wird, auch wenn ein Gen, das nicht mit dem RNAi-Signalweg in Verbindung steht, als Ziel gewählt wird. Dies ist wichtig im Zusammenhang mit Untersuchungen von gentechnisch veränderten Organismen (GVOs). Wir schlagen vor, dass zukünftige Studien transkriptomische Veränderungen, die durch RNAi-induzierende Techniken verursacht werden, von tatsächlichen regulatorischen Veränderungen unterschieden werden müssen.

Anschließend adaptierten wir bestehende Epigenomik-Analysetools, um die erste umfassende epigenomische Charakterisierung der Nukleosomenpositionierung und Histonmodifikationen des *Paramecium* MAC durchzuführen. Wir identifizierten präzise positionierte Nukleosomen, die in 3'-Richtung von der Transkriptionsstartstelle verschoben sind. Der GC-Gehalt scheint in *cis* die Positionierung von Nukleosomen, Histonmarkierungen (H3K4me3, H3K9ac und H3K27me3) und Pol II in dem AT-reichen *Paramecium*-Genom vorzugeben.

Wir verwendeten eine Technik zur Segmentierung des Chromatinzustands mittels der Position von Nukleosomen und Histonmarkierungen, deren Ergebnis Gene mit aktiven, repressiven und bivalenten Chromatinzuständen aufzeigte. Außerdem konstruierten wir ein regulatorisches Assoziationsnetzwerk aus allen oben genannten Daten, indem wir die sogenannte Sparse Partial Correlation Network Methode verwendeten. Mit unserer Analyse fanden wir Gene, deren Expression positiv mit H3K27me3 assoziiert ist, wohingegen diese Histonmarkierung in anderen Organismen üblicherweise mit einer negativen Genexpression in Verbindung gebracht wird.

Weiterhin entwickelten wir einen Random forests Klassifikator zur Vorhersage der Genexpression unter Verwendung von genetischen (Genlänge, Intron-Frequenz, etc.) und epigenetischen Merkmalen. Unser Modell hat eine Testgenauigkeit (PR-AUC) von 0,83. Bei der Evaluierung verschiedener Gruppen von Merkmalen haben wir festgestellt, dass genetische Merkmale genauso viel zu der Vorhersage der Genexpression beigetragen haben wie die epigenetischen. Wir benutzten die sogenannten Shapley local feature explanation values, die nahelegen, dass hohe H3K4me3, hohe Intron-Frequenz, geringe Genlänge, hohe sRNA und hoher GC-Gehalt die wichtigsten Elemente für die Bestimmung des Genexpressionsstatus sind.

In dieser Arbeit haben wir neuartige Werkzeuge entwickelt und verschiedene bioinformatische und maschinelle Lernmethoden eingesetzt, um die regulatorischen Mechanismen des (Epi-)Genoms von *Paramecium* zu charakterisieren.

Zusammenfassung

Einführung

Das Verständnis der menschlichen Genregulation ist für die Verbesserung der Gesundheitsversorgung von größter Bedeutung. Mehrere Konsortialprojekte wie die Encylopedia of DNA Elements (ENCODE) und das International Human Epigenome Consortium (IHEC) haben unser Verständnis von Transkriptomik und Epigenomik enorm verbessert. Forscher haben zahlreiche integrative Datenanalysetools entwickelt, um die Projekte der Konsortien zu ergänzen. Allerdings sind die vorhandenen Tools auf menschliche Daten oder wenige andere Modellorganismen, wie Maus und Drosophila, zugeschnitten. Oft sind diese Tools nicht direkt von Forschern anwendbar, die an Nicht-Modellorganismen arbeiten. Die Forschung an Nicht-Modellorganismen ist entscheidend für die Beantwortung mechanistischer biologischer und evolutionärer Fragen.

Paramecium tetraurelia ist ein freilebender einzelliger Eukaryote (Zellen mit eingekapselter DNA). Sie werden als Modellorganismus verwendet, um unsere Grundlagen in der Zell- und Evolutionsbiologie zu erweitern, da sie mehrere ungewöhnliche Eigenschaften aufweisen. Ihre große Zelloberfläche ($50-300\mu m$) besteht aus Basalkörpern und Zilien (haarähnliche Auswüchse), was morphogenetische Studien ermöglicht. Diese Studien zeigten die Existenz einer nichtmendelschen zytoplasmatischen Vererbung der Zilienanordnung (Beisson and Sonneborn, 1965). *Paramecium* weist mehrere Arten der Fortpflanzung auf. Sie können sich selbst befruchten (Autogamie) oder eine Konjugation mit anderen Paarungspartnern eingehen und sich ungeschlechtlich fortpflanzen. Die verschiedenen Paarungstypen (oder "Geschlechter") (Sonneborn, 1947) und die Expression von Oberflächenantigenen (Epstein and Forney, 1984) sind weitere Beispiele für zytoplasmatische Vererbung. Die Expression der Oberflächenantigene ist gegenseitig exklusiv, d.h. es wird jeweils nur ein Antigen exprimiert. Die Expression eines spezifischen Oberflächenantigens wird als Serotyp bezeichnet. Die Serotypen zeigen eine Anpassung an die Umgebung, z. B. kann die Temperatur eine Verschiebung von einem Serotyp zum anderen bewirken (Matsuda and Forney, 2005).

Ein weiteres interessantes Merkmal von *Paramecium* ist ihr Kerndimorphismus. Sie tragen zwei Keimbahn-Mikrokerne (MIC) und einen somatischen Makronukleus (MAC). Der MIC ist diploid und transkriptionell inaktiv. Der MAC ist transkriptionell aktiv und weist Polyploidie (800n) auf (Beale and Preer, 2008). Der ursprüngliche MAC zerfällt nach jedem sexuellen Reproduktionszyklus, wenn sich eine Mutterzelle teilt und vier Tochterzellen bildet. Ein neuer MAC, der sich aus der befruchteten zygotischen MIC durch genomische Umlagerung entwickelt, wird als Entwicklungs-MAC (Van Houten, 2019) bezeichnet. Es ist bekannt, dass kleine RNA (sRNA) und andere nicht-kodierende

RNA (ncRNA) die genomischen Umlagerungen eines Entwicklungs-MAC epigenetisch kontrollieren (Beisson et al., 2010). Während der ungeschlechtlichen Vermehrung (oder des vegetativen Wachstums) teilt sich eine Zelle in zwei Tochterzellen. Die Teilung von MIC und MAC erfolgt auf mitotische bzw. amitotische Weise (Simon and Plattner, 2014). Über ein Jahrzehnt der Forschung am MAC von vegetativem *Paramecium*, zeigte die Rolle von sRNAs bei der Kontrolle der Serotyp-Expression (Marker et al., 2010).

Die Hauptnahrung von *Paramecium* sind Bakterien. Dies ermöglicht es den Forschern, genetisch veränderte Bakterien, die eine doppelsträngige RNA (dsRNA) tragen, an *Paramecium* zu füttern und die genetischen Veränderungen zu untersuchen, die dadurch auftreten. Mehrere Studien nutzten diese Fütterungstechnik, um RNA-Interferenz (RNAi) Signalwege in *Paramecium* zu verstehen (Galvani and Sperling, 2002). Der RNAi-Signalweg hängt von der RNA-abhängigen RNA-Polymerase (RDR) und dem Enzym Dicer (DCR) ab, um kurze interferierende RNAs (siRNAs) zu erzeugen (Carradec et al., 2015; Marker et al., 2014). Wenn die Dicer-Spaltung in regelmäßigen Abständen auf einer dsRNA stattfindet, wird dies als Phasing bezeichnet und die resultierenden siRNAs werden als phased-siRNAs bezeichnet. Diese siRNAs könnten wiederum die Genexpression regulieren, entweder auf transkriptioneller oder auf posttranskriptioneller Ebene (Moazed, 2009; Zhang, 2009). Während wir wissen, wie die exogene dsRNA, die *Paramecium* zugeführt wird, zu siRNAs prozessiert wird, kennen wir die endogene Zusammensetzung der kleinen RNA von vegetativem *Paramecium* nicht.

Die regulatorische Rolle der nicht-kodierenden DNA, der intergenen Regionen und Introns, ist in vielen Organismen gut untersucht. Die intergenen Regionen beherbergen regulatorische Elemente wie Promotoren, Enhancer und Silencer, die verschiedene epigenetische Modifikationen wie DNA-Methylierung und Histon-Methylierung aufweisen. Diese regulatorischen Elemente und Introns kontrollieren zusammen die Genexpression in verschiedenen Organismen (Elkon and Agami, 2017). Die MAC-Genom-Annotation von *Paramecium* ergab eine hohe Proteincodierungsdichte von 80%, die die höchste unter den freilebenden Eukaryoten ist. Daraus ergeben sich kurze intergene Regionen von durchschnittlich nur 352 bp, was die Frage aufwirft, wie das Epigenom des *Paramecium* MAC organisiert ist.

In dieser Arbeit haben wir es uns zum Ziel gesetzt, neue Werkzeuge zu entwickeln und bestehende Software für die Analyse von transkriptomischen und epigenomischen Daten von *Paramecium* anzupassen.

RAPID: Ein automatisiertes Werkzeug zur Analyse kleiner RNAs

Die zahlreichen existierenden sRNA-Analysetools konzentrieren sich in erster Linie auf die Vorhersage und Analyse von micro RNAs (miRNAs) und piwi-interacting RNAs (piRNAs). Sie sind oft nicht in der Lage, systematisch und automatisiert mehrere Proben zu vergleichen und geeignete Normalisierungsstrategien für die siRNA-Analyse zu verwenden.

Wir haben ein automatisiertes Werkzeug zur eukaryotischen siRNA-Analyse entwickelt, genannt RAPID. Unser Tool erfasst diverse siRNA-Charakteristika aus sRNA-Sequenzierungsdaten und bietet leicht navigierbare HTML Visualisierungen. Zu den siRNA-Merkmalen, die RAPID erfasst, gehören strangspezifische Reads und nicht-templierte Nukleotide. Wir haben auch eine Normalisierungstechnik eingeführt, um den Vergleich von siRNA-basierten Gen-Knockdown-Proben zu erleichtern. Wir haben zudem eine Software zur Analyse der differentiellen Expression, DESeq2, integriert. RAPID ist für die öffentliche Nutzung als Conda-Paket und über GitHub (`https://github.com/SchulzLab/RAPID`) verfügbar. Die Nützlichkeit und Benutzerfreundlichkeit von RAPID lassen sich an den über 16.000 Conda-Downloads zum Zeitpunkt der Erstellung dieser Arbeit ablesen.

Genomweite Analyse von RNAi-Mechanismen in *Paramecium*

Zusätzlich zu RAPID entwickelten wir eine Pipeline zur Charakterisierung des ersten genomweiten sRNA-Profils vom vegetativem MAC in *Paramecium* mit vier verschiedenen Serotypen (51A, 51B, 51D, 51H). Wir identifizierten 1.618 endogene kurze interferierende RNAs (endo-siRNAs), die von proteinkodierenden Genen produziert werden. Wir haben keine Mikro-RNAs entdeckt. Des Weiteren fanden wir heraus, dass die Mehrheit der endo-siRNAs (973 von 1.618) in den vier analysierten Serotypen vorkommt, die als Gene associated with small RNA clusters (GSRCs) bezeichnet werden. Die Mehrheit der GSRCs zeigte eine positive Korrelation mit der mRNA-Expression. Die positive Korrelation deutete darauf hin, dass die Endo-siRNAs in *Paramecium* nicht strikt in *cis* agieren, um ihre Eltern-mRNA zu inhibieren, im Gegensatz zu vielen anderen Organismen (Moazed, 2009; Zhang, 2009). Wir haben auch die Phasenlage der siRNAs vorhergesagt, die durch den RNA-Interferenz (RNAi) Signalweg reguliert wird, welche wiederum durch die RDR-Enzyme (RDR1 und RDR2) vermittelt wird.

Des Weiteren untersuchten wir mit Hilfe von RAPID die Aberrationen von endo-siRNAs und ihre jeweiligen transkriptomischen Veränderungen, die durch einen RNAi-Signalweg verursacht werden, der durch die Zuführung von sRNAs gegen ein Zielgen ausgelöst wird. Wir fanden heraus, dass das Transkriptom der sRNAs verändert wird, auch wenn ein Gen, das nicht mit dem RNAi-Weg in Verbindung steht, als Ziel gewählt wird. Dies ist wichtig im Zusammenhang mit Untersuchungen von gentechnisch veränderten Organismen (GVO), da mit RNAi behandelte Organismen als GVO-frei gelten und zunehmend zur Bekämpfung von Viren- und Schädlingsresistenzen in Bakterien und Pflanzen eingesetzt werden. Wir schlagen vor, dass zukünftige Studien transkriptomische Veränderungen, die durch RNAi-induzierende Techniken verursacht werden, von tatsächlichen regulatorischen Veränderungen unterschieden werden müssen.

Wir identifizierten auch mehrere sRNAs in den intergenen und anderen nicht-kodierenden Regionen des MAC-Genoms. Wir stellen die Hypothese auf, dass viele der sRNAs, insbesondere die aus nicht-kodierenden Regionen, wahrscheinlich an der Regulierung der Polyploidie des MAC von *Paramecium* beteiligt sind. Allerdings sind zukünftige Experimente notwendig, um eine solche Hypothese zu bestätigen. Während wir die erste genomweite Analyse der sRNAs mit RAPID und anderen Werkzeugen dokumentiert haben, sind ihre Regulationsmechanismen noch unklar.

Epigenomische Eigenschaften des makronukleären Genoms von *Paramecium*

Die Annotation des vegetativen MAC-Genoms von *Paramecium* ergab eine hohe Proteincodierungsdichte von 80%, die höchste unter den frei lebenden Eukaryoten. Daraus ergeben sich kurze intergene Regionen von nur 352 bp, was die Frage aufwirft, wie der MAC von *Paramecium* reguliert wird. Daher verlagerten wir unseren Fokus auf das Verständnis der epigenomischen Eigenschaften des MAC, die bis dahin noch nie charakterisiert worden waren. Mit Hilfe von Nukleosomen-Positionierungssoftwares identifizierten wir präzise positionierte Nukleosomen, die in 3'-Richtung der Transkriptionsstartstelle (TSS) verschoben sind. Wir fanden heraus, dass die Introns von Nukleosomen flankiert werden, was darauf hindeutet, dass sie eine Rolle bei der Regulierung der Spleiß-Effizienz spielen.

Die Nukleosomen und die epigenetischen Markierungen (H3K4me3, H3K9ac und H3K27me3) befanden sich entlang der Genstruktur und waren in den nicht-kodierenden Regionen um die Gengrenzen herum weniger angereichert. Sie waren auch direkt proportional zu verschiedenen Genexpressionsgruppen, einschließlich H3K27me3, das in anderen Organismen oft mit stillen Genen assoziiert ist (Bannister and Kouzarides, 2011). Der GC-Gehalt scheint in *cis* die Positionierung von Nukleosomen, epigenetischen Markierungen und des RNA-Polymerase-Enzyms (Pol II) im AT-reichen *Paramecium*-Genom zu beeinflussen. Wir verwendeten einen Ansatz zur Segmentierung des Chromatinzustands (ChromHMM) auf Nukleosomen und Histonmarkierungen, der Gene mit aktiven, repressiven und bivalenten Chromatinzuständen aufzeigte. In multizellulären Organismen sind Gene in bivalenten Domänen, die sowohl H3K4me3- als auch H3K27me3-Markierungen tragen, mit der Regulation der Zelldifferenzierung während der Entwicklung durch das Pausieren der Genexpression verbunden (Voigt, Tee, and Reinberg, 2013; Sen et al., 2016; Blanco et al., 2020). Unter den Genen mit bivalenten Domänen in *Paramecium* fanden wir jedoch keine Anreicherung von pausierenden Genen. Daher ist die Rolle der bivalenten Domänen in *Paramecium* noch unklar.

Statistische Analyse der Regulation der makronukleären Genexpression in *Paramecium*

Wir konstruierten ein genregulatorisches Assoziationsnetzwerk, indem wir die sogenannte Sparse Partial Correlation Network Methode auf verschiedene epigenetische Markierungen, Nukleosomen, sRNA und Genexpressionsdaten anwandten. Wir beobachteten gemeinsame positive Assoziationen von H3K4me3 mit mRNA und H3K9ac mit H3K4me3. Wir beobachteten auch eine negative Assoziation für H3K27me3 mit mRNA, insbesondere für die pausierten Gene. Allerdings zeigten die Gene, die nur H3K27me3-Markierungen trugen, eine positive Assoziation für H3K27me3 und mRNA, was die Rolle von H3K27me3 in *Paramecium* weniger klar macht.

Zusätzlich verwendeten wir Methoden des maschinellen Lernens, um inhärente Genexpressionsmuster zu verstehen. Wir konstruierten einen auf überwachtem Lernen basierenden Random forest Algorithmus, um die Genexpression als hoch oder niedrig zu klassifizieren, indem wir genetische (Genlänge, Intronfrequenz, GC-Gehalt und intergene Länge) und epigenetische Merkmale (Histonmarkierungen, Pol II, MNase und sRNA) konstruierten. Unser Modell hat eine Testgenauigkeit (PR-AUC) von 0,83. Bei der Auswertung verschiedener Gruppen von Merkmalen haben wir festgestellt, dass die genetischen Merkmale genauso viel zu der Vorhersage für die Genexpression beigetragen haben, wie die epigenetischen. Weiterhin haben wir die sogenannte Shapley additive explanations (SHAP) Technik auf unser Random forests Modell angewendet, um allgemeine Genexpressionsmuster abzuleiten. Folglich berichten wir, dass die fünf wichtigsten Merkmale eines hochexprimierten Gens eine hohe H3K4me3-Häufigkeit in seinem Genkörper, eine hohe Intronfrequenz, eine geringe Genlänge, ein hoher sRNA-Gehalt und ein hoher GC-Gehalt sind. Der vermutete Zusammenhang zwischen der Intronfrequenz und der Regulierung der Genexpression, der durch die SHAP-Analyse aufgedeckt wurde, ist bisher noch nie für *Paramecium* berichtet worden.

Schlussfolgerung

In dieser Arbeit haben wir ein neuartiges Werkzeug, RAPID, entwickelt, um eine generische eukaryotische siRNA-Analyse durchzuführen, und es als Conda-Paket und über GitHub (`https://github.com/SchulzLab/RAPID`) öffentlich verfügbar gemacht. RAPID ist jedoch keine vollständige Lösung für die siRNA-Analyse. Einige der möglichen Erweiterungen von RAPID beinhalten die Erfassung von Attributen auf Sequenzebene, wie z. B. die Erstellung von Sequenzlogos von konservierten siRNA-Sequenzen oder die Identifizierung von siRNA-Zielregionen.

Wir haben eine Pipeline entwickelt, um das erste genomweite endo-siRNA-Profil vom vegetativem MAC in *Paramecium* zu charakterisieren und dessen Regulationsmechanismen besser zu verstehen. Wir zeigten die transkriptomischen und endo-siRNA-Aberrationen, die durch die RNAi-induzierende Technik, die Fütterung, ausgelöst werden. Wir berichten zum ersten Mal über die epigenomischen Eigenschaften des MAC von *Paramecium.* Wir haben verschiedene Methoden der Bioinformatik und des maschinellen Lernens eingesetzt, um die regulatorischen Mechanismen des (Epi-)Genoms von *Paramecium* zu charakterisieren.

Während wir die erste genomweite Analyse der sRNAs mit Hilfe von RAPID und anderen Werkzeugen dokumentiert haben, sind ihre Regulationsmechanismen noch unklar. Zukünftige Analysen sind erforderlich, um die Ziele der sRNAs zu identifizieren. Unsere Ergebnisse deuten darauf hin, dass die sRNAs wahrscheinlich nicht direkt auf mRNAs abzielen, sondern in ein komplexes Netzwerk von sRNA-Protein-Interaktionen eingebunden sind, um das Genom zu regulieren. Es ist wichtig zu beachten, dass die *Paramecium*-Zellen in einer Kultur in ihrem vegetativen Entwicklungszyklus nicht synchronisiert sind. Das bedeutet, dass sich verschiedene Zellen in unterschiedlichen epigenomischen/transkriptomischen Zuständen befinden. Dadurch können die Ergebnisse nur einen Durchschnitt der individuellen polyploiden MAC abbilden. Um ein besseres Verständnis der verschiedenen Chromatinzustände zu erhalten, wären Einzelzellmessungen nötig. Allerdings sind die experimentellen Methoden zur Gewinnung von Einzelzelldaten in *Paramecium* derzeit nicht verfügbar.

Wir glauben, dass unser Beitrag das Verständnis der kleinen RNA-omischen und der epigenomischen Regulation der Genexpression in den Makronuklei von *Paramecium* verbessert hat.

Acknowledgements

During the last four and half years of my PhD, I have had the pleasure of interacting with the following people (in alphabetical order of first names). They have all contributed, in different capacities, in my journey towards attaining a PhD. Thank you all for your support!

Abdulrahman Salhab, Andreas Kramer, Angelica Zepeda, Anna Hake, Arno Thibau, Avy Ellis, Azim Dehghani Amirabad, Claudia Herfurth, David John, David Porubsky, Debanjan Mukherjee, Deepak Palaksha, Dorith Kramer, Fabian Kern, Fabian Müller, Florian Neukamm, Francis Cao, Franziska Drews, Geetanjali Vohra, Georg Friedrich, Giammarco Nalin, Goutam Bhatt, Guruprasad Hegde, Heike Neuss, Jana Ebler, Joachim Büch, Jonas Fischer, Julian Wagner, Karl Nordström, Kashyap Popat, Kerstin Sharaka, Kerstina Scherbaum, Lara Schneider, Lena Hoffmann, Linda Sulzmann, Lisa Handl, Lukas Tombor, Madhupriya Murugan, Marcello Pirritano, Markus List, Maryam Ghareghani, Michael Scherer, Mikko Rautiainen, Miriam Cheaib, Mona Linn, Natalia Weis, Neha Agarwal, Nirmal kumar Ramadoss, Nora Speicher, Oren Neumann, Polina Quaranta, Ram Mahalingam, Ramachandra Bhaskara, Ramanuj Ram, Ruth Schneppen-Christmann, Sanjay Srikakulam, Sarvesh Nikumbh, Satish Kumar Verma, Shilpa Garg, Shreeram Ponpathirkootam, Sonali Chaudhury, Srinath Madhwaprasad, Sujana Ram, Suresh Krishna Thiagarajan, Tejaswini Ramachandra, Tim Kehl, Tobias Marschall, Tomas Bastys, Valentina Galata, Vidya Oruganti, and Vikram Narayanan.

I would like to specially thank the various support team members of the the Max Planck Institute for Informatics (MPII, Saarbrücken, Germany), the Cluster of Excellence on Multimodal Computing and Interaction (MMCI, Saarbrücken, Germany), and Goethe University Hospital (Frankfurt am Main, Germany) for their support in ensuring the smooth functioning of the IT systems, and all other administrative tasks. Also, I would like to thank all the other friends and colleagues from the MPII, the Center for Bioinformatics (University of Saarland, Saarbrücken, Germany), and the Institute for Cardiovascular Regeneration (Frankfurt am Main, Germany). I am obliged to extend my gratitude to the members of the public discussion forums, *e.g.* Stack Overflow, for providing inspiring ideas and solutions, whenever I was stuck.

I am indebted to my supervisor, Prof. Dr. Marcel H. Schulz, not only for his patient and relentless guidance, but also for being a role model in numerous aspects. Personally, I will keep admiring and aspiring to reach his level of inquisitiveness, and his style of life-work balance. I would also like to extend my gratitude to Prof. Dr. Martin C. Simon, who has been incredibly patient in bearing my naivety and clarifying my biological questions. The perseverance of Martin is something I aspire to maintain through out my life.

My special thanks to Dennis Hecker, Fatemeh Behjati, Florian Schmidt, Marcel Schulz, and Nina Baumgarten for proof reading my thesis. I would also like to thank the PhD committee members for agreeing to invest their time and effort for evaluating my thesis.

My day-to-day life would not have been a smooth sailing ship with out the following people, who provided me with scientific, technical, mental and moral support: Anna Hake, Charlotte Fürstenberger, Dennis Hecker, Dilip Durai, Fatemeh Behjati, Florian Neukamm, Florian Schmidt, Marcel Schulz, Matthias Döring, Mikhail Zhegalin, Nikoletta Katsaouni, Nina Baumgarten, Peter Ebert, Prabhav Kalaghatgi, Ranjan Kumar Maji, and Veronica Larcher. They will always have a special place in my heart.

I would also like to thank my extended family (specially the nephews and nieces), for ensuring a gala time whenever I had the opportunity to talk with them or visit them in India. Finally, I am not the person who I am today, without the constant support and guidance of my parents. I possibly cannot find the right words to thank them. Hence, this thesis is dedicated to my loving mom, Nalini, and, dad, Karunanithi.

Contents

List of Figures

List of Tables

List of Abbreviations

bp	base pair
kbp	kilo base pair
nt	nucleotide
DNA	Deoxy ribonucleic acid
RNA	Ribonucleic acid
cDNA	complementary DNA
NGS	Next Generation Sequencing
ncRNA	non-coding RNA
sRNA	small RNA
mRNA	messenger RNA
tRNA	transfer RNA
rRNA	ribosomal RNA
UTR	untranslated region
miRNA	micro RNA
siRNA	short interfering RNA
piRNA	piwi-interacting RNA
snoRNA	small nucleolar RNA
taRNA	trans-acting RNA
HM	histone modification/marks
PCR	polymerase chain reaction
USD	United States Dollar
RNAi	RNA interference
WT	Wildtype

ENA	European Nucleotide Archive
SRC	Small RNA Cluster
GSRC	Genes associated with Small RNA Cluster
TPM	Transcripts Per Million
GMO	Genetically Modified Organism
IGV	Integrative Genomics Viewer
MAC	macronucleus
MIC	micronucleus
EEJ	exon-exon junction
EIJ	5'-exon-intron junction
IEJ	3'-intron-exon junction
TGS	Transcriptional Gene Silencing
PTGS	Post Transcriptional Gene Silencing
ssRNA	single stranded RNA
dsRNA	double stranded RNA
RDR	RNA Dependent RNA polymerase
TSS	Transcription Start Site
TSSR	Transcription Start Site Region
SS	start-to-start
SE	start-to-end
TES	transcription end site
CoV	coefficient of variation
MDS	Multi Dimensional Scaling
PCA	Principal Component Analysis
NMO	Non model organisms
MNase	micrococcal nuclease
ChIP	chromatin immunoprecipitation
GO	gene ontology
SPCN	sparse partial correlation network

TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
PR-AUC	area under the precision recall curve
CV	cross validation
HMM	hidden markov model
SHAP	Shapley additive explanations
DE	Differential Expression
CRISPR	clustered regularly interspaced short palindromic repeats
ENCODE	encyclopedia of DNA elements

In dedication to my loving parents,
Nalini and Karunanithi …

Chapter 1

Introduction

The existence of inheritable units (now called genes) was reported by Gregor Mendel in 1865. However, it was not until 1944 these inheritable units were discovered to be Deoxy ribonucleic acid (DNA). This was arguably the defining moment of modern biology and genetics. This discovery led to the characterisation of the double helical DNA strands, and subsequently the central dogma of biology, which outlined how a protein coding gene is formed (McCarty, 2003; Cobb, 2017). Figure 1.1 depicts the steps involved in the central dogma of biology. DNA is transcribed into Ribonucleic acid (RNA) by the process of transcription. RNA further acts as a template to synthesise protein through a process termed translation. Proteins perform most of our bodily functions. Any mishap in these steps can lead to diseases. Hence, researchers try to understand the functioning of all the genes, which can systematically help us cure diseases.

History shows us that the understanding of biology was a crucial step in the development of human health (Sallam, 2010). The discovery of antibiotics is one such instance, which saves millions of deaths. With a motto of improving human health, the successful human genome project was designed. The aim was to characterise all the genes, understand the function of all the genes, and by extension live disease free. The first draft of the human genome sequence was published in 2003 (Collins, Morgan, and Patrinos, 2003). Soon after that, we started unveiling the curtains of the functional role non-(protein) coding DNA through the 2012 encyclopedia of DNA elements (ENCODE) project (Pennisi, 2012). In addition, the past decades brought epigenetics to the forefront of research. Epigenetics is the "study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" (Felsenfeld, 2014).

The continual development of new technologies helps us to unravel new biology and translate them to disease prevention or treatment. For instance, next generation sequencing (NGS) methods were extensively employed in the ENCODE project. We employ similar NGS techniques in clinics today to sequence a patient's genome. Thanks to the lowering costs of NGS, we can sequence hundreds of patients genomes' today for a fraction of the cost of the human genome project (Di Resta et al., 2018). Sequencing a patient's genome is crucial to determine the disease causing changes in DNA (variant/mutation) in coding and non-coding regions of their DNA (Zhang and Lupski, 2015; French and Edwards, 2020). Identifying a patient's disease variants translates to development of personalised therapies (Momozawa and Mizukami, 2021). Such personalised

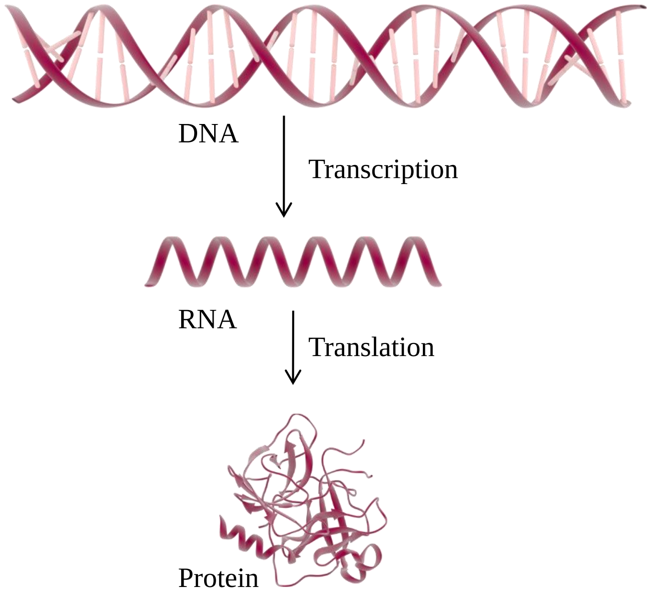


FIGURE 1.1: The central dogma of biology: The hereditary unit, double stranded DNA, through transcription forms a single stranded RNA. In the next step, translation, the RNA molecule gets translated to protein, which performs various bodily function. This figure is created by me with the help of DNA/RNA images obtained in 2016 (*Somersault1824*).

therapies are already available for diseases like cancer (Krzyszczyk et al., 2018) and to treat human immunodeficiency virus (HIV) infection (Lengauer, Pfeifer, and Kaiser, 2014).

The most recent technological addition, in the last decade, which facilitates personalised therapies, is the CRISPR-Cas genome editing tool. The CRISPR-Cas system is capable of making precise cuts on both DNA strands, for example, where a disease causing variant is identified. Following the cut, the DNA repair system takes over, during which the disease causing variant can be modified or new genetic information can be inserted (Charpentier and Doudna, 2013), with up to 38% success. Applications of CRISPR-Cas can range from treating blood disorders, like thalassemia, to hereditary blindness (Ledford, 2020). However, this editing system is not a one-stop solution. A recent study, published in 2019, introduced an improvised method, prime editing, which has enhanced precision than the CRISPR-Cas system. The prime editing technique cuts only one of the DNA strands and does not rely on the cell's repair system. The study claims, prime editing, "could correct up to 89% of known genetic variants associated with human diseases" (Anzalone et al., 2019). I believe, we are slowly navigating towards a future where almost every disease can be cured, with personalised therapies.

1.1 Role of model organisms

The 20^{th} century saw a rise in the use of model organisms to understand biological mechanisms. They were systematically chosen to facilitate experimental work, with properties including small size, fast growth, short developmental (gestation) time, and tractable. The most commonly used model organisms are the bacterium *Escherichia coli*, the nematode worm *Caenorhabditis elegans*, the mustard plant *Arabidopsis thaliana*, the fruit fly *Drosophila melanogaster*, the zebra fish *Danio rerio* and the mouse *Mus musculus* (Fields and Johnston, 2005). Let us have a look at the following two examples on the resourcefulness of model organisms.

The landmark discovery of DNA was, in part, motivated by the infamous Griffith experiment (O'Connor, 2008). Frederick Griffith experimented on the fatality of mice, by injecting them with different strains of pneumonia causing bacteria *Streptococcus pneumoniae* (Griffith, 1928). His experiments proved the existence of a transformative substance in *S. pneumoniae* cells. These results motivated Oswald Avery to isolate the transformative material, DNA (Avery, Macleod, and McCarty, 1944).

The aforementioned CRISPR-Cas system was also a product of a model organism research. In 1987, researchers were investigating an alkaline phosphatase enzyme coding gene, iap, in *E. coli.* A sequence analysis of the iap gene revealed short repeat sequences, which were flanked by short and unique DNA segments, of unknown biological significance (Ishino et al., 1987). These regions of unknown significance are today referred to as clustered regularly interspaced short palindromic repeats (CRISPR) (Charpentier and Doudna, 2013). CRISPR-Cas system's efficiency was also shown using different model organisms like zebra fish (Hwang et al., 2013), *S. pneumoniae*, and *E. coli* (Jiang et al., 2013).

Almost all of what we know about the basic cellular and molecular aspects were found with the help of model organisms (Russell et al., 2017). However, the rise of large consortia projects, like the 100,000 genomes project, churn out gigantic human data sets. Alongside advancements in computational power, and artificial intelligence methods, we may be able to solve the still open fundamental questions of biology with human data. This raised a debate on the resourcefulness of model organisms.

Fields and Johnston claim that in a few decades time, likely, all fundamental biological questions of simple model organisms will be answered. However, they suggest the use of model organisms will persist with changes to their application. They argue that model organisms will still be the experimental play ground to test new mechanistic hypothesis; develop novel technology; test new drugs; and serve as disease models to understand disease pathways (Fields and Johnston, 2005). Not everyone completely agrees with Fields and Johnston. Hunter argues that to test new drugs the relevance of model organisms may be questionable. There have been cases where drugs tested on model organisms, including primates, showed promising results, and yet invoked high allergic responses in humans. Such examples are used to question the relevance of model organisms in drug testing. Hunter adds, the advancements in stem cell research, human tissue culture, and the development of organoids challenge the need for model organisms in drug development. He says, the use of model organisms will eventually be phased out, once the fundamental mechanistic questions are solved (Hunter, 2008).

The same technological advances, which threaten the need for model organisms, have paved the way for an increase in the study of Non model organisms (NMO). NMOs are not so commonly used model organisms. While model organisms were simply a convenience tool, NMOs help us expand our fundamental cellular understandings, and address evolutionary questions. Several species of the unicellular organism, *Volvox*, lives in colonies which acts as our bridge to answer the evolution of multi-cellular life (Herron, 2016). Killifish (*Nothobranchius furzeri*), and naked mole-rat (*Heterocephalus glaber*) are two NMOs, which improve our understanding of ageing and potentially even stop ageing in humans (Kim, Nam, and Valenzano, 2016; Buffenstein and Ruby, 2021).

1.2 Paramecium as a model organism

One of the NMOs is *Paramecium*, a free-living unicellular eukaryote (cells with encapsulated DNA). They gained traction among researchers for several reasons. Their large cell surface ($50 - 300\mu m$) is composed of basal bodies, and cilia (hair like outgrowth), which enables morphogenetic studies. These studies showed the existence of non-mendelian cytoplasmic inheritance of ciliary arrangements (Beisson and Sonneborn, 1965). Other evidence of cytoplasmic inheritance were observed in the different mating types (or "genders") (Sonneborn, 1947), and the expression of surface antigens (Epstein and Forney, 1984). The surface antigen expression shows adaptation to the environment, for instance, temperature (Matsuda and Forney, 2005). *Paramecium* also exhibits several modes of reproduction. They can self fertilise (autogamy) or engage in conjugation with other mating pairs, and can reproduce asexually. Adding more curiosity towards *Paramecium*, is their nuclear dimorphism. They carry two germline micronucleus (MIC), and a somatic macronucleus (MAC). The MIC is diploid and transcriptionally inactive. The MAC is transcriptionally active and exhibits polyploidy (800n). The MAC disintegrates after every sexual cycle. A new MAC develops from the fertilised zygotic MIC, as a result of genomic rearrangements. Small RNA (sRNA), and other non-coding RNA (ncRNA) mediated epigenetic control is known to exercise the genomic rearrangements of a developmental MAC (Beisson et al., 2010). During asexual reproduction (or vegetative growth), a vegetative cell separates into two daughter cells. Over a decade of research on the MAC of vegetative *Paramecium*, showed the role of sRNAs in controlling surface antigen gene expression (Marker et al., 2010). The prime diet of *Paramecium* is bacteria. This allows researchers to feed genetically modified bacteria to them, and study the genetic changes caused. Several studies exploited this method to understanding their RNA interference (RNAi) pathways (Galvani and Sperling, 2002; Carradec et al., 2015; Marker et al., 2014).

1.3 Outline of the thesis

In this thesis, we make several contributions to the communal understanding of gene regulatory mechanisms in the vegetative MAC of *Paramecium tetraurelia*. We created methods and tools to investigate the properties of sRNAs. We took an integrative approach to understand the epigenomic orchestration of gene expression, by employing multitude of bioinformatics and statistical/machine learning methods on multi-omics data.

Chapter 2 introduces the reader to the biological and statistical concepts, and also briefly describe the bioinformatics methods used in this thesis. In Chapter 3, we describe an automated sRNA analysis tool, RAPID, which we developed. We utilise RAPID and other existing bioinformatics methods, to characterise the endogenous sRNA landscape of *P. tetraurelia* and shed light on their biogenesis and function in Chapter 4. The MAC genome annotation of *P. tetraurelia* revealed their high protein coding density of 80%, highest among the free living eukaryotes. This results in short intergenic regions of mere 352 base pairs, raising the question: Where are the regulatory controls of located in *Paramecium*'s MAC? We answer this question, at least in part, in Chapter 5. In the next chapter 6, we apply network construction and statistical learning methods on the data from the chapters 4 and 5 to understand the general patterns of gene expression in the vegetative *Paramecium*'s MAC. The final Chapter 7 summarises our findings. We also delineate the individual contributions of all our collaborators involved in the projects at the end of Chapters 3 to 6.

Chapter 2

Background

This chapter will briefly introduce the reader to the biological concepts and computational methods used in this thesis. It also provides a brief overview of the basic biology of *Paramecium tetraurelia.*

2.1 Eukaryotic gene expression

Organelles are defined structures in a cell which performs several functions. Some examples include mitochondria - the power house of a cell - and nucleus. All organisms, which possess membrane bound organelles are called eukaryotes. An eukaryotic nucleus enveloped by a nuclear membrane contains the hereditary genetic material, DNA.

DNA is made of two helical strands composed of four nucleotides (or bases) adenine (A), guanine (G), thymine (T), and cytosine (C). These two strands are called sense and antisense strand running in the direction of $5'$ (read as 5 prime) to the $3'$, and *vice versa*, respectively. The sense and antisense strands are often referred to as forward and reverse or plus and minus strands, respectively. The two strands are complementary to each other and are held together due to the chemical pairing rules of these nucleotides. Nucleotide A always pairs with T, and nucleotide G pairs with C (Watson et al., 2003).

We introduced the central dogma of biology in Figure 1.1 showing that a DNA gets transcribed to RNA, through transcription, and RNA gets translated to protein. This process is also called gene expression (Watson et al., 2003).

2.1.1 Transcription

RNA is a single stranded molecule made of the same nucleotides as of DNA, except for thymine (T) which is replaced by an uracil (U). Figure 2.1 illustrates the process of transcription. Transcription always occurs in the direction of $5' - 3'$, and is mediated by an enzyme called RNA polymerase. Hence, the antisense DNA strand (whose orientation is in $3' - 5'$) acts as the template strand for transcription. RNA polymerase enzyme binds to DNA sequence, upstream of a gene, called promoter in order to start the transcription and create the RNA transcript.

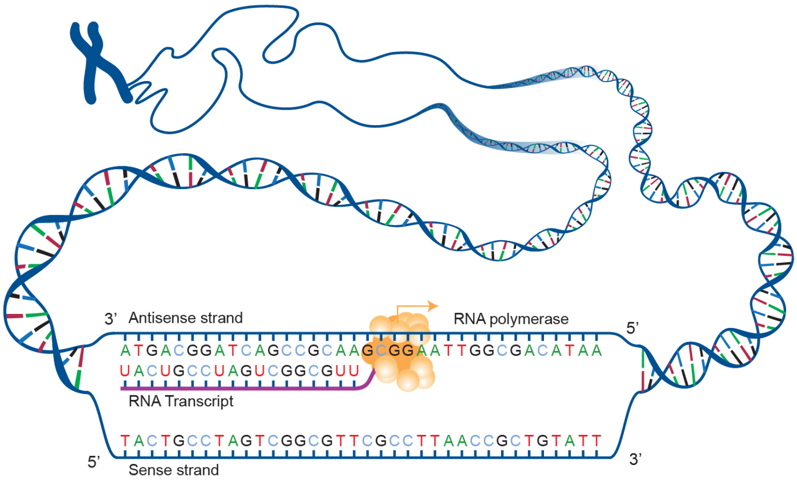


FIGURE 2.1: Transcription: DNA unwinds such that the antisense strand $(3' - 5')$ is accessible to the RNA polymerase enzyme to form a single stranded RNA transcript. This figure is a courtesy of the National Human Genome Research Institute.

2.1.2 Splicing

The DNA of a gene consists of few structural elements namely the exons, introns, and the untranslated regions (UTRs), as shown in Figure 2.2. The UTRs in the left and right end of a gene are called the $5'$ and the $3'$ UTR respectively. Exons are the only parts of DNA which eventually codes for proteins. All other structures introns, UTRs, and regions in between genes (intergenic DNA) are referred as non-coding DNA.

In the nucleus, the newly formed RNA transcript contains all these structural elements as well. Introns are excised from this RNA transcript through splicing. In many organisms, a combination of different exons can also get excised along with introns to create different proteins from the same gene. This process is called alternative splicing. Following splicing, the spliced transcript undergoes some post-transcriptional modifications, like $5'$ capping and poly-A tailing, to form a messenger RNA (mRNA) (Figure 2.2). The mRNA further gets transported out of the nucleus to get translated as protein. The UTRs, and the post transcriptional modifications are crucial for the stability of the mRNA, which in their absence would get degraded quickly (Mignone et al., 2002).

2.1.3 Translation

During translation, an mRNA molecule is processed by ribosomes to produce proteins. The ribosomes are made of two major sub units: a small 40S unit and a big 60S unit. The 40S unit traverses through the mRNA, and reads the mRNA as words of three nucleotides (three-letter code or codon). The 60S unit identifies the transfer RNAs (tRNAs), carrying the aminoacids respective

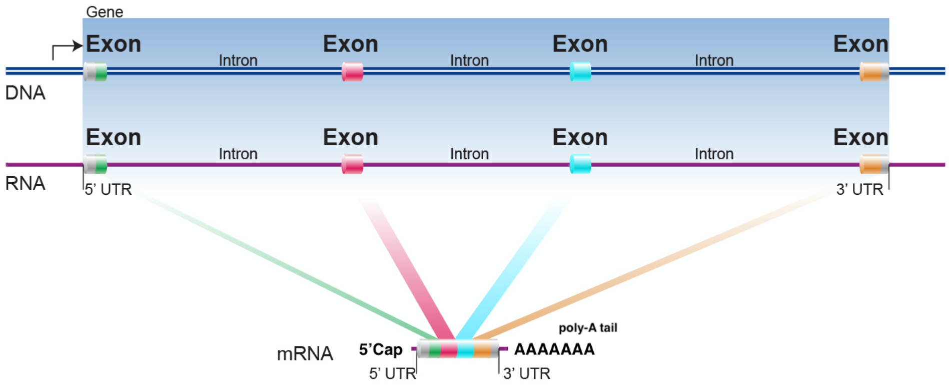


FIGURE 2.2: Gene structure: The different structural elements of a gene is illustrated. The RNA transcript contains all the structural elements found in a DNA. Through splicing the intron regions are eliminated to form an mRNA. This figure is a courtesy of the National Human Genome Research Institute.

to the codon, and synthesises a polypeptide chain of amino acids. Once the polypeptide chain is synthesised, they fold into a minimal energy state called proteins (Watson et al., 2003).

There are 64 possible codons from the four nucleotides, and there are only 20 amino acids. A mapping of these codons to the respective aminoacids is called the genetic code of the organism. Apart from the 20 amino acids, the codons also contain the start and stop signals to indicate the ribosomes on where to initiate and terminate the translation process. The genetic code of humans is shown in Figure 2.3.

2.2 Non-coding DNA

By definition, the non-coding DNA does not code for proteins. However, they are crucial in regulating the gene expression. For instance, in many organisms, introns regulate alternative splicing and enhance gene expression (Jo and Choi, 2015). Promoters, enhancers, and silencers are few other regulatory elements harbored from non-coding DNA. Promoters are non-coding DNA immediately upstream of a gene, where the RNA polymerase binds to initiate the RNA transcription (Maston, Evans, and Green, 2006). Enhancers and silencers are often found several 1000 base pairs (bps) up or downstream of a gene. Proteins can bind to these enhancers (or silencers) to activate (or repress) the transcription of one or more genes (Pennisi, 2012). Not all non-coding DNA act as binding sites, many of them undergo transcription to produce ncRNAs. The ncRNAs can be broadly classified as long- and short- ncRNAs.

The short ncRNAs are less than 200 nucleotide (nt) in length, which perform a wide range of functions. For instance, the tRNAs and ribosomal RNAs (rRNAs) help in translating the mRNA to a protein. Other short ncRNAs like

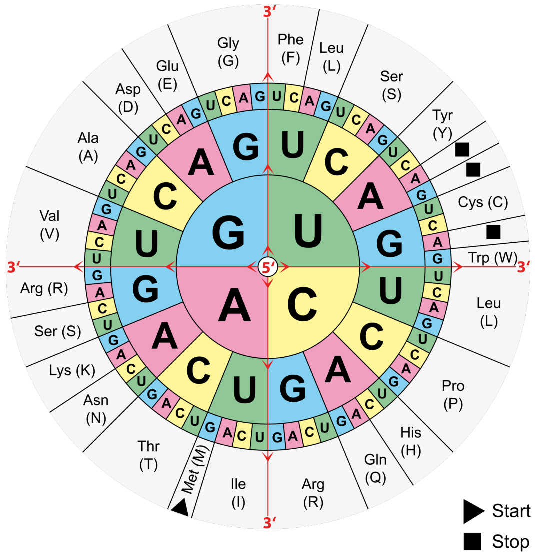


FIGURE 2.3: Genetic code of humans represented in a circular fashion. The codons should be read in the direction of the arrows (inside to outside), to know their respective aminoacid. This figure is available for public use from Wikipedia (Wikimedia, 2021).

micro RNA (miRNA) and short interfering RNA (siRNA) can block protein production, by interfering in the translation process. More information on that will follow in Section 2.4.

Long ncRNAs are more than 200 nt in length. They are known to be involved in DNA repair, when the DNA is damaged due to extensive exposure to ultra violet rays, heat shock, *etc* (Ratti et al., 2020). They are also known to control gene expression, indirectly, by yielding short ncRNAs (Wilusz, Sunwoo, and Spector, 2009). However, the common mode for long ncRNAs to regulate gene expression is by partnering up with chromatin remodelling enzymes (Moran, Perera, and Khalil, 2012; Han and Chang, 2015). But, what is a chromatin?

2.3 Organisation of DNA

The eukaryotic DNA is efficiently organised in to chromosomes such that they can be packed in to a nucleus of few microns in size. The chromosomes are made of smaller sub-units called chromatin fibers. A chromatin fiber consists of DNA packaged around proteins called histones. Figure 2.4 shows a schematic view of the organisation of DNA.

The histone proteins are composed of a core unit and a flexible tail with few amino acids. There are different kinds of histone proteins: H1, H2A, H2B, H3 and H4. A complex of eight histone proteins, two H2A, two H2B, two H3 and two H4, is called a histone octamer. When a 146 bp long DNA molecule wraps around this histone octamer, it is termed a nucleosome. Multiple nucleosomes are connected like "beads on a string" by linker DNA of length up to 80 bp.

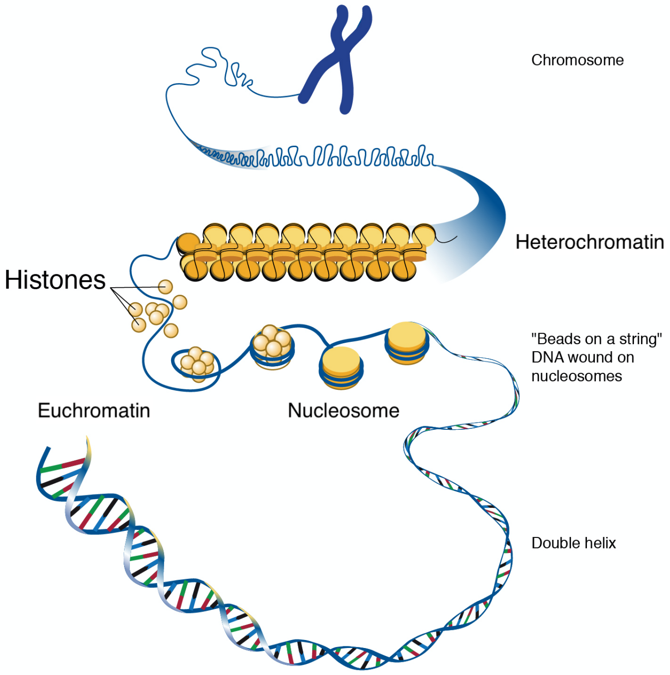


FIGURE 2.4: An illustration of the organisation of DNA. DNA is packaged with histone protein in different stages of compaction: euchromatin, heterchromatin and chromosome. This figure is a courtesy of the National Human Genome Research Institute.

The H1 protein is quintessential for further compaction of DNA, as it further coils the nucleosomes into chromatin fibers.

There are two types of chromatin fibers: heterochromatin and euchromatin. Heterochromatin is a highly condensed form of chromatin where the nucleosomes prevent the accessibility of DNA, and hence blocking transcription. On the contrary, euchromatin is loosely packed DNA, which enables the latter to be accessible by DNA binding proteins, like RNA polymerase.

2.3.1 Chromatin modifications

Cells can efficiently transition between euchromatic and heterochromatic states in order to control gene expression (Figure 2.5). This regulation is achieved through chromatin modifications or DNA methylation. The methylation of some cytosine bases in the DNA are known to be associated with heterochromatin formation, mediated by the enzymes called DNA methyl transferases. These modifications are inheritable through several mechanisms like genomic imprinting, and are called epigenetic modifications, as they are not directly changing the genetic sequence.

The chromatin modifications which occur on the histone proteins are called histone modification/marks (HM), which include modifications like methylation, acetylation and phosphorylation. The nomenclature for describing a HM is to state the histone protein whose tail is modified, followed by the one letter

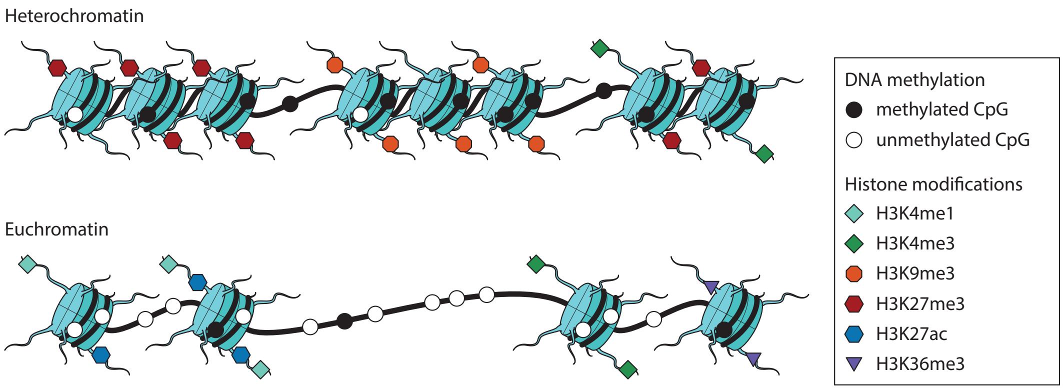


FIGURE 2.5: An overview of the epigenomic modifications, which can influence the eu- and hetero- chromatin formation, controlling the gene expression. The green cylinders represent histones, wrapped with DNA (like a black thread). The different modifications are shown on the legend. This figure is a courtesy of Fabian Müller (Müller, 2017).

code of the amino acid and then the modification type. For instance, H3K27me3 refers to the trimethylation of the aminoacid lysine occurring at the 27th position on the tail of histone protein H3.

Histone methyl transferases, histone acetyl transferases are some of the enzymes, which are responsible for methylation and acetylation, respectively. On the other hand, histone deacetylases and demythlases are responsible for removing the methylation and acetylation, respectively. These enzymes interact with several other proteins, and ncRNAs to regulate the chromatin modifications.

By regulating the transition of chromatin states, HMs are indirectly regulating gene expression. Some of the well studied HMs in several organisms, which we also discuss in this thesis, are H3K4me3, H3K27me3 and H3K9ac. The commonly known association of the HMs and gene expression are as follows: H3K4me3 and H3K9ac are associated with active gene expression; H3K27me3 is associated with not expressed (or silent) genes (Bannister and Kouzarides, 2011).

2.4 RNA interference

RNA interference (RNAi) is another widely known mechanism to regulate gene expression either at transcriptional or post-transcriptional level, using small RNA molecules in eukaryotes (Figure 2.6). In order to silence the gene expression, when the RNAi process triggers DNA or chromatin modifications, it is termed Transcriptional Gene Silencing (TGS). When the RNAi process degrades an mRNA or inhibits the translation of an mRNA to protein, through a cascade of biomolecular processes, it is called Post Transcriptional Gene Silencing (PTGS) (Moazed, 2009; Zhang, 2009).

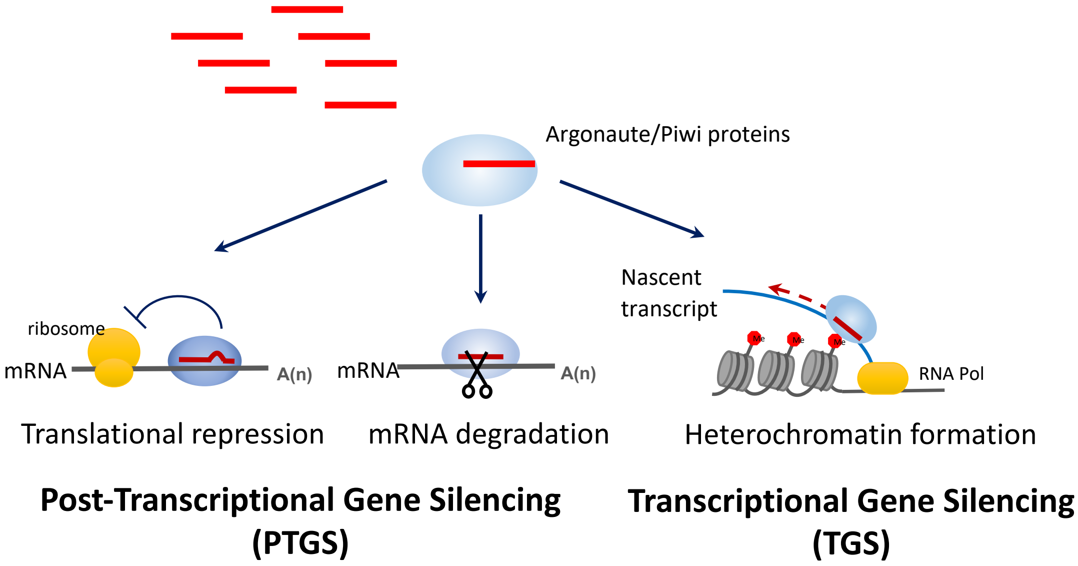


FIGURE 2.6: A representation of the transcriptional, and post-transcriptional gene silencing mechanisms. This figure is a courtesy of *Martin Simon*.

2.4.1 Small RNAs

Small RNAs are usually $18 - 30$ nt in length. There are several members of the small RNA family including micro RNA (miRNA), short interfering RNA (siRNA), piwi-interacting RNA (piRNA), small nucleolar RNA (snoRNA), and trans-acting RNA (taRNA). The miRNAs and siRNAs are the two well studied members of the sRNA family in the context of RNAi (Carthew and Sontheimer, 2009). There are diverse pathways of biogenesis for miRNAs and siRNAs, in different organisms (Meister and Tuschl, 2004). Regardless of their biogenesis and functional diversity, the starting point is always a double stranded RNA (dsRNA).

2.4.2 Short interfering RNAs

A dsRNA can be produced by different mechanisms: (i) a single stranded RNA (ssRNA) can get converted to a dsRNA by the enzymatic activity of RNA Dependent RNA polymerase (RDR) (Figure 2.7A) or (ii) bidirectional transcription (Figure 2.7B). This dsRNA can be processed by Dicer (DCR) enzyme into siRNAs, called primary siRNAs (1^o). The primary siRNAs can load onto proteins (called argonaute) and degrade a target mRNA. There are different ways to degrade a target mRNA. For example, in plants, the primary siRNAs cleaves a targeted mRNA. The cleaved mRNA is then converted in to dsRNA by RDR, which is further diced by Dicer to produce secondary siRNAs (2^o). These secondary siRNAs can further target another mRNA, similar to primary siRNAs. If the Dicer cleaving occurs at regular intervals in a dsRNA, it is called phasing and the resulting siRNAs are called phased-siRNAs. On the contrary, *C. elegans* follows a dicer independent mechanism, using RDR to

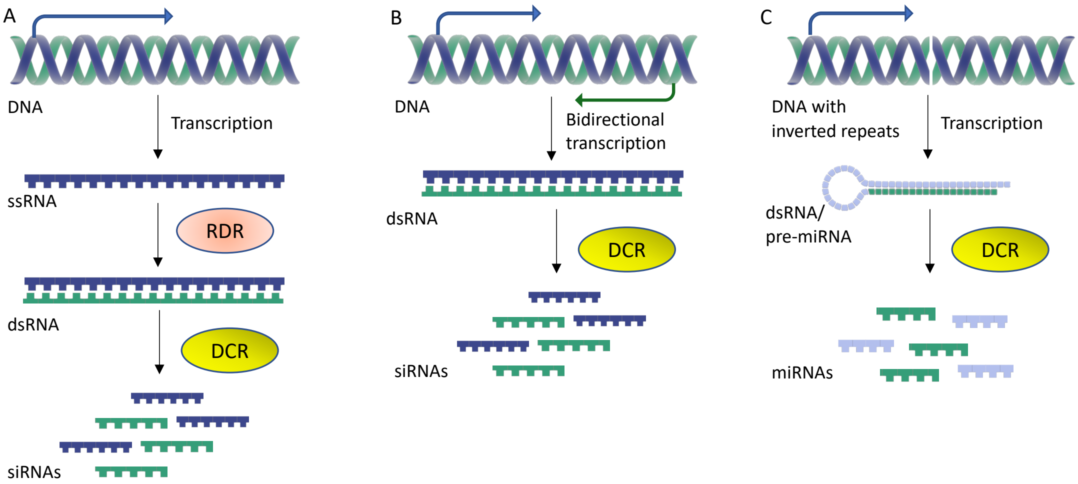


FIGURE 2.7: A representational view of the siRNA (A,B) and miRNA (C) biogenesis. A) A ssRNA transcribed from a gene, becomes dsRNA by RDR activity which further gets processed by an enzyme Dicer (DCR) to produce siRNAs. B) A dsRNA is produced from bidirectional transcription, further processed by DCR as siRNAs. C) Transcription of inverted repeats resulting in a dsRNA, processed by DCR producing miRNAs. This figure is created by me with the help of DNA/RNA images obtained in 2016 (*Somersault1824*).

directly produce 2^o siRNAs from the 1^o siRNA targeted mRNA (Allen et al., 2005; Baulcombe, 2007). An interference mechanism is described as acting in *cis*, if the same gene, which produced the initial dsRNA is silenced; otherwise, it is called *trans*-acting.

2.4.3 Micro RNAs

The dsRNA source for a miRNA is derived often from a single stranded RNA produced from a near complementary 20-50 bp inverted repeats, which folds in to a dsRNA, forming a hairpin loop (Figure 2.7C). This dsRNA with hairpin loop is called pre-miRNA, which gets further processed by Dicer (or Drosha) (Li and Patel, 2016) to form miRNAs. The miRNAs are on average 22 nt in length, and are involved in several regulatory pathways. As the functional role of miRNA is a broad topic, and as this thesis does not focus on miRNA, but siRNAs, I merely point the reader to a recent review of miRNAs (Gebert and MacRae, 2019).

2.5 Sequencing methods

The most commonly used method in the recent decades to measure gene expression and epigenetic modifications is sequencing.

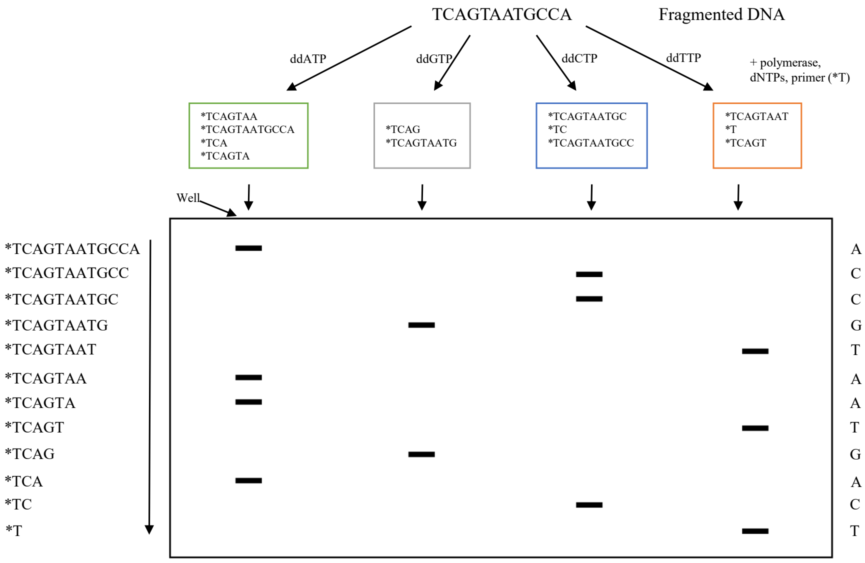


Figure 2.8: A schematic illustration of Sanger sequencing is shown. The fragmented DNA to be sequenced is subjected to four different mixtures. Each mixture performs multiple cycles of PCR with the available enzymes, nucleotides and four different di-deoxy forms of nucleotides. Further, they are subjected to electrophoresis, where the synthesised fragments are size separated. On the left, the synthesised fragments are shown. On the right, the sequenced read is shown. This figure is adapted from ATDBio, 2016.

2.5.1 DNA sequencing

The process of unraveling the order of nucleotide arrangement in a DNA is DNA sequencing. One of the oldest such technique is Sanger sequencing, which is shown in Figure 2.8. The basic idea is sequencing by synthesis, *i.e.* to synthesise the complementary strand of the DNA sequence of interest.

First, the DNA sequence of interest is extracted through chemical methods, and is fragmented into multiple pieces. Subsequently, these fragments are added in to four different mixtures. All mixtures contain the DNA polymerase enzyme, nucleotides, and a primer. In addition, to each mixture only one of the di-deoxy forms of the four different nucleotides is added. This mixture is sufficient to start a polymerase chain reaction (PCR).

In a PCR, the primer gets hybridised to the DNA fragment, upon which the DNA polymerase starts synthesising the complementary strand of fragmented DNA. However, if a di-deoxy nucleotide gets incorporated the PCR stops. The PCR is carried out for multiple cycles, and each cycle stops at different positions of the complementary strand synthesis.

In the next step, the synthesised DNA strands are subjected to gel electropheresis. Gel electrophoresis is a simple technique where DNA, upon application of mild electric current, gets separated according to their molecular weight (in our case, decreasing length of the synthesised DNA). Figure 2.8 shows a typical electrophoresis setup. The four wells in the gel are loaded with the synthesised

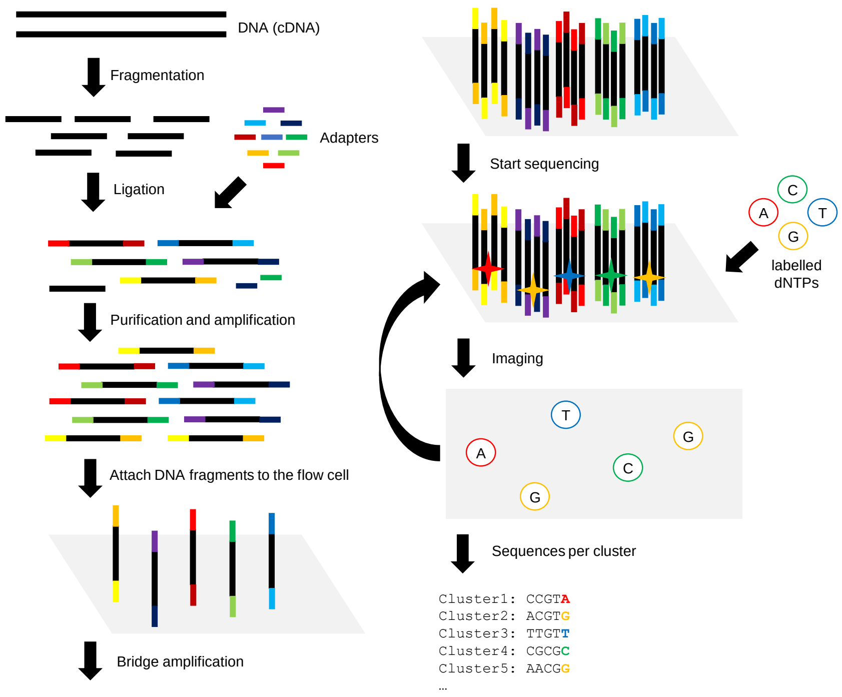


FIGURE 2.9: A schematic representation of Illumina's NGS methodology. Following purification of the cDNA fragments ligated with adapters, the library is loaded on the flow cell. In the flow cell, bridge amplification is performed to enable formation of cluster of single stranded DNA reads. Subsequently, sequencing by synthesis is performed in massively parallel manner producing the sequence reads. This figure is a courtesy of *Florian Schmidt*

DNA strands from the four different mixtures carrying the different di-deoxy nucleotides. The wells are named after the respective di-deoxy nucleotides. After the electrophoresis is complete, the DNA sequence of interest is obtained by reading the name of the corresponding well wherever a fragment is observed in the gel. These read outs are referred as reads.

In order to sequence the human genome, a similar approach called shotgun sequencing was employed. It took approximately 13 years to complete the human genome project with an approximate cost of 300 million United States Dollar (USD). With the technological advancements we have today, a human genome can be sequenced in three days at an approximate cost of 1000 USD. These advancements are broadly categorised as Next Generation Sequencing (NGS). I will only briefly mention how one of the NGS techniques work, in the Section 2.5.2. For a detailed review on the latest NGS methods, I would like to direct the reader to these review articles (Heather and Chain, 2016; Shendure et al., 2017).

2.5.2 RNA-seq

The quantity of RNA found in a cell (or sample) corresponds to the respective gene's expression. The most common technique to measure RNA is RNA-seq. If the entire RNA content of the organism is sequenced, it is termed transcriptome.

After biochemically isolating the RNA, different types of RNA, like small RNAs and mRNAs, can be isolated with specific protocols. For instance, to isolate mRNAs researchers often employ a poly-A tail enrichment or rRNA depletion (Stark, Grzelak, and Hadfield, 2019).

A generic RNA-seq involves two major steps (i) library preparation and (ii) sequencing by sequencers provided by different companies like Illumina, Roche, etc. I'll discuss an Illumina based workflow below, which is also depicted in the Figure 2.9. First, RNA is converted into a complementary DNA (cDNA), which is subsequently fragmented into small pieces. Adapters are added to both $5'$ and $3'$ ends of the fragmented cDNA. Next, a PCR is performed to amplify the number of copies of these cDNA. The library preparation is complete, after a clean up of the free floating adapters or a size selection of fragments from the PCR enriched fragments.

In order to perform sequencing, the library is loaded on to a flow cell. The flow cell is like a glass slide whose surface is bound with complementary oligonucleotides of 15-30 bp in length, to which the adapters in the library gets attached. Further, the bound DNA fragments are amplified by a bridge amplification technique, which produces clusters of identical single stranded DNA sequences. Following amplification, Illumina performs sequencing by synthesis technique. One of the prime differences to the approach, described earlier in Section 2.5.1, is the use of fluorescent labelled nucleotides as reversible terminators instead of the di-deoxy nucleotides. This enables the use of image processing techniques to infer the added nucleotides in each sequencing cycle. Depending on the experimental conditions, we will end up with millions of fragmented reads from the sequencer.

2.5.3 MNase-seq

For a gene to be expressed, their corresponding nucleosomes should be accessible for transcription. Leveraging this principle, researchers created the micrococcal nuclease (MNase) sequencing method, which can help us study the properties of DNA associated with nucleosomes.

A schematic overview of this method is shown in Figure 2.10. The histone proteins and the DNA surrounding it are first cross-linked using specific chemicals. Next, they are digested by the micrococcal nuclease enzyme, which degrades all DNA that are not linked to proteins. Following the MNase digestion, reverse cross-linking retrieves only the DNA, which were bound to the histone proteins (Cusick et al., 1981). Subsequently, these DNA reads can be prepared as a library and are subjected to a sequencing technique, similar to what was described in Section 2.5.2. In order to control for experimental bias, a control MNase-seq is performed with a very mild quantity of MNase. The reads from the control are often subtracted from the MNase-seq experiment, or, a ratio of reads from the MNase-seq over the control is used in downstream analyses.

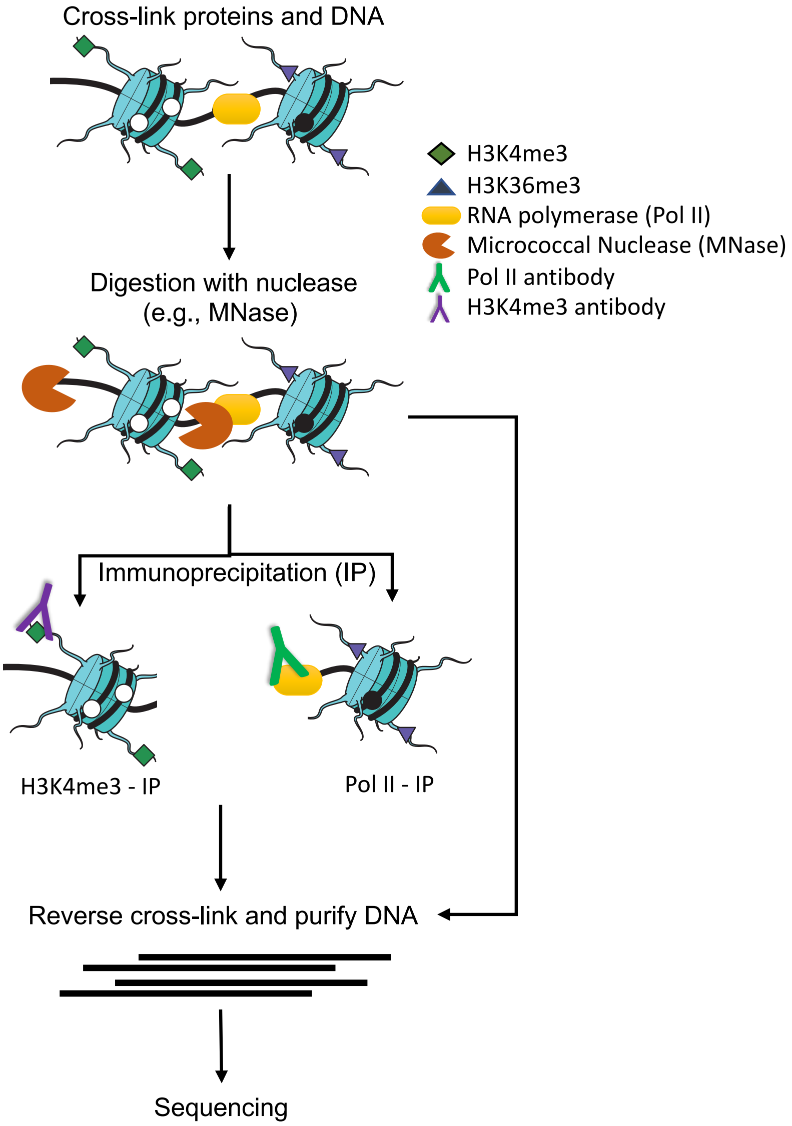


FIGURE 2.10: An overview of the steps involved in MNase-seq and ChIP-seq is shown. Both approaches start with cross-linking DNA with proteins bound to DNA followed by digesting the unbound DNA. For MNase-seq, sequencing is carried out on the purified DNA obtained from reverse cross-linked protein-bound DNA. For ChIP-seq, an additional immunoprecipitation step is carried out to precipitate the protein/histone mark of interest with the corresponding antibody, before proceeding to sequencing. The green cylinders represent histones, wrapped with DNA (like a black thread). The different modifications, the RNA polymerase enzyme Pol II, and the respective antibodies are shown on the legend. This figure is created by me with the nucleosomes obtained from Fabian Müller (Müller, 2017).

2.5.4 ChIP-seq

MNase-seq helps us sequence all the DNA, which are bound with histone proteins. However, there are multiple epigenetic modifications, which controls the gene expression (see Section 2.3.1).

Chromatin immunoprecipitation (ChIP)-seq is widely used to study epigenetic modifications, and other proteins which bind with DNA, like transcription factors. ChIP-seq is simply sequencing of DNA, which are precipitated with antibodies specific to our protein of interest. Antibodies are Y-shaped proteins, which can bind with other specific proteins.

Figure 2.10 shows a schematic overview of ChIP-seq. Similar to MNase-seq, the proteins are first cross-linked with DNA. Following which the unbound DNA (free or any protein binding) can be digested with nucleases (e.g., MNase). The remaining protein bound DNA is subjected to antibodies, which target the protein of our interest. For instance, we can introduce antibodies designed to precipitate only DNA bound with H3K4me3 modifications. This step is called immunoprecipitation (IP). Following IP, reverse-crosslinking will yield only the DNA, which can subsequently be sequenced.

RNA polymerase is a crucial enzyme, which transcribes the DNA to RNA. There are several types of RNA polymerases. RNA polymerase II (or Pol II) is the enzyme, which is involved in transcribing protein coding genes. If needed, they can pause the transcription process, thereby effectively regulating gene expression. In order to study such genes, which are regulated using the pausing mechanism, the genomic loci bound with Pol II enzyme needs to be known. To that end, we can use ChIP-seq with a Pol II specific antibody (Figure 2.10).

While performing a ChIP-seq experiment, a control is preferably performed along side. The control is often an antibody not specific to our protein of interest. Instead of a control antibody, researchers also often use input DNA as control. The input DNA is the DNA isolated from cells before performing any kind of enzyme digestion or antibody selection. The reads from the control experiments are often subtracted or only a ratio of reads from the ChIP-seq experiment over the control is used in downstream analyses.

2.6 NGS data processing tools

Once the sequencing reads are obtained in the infamous FastQ format, there are a few standard steps as shown in the Figure 2.11, which is by no means a *one-size fits all* pipeline. Nevertheless, I will briefly describe these steps and the respective publicly available tools we used as part of this thesis.

2.6.1 Preprocessing

For all the sequencing data sets used in this thesis, we performed the same preprocessing steps mentioned in Figure 2.11. First, a quality check is performed using the fastqc (Andrews et al., 2015) tool, where one can check the basic statistics like average read length, their sequence quality distribution, presence of adapters, GC content, etc. Subsequently, the adapters shall be removed, and

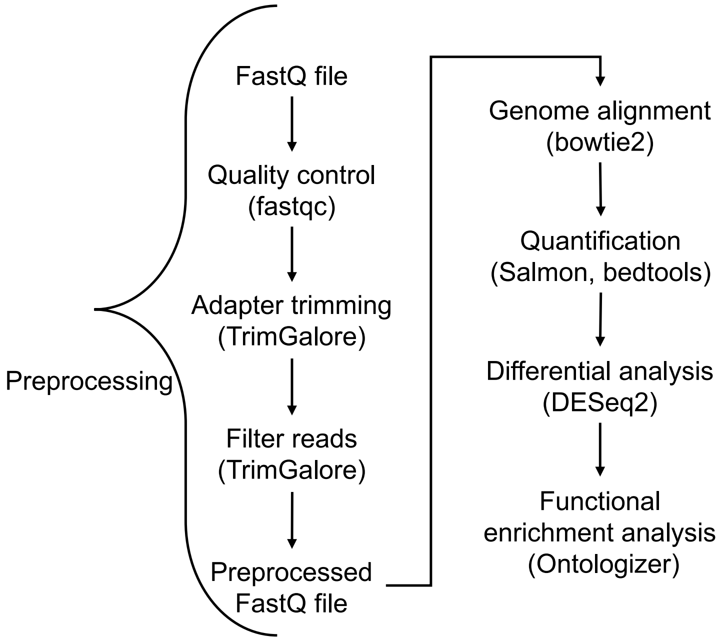


FIGURE 2.11: An example NGS data processing workflow is shown.

if needed a read size based filtering can be performed using TrimGalore (Martin, 2011).

2.6.2 Alignment

Following quality control, the sequenced reads are often aligned to the genome of the respective organism, using alignment tools like bowtie2 (Langmead and Salzberg, 2012). However, for RNA-seq data sets from organisms with splicing, special splice-aware aligners, like STAR (`https://github.com/alexdobin/STAR`), should be used.

2.6.3 Quantification

Subsequent to alignment, one needs to quantify the number of reads aligning to different genomic loci of interest. One of the most commonly used tools for this purpose is bedtools (Quinlan and Hall, 2010). The authors of bedtools befittingly describe it as following: "Collectively, the bedtools utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks".

There are other specific purpose driven tools, like Salmon (Patro et al., 2017) for RNA quantification. The aforementioned alignment step can be skipped, using Salmon, and the RNA reads mapping to each gene/transcript are quantified in a time-efficient manner.

Quantification metrics

There are two commonly used quantification metrics in this thesis (i) read counts and (ii) Transcripts Per Million (TPM). Read counts are simply the number of aligned reads in a particular genomic loci. However, read counts is not an appropriate measure if we want to compare different samples/loci, as they can have differences caused due to sequencing depth and the length of the genomic loci. The TPM metric attempts to normalise for these differences. Let us say a genomic loci, i, has R_i number of aligned reads and l_i is the length of the genomic loci (in kilo base pair (kbp)). T is the total number of reads sequenced. TPM of i can be calculated as

$$TPM_i = \frac{R_i}{l_i} * \frac{10^6}{RPK}, \tag{2.1}$$

where reads per kilobase (RPK) is given by,

$$RPK = \sum_{i \in I} \frac{R_i}{l_i}, \tag{2.2}$$

with I being the set all the quantified genomic loci. As the sum of TPM values will be identical between all samples, it makes comparison between samples more appropriate than read counts.

2.6.4 Differential analysis

After quantification of reads, another common analysis is to compare them between multiple samples or cell types or different experimental conditions. For instance, which genes expression levels are significantly different in heart and lung. To this end, a Differential Expression (DE) analysis is performed, using tools like DESeq2 (Love, Huber, and Anders, 2014), to identify genes which are differentially expressed with statistical significance.

2.6.5 Functional enrichment analysis

Once we identify some differentially expressed genes, the next logical question, to investigate, is the function of the differentially expressed genes. To this end, researchers use functional enrichment analysis.

Gene ontology (GO) is a hierarchical set of terms, which annotate biological process, cellular component and molecular function of all known genes. A functional enrichment analysis aims to identify statistically overrepresented GO terms associated with a given list of genes. While there are numerous tools out there to perform GO analysis, we resorted to using Ontologizer (Bauer et al., 2008).

2.7 Small RNA analysis tools

There are a myriad of sRNA analysis tools available, which can be broadly categorised as (i) prediction tools and (ii) analysis tools. These tools differ a lot in their portfolio and user-friendliness.

2.7.1 Prediction tools

As mentioned earlier there are diverse types of sRNA like miRNA, siRNA, piRNA, etc. The prediction tools use the NGS data sets to predict novel small RNA loci. Such methods often employ: sequence comparisons with known sRNA databases; or computational approaches to calculate the sequence and structural properties of sRNAs in order to predict novel sRNAs. Some examples of such tools include miRDeep2 (Friedländer et al., 2012), and ShortStack (Johnson et al., 2016).

miRDeep2, as the name suggests, is predominantly aimed at accurately predicting miRNAs and has been shown to have high accuracy in several organisms. It predicts known and novel miRNAs from any sRNA data, using a multi-tiered algorithm, which probes for hairpin loops (a signature of miRNAs), miRNA secondary structure and stability. They are also capable of distinguishing pre- and mature miRNAs. While there are several competing tools, the high accuracy of miRDeep2 makes it a valuable resource to be part of several analysis tools like Oasis (Capece et al., 2015).

ShortStack discovers *de novo* sRNA clusters (genomic loci) in any given sRNA data set. ShortStack identifies islands of sRNA reads, which are contiguous and have a user-defined number of aligned sRNA reads. Another user-defined value, padding, controls when two contiguous islands should be merged. These crucial parameters enable the user to account for known sRNA characteristics of an organism. The padded islands of sRNA are defined as the *de novo* sRNA clusters. There are more steps in refining these sRNA clusters, which are not discussed here, but can be read in their publication (Johnson et al., 2016). ShortStack is one of the few tools, which does not focus primarily on miRNA or piRNAs, but provides generic sRNA characterisation along side. We employed ShortStack on our data sets to identify novel sRNA producing loci, which will be discussed in the Chapter 4.

2.7.2 Analysis tool

The analysis tools are meant to aid researchers in analysing properties of sRNA from the NGS data. By definition, such tools aggregate statistical properties of sRNAs from the NGS data and provide graphical illustrations. These analysis tools are often an automated or semi-automated workflow of different computational tools, similar to Figure 2.11, but optimised to handle different types of sRNA. Some examples of such tools include sRNAtoolbox (Rueda et al., 2015), and piPipes (Han et al., 2015).

sRNAtoolbox comprises of 8 different tools to profile the expression of miRNAs; to visualise aggregate miRNA statistics; to identify differential miRNA expression; to visualise the miRNA alignments in a genome browser; to perform functional enrichment analysis; and to predict miRNA targets. In essence, it has a suite of downstream analysis workflows to analyse miRNAs of interest.

piPipes is another analysis pipeline aimed at extracting useful information related to only piRNAs, as they possess considerable differences to miRNAs. For instance, the ping-pong signature, where 10 nucleotides in the $5'-$end of $23 - 36$ nt piRNAs show sequence complementarity in the sense and antisense reads, is seen only in the piRNAs. Apart from identifying these signature, piPipes also provides an illustrative summary of the results. The diversity of such analysis tools, and their diverse abilities are summarised in Chapter 3.

2.8 Chromatin analysis tools

Thanks to the research communities around the world, we have plentiful tools for chromatin analysis. In this thesis, we predominantly used (i) Sparse partial correlation network (SPCN) (Lasserre, Chung, and Vingron, 2013), (ii) DANPOS (Chen et al., 2013), (iii) deepTools (Ramírez et al., 2016), and (iv) ChromHMM (Ernst and Kellis, 2017). I'll briefly describe below their functionalities, in the context of this thesis.

2.8.1 Sparse partial correlation network analysis

Researchers often resort to Pearson correlation analysis in order to understand the relationship between different types of data. For instance, how does H3K4me3 correlate with mRNA expression, or how does H3K4me3 and H3K27me3 correlate. Pearson correlation coefficient (r) gives the linear relationship between two variables. For a data set of n samples, with paired x and y values, $\{(x_1, y_1), ..., (x_n, y_n)\}$, we can calculate the Pearson correlation coefficient using the Equation 2.3. The coefficient takes the range of $[-1, 1]$. A positive correlation would suggest that x and y are directly proportional, and a negative correlation would suggest they are inversely proportional.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.3}$$

While correlations can show the relationship between two variables, they do not reveal causality and are susceptible to confounding factors. For instance, consider the observation that an increase in ice cream sales on the beach correlates with shark attacks. This correlation clearly cannot be a causation. The confounding factor here is, a third variable, warm temperature. When the temperature is warm, people go to swim in the sea, hence the increase in ice cream sales on beach and shark attacks. The variables ice cream sales and warm temperature is said to be collinear. Of course, there are other hidden (often unmeasured) confounding factors for the sharks to swim near the shore, like ocean currents, breeding season of prey fish, etc. If those variables are also measured, they are likely to be multicollinear and can confound the causal analysis of shark attacks. Similar to the ice cream sales example, biological data can also be multicollinear. This challenges the causal interpretation of correlation between two observed variables. However, we can resort to partial correlation.

Partial correlation removes the effect of an observed confounding variable from the two variables, whose partial correlation is being calculated. We use the sparse partial correlation network (SPCN) method (Lasserre, Chung, and Vingron, 2013), which is improvised up on the partial correlation to construct an undirected network of associations. The SPCN method is a statistically powerful technique, when there are large number of samples (*e.g.* genes), and few variables (*e.g.* epigenetic marks). In the SPCN network, an edge is drawn between two nodes (variables), only if a significant partial correlation coefficient is observed between them. The first step involved in SPCN construction is to calculate a partial correlation matrix, P, as shown below (Algorithm 1).

```
Data: Data matrix $X$, with $D$ variables ($X = X_1, ..., X_D$), and $n$
      samples
Result: Partial correlation matrix, P where
        $P_{ij} = Cor(X_i, X_j | X \backslash \{X_i, X_j\})$; and $Cor$ is the correlation
for $d$ = 1, ..., $D$ do
    Rank $d^{th}$ column of data matrix X
    Replace data in the $d^{th}$ row, with the corresponding rank
end
Calculate a covariance matrix for the rank transformed $X$
Apply matrix inversion ($\bigwedge$) on the covariance matrix
The row and column normalised partial correlation matrix, P is
  obtained by $P_{ij} = -\frac{\bigwedge_{ij}}{\sqrt{\bigwedge_{ii}\bigwedge_{jj}}}$
```

Algorithm 1: Pseudocode to construct a partial correlation matrix (Lasserre, Chung, and Vingron, 2013).

In the next step, a cross validation approach is applied on the data set, by first splitting them into 10 subsets. Keeping aside one subset (t), rest of the data is used to construct, M, a partial correlation matrix (Algorithm 1). In M, the partial correlations are ranked according to their multiple hypothesis testing corrected, q-values. The partial correlations in M are modified as zero, if their respective $q-$values are above a set threshold. Hence, sparseness is introduced to the partial correlation matrix, M. Further, using the subsets (except t), we aim to predict a variable, say X_i, using a linear function of the other variables who have non-zero entries in M. We calculate the predictive error of the linear function using the subset, t. We optimise the $q-$value threshold, such that the predictive error is minimal. These steps are repeated 10 times, with each time using a different subset as t.

Further, the individual entries of the actual partial correlation matrix, P, is set to zero, if the respective entries in at least three of the ten M sparse matrices are zero. A graphical version of the resulting P is the sparse partial correlation network, which has edges only between the nodes whose partial correlation values are non-zero.

Although the SPCN network reveals statistically significant associations (or edges), it is limited to the variables analysed. However, biology is usually more complex, and many unmeasured variables are likely to be involved in an association network. If new variables are measured and included in the analysis, the SPCN associations will be susceptible to change.

2.8.2 DANPOS

DANPOS is primarily a nucleosome (or MNase-seq) analysis software, which accounts for factors like nucleosome position shifts while calculating a nucleosome occupancy peak. A peak is defined as a genomic locus, where there is an enrichment of the nucleosome reads in relation to the control experiment. Although, primarily aimed at nucleosome analysis, DANPOS also contains programs tailored to identify peaks from histone marks or other protein based ChIP-seq. In addition, DANPOS provides a profiling tool to analyse the distribution of chromatin features in genomic loci of interest.

2.8.3 deepTools

deepTools is an user-friendly tool suite, primarily designed for ChIP- and MNase-seq data analysis. They feature a range of tools to perform quality control, normalise multiple data sets, and comparative visualisation of genomic features across multiple data sets. For instance, normalisation can be performed against the control experiments using the *bamCompare*. In addition, multiple data sets can be normalised against each other using different scaling or normalisation methods using *multiBamSummary*. Further, as a quality control measure, experimental replicates can be checked for how well they correlate with each other using *plotCorrelation*. The two comparative visualisation features, *plotHeatmap* and *plotProfile*, can be used to compare how reads are distributed across different genomic features like exons, introns, UTRs, etc.

2.8.4 ChromHMM

Researchers often investigate several histone marks at a time. While tools, like DANPOS, can provide insights into individual histone mark distribution across genomic loci, a combinatorial approach is not feasible with such tools. Although, the SPCN approach explains an undirected network epigenetic marks, they cannot directly provide genomic feature level granularity of the networks.

ChromHMM is a probabilistic model, based on multivariate hidden markov model (HMM). It infers a user-defined number of chromatin states, for any given set of chromatin mark data. First, the genome is binned into small windows (default 200 bp), and the aligned reads in these windows are binarised (present/absent) based on a Poisson background distribution. Following which, the bins are categorised into different states, where a state represents the availability of the combination of one or more chromatin associated data. This process is termed chromatin state segmentation.

ChromHMM relies on the user to set the number of hidden states K to be identified. The emission distribution of a hidden state, is the probability distribution of the combination of one or more chromatin associated data. ChromHMM models the emission distribution with a product of independent Bernoulli random variables (Ernst and Kellis, 2017). The mathematical notion can be described as the likelihood of the observed data to be in the hidden states. Formally,

- k is a hidden state among the number of possible hidden states ($k = 1, ..., K$).
- m is the chromatin associated data (*e.g.* H3K27me3) among all the available chromatin associate data ($m = 1, ..., M$).
- $p_{k,m}$ is the emission parameter showing the probability of the chromatin associated data m to be present in a hidden state, k.
- c is a chromosome of all available chromosomes ($c = 1, ..., C$), and
- c_t is the t^{th} 200 bp interval (or user-defined) on chromosome c, where $t = 1, ..., T_c$ and T_c is all the non-overlapping 200 bp intervals on chromosome c.
- $v_{c_t,m}$ denotes the availability (1 or 0) of the chromatin associated data (m) at an interval c_t.
- At an interval c_t, a combination of different chromatin associated data is denoted as $v_{c_t} = (v_{c_t,1}, ..., v_{c_t,m})$.
- The transition probability from state i to j, where $i, j \in (1, ..., K)$, is denoted as b_{ij}.
- c_1 is the first 200 bp interval on a chromosome. The probability of c_1 being in the state i, is denoted as a_i, where $i \in (1, ..., K)$.
- s_c is the hidden sequence of states in chromosome c, such that $s_c \in S_C$, where S_C is the set of all possible of sequence of states.
- s_{c_t} is the state on chromosome c at a 200 bp interval t for the sequence of states s_c.

$$P(v|a,b,p) = \prod_{c\in C}\sum_{s_c\in S_C} a_{s_{c_1}} \left(\prod_{t=2}^{T_C} b_{s_{c_{t-1}},s_{c_t}}\right)\prod_{t=1}^{T_C}\prod_{m=1}^{M} p_{s_{c_t,m}}^{v_{c_t,m}}(1-p_{s_{c_t},m})^{(1-v_{c_t,m})} \tag{2.4}$$

For the parameters (a, b, and p) defined above, the likelihood of the observed data v is defined as the Equation 2.4. ChromHMM uses the Baum-Welch algorithm to perform the expectation-maximisation while training the HMM parameters.

In addition to defining the hidden states, ChromHMM also facilitates an enrichment analysis of the identified chromatin states in different genomic features of interest. For example, a chromatin state with a combination of H3K4me3 and H3K36me3 is identified by ChromHMM. This can be called an active state, as both of those histone marks are associated with active gene expression. ChromHMM can also report us whether such an active state is enriched in exonic or intronic regions.

While ChromHMM can identify different states, the biological interpretation have to be handled by the researchers. Similarly, the researchers have to experiment with the different number of states for ChromHMM, as there is no rule of thumb.

2.9 Statistical learning methods

Statistical learning is a set of computational/statistical methods used to understand the relationship of different variables in a data set. Broadly, they are classified as supervised and unsupervised methods. Supervised methods aims to predict an output from a set of input variables. To achieve this the supervised methods use a data set of inputs with known outputs. For instance, given the age and maximum heart rate of a patient, can we predict/classify whether the patient has a risk of heart disease or not. Unsupervised methods does not use data sets with any known outputs, but can still reveal relationships or structures in data. For instance, given a set of patients gene expression values from heart, can we identify groups of genes/patients with similar expression patterns. We predominantly use only supervised classification methods in this thesis. Hence, the following sections will brief about some of those methods. For a detailed description of these methods, I would like to refer the reader to the Introduction to Statistical Learning book (James et al., 2013).

2.9.1 Supervised classification

In a supervised classification problem, there are a set of quantitative/qualitative variables called predictors (or features) and their respective categories, a qualitative response variable called class. The process of creating a statistical model to predict the class of the predictors is training a classifier.

Formally, a data set is represented by a matrix $X \in \mathbb{R}^{n \times p}$ and vector $Y \in \mathbb{R}^{n}$. Every row in this data set is a sample. Each sample is a (X_{ij}, Y_i) pair, $i \in (1, ..., n)$ and $j \in (1, ..., p)$, where n represents the number of samples and p represents the number of predictors. The relationship between X and Y, is represented as

$$Y = f(X) + \epsilon. \tag{2.5}$$

In equation 2.5, f is a function of X and ϵ is a random error term independent of X. But, $f(X)$ is unknown, *i.e.* we do not have the actual function that derives Y from a given X. In reality, the task of finding Y is difficult. Hence, Y has to be predicted, with the function $f(X)$. As we do not know f, we estimate

Predicted / Actual	**Yes**	**No**
Yes	True Positive (TP)	False Positive (FP)
No	False Negative (FN)	True Negative (TN)

TABLE 2.1: An illustrative confusion matrix for a hypothetical model classifying whether a gene is expressed or not. Yes - Positive class; No - Negative class

$\hat{f}$, with the available samples (X, Y) to predict $\hat{Y}$ as

$$\hat{Y} = \hat{f}(X). \tag{2.6}$$

This process of estimating the function f, with available samples (X, Y) is called training. The aim is to estimate a function, $\hat{f}$, such that it fits the data as close as possible to the actual f (James et al., 2013). The methods, used in this thesis to estimate $\hat{f}$, are discussed in the subsequent sections.

Model assessment metrics for classification

There are several metrics to evaluate the performance of a classifier. Let us say, hypothetically, we have a classifier, which is trained on some epigenetic marks (predictors or features) to classify whether a gene is expressed or not (class). If the gene is (not) expressed, we call it positive (negative) class. As we have the actual class information for the training data, a common way to evaluate the classifier is to create a confusion matrix by comparing the predictions against the actual class (Table 2.1). True Positive (TP) and True Negative (TN) are the samples, which are correctly predicted as positive and negative class, respectively. The False Positive (FP) and False Negative (FN) are the samples, which are erroneously classified as positive and negative class, respectively.

With the confusion matrix as basis, several other derived measures are often used. These measures include accuracy (equation 2.7), precision (equation 2.8), and recall (equation 2.9). Accuracy reports the ratio of the number of correctly classified predictions among all samples. Precision reveals how many of the predicted positive class is correctly classified. Whereas the recall denotes how many of the samples with actual positive class is correctly classified. All of these measures are in the value range of $[0, 1]$, with 1 being the best.

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

$$pre = \frac{TP}{TP + FP} \tag{2.8}$$

$$rec = \frac{TP}{TP + FN} \tag{2.9}$$

Another commonly used performance metric is the area under the precision recall curve (PR-AUC). The precision and recall values are plotted as a curve, and the area under this curve is the PR-AUC. An area of 1 indicates a perfect classifier with all samples being correctly classified. This measure is especially useful in case of imbalanced data sets (Davis and Goadrich, 2006).

Cross validation

The model assessment metrics are useful measures. However, if such measures are applied on all the training data, we will not know how the model performs on a new sample data (never seen before by the model). In other words, we would need to ensure the model is simply not memorising the training data and achieving higher performance. Hence, resampling methods like cross validation (CV) is used. The idea is to first split the available data randomly into $k-$folds. Next, using the data from $k-1$ folds the model is trained. This model's performance is calculated, say using precision, only for the data from the k^{th} fold. This process is repeated k-times, with each time setting aside a different fold of data for testing. The final performance of the model is reported as an average of the performance metric from the k iterations.

2.9.2 Decision trees

The basic idea behind decision trees is to split the predictor space into multiple regions, such that the regions identified have the majority of the points from a single class. An illustration of a decision tree is shown in Figure 2.12.

Let us consider a toy example. We want to predict the risk of heart disease in a patient based on two predictors. The two predictors are age (X_1), and maximum heart rate (X_2). We divide the predictor space (X_1 vs X_2) of our training data, into J distinct non-overlapping regions ($R_1, R_2, ..., R_J$). An optimal value of t is identified along the predictor axes X_1 or X_2, in order to identify the non-overlapping regions, such that many points in the region belong to the same class. This can be represented as a tree, as shown in Figure 2.12B. The optimal t values act as the decision nodes, and the regions are the leaf nodes. In order to make a prediction, say for a new patient, the decision tree is traversed from the root node till reaching a leaf node. The new sample is assigned the same class as the majority of observations in the leaf node. In order to find the non-overlapping regions with majority of points from the same class, often the Gini index (Equation 2.10) is used. Gini index calculates the total variance across all classes as,

$$G = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk}), \tag{2.10}$$

where K is the total number of classes. $\hat{p}_{jk}$ is the proportion of training data points belonging to k^{th} class in the j^{th} region. Gini index takes a range of $[0, 1]$. A Gini index of 0 represents a pure node with all samples belonging to the same class.

The advantage of decision trees is their ease of interpretation, which mimics human decision making. However, the disadvantage is the high variance. When a decision tree yields different results, for different random subsets of data, then the tree is said to have high variance.

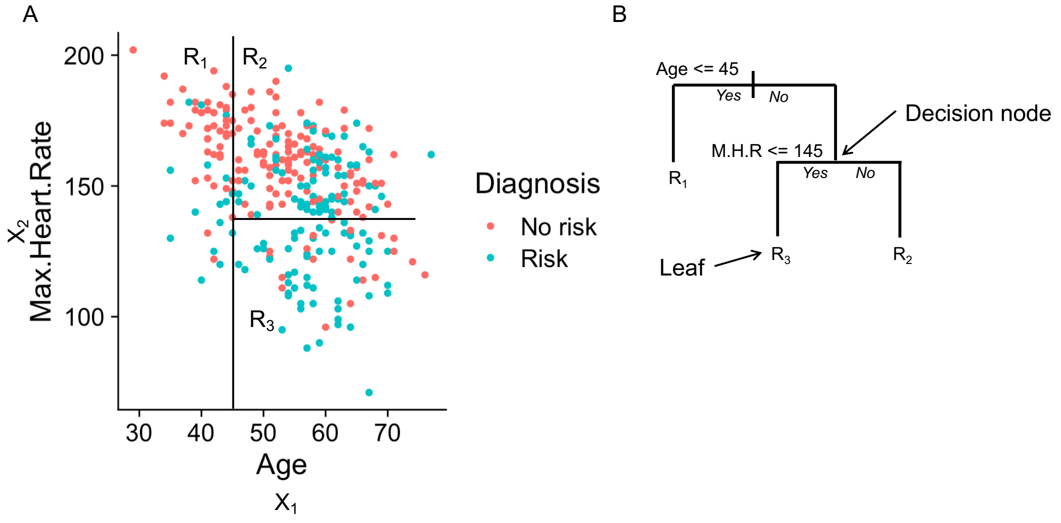


FIGURE 2.12: (a) The predictor space stratification is shown here. X_1 and X_2 are the predictors. R_i are the non-overlapping regions splitting the predictor space, where $i \in 1, 2, 3$. (b) A tree representation of the stratification is shown. This figure is adapted from the book - An Introduction to Statistical Learning (James et al., 2013)).

2.9.3 Bagging

Bagging aims to reduce variance in statistical methods, by averaging over a set of values. The application of bagging on decision trees is illustrated in the Figure 2.13. First, the available training data is bootsrapped, *i.e.* randomly sampled with replacement to create multiple (say B) subsets. For each of these B subsets, a decision tree is built. If we want to predict the class of a new sample, B predictions are obtained from B decision trees. The final prediction is the majority vote from all the individual predictions made by the B trees.

While bagging reduces the variance in decision trees, it suffers from another problem. The B trees created are highly correlated, as the same set of predictors are considered while constructing each tree. If there is a strong predictor among all the available predictors, the B trees will have likely used this predictor in all the trees. This needs to be avoided, as it may cause a model to fail in real time, for example, if the strong predictor value is not available.

2.9.4 Random Forests

The Random Forests model (Figure 2.14) also employs the bagging approach, but aims to decorrelate the trees. Here, while constructing each decision node of a decision tree from a bagged data set, only a random subset m of the available p predictors are considered. The rest of the steps are same as bagging. The recommended choice for the number of predictors in the subset, $m \approx \sqrt{p}$.

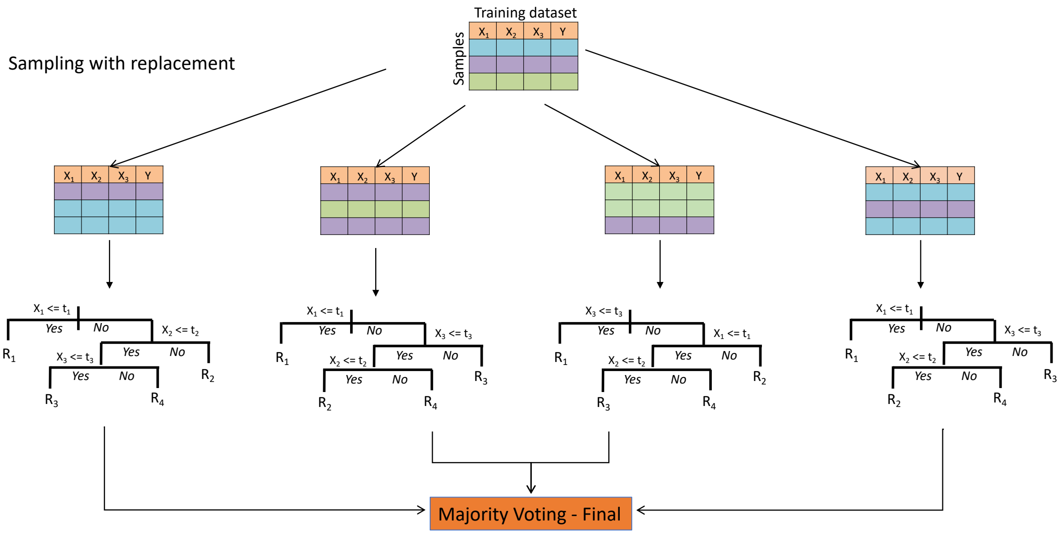


FIGURE 2.13: The bagging methodology is depicted here. The squares in the first lane are the actual training data set. Different colored rows show individual training samples and each column represents a predictor. The second lane shows the bootstrapped subsets (sampling with replacements). The third lane shows the corresponding decision trees for each bagged data set. The last row depicts that, to make a final prediction for a new sample, majority votes of individual predictions are considered.

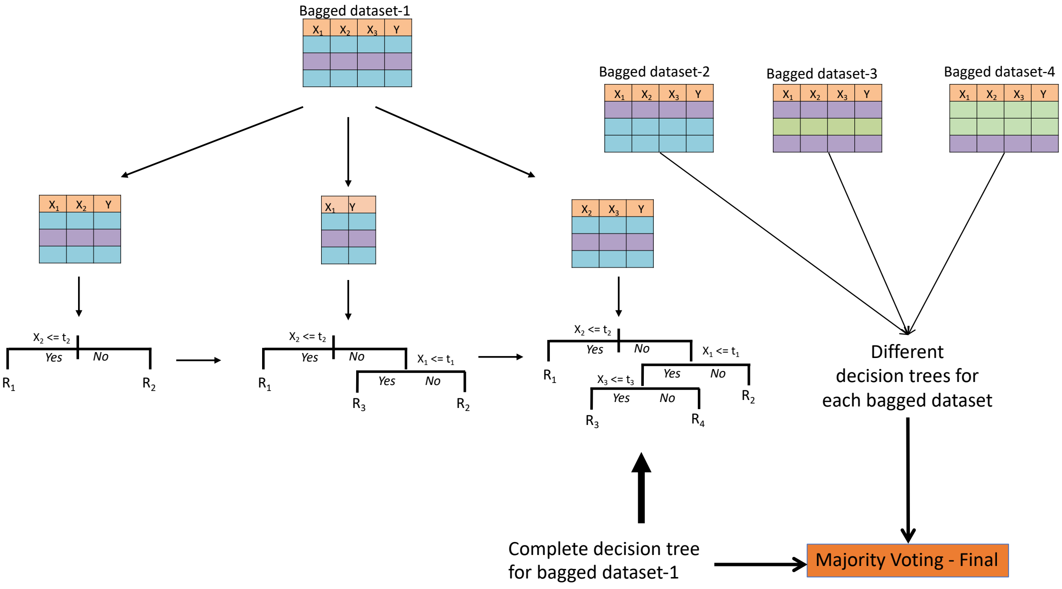


FIGURE 2.14: The Random Forests classifier construction is shown here. We start with a bagged data set (similar to Figure 2.13. While constructing the decision tree, before making each decision node, we consider only a random subset of predictors. The final prediction of a new sample is done using majority voting of individual predictions.

2.9.5 Artificial neural networks

Neural networks are a class of statistical learning algorithms heavily inspired from the biology of the brain. The brain is made up of millions of neurons, which are the learning units. We learn from experimenting with the real world. For instance, "I should not touch a hot plate" is a decision, which a kid learns after accidentally touching the hot plate few times.

Figure 2.15 shows a simple artificial neural network consisting of three layers, namely, input layer, hidden layer and output layer. The input layer has nodes, which are essentially the predictors, and the hidden layer consists of the neurons (or perceptrons). The output layer is, in principle, the class predictions. A neuron, connected with the input layer, is a mathematical operation.

Each input is randomly initialised with a weight. Following which, the neurons calculate the sum of the linear combination of the input values (predictors) and the randomly initialised weights for each input node. Further, the neurons apply an activation function and transmits the value to the output layer. The activation functions are usually non linear, in order to smooth the output. This output is, in essence, the probability of the input sample being classified to a particular class. The model learns to improve its performance with enough training data, by experimenting with the weights of the input nodes. This type of simple artificial neural networks are also known as feed forward networks (Goodfellow, Bengio, and Courville, 2016). When the number of neurons and the hidden layers used are manifold, such models are termed deep neural networks.

2.9.6 Model interpretation

A statistical model is said to be interpretable, if humans can understand the reasoning behind the decision (prediction) of a model. For example, decision trees are interpretable, as it mimics human decision making. The need for interpreting the statistical models have risen tremendously, with the rise of the so called black-box models like deep neural networks. Such sophisticated black-box models have high performance, but loose their interpretability. Model interpretability is essential, as it can help us ensure the models are fair, robust, and trustworthy (Doshi-Velez and Kim, 2017).

There are two types of model interpretation: (i) when a model explains how often a particular feature is used while making a prediction, termed global interpretation and (ii) when a model explains one sample (or data instance) at a time, termed local interpretation. Both global and local interpretations are useful. However, the importance of the local interpretations are in the rise with statistical learning applications being increasingly used in medical applications. For instance, the statistical model should clarify the doctor, based on which information (feature), it diagnosed (or classified) the patient to have a heart disease or cancer. Without such granular information, the doctor would not be able to prescribe the right course of treatment. Please note that, I only briefly discuss the interpretable method used in this thesis. Interpretability is an active area of research, and I would like to refer the reader to the Interpretable Machine Learning book (Molnar, 2019).

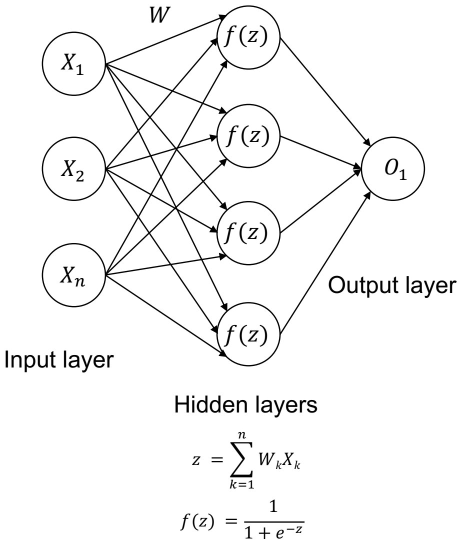


FIGURE 2.15: A simple artificial neural network is illustrated. The input layer contains the predictors. Each input is associated with a weight, and connected to a neuron in the hidden layer. The neurons perform the mentioned mathematical operation and applies a non-linear activation function (*e.g.* $f(z)$). The output layer returns the prediction values.

Shapley additive explanations (SHAP) values

Shapley additive explanations (SHAP) is an interpretability method, which aims to provide local interpretations. SHAP works based on Shapley values, which were originally developed, in game theory research, by Lloyd Shapley (Shapley, 1953). In a game played by a group of people, when the game is complete, the Shapley values help us to fairly distribute the reward according to the contributions of all the players. The contributions of the players are also termed coalitions. In other words, the average marginal contribution of a player among all possible coalition of players is called Shapley value.

In the setting of a statistical model, the prediction task is the game and each predictor (or feature) is a player of the game. The reward is the actual prediction of a particular sample (or data instance). Using Shapley values, we can calculate what is the average contribution of each player towards the reward, among all possible coalition between players. First, for a data instance, we calculate its prediction probability to be in any class with a combination of subset of features. Then, we remove one random feature from this subset and calculate the prediction probability again. The difference in prediction probability, when the random feature is removed, is the marginal contribution (or importance) of the removed feature to the overall prediction probability. We repeat these steps for all possible combinations of feature subsets. The average of all the marginal contributions to all possible feature subset combinations of a feature is the Shapley value of that feature. A pseudocode to estimate Shapley value for a single feature is shown in Algorithm 2.

```
Data: Data matrix X, data instance of interest x, number of iterations
      M, machine learning model f, and feature of interest j
Result: Shapley value for the feature j of data instance x
for $m$ = 1, ..., $M$ do
    draw random instance $z$ from data matrix X
    choose a random permutation, $o$, of the feature values
    Order instance x: $x_o = (x_1, ..., x_j, ...x_p)$
    Order instance z: $z_o = (z_1, ..., z_j, ...z_p)$
    Construct new instance with feature j:
      $x_{+j} = (x_1, ..., x_{j-1}, x_j, z_{j+1}, ...z_p)$
    Construct new instance with out feature j:
      $x_{-j} = (x_1, ..., x_{j-1}, z_j, z_{j+1}, ...z_p)$
    Compute marginal contribution: $\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$
end
Compute Shapley value as the average: $\phi_j(X) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$
```

Algorithm 2: Pseudocode to calculate Shapley values for a single feature. This pseudocode is adapted from the Interpretable Machine Learning book (Molnar, 2019).

With increasing number of features, Shapley estimation can be extremely time consuming. Hence, model specific approaches have been introduced. For instance, TreeExplainer is a time efficient Shapley value estimator designed to address tree based models like Random Forests (Lundberg et al., 2020). We make use of this estimator in Chapter 6.

2.10 *Paramecium tetraurelia*

Paramecium tetraurelia is a non-model organism. As mentioned in the introduction, this thesis is about understanding the regulatory landscape of the transcriptome of *Paramecium.* Hence, this section is dedicated to more details about the fascinating biology of *Paramecium.*

2.10.1 Cell biology

Figure 2.16 shows the structural components of the free-living unicellular eukaryote, *Paramecium.* Paramecium is one of the first micro organisms to be observed in a microscope in the late 17th century (Van Houten, 2019). These $50 - 300\mu m$ cells are covered with hair like structures, called cilia. The cilia helps the *Paramecium* to swim, and sweep their food closer to the oral groove.

The oral groove directs the food to the mouth in order to ingest other micro organisms, like amoeba, bacteria, *etc.* The ingestion happens through phagocytosis, where the outer cellular membrane engulfs the food and forms a food vacuole. As the food vacuole moves towards the anal pore to get egested, the enzymes from the cytoplasm help digest the food. In addition, they also have

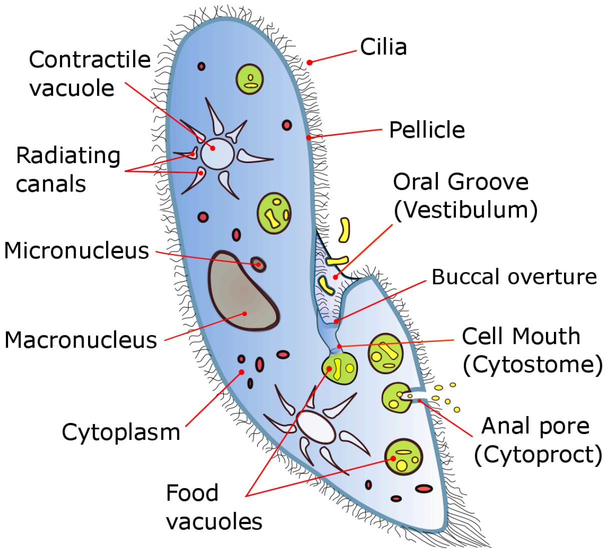


FIGURE 2.16: An illustration of the cellular structures of *Paramecium*. This figure is available for public use from Wikipedia (Wikimedia, 2017).

special structures called the radiating canals, to collect excess water and cytoplasmic waste materials. The radiating canals empty the excretory products into the contractile vacuole, which helps the osmoregulation (water and ionic content of the cell) (Beale and Preer, 2008).

Paramecium exhibits nuclear dimorphism with two germline micronuclei (MIC), and a somatic macronuclei (MAC). The MIC is diploid (2n) and transcriptionally inactive, while the MAC is transcriptionally active and exhibits polyploidy (800n). Ploidy refers to the number of paired chromosomes. While the ploidy information is estimated for *Paramecium*, the exact number of chromosomes remains unknown (Aury et al., 2006).

2.10.2 Life cycle

Paramecium exhibits several modes of reproduction. They can self fertilise (autogamy) or engage in conjugation with other mating pairs, and can reproduce asexually.

Asexual reproduction

Figure 2.17 shows the asexual (vegetative) mode of reproduction. Asexual reproduction is the process when an "adult" *Paramecium* splits itself into two daughter cells. It is also termed vegetative growth. During the early stage of the vegetative cycle, the cells start elongating, which will eventually break through cytokinesis to form two daughter cells. In the mean time, the two germline MIC duplicates itself by a process called mitosis, meaning they are genetically identical copies. The MAC gets elongated amitotically, and gets split between the two daughter cells (Van Houten, 2019). The mechanisms behind

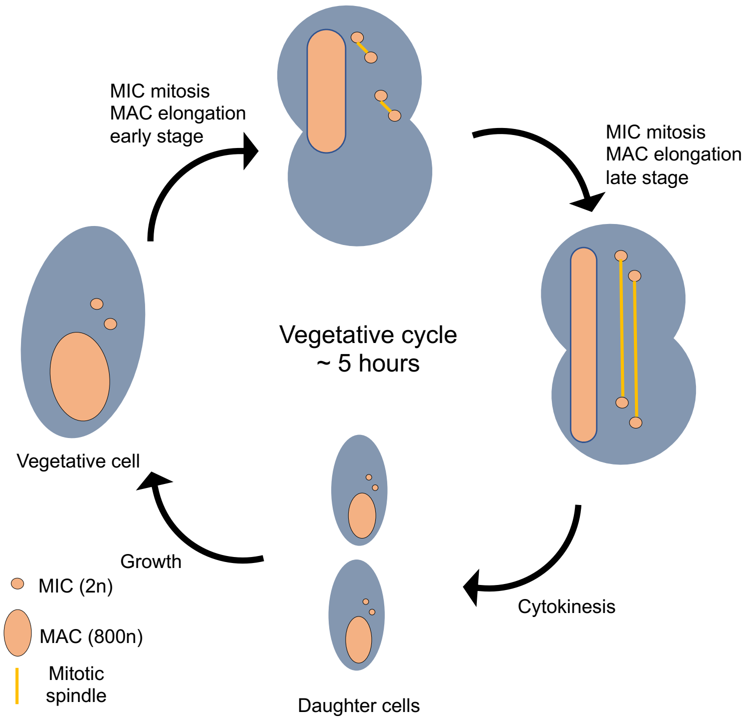


FIGURE 2.17: The asexual (or vegetative) reproduction cycle in *Paramecium* is shown. This figure is adapted from (Van Houten, 2019).

how the vegetative MAC maintains the 800n polyploidy, albeit dividing amitotically, is not well understood. Please note that "adult" is used loosely, as a vegetative cycle lasts only about 5 hours and by the end of it, the *Paramecium* cells are already fit to enter the next asexual cycle. Owing to the nature of reproduction, *Paramecium* does not undergo chronological aging. However, they have been shown to lose their vitality. That means, they are unable to undergo reproduction after about 200 vegetative cycles (Aufderheide, 1986).

Sexual reproduction

Paramecium enters sexual reproduction, when they experience stressful conditions like starvation. They have different mating types (more loosely genders), and genetic screens have shown that only certain mating types are compatible with each other. Some species of *Paramecium* have up to 23 mating types (CHEN, 1946). Under stress, the cells can either mate with the right mating type, or undergo autogamy (self-fertilisation) in the absence of the right mating type (Van Houten, 2019).

Figure 2.18 shows the sexual mode of reproduction. First the mating pairs enter into conjugation, by forming a cytoplasmic bridge. The cells undergo nuclear reorganisation through meiosis. Meiosis results in four haploid (1n) MIC of which three of them disintegrate. The remaining one MIC divides, by mitosis. Each cell in the conjugation pair now exchanges one of their MIC with the other through the cytoplasmic bridge. After this exchange, the conjugation is broken.

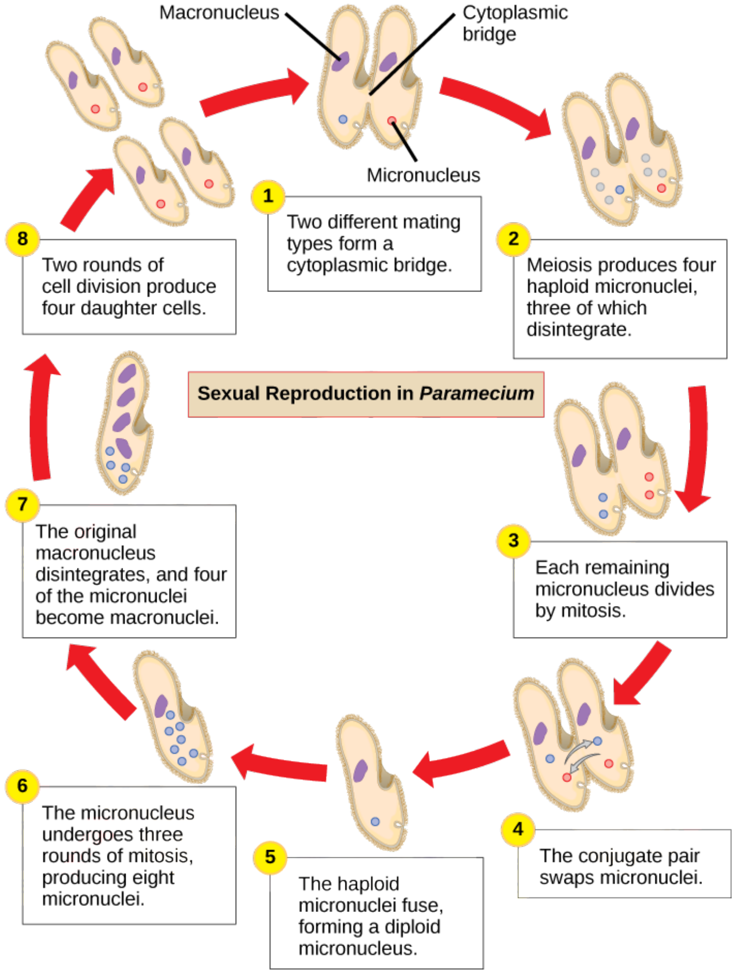

Macronucleus
Cytoplasmic bridge
Micronucleus
1
Two different mating types form a cytoplasmic bridge.
2
Meiosis produces four haploid micronuclei, three of which disintegrate.
Sexual Reproduction in *Paramecium*
3
Each remaining micronucleus divides by mitosis.
4
The conjugate pair swaps micronuclei.
5
The haploid micronuclei fuse, forming a diploid micronucleus.
6
The micronucleus undergoes three rounds of mitosis, producing eight micronuclei.
7
The original macronucleus disintegrates, and four of the micronuclei become macronuclei.
8
Two rounds of cell division produce four daughter cells.

FIGURE 2.18: The sexual reproduction cycle in *Paramecium* is shown. This figure is available for public use (Carter and Learning, 2021).

Now, the cells fuse their two haploid MIC to form a diploid MIC. Following that, the diploid MIC undergoes three rounds of mitosis to produce eight diploid MIC. At this stage, the original (or parental) MAC disintegrates, and four of the eight diploid MIC transform into MAC. Subsequently, two rounds of cell division produces four daughter cells. Currently sRNAs, and other ncRNA mediated epigenetic control mechanisms are known to regulate the genomic rearrangements of the developmental MAC (Beisson et al., 2010).

2.10.3 Serotypes in *Paramecium*

Antigens are protein or complex cellular sugar material, which can trigger an immune response in an organism to produce antibodies (Wichterman, 1986). As these proteins, in *Paramecium*, include the cilia and the cell surface, they are known as surface antigens. These surface antigens are from a multigene family, which are known to express more than 11 surface antigen genes. The surface antigen gene expression is mutually exclusive, *i.e.,* only one antigen is expressed at a time. The expression of a specific surface antigen is called a serotype. In *Paramecium*, these serotypes are named as a combination of the laboratory stock of paramecium (stock 51), and the expressed surface antigen. For instance, serotype 51A is the expression of the surface antigen gene A in the stock 51 *Paramecium.* Among the different *Paramecium* serotypes (namely 51A, 51B, 51D, and 51H), it has been shown that their entire transcriptomic profile is altered (Cheaib et al., 2015). It has also been shown that a shift in the serotype can occur with changes in environmental conditions like temperature (Simon, Marker, and Schmidt, 2006). Over a decade of research on the MAC of vegetative *Paramecium*, showed the role of sRNAs in controlling serotype expression (Marker et al., 2010). However, the complexity of sRNA molecules that are active in the *Paramecium* MAC and their function are poorly understood.

Chapter 3

RAPID: An automated small RNA analysis tool

This chapter summarises our work published in a peer-reviewed article (Karunanithi, Simon, and Schulz, 2019).

3.1 Motivation

Advancements in NGS technology, and lowering costs have enabled researchers to unravel novel biological mechanisms. We have seen in Chapter 2, the diverse regulatory roles of different classes of sRNAs like siRNA. The interest in siRNA research has grown in the past decades, because of their therapeutic potential (Patwardhan et al., 2017). Application of siRNA in drug discovery and therapy demands a better understanding of siRNA biogenesis and behavior. Several sequence level properties, like read length and strand of origin, are crucial in discerning the siRNA biogenesis mechanisms and function of siRNA. The sheer number of publicly available tools show the importance and complexity of analysing the diverse classes of sRNA data sets.

There are two broad categories of sRNA analaysis tool based on their function: (i) prediction tools and (ii) analysis tools. The prediction tools employ diverse computational strategies to predict different classes of sRNA like miRNA, piRNA, etc. Some examples of these tools include Shortstack (Johnson et al., 2016), miRDeep2 (Friedländer et al., 2012), iMir (Giurato et al., 2013), Piano (Wang et al., 2014), etc. The analysis tools perform annotation and gene ontology (GO) enrichment of the sRNAs. Some examples of such analysis tools are miRTools2 (Wu et al., 2013), iSmart (Panero et al., 2017), and CPSS (Wan et al., 2017).

The existing analysis tools are often hard coded to work only on certain organisms like humans or mouse. This poses a challenge to researchers of uncommon model organisms, as there is little user flexibility in the existing tools. Although, few tools like sRNAtoolbox (Rueda et al., 2015), Oasis (Capece et al., 2015), and ncPRO-Seq (Chen et al., 2012) provide user-flexibility, they lack insightful graphical analysis. They are often not adept in systematic, and automated multiple sample comparison employing appropriate normalisation strategies for sRNA analysis. In essence, the myriad of existing sRNA analysis tools are less flexible or they do not capture functionalities which are crucial to understand siRNA biogenesis and functions. In Table A.1 we provide a comparison of non-exhaustive list of sRNA analysis tools and their abilities to address eukaryotic siRNA properties.

3.2 Project objectives

In this context, we wanted to create a generic siRNA analysis tool which enables

1. in depth analysis of eukaryotic siRNA,
2. automated visualisation of diverse siRNA properties, and
3. multiple sample comparisons with appropriate normalisation methods.

3.3 Tool description

Keeping our objectives in mind, we developed an offline sRNA analysis tool: Read Alignment, Analysis, and Differential PIpeline (RAPID). This section describes the different modules of our tool, shown in Figure 3.1.

3.3.1 Basic module

The first RAPID module is *rapidStats*, which invokes Bowtie2 (Langmead and Salzberg, 2012) to perform sequence alignment, with user defined options to remove contaminants. Following alignment, RAPID quantifies several statistics such as read length distribution, soft-clipped nucleotides, strandedness, and nucleotide content. RAPID can also use alignment files (BAM/SAM) created by other aligners. We efficiently process the alignments, and capture the aforementioned statistics using SAMtools (Li et al., 2009), BEDtools (Quinlan and Hall, 2010), and custom Perl, Shell, and R scripts. The statistics captured by this module serve as input for other modules.

3.3.2 Normalisation module for multi-sample comparison

The second RAPID module, *rapidNorm*, aims to facilitate an unbiased comparison of genes or regions across multiple siRNA samples. Other than the sequencing depth itself, siRNA studies pose an additional challenge during normalisation. For instance, to understand RNA interference (RNAi) mechanisms and how the siRNA homeostasis is maintained, often a gene or siRNA region is knocked down. One such knockdown strategy is to introduce large amounts of siRNAs, called primary siRNAs, against the knockdown gene or any siRNA region. Consequentially, secondary siRNA production is triggered by the primary siRNAs. These primary and secondary siRNAs, which are also sequenced, can add up to millions of reads in the total library size. More information on RNAi can be read in Chapter 2.

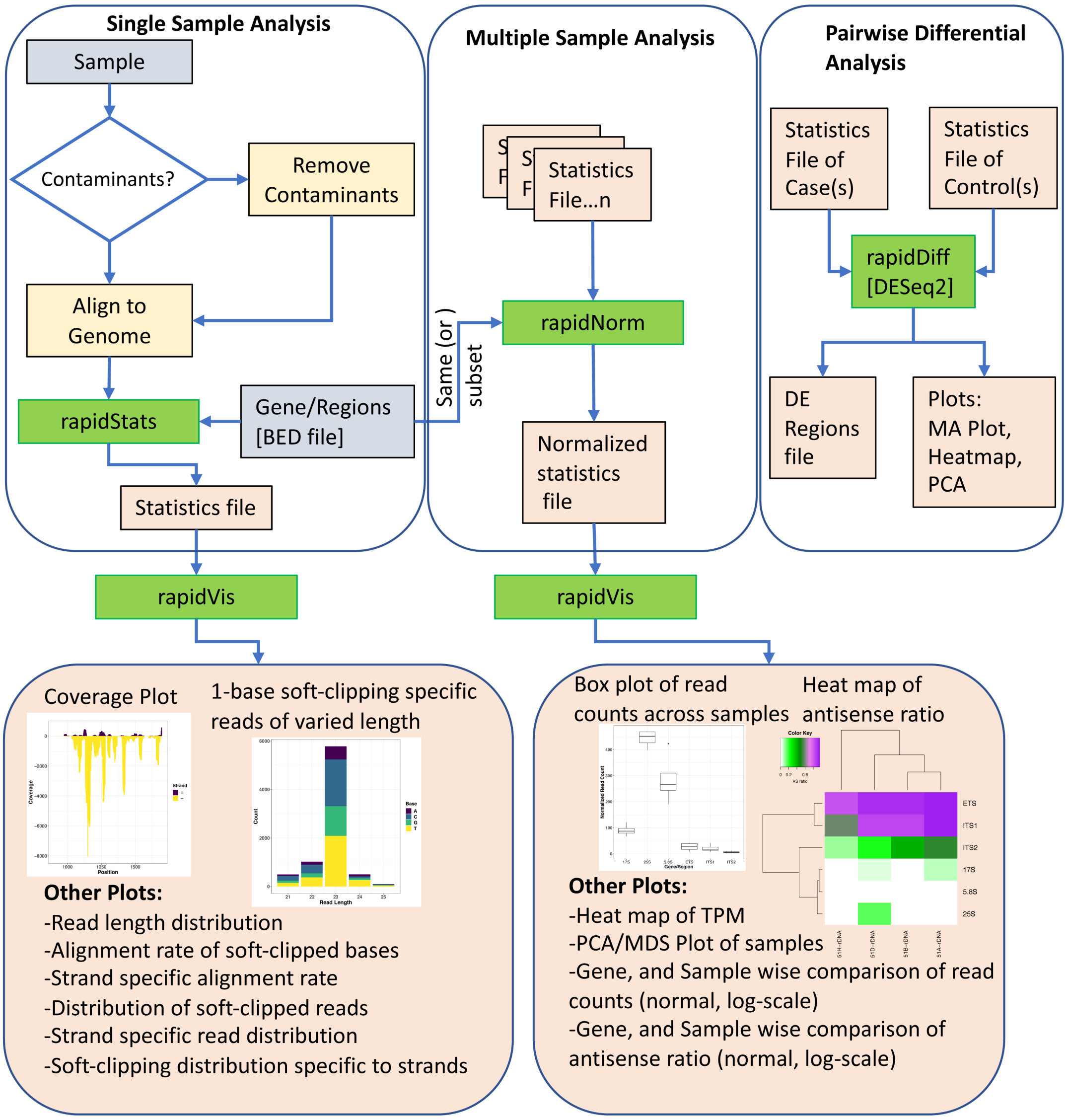

Single Sample Analysis
Sample
Contaminants?
Remove Contaminants
Align to Genome
rapidStats
Gene/Regions [BED file]
Statistics file
Same (or) subset
Multiple Sample Analysis
Statistics File…n
rapidNorm
Normalized statistics file
Pairwise Differential Analysis
Statistics File of Case(s)
Statistics File of Control(s)
rapidDiff [DESeq2]
DE Regions file
Plots: MA Plot, Heatmap, PCA
rapidVis
rapidVis
Coverage Plot
1-base soft-clipping specific reads of varied length
Other Plots:
-Read length distribution
-Alignment rate of soft-clipped bases
-Strand specific alignment rate
-Distribution of soft-clipped reads
-Strand specific read distribution
-Soft-clipping distribution specific to strands
Box plot of read counts across samples
Heat map of antisense ratio
Other Plots:
-Heat map of TPM
-PCA/MDS Plot of samples
-Gene, and Sample wise comparison of read counts (normal, log-scale)
-Gene, and Sample wise comparison of antisense ratio (normal, log-scale)
Color Key
ETS
ITS1
ITS2
17S
5.8S
25S
Position
Coverage
Strand
Read Length
Count
Base
A
C
G
T
Gene/Region
Normalized Read Count

FIGURE 3.1: The pipeline of our tool RAPID is depicted. Green boxes are executables. Blue, and orange boxes represent input, and output files respectively. The executable RAPID modules are: (i) *rapidStats* module performs reference alignment and quantifies the expression of user-defined genes and/or regions. (ii) *rapidNorm* facilitates sample (or gene) wise comparison of genes/regions (or samples) after appropriate normalisation. (iii) The *rapidVis* module provides multiple visualisations representing the information obtained from *rapidStats* and *rapidNorm*. Selective screenshots from the output of our case studies are shown in the boxes. (iv) *rapidDiff* is the differential expression analysis module implementing DESeq2.

To our knowledge, there are no normalisation methods specialized for knockdown based siRNA studies. However, many methods have been proposed to normalise mRNA-seq data, which can be broadly categorized in two classes: (i) total count scaling (TCS) methods and (ii) methods which utilize quantities like median log-fold change, among all genes between mRNA-seq experiments. To be able to use the latter methods, siRNA loci annotation should be available, and should assume that most of the siRNA loci between samples are not differentially expressed.

In model organisms like *Paramecium tetraurelia*, little is known about the localisation, and expression variability of endogenous siRNA loci. Hence, the second class of methods may not be applicable. However, the disadvantage of TCS methods is that the used normalisation factors were shown to be biased by highly expressed genes in the data set (Dillies et al., 2013). In case of knockdown samples, TCS methods will be heavily skewed because of the millions of primary, and secondary siRNAs associated with the knockdown gene or region.

In mRNA-seq data, a variant of the TCS method (Sultan et al., 2008) was introduced, where normalisation is achieved by scaling through a factor that estimates the difference in the number of reads mapped between samples. Similarly, we propose a variant of the TCS method for knockdown based siRNA studies. We term this variant as KnockDown Corrected Scaling (KDCS) method, where we remove from the estimated total library size, all reads that map against the knockdown genes, this quantity is denoted K below. Assume read count R for a region of interest that we want to compare between samples. T is the total number of reads mapping to the genome, and K is the number of reads mapping to the knockdown gene. We compute the normalised read count $\hat{R}$:

$$\hat{R} = R \cdot \frac{M}{T - K}, \tag{3.1}$$

where M is the maximum over all values $(T_1 - K_1), ..., (T_n - K_n)$ over all n samples.
RAPID uses the KDCS method, by default. Hence, in the absence of knockdown genes, the normalisation works as the normal TCS method. However, in order to provide flexibility with the choice of normalisation for knockdown free analysis, we have also incorporated size factor-based normalisation from DESeq2 (Love, Huber, and Anders, 2014). If an user can safely assume that most of the genes or regions between samples are not differentially expressed then they can use the DESeq2 normalisation. In essence, *rapidNorm* reports normalised values for all the statistics captured in the *rapidStats* step when the user has multiple samples to analyse. We demonstrate the use of our KDCS normalisation method in Section 3.4.

3.3.3 Visualisation module

For a better interpretation of sequencing data, we need smart visualisations. The *rapidVis* module of our tool can automatically generate visualisations of the statistics captured in the previous modules. Using Rmarkdown (`http://rmarkdown.rstudio.com`), RAPID generates easily navigable HTML reports.

This module contains two modes: *statistics* and *comparison* mode. The statistics mode takes as input the *rapidStats* output file, and provides various single category plots detailing on the distribution of read length, strandedness, soft-clipped nucleotides, and coverage plots for each gene/region analyzed. In addition, this report also provides combinations of the aforementioned properties. For instance, how does strandedness differ across different read lengths. The comparison mode accepts the *rapidNorm* analysis output file, and equips the user with qualitative reports (Heatmaps, Principal Component Analysis (PCA), Multi Dimensional Scaling (MDS)) of samples. All plots are shown both in normal and log scale such that the user can directly incorporate them into publications.

3.3.4 Differential analysis module

Differential Expression (DE) analysis is one of the common downstream analysis in comparative studies. RAPID offers the user with this functionality by incorporating the DESeq2 package. Upon invoking the *rapidDiff* module, raw counts are utilized from the output of the *rapidStats* module to perform DE analysis, with default parameters of DESeq2. Results of the DE analysis include intuitive plots (such as MA Plot, Heatmap, PCA) and the list of DE genes/regions.

3.3.5 Usage and availability

We have made available a conda recipe for RAPID through the bioconda channel (Grüning et al., 2018). This ensures ease of installation by avoiding dependencies. We strongly recommend using RAPID from `https://anaconda.org/bioconda/rapid` as a conda recipe. Nevertheless, it can also be freely accessed from `https://github.com/SchulzLab/RAPID`.

3.4 Case Study

One of the unique features of RAPID is the KDCS normalisation that can correct for the excess of sRNAs introduced in knockdown experiments in experimental approaches utilized in many diverse organisms. In this section, we discuss an use case to demonstrate the effectiveness of KDCS normalisation. An in-depth use case based documentation is provided at `https://rapid-doc.readthedocs.io/en/latest/`.

3.4.1 Problem setup

A knockdown study on *P. tetraurelia* investigates the molecular mechanisms of different sets of trans-acting RNAi components (Götz et al., 2016). ICL is a gene in *P. tetraurelia*, which is not involved in the RNAi machinery. Hence, ICL is knocked down as a control in the original study by introducing millions of primary siRNAs against ICL (see 4.1.2). We downloaded, and preprocessed the five publicly available ICL knockdown data sets (NCBI Accession ID: PRJEB13116) from this study. In these data sets, we wanted to quantify and compare their sRNA read counts of four example sRNA regions (which are in the original study different constructs of the ND169 gene).

3.4.2 Methodology

As the five data sets are knockdowns of the same ICL gene, we expect that all data sets behave the same. They are simply biological replicates of the same system. To compare the sRNA accumulation of the example sRNA regions across these five data sets, we need to normalise them for sequencing depth and knockdown. As mentioned ealier, very little is known about the localization and expression variability of endogenous small RNA loci in *P. tetraurelia*. This may violate the assumption of normalisation methods such as DESeq2, which assumes that majority of regions remain unchanged in different samples. Nevertheless, we applied DESeq2 normalisation, our KDCS normalisation, and an often used TCS normalisation on our case study data.

Normalisation should reduce the variance in read count per region. Hence, to evaluate the different normalisation methods, we used coefficient of variation (CoV). CoV is the ratio of standard deviation to mean of the data set. For a gene or region of interest, i, with n samples, CoV is represented as

$$CoV_i = \frac{\sigma_i}{\mu_i}, \tag{3.2}$$

where σ_i and μ_i are the standard deviation and mean of the gene or region of interest i in the n samples, respectively. A smaller CoV suggests a better performance of a normalisation method, as normalisation should reduce the variance.

3.4.3 Results

Figure 3.2 shows the CoV values of the raw, and normalised sRNA read counts, for four example regions that had been studied by Götz et al., 2016. We can observe from Figure 3.2, that the KDCS method performs better in all the regions, compared to the generic TCS method. It also achieves as good or better than the DESeq2 normalisation for this example. All normalisation approaches are better than using no normalisation, which strongly argues for their use. This experiment suggests that our KDCS method is a better alternative to the TCS method and is applicable when few regions are known.

3.5 Conclusion

We created an automated small RNA analysis pipeline, RAPID. It is an offline, open-source, and user-friendly tool designed to simplify eukaryotic siRNA data analysis. RAPID is not an exhaustive analysis or annotation pipeline. With an available set of siRNA (or other sRNA) loci, our tool can be used to analyze

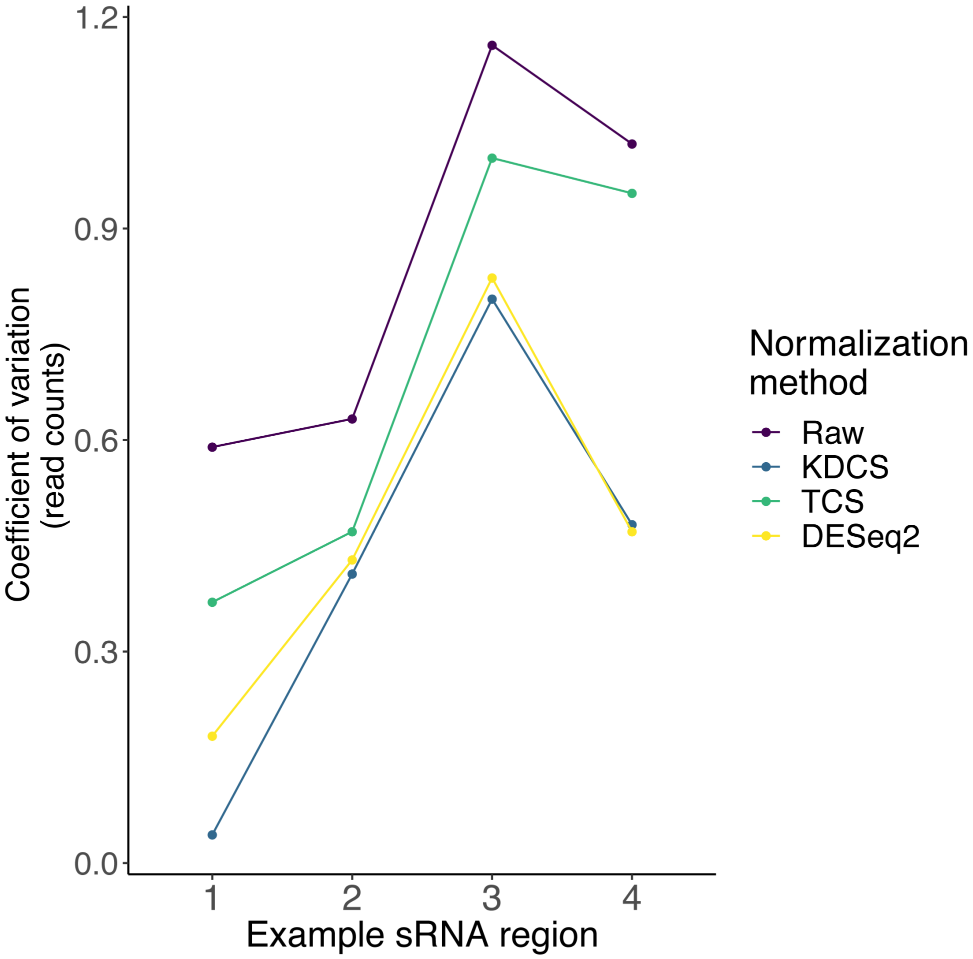


Figure 3.2: The effect of different normalisation methods on sRNA regions (x-axis) studied in Götz et al., 2016 are assessed using the coefficient of variation (y-axis; lower is better) of the read counts obtained from RAPID. Raw - No normalisation; KDCS - KnockDown Corrected Scaling; TCS - Total Count Scaling; DESeq2 - size factor-based normalisation from DESeq2

single or multiple sRNA samples at ease with the aid of different normalisation techniques. The diverse set of visualisations generated by RAPID will enhance the understanding of any sRNA-based study. RAPID is available for free use and can be used over the command line. It is available at the GitHub repository `https://github.com/SchulzLab/RAPID`. A detailed user tutorial can be accessed from this repository. The resourcefulness, and user-friendliness of RAPID can be demonstrated by the 16,000+ conda downloads, at the time of writing this thesis.

3.6 Contributions

The basic design of the pipeline, and the KDCS normalisation method was done by Prof. Dr. Marcel H. Schulz. I restructured, added improvements on all the modules, especially the visualisations, created a conda package, and performed the use cases. The features we capture in *rapidStats* module were advised by Prof. Dr. Martin Simon. A detailed list of my contributions in the development of RAPID can be seen at our GitHub page `https://github.com/SchulzLab/RAPID/graphs/contributors`.

Chapter 4

Genome wide analysis of RNAi mechanisms in *Paramecium*

This chapter summarises our work published in two peer-reviewed articles (Karunanithi et al., 2019; Karunanithi et al., 2020). Parts of this chapter have also been orally presented in two conferences: (i) Ciliate Molecular Biology conference 2018 in Washington DC, USA, and (ii) GDRE 2017: Conference on Paramecium [Epi]genome Organization, Dynamics and Evolution, Nohfelden, Germany.

4.1 Background

RNA interference (RNAi) is widely known to regulate gene expression either at transcriptional or post-transcriptional level, using small RNA molecules in eukaryotes. In order to silence gene expression, when RNAi triggers DNA or chromatin modifications it is termed Transcriptional Gene Silencing (TGS). When a RNAi process degrades an mRNA or inhibits the translation of an mRNA to protein, through a cascade of biomolecular processes, it is called Post Transcriptional Gene Silencing (PTGS) (Moazed, 2009; Zhang, 2009). Please refer to Chapter 2 for more information on RNAi.

4.1.1 RNAi plays defense

RNAi is known to act as a natural defense mechanism against viral infections in plants. Studies have also shown the existence of RNAi-dependent small RNAs acting against viral infections in mammals (Li et al., 2013; Maillard et al., 2013). Recently the United States Food and Drug Administration even approved a first RNAi based drug, *patisiran*, to treat people with polyneuropathy. Nevertheless, there is still a lot to be understood at a systemic level on the functioning of RNAi. There are several more such drugs being tested in clinical trials. This impressive feature of using RNAi in medicine, took years of effort from researchers to artificially trigger RNAi (Setten, Rossi, and Han, 2019).

4.1.2 RNAi as a molecular tool

RNAi can be triggered in cells or species when they consume artificially introduced regulatory dsRNA through their food or from the environment (Whangbo and Hunter, 2008). This artifical dsRNA is called exogenous RNA. This method

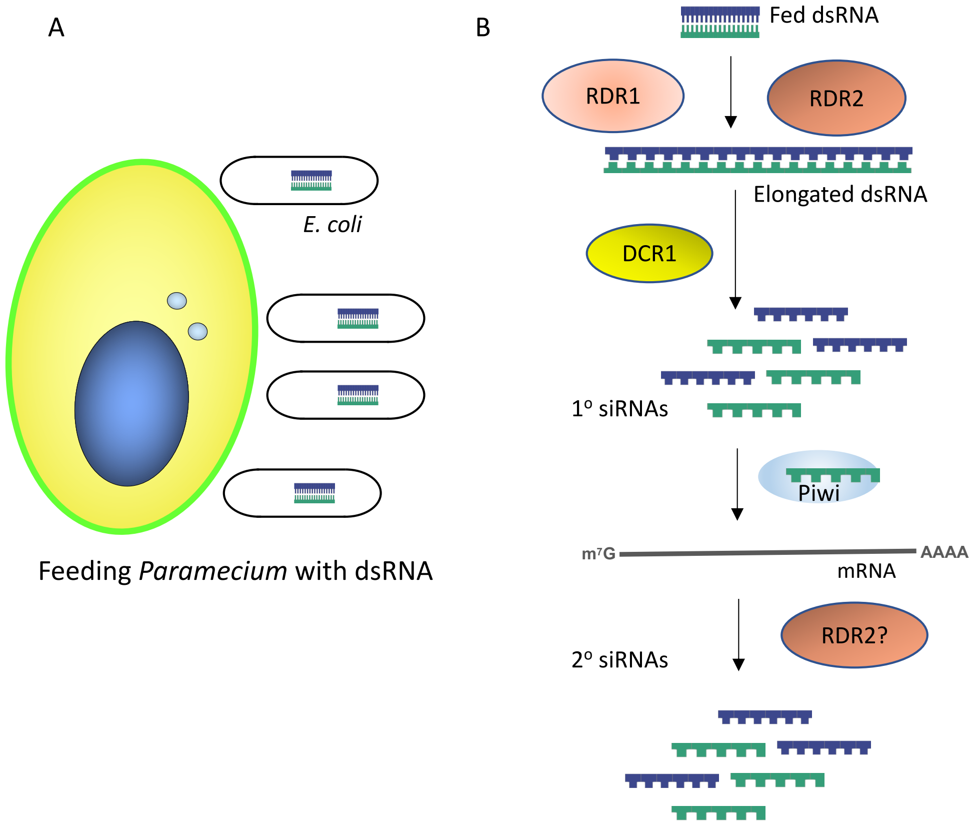


FIGURE 4.1: RNAi feeding pathway in *Paramecium*: A) A representation of feeding in *Paramecium* is shown. Engineered *E.coli* expressing dsRNA of interest is fed to Paramecium. B) The fed dsRNA is processed by different enzymes to produce primary (1^o) and secondary (2^o) siRNAs. This figure is created by me with the help of DNA/RNA images obtained in 2016 (*Somersault1824*).

has been first made available in *Caenorhabditis elegans* (Fire et al., 1998). In *Paramecium*, RNAi is achieved by feeding the *Paramecium* cells (Figure 4.1A) with the *Escherichia coli* bacteria carrying engineered dsRNA (Galvani and Sperling, 2002).

4.1.3 Exogenously induced RNAi pathways in *Paramecium*

Figure 4.1B shows a representation of the exogenous RNAi processing pathway in *Paramecium*. When an exogenous dsRNA is introduced in *Paramecium*, two RDR enzymes (RDR1 and RDR2) amplify it, and then use Dicer (DCR1) to produce 23 nt primary (1^o) siRNAs. The mechanism of secondary (2^o) siRNA production in *Paramecium* is unclear. However, the 2^o siRNAs are less abundant than 1^o siRNAs (Marker et al., 2010; Carradec et al., 2015; Marker et al., 2014). While we know how the exogenous RNA fed to *Paramecium* is processed, we do not know the endogenous small RNA composition of vegetative *Paramecium*.

4.2 Research objectives

In this context, we defined the following objectives:

1. Define a genome wide small RNA profile of vegetative *Paramecium*
2. Identify the possible sources, and mechanism of action of the defined small RNAs
3. Clarify if environmental RNA, introduced by feeding technique, affect small RNA profile and transcriptome

4.3 Data and methodology

All the methods related to the work discussed in this chapter are described here. The biological steps involved in creating our data sets are not discussed, as it is beyond the scope of this thesis. However, the background chapter includes, in brief, the idea behind many of these biological methods.

4.3.1 Data set description and retrieval

We categorize the data sets in this work into two groups: (i) Cluster definition data, and (ii) Analysis data.

Cluster definition data

In order to characterize small RNA clusters from the Wildtype (WT) serotypes 51A, 51B, 51D, and 51H, we sequenced sRNA data sets, two replicates each.

Analysis data

Wild type: We obtained the mRNA expression data from a recent study of our collaborators for the WT serotypes 51A, 51B, 51D, and 51H of *P. tetraurelia* (Cheaib et al., 2015), three replicates each (European Nucleotide Archive (ENA) Accession: PRJEB9464). For the same biological replicates, we sequenced sRNA data sets (four WT serotypes, three replicates each) in order to have paired sRNA and mRNA expression data.

RdRP mutants: We created two mutant strains of WT serotype 51A, namely 51A-RDR1 and 51A-RDR2 lacking the enzymes RDR1 and RDR2, respectively. We sequenced three replicates of sRNA and mRNA data for these RdRP mutants. These data sets are publicly accessible at ENA (Accession: PRJEB25903). Details on how these mutants are created is beyond the scope of this thesis. However, this information can be found in our publication (Karunanithi et al., 2019).

RNAi knockdown: Using the RNAi by feeding technique, we knocked down the genes ND169, and DCR1 of WT serotypes 51A (and 51B), resulting in knocked down strains namely 51A-ND169 (and 51B-ND169) and 51A-DCR1 (and 51B-DCR1), respectively. We performed sRNA (two replicates each) and mRNA (three replicates each) sequencing on RNAi knockdown samples for both 51A and 51B serotypes. These data sets can be accessed at ENA (Accession: PRJEB33364). The feeding fragments used for the gene ND169 (GeneID: PTET.51.1.G0210080) and DCR1 (GeneID: PTET.51.1.G0700179) are scaffold51_21 : 137857 - 138267 and scaffold51_70 : 312063 - 313251, respectively.

4.3.2 Sequencing data preprocessing

All sequencing data sets were trimmed for adapter sequences using Trim Galore (`https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`) which employs cutadapt (Martin, 2011) using a stringency cutoff of 10.

Small RNA data preprocessing

Small RNA reads of length shorter than 21 nt were removed from our analysis, as they are potential RNA degradation products.

4.3.3 Small RNA cluster definition

The replicates of the cluster definition data sets, after preprocessing, were merged and then aligned using the alignment module from ShortStack (version 3.4) (Johnson et al., 2016), with default parameters. Reads were aligned to the *P. tetraurelia* MAC genome (version 2;stock 51). By downsampling we ensured the alignments had equal number of reads in all samples, in order to avoid sequencing depth biasing the cluster identification. These downsampled alignments of each WT serotype were subjected to the cluster calling module of ShortStack to identify small RNA clusters. We used a minimum alignment coverage (*mincov*) of 20 alignments, with a padding (*pad*) of 100 bp. The padding parameter merges distinct clusters within 100 bp as a single cluster. The identified clusters from each WT serotype were unified as one consistent set of clusters, called Small RNA Clusters (SRCs), enabling unbiased comparison across serotypes. Unification of clusters was performed using mergeBed (from BEDtools v2.23; default parameters) (Quinlan and Hall, 2010), which unified clusters with at least 1 bp overlap.

4.3.4 Boundary modification of SRCs and definition of endo-siRNAs

We investigated them in the Integrative Genomics Viewer (IGV) (version 2.3.91) (Robinson et al., 2011) against the MAC genome annotation of *P. tetraurelia* (Stock 51; version 2). We identified several SRC loci with non-specific gene boundaries (see B.1). For instance, one SRC locus could overlap with more than one gene, just by few bps as the average intergenic distance of *P. tetraurelia* genome is only 352 bp. As we wanted to compare small RNA accumulation with gene expression data, SRC loci with non-specific gene boundaries would cause bias in such downstream analyses. Hence, we first removed SRC loci which did not overlap with any protein-coding gene. Of the remaining SRC loci, we modified the ones with non-specific gene boundaries, by subjecting a SRC-gene overlap to two conditions; (i) if the gene was covered by more than 80% and the SRC locus was covered at least by 20%, then the SRC's boundary was limited to the gene's boundary. This removed non-specific gene overlaps. However, this condition also removed genes which can genuinely overlap with multiple SRCs to achieve some biological function. In order to account for such cases, another condition was introduced (ii) if the gene was covered at least by 10% and the SRC was covered by more than 80%, such SRCs were retained without any boundary changes. This filtered and boundary modified SRC loci are termed endo-siRNAs, short for endogenous small interfering RNAs. We have used the SRCs only to show their general properties, phasing characteristics, and their genome distribution. All other analyses utilizes endo-siRNAs.

4.3.5 Quantification of sRNA reads

We quantified the sRNA reads using the RAPID software for all SRCs and endo-siRNAs. RAPID was run with default parameters, which only considers error-free alignments but allows multi-mapping reads (-k 100; -k is the bowtie2 parameter controlling the number of multi-mapping reads to be reported in alignments). We use the mean of replicates in all analyses, unless mentioned otherwise. The quantification of SRCs were used only to describe their general characteristics and to study their genome wide distribution. All other quantitative analysis were limited to the endo-siRNAs.

4.3.6 Normalization of sRNA reads

For all comparative analyses of sRNA data sets, we used the normalized counts using the TCS method implemented in RAPID. Further, we converted these values to TPM using RAPID, which we also refer to as sRNA accumulation. We utilized the KDCS method of RAPID while normalizing the sRNA read counts of RNAi knockdown samples to adjust for the feeding associated small RNAs. Analyses pertaining to individual wildtype serotypes use only the SRCs with a TPM value greater than one, denoted *serotype specific SRCs*.

4.3.7 Quantification of mRNA expression

We quantified the mRNA expression of all the *P. tetraurelia* transcripts from the MAC genome annotation (version 2; stock 51), using Salmon (version 0.8.2; default parameters) (Patro et al., 2017). We used the mean of replicate expression, in TPM units, unless mentioned otherwise. We excluded the feeding regions, and 100 bps upstream and downstream to account for alignment artifacts, from the transcripts for the respective RNAi knockdown samples.

4.3.8 Clustering of expression data

We created heat maps of the endo-siRNA and mRNA expression data and performed hierarchial clustering of samples with complete linkage using an euclidean distance measure. We made use of the R package gplots (version 3.0.1.1).

4.3.9 Comparative analysis of sRNA and mRNA in wild-type

We calculated Pearson correlation between mRNA expression and sRNA accumulation using all WT serotype replicate measurements. We had two special cases while associating endo-siRNAs with mRNAs. First case is when a gene encompassed multiple endo-siRNAs, we used the sum of all such endo-siRNA's accumulation to correlate with the respective gene's mRNA expression. Second case is when an endo-siRNA mapped to multiple genes, we used that endo-siRNA's accumulation to correlate with each gene's mRNA it mapped to.

Quantification of sRNA in exon-exon junctions and introns

To investigate the source of sRNA, we quantified the accumulation of sRNA in the exon-exon junction (EEJ), and introns. We obtained the list of introns from the ParameciumDB (`https://paramecium.i2bc.paris-saclay.fr/`) (Arnaiz et al., 2007). Using the exon information from the MAC genome annotation (version 2; stock 51), we defined an EEJ as 18 bps upstream and downstream of an exon-exon boundary.

4.3.10 Phasing prediction

We predicted phased loci, from our downsampled sRNA alignments, using the (P-score) method (Howell et al., 2007). The genome is scanned in windows of size 253 bp considering 11 registers of length 23. Each register is made of 23 bins; one phased and 22 non-phased bins. We used the P-score to calculate the enrichment of sRNA reads on both strands in phased registers. If a window (loci) has a minimum of 20 reads, at least 3 (out of 22) distinct phased bins in that window has sRNA reads for each strand, and a P-score > 10 then that window is predicted as phased.

4.3.11 Annotation of SRCs and endo-siRNAs

Using the genome annotation file of *P. tetraurelia* (Stock 51; version 2) downloaded from the parameciumDB and BEDtools (intersectBed; version 2.23) (Quinlan and Hall, 2010), we annotated the identified *serotype specific SRCs* in to different genomic categories. Annotations mapping to the number of protein-coding genes were handled separately using the endo-siRNAs in order to avoid the non-specific gene boundaries described above.

Pseudogene annotation

Pseudogenes are non-functional copies of functional genes in a genome. As pseudogenes were not part of the *P. tetraurelia* MAC genome annotation, we used the PseudoPipe (Zhang et al., 2006) software to predict pseudogenes. PseudoPipe performs a comprehensive homology search of the protein coding genes in a genome followed by several filtering criteria like homology score, intron-exon structure, and frameshift mutations. We used the default parameters, except for one change. *P. tetraurelia*'s genetic code differs from other organisms. Specifically the codons UAA and UAG codes for the aminoacid glutamine instead of acting as a stop codon. So, we adapted the software's tblastn (-D 6) step of the software to accommodate the modified genetic code.

4.3.12 Differential expression analysis

Using the WT samples as control, we performed a differential expression (DE) analysis of endo-siRNAs raw read counts for each RNAi knockdown and RdRP mutant samples. We had the same setup for mRNA samples, except that we used the raw mRNA read counts obtained from HTSeq (version 0.9.0) (Anders, Pyl, and Huber, 2014). We used the R/Bioconductor package DESeq2 (version 1.18.1) (Love, Huber, and Anders, 2014) to perform the DE analysis. Following the DE analysis, endo-siRNAs (or mRNAs when applicable), with a false discovery rate lesser than 0.05 (FDR<0.05) were considered as differentially expressed with statistical significance.

4.3.13 Off-target analysis

We created all possible 23−mers from the feeding regions, used in the RNAi knockdown samples of both ND169 and DCR1 genes, as well as their reverse complement 23−mers. We aligned these 23−mers against the rest of the *P. tetraurelia* MAC genome (version 2; stock 51). We used the bowtie2 (Langmead and Salzberg, 2012) aligner to perform local alignments (−−local) and report up to 100 distinct alignments for each read (−k 100). Further, we identified the genes overlapping with a unique exact match from these alignments.

4.3.14 Gene ontology enrichment

Using the Gene ontology (GO) association file downloaded from ParameciumDB, we performed GO enrichment analysis using Ontologizer (version 2.0) software (Bauer et al., 2008) with default parameters except for using the parent-child union method, and Benjamini-Hochberg correction method for multiple testing. We considered GO terms with a multiple testing corrected p-value < 0.05 as statistically significant. We used all *P. tetraurelia* genes as the population set.

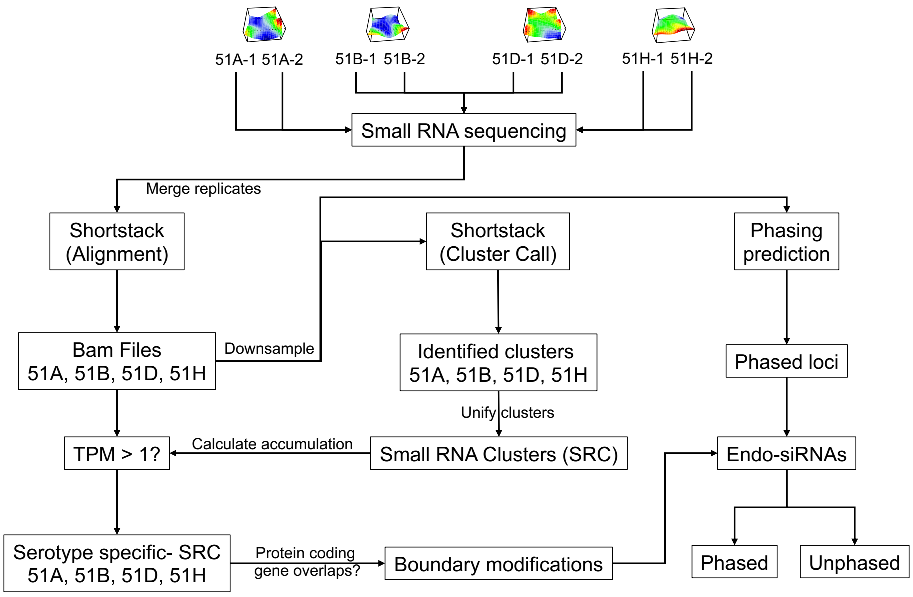


FIGURE 4.2: Overview of the small RNA cluster definition workflow. The first row visualises the different WT serotypes according to a transcriptome analysis (Cheaib et al., 2015)

4.3.15 Nomenclature

- Results of all sequencing experiments are termed samples. Eg. 51A is a wildtype sample

- The abbreviations of wildtype samples are WT (Eg. 51A); RDR1/RDR2 mutant samples are RDR1/RDR2 (Eg. 51A-RDR1); ND169/DCR1 knocked down samples are ND169/DCR1 (Eg. 51B-ND169)

- When an experiment is repeated more than once for a sample, we get replicates. Replicates are enumerated following the sample name. Eg. 51B-ND169-1

4.4 Results and discussion

4.4.1 Definition of small RNA clusters (SRCs)

Paramecium undergoes diverse transcriptomic alterations in different WT serotypes (51A, 51B, 51D, 51H) as shown in the heat maps of Figure 4.2. A genome wide small RNA profile of Paramecium had never been documented. So, we set out to characterize a genome wide profile of small RNAs in these four WT serotypes. Using the workflow shown in Figure 4.2, we identified 2602 SRCs, with none of them predicted as canonical miRNAs by the ShortStack algorithm. We observed that the majority of the SRCs are between 100 and 1000 bps in length (Figure 4.3A), with a predominant sRNA length of 23 nt (Figure 4.3B). We quantified the expression of 2602 SRCs in all WT serotypes, and observed that

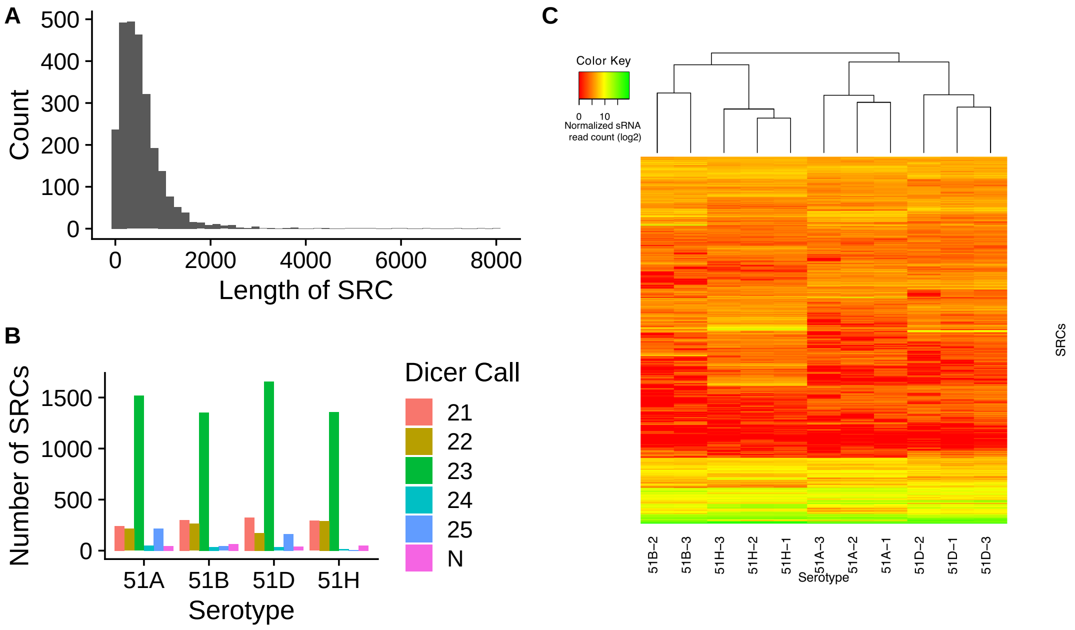


FIGURE 4.3: Characteristics of the SRCs: A) Length distribution of SRCs. B) Number of serotype specific SRCs (y-axis) detected in the WT serotype samples (replicates were merged), stratified according to the predominant small RNA length (dicer call), where N means that no predominant length could be found. C) Heatmap of normalized sRNA read counts after hierarchical clustering (Euclidean distance measure) of the SRCs (rows) for all WT serotype replicates (columns).

the individual serotypes clustered according to the expression of SRCs (Figure 4.3C). This suggested us that individual serotypes are distinguishable based on the expression of SRCs. Hence, we used a cut-off on the expressed small RNA accumulation (TPM>1), to define 2236, 2058, 2393, and 2012 SRCs as serotype specific SRCs in WT serotypes 51A, 51B, 51D, and 51H, respectively.

4.4.2 Majority of SRCs are in protein-coding genes

We investigated which genomic regions produce small RNAs, by overlapping the identified SRCs against the genome annotation of *Paramecium*. The genome annotation included protein-coding, and other non-coding RNA like tRNA, 5S rRNA, snoRNA, snRNA. As pseudogenes were not part of the publicly available annotation at ParameciumDB, we predicted pseudogenes using Pseudopipe software (see 4.3.11) and included them in our analysis. We show the number of serotype specific SRCs that overlap with distinct annotated genomic regions in Figure 4.4A. We found that the majority of SRCs ($\approx$ 1300) overlap with protein-coding genes. As we were interested in analysing their source and function in downstream analyses, we investigated some of them in IGV browser. We found that many SRCs had non-specific gene boundaries, which we rectified (see

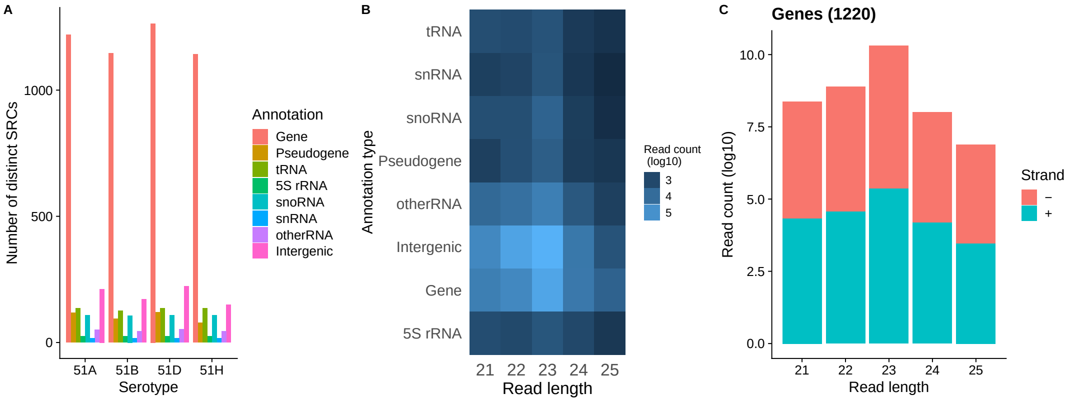


FIGURE 4.4: Annotation of SRCs: A) Serotype specific SRCs expressed in the wildtype serotype (replicates were merged) samples were overlapped with annotated regions. Each annotated element is counted only once (distinct counting) and the number of elements of the different types (colors) is shown on the y-axis for all 4 WT serotypes. B) sRNA accumulation (log10 TPM, color scale) in SRCs overlapping different genomic annotations (rows) and restricted to small RNA length (x-axis) for 51A serotype. C) Length distribution of sense (green) and antisense (red) sRNAs mapping to protein-coding genes in 51A serotype.

4.3.4). After boundary modifications, in total, we retained 1618 serotype specific SRCs overlapping with protein-coding genes, which we called endo-siRNAs (short for endogenous small interfering RNAs). We observed that only few SRCs were found in various non-coding RNA loci. However, when we account for the fact that only ≈ 3% of the *Paramecium* genome is non-coding, we observed that almost all of the non-coding RNA loci produce small RNAs. For instance, in serotype 51A, 1,220 SRCs were in genes (≈ 3% of 40,460), 117 in pseudogenes (≈ 5% of 2435 pseudogenes), 212 in intergenic regions (≈ 0.4% of 39,156), 135 in tRNAs (≈ 68% of 198 annotated tRNAs), 16 in snRNAs (100%), 24 in 5S-rRNAs (≈ 96% of 25 5S rRNAs), 108 in snoRNAs (≈ 76% of 142 annotated snoRNAs) and 50 in the category of other RNAs with diverse functions (≈ 7.2% of 689 other RNA loci). Further, we investigated the read length and strand distribution of small RNAs across all annotation types. Figure 4.4B shows that 23 nt is the predominant length across all annotation types in WT serotype 51A. Figure 4.4C shows the presence of both strands of small RNA reads, across different read lengths in the genes loci of 51A. Appendix Figure B.4 shows that the same information holds true for all other WT serotypes, and annotation. The predominance of 23 nt sRNAs, in both strands suggests that (i) they are not mRNA degradation products, and (ii) they are likely products of the RNA interference machinery.

4.4.3 Common set of protein-coding genes with endo-siRNAs

We were curious to find out whether there is a common set of protein-coding genes which overlap with our endo-siRNAs among all four WT serotypes. Using the protein-coding genes which overlap with our endo-siRNAs in each serotype, we created a set-intersection plot (popularly called UpSet plot) shown in Figure 4.5A. We observed that 973 protein-coding genes were found to be common among all four WT serotypes, which overlap with our endo-siRNAs. These are termed, Genes associated with Small RNA Clusters (GSRCs). A GO enrichment analysis on these GSRCs showed enrichment of wide range of biological processes including gene expression, translation, structural molecular activity, and cellular biosynthetic processes. This suggested that GSRCs play a crucial role in the functioning of *Paramecium*.

4.4.4 mRNAs are a predominant source of small RNAs in GSRCs

We queried the mRNA expression data to see if there are differences between our GSRCs and other genes. Figure 4.5B shows a box plot of mRNA expression among GSRCs and other genes, for all serotypes. In all serotypes, GSRCs had a higher median expression compared to other genes, found to be statistically significant (Wilcoxon test, $P < 0.05$). Further, we wanted to investigate whether mRNAs of the GSRCs can be a source of endo-siRNAs. To this end, we analysed the small RNA content of 708 GSRCs, which had at least one exon-exon junction (EEJ). Figure 4.5C shows a box plot of the total sRNA read counts in the EEJs and introns for the 708 GSRCs. Except in 51H serotype, we do not observe any sRNA reads in introns. In all serotypes, the sRNA content, in comparison to introns, is found to be higher in EEJs with statistical significance (Wilcoxon test, $P < 0.05$). This suggests clearly that endo-siRNAs are mRNA products of GSRCs.

Next, we investigated whether there is a correlation of mRNA and sRNA content of the 973 GSRCs. Figure 4.5D shows the distribution of gene wise Pearson correlation coefficient values. We observe both positive and negative Pearson correlation coefficient values. Following multiple testing correction, only $\approx 8\%$ of the correlations were statistically significant ($FDR < 0.05$). In the absence of a clear trend for all GSRCs, we found that 71 and 3 GSRCs showed statistically significant positive and negative correlation values, respectively. Based on GO enrichment analysis, we did not find any unique functions enriched for these set of positively and negatively correlated genes. Nevertheless, the positive correlations with endo-siRNAs suggest, that the majority does not act in *cis* to silence their parent mRNA, unlike in typical miRNA-based RNAi.

4.4.5 Phasing of small RNA occurs in *Paramecium*

We hypothesised, based on the 23 nt small RNA predominance in both strands, that the RNAi machinery is involved in production of endo-siRNAs. To test

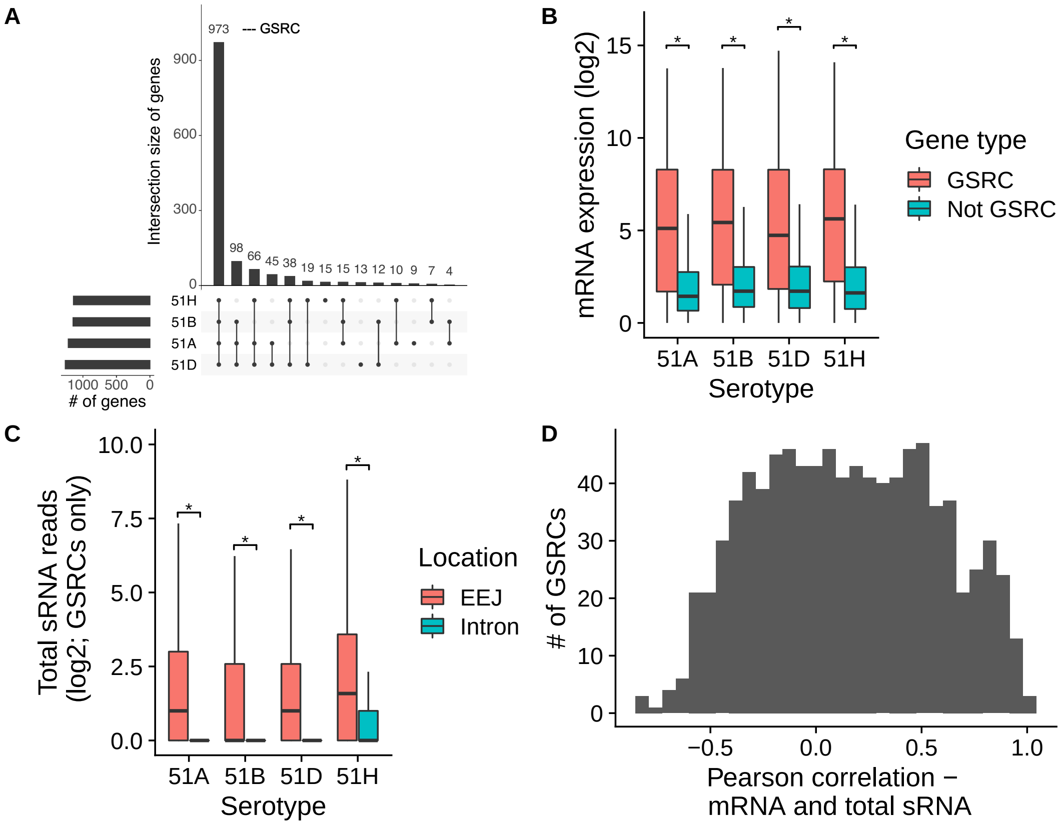


FIGURE 4.5: Genes associated with SRCs: A) Set intersection plots for endo-siRNAs overlapping with protein-coding genes across serotypes. Genes consistently overlapping in the 4 wildtype serotypes, are called Genes associated with SRCs (GSRCs). B) Boxplot of mRNA expression (y-axis, log2 TPM) in the 4 wildtype serotypes of GSRCs (red) and other expressed genes (green). C) Boxplot of total sRNA reads (y-axis; log2) in the exon-exon junctions (EEJ; red), and introns (green) of GSRCs. D) A histogram of the number of GSRCs (y-axis) plotted against the Pearson correlation of mRNA expression and total sRNA (x-axis) of all WT replicates.

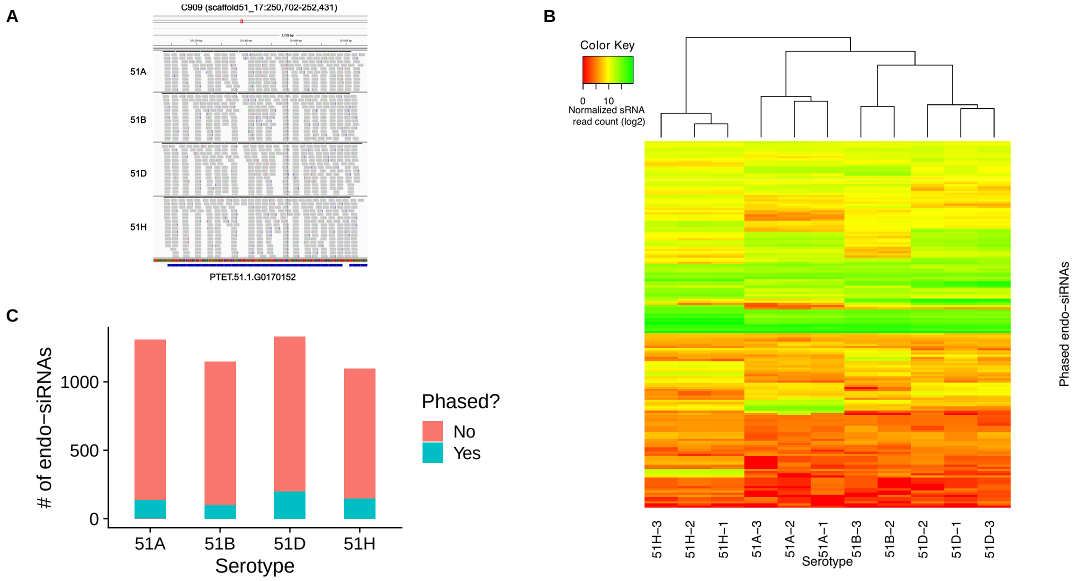


Figure 4.6: Phased small RNA loci : A) Example IGV screenshot of identified phased cluster, C909, annotated as a gene (ID: PTET.51.1.G0170152). B) Heatmap of normalized sRNA read counts after hierarchical clustering (Euclidean distance measure) of the phased endo-siRNAs (rows) for all WT serotype replicates (columns). C) Total number of phased endo-siRNAs observed in all WT serotypes.

that hypothesis, we predicted phased small RNA loci from our data and overlapped them with our endo-siRNA loci (see 4.3.10). A visual example of a phased endo-siRNA locus is shown in Figure 4.6A as an IGV screenshot. Figure 4.6B shows a clustered heat map of the small reads in phased endo-siRNAs across different serotypes. We observed that serotypes clustered based on the phased endo-siRNA read counts, demonstrating again the diversity of endo-siRNA abundance in different serotypes. Further, we overlapped the phased loci with our endo-siRNAs. Figure 4.6C shows the fraction of phased and unphased endo-siRNAs. While 51D serotype had the highest number of phased endo-siRNAs, all serotypes had at least $\approx 10\%$ of endo-siRNAs as phased. Further we found that only $\approx 12\%$ of GSRCs were phased (not shown in plots). If the known RNAi machinery were to be the only source of endo-siRNAs in our case, we should have seen a phasing of all endo-siRNA or GSRCs. The contrary suggests that, different endo-siRNAs are produced through different mechanisms.

4.4.6 Phased endo-siRNAs depend on two RDR enzymes

The RDR enzymes (RDR1 and RDR2) are part of the RNAi machinery. Their role in siRNA biogenesis after processing the exogenous RNA (environmental RNA) is well documented by previous works in *Paramecium* (Carradec et al., 2015; Marker et al., 2014). We have seen in the previous sections, the RNAi machinery plays a role in creating the endo-siRNAs. In order to confirm that, we created RDR1 and RDR2 mutant data sets of wildtype *Paramecium* expressing serotype 51A. The total small RNA reads of the endo-siRNAs (Figure 4.7A), and the mRNA expression (Figure 4.7B) in our mutant data sets (51A-RDR1, and 51A-RDR2) are statistically significantly lower than the respective wildtype serotype 51A (Wilcoxon test, $P < 0.05$). This indicates clearly that both RDR1 and RDR2 play a role in the biogenesis of endo-siRNA as well. Further, we analysed the fold changes of the mutants over the wildtype samples. The total sRNA accumulation (Figure 4.7C) fold change of phased endo-siRNAs is statistically significantly lower than the unphased endo-siRNAs. This suggests that the phased endo-siRNAs are dependent on the RDR enzymes, and other endo-siRNAs are likely to have a different biogenesis mechanism. On the contrary, when we analysed the mRNA expression fold change (Figure 4.7D), we observed that the fold change of phased mRNAs (*i.e.* genes which overlap with phased endo-siRNAs) are statistically significantly higher expressed than the unphased mRNAs. This observation suggests that phased endo-siRNAs tend to act negatively in *cis* to silence the parent mRNA. However, we do not see a perfect negative correlation of all the phased endo-siRNAs and respective mRNAs.

4.4.7 Drastic alterations of the endo-siRNA repertoire in control feeding

While inducing RNAi by feeding (see 4.1.2), researchers commonly feed against a gene unrelated to the phenomenon under study and use that as control (even treat it as similar to wildtype). However, the genome wide changes caused due to feeding have never been documented in *Paramecium*. We exploited our endo-siRNA data to study the genome wide changes caused by feeding in two wildtype serotypes (51A, and 51B). In both these serotypes, we created two knocked down data sets of two genes and sequenced their small RNA, and mRNA (discussed in next section). The two genes knocked down by feeding are namely (i) ND169, a gene known to be unrelated to the feeding pathway, as control and (ii) DCR1, the dicer gene known to be involved in the feeding pathway, which processes the dsRNA into siRNA. Figure 4.8 (A and B) shows a heat map of the normalized endo-siRNA read counts in wildtype and knocked down samples of serotype 51A and 51B, respectively. We observed that different subsets of endo-siRNAs undergo changes in their abundance in different knockdown samples. The wildtype samples clustered separately from the rest of the knocked down samples. Surprisingly, ND169 knocked down samples clustered together with the DCR1 knocked down samples. As ND169 is thought to be not involved in feeding pathways, we expected ND169 knocked down samples to cluster with the wildtype samples. This suggests that samples with control feeding are not similar to wildtype samples. We show the quantitative changes between wildtype and knocked down samples using the abundance of endo-siRNAs in all samples (replicates were merged) in Figure 4.9 (A and B). We observed a statistically significant reduction (Wilcoxon test; $P-$value < 0.05) of endo-siRNAs in the

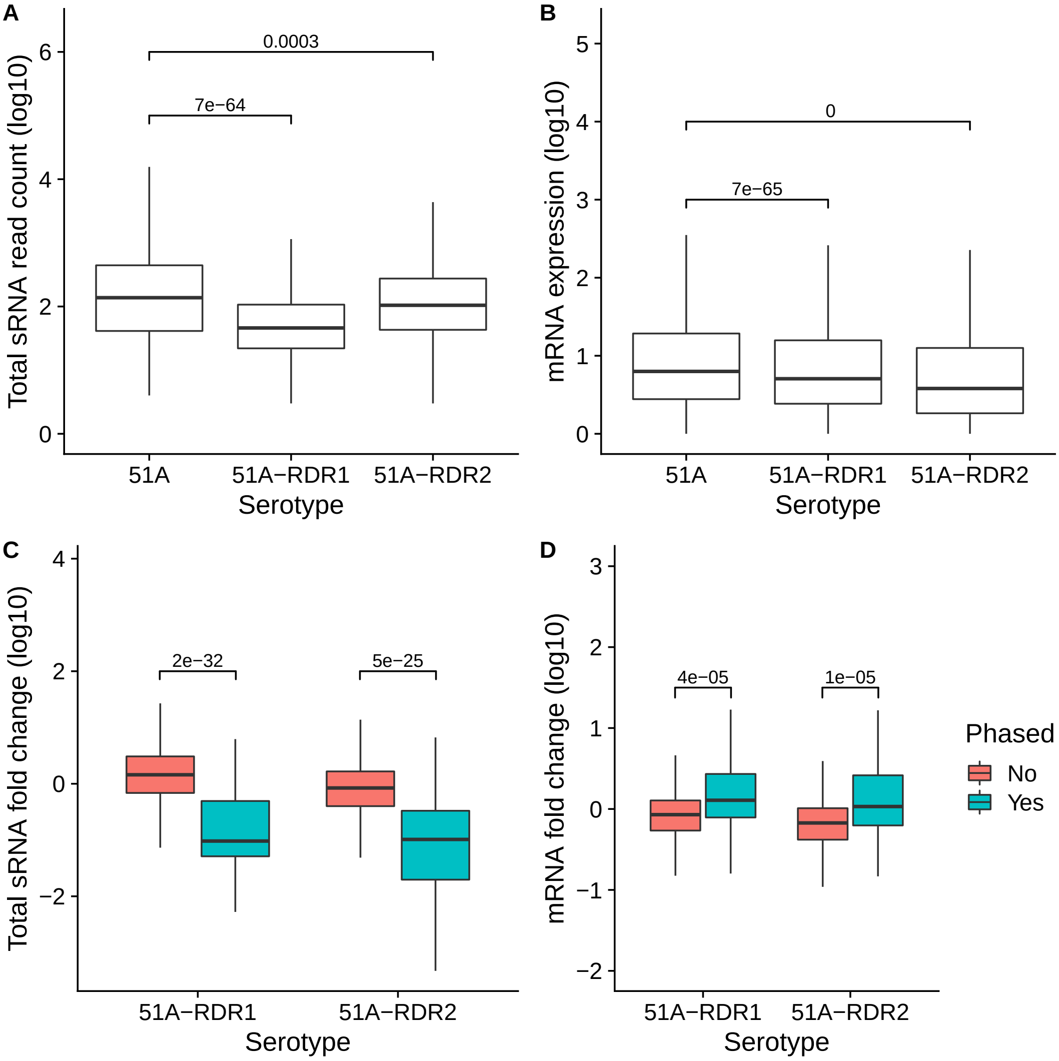


FIGURE 4.7: RDR mutant analysis: A) Box plots of the total small RNA read counts (y-axis; log10) from the WT, and mutant samples (51A-RDR1, and 51A-RDR2). B) Same as (A), but showing the mRNA expression (y-axis; log10 TPM). C) Boxplots of total sRNA fold change (y-axis; log10 mutant/WT) in each mutant is shown. Boxplots are grouped based on whether an endo-siRNA is predicted as an unphased, or phased locus. (B) Same as (A), but shows the mRNA fold change (y-axis; log10 mutant/WT). The P-values indicated are from a two-tailed Wilcoxon test.

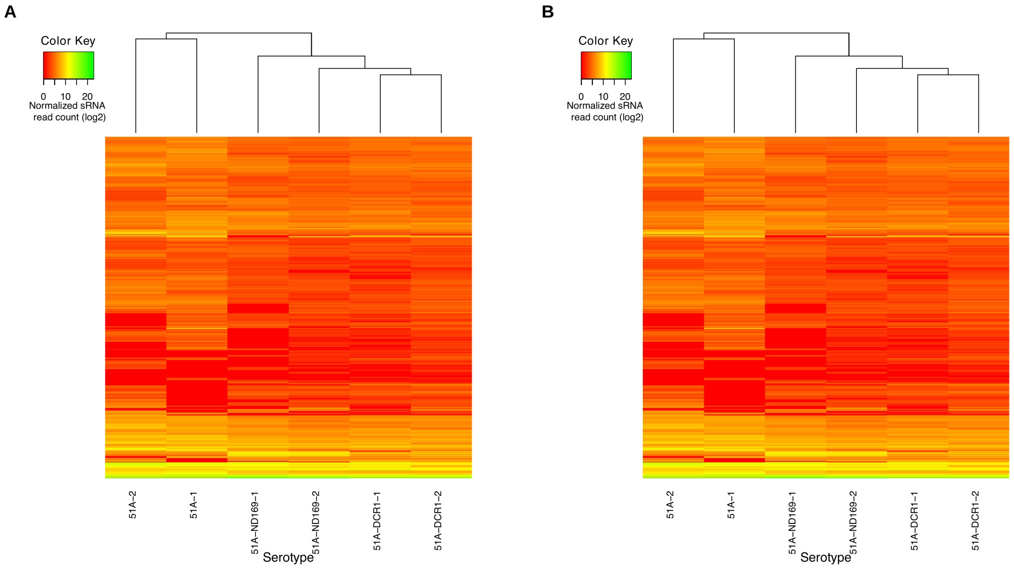


FIGURE 4.8: Endo-siRNAs in knockdown: Heatmap of normalized sRNA read counts after hierarchical clustering (Euclidean distance measure) of the 1,618 endo-siRNAs (rows) for all replicates (columns) respective to the serotype 51A (A) and 51B (B) is shown.

knocked down samples compared to the wildtype in both serotypes 51A and 51B. In addition, the difference between the median sRNA read counts is less among the knocked down samples. These observations further strengthen our claim that control feeding is not similar to wildtype. We performed a differential expression analysis of the endo-siRNAs between the wildtype and knocked down samples. Using the differentially expressed genes in each knocked down sample, which were statistically significant ($FDR < 0.05$), we created set intersection plots for both serotypes separately (Figure 4.9 C and D). In serotypes 51A and 51B, the DCR1 feeding samples have 371 and 367 DE endo-siRNAs, respectively. Of them approximately 70% of the endo-siRNAs (257 in 51A; 254 in 51B) are differentially expressed in ND169 feeding as well, which suggests they are a common response to the knock down experiments (or the feeding of exogenous dsRNA).

We performed a GO enrichment analysis to investigate whether genes associated with the DE endo-siRNAs have any overrepresented GO terms. We identified diverse functions and processes associated with these genes. Following are some of the significantly enriched terms: cofactor metabolic process, pteridine-containing compound metabolic process, single-multicellular organism process, multicellular organism process, developmental process, and others. These results suggest that feeding interferes with a diverse set of pathways irrespective of the feeding gene.

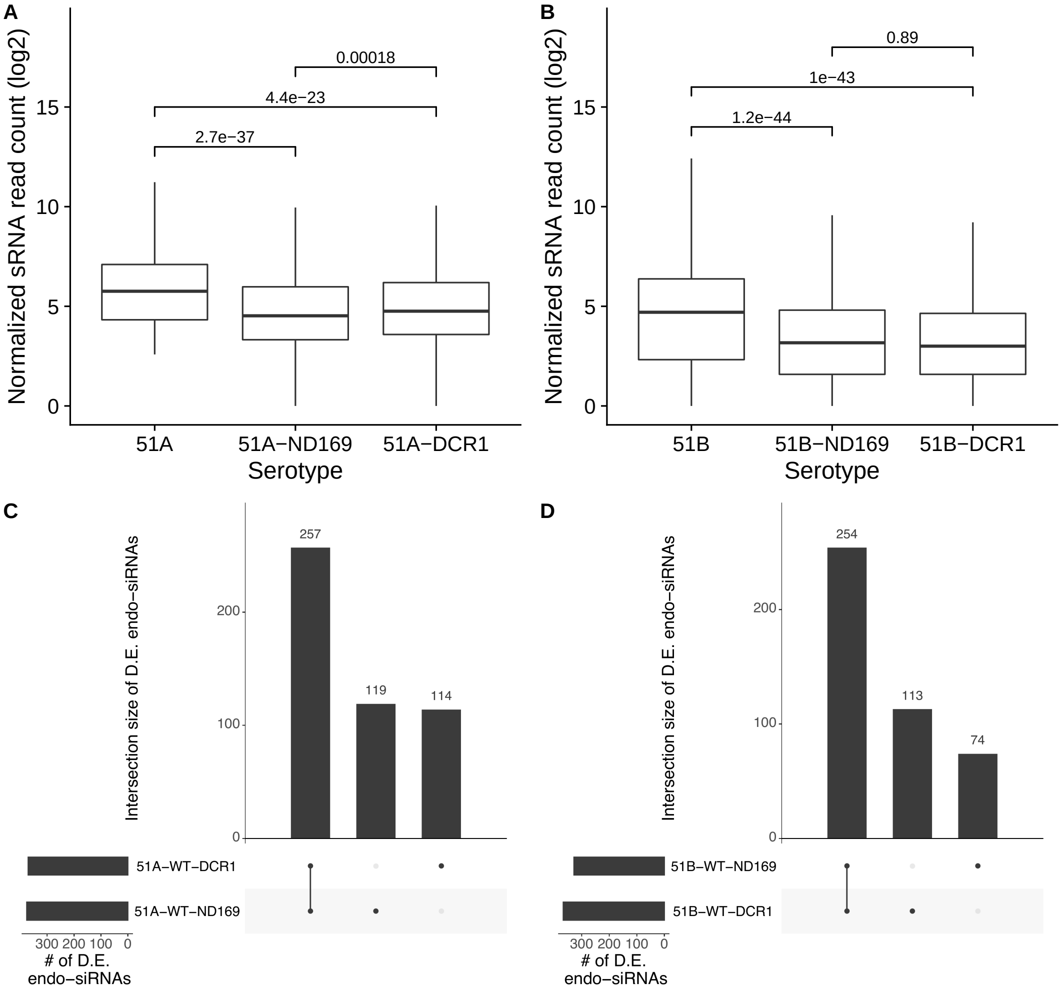


FIGURE 4.9: Quantification and differential expression of the 1,618 endo-siRNAs: (A and B) Boxplots of the 1,618 normalized endo-siRNA read counts (y-axis; log2) of serotype (51A and 51B, respectively) and their knockdowns (ND169 and DCR1). The P-values indicated are from a two-tailed Wilcoxon test. (C and D) Set intersection plots of differentially expressed endo-siRNA clusters in the knockdown serotype against each WT serotype for 51A, and 51B, respectively.

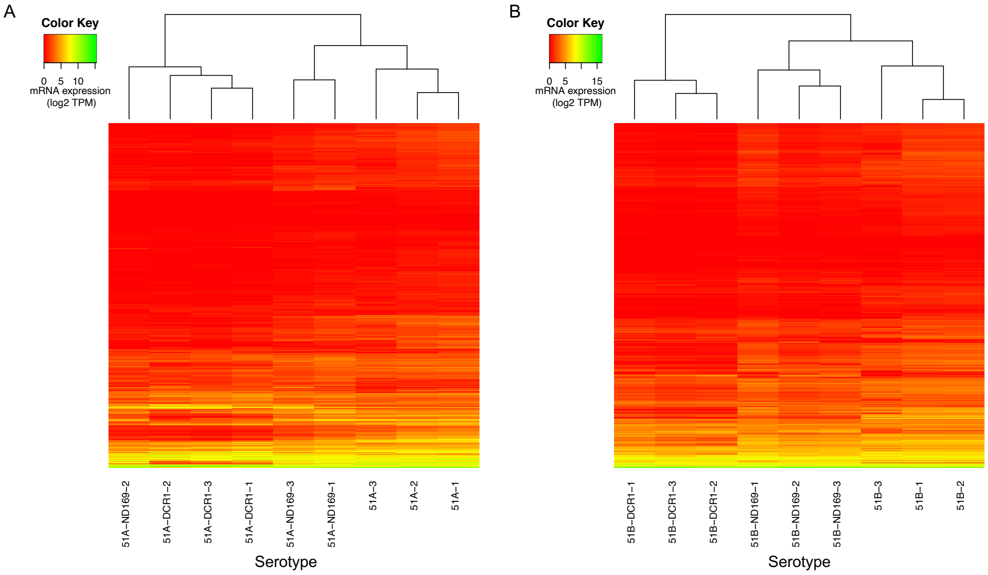


Figure 4.10: Knockdown effects on the transcriptome: Heatmap of mRNA expression after hierarchical clustering (Euclidean distance measure) of all the mRNAs (rows) for all replicates (columns) respective to the serotype 51A (A) and 51B (B) is shown.

4.4.8 Feeding technique deregulates gene expression

In order to investigate the effect of feeding on gene expression, we performed the same set of analysis as described in the previous section, but for the mRNA expression data. A heatmap of the gene expression values (Figure 4.10 A and B) shows the clustering of replicates based on mRNA expression. We observed that there are large changes in the mRNA transcriptome after ND169 and DCR1 feeding. We observed that wildtype samples clustered separately, yet again, in both serotypes. However, the ND169 feeding replicates (except one replicate in serotype 51A) clustered relatively closer to the wildtype replicates than what we observed in endo-siRNA analysis (Figure 4.8 A and B). To observe the quantitative differences among wildtype and feeding samples, we checked the distribution of mRNA expression (Figure 4.11 A and B). We observed a statistically significant reduction (Wilcoxon test; $P-$value < 0.05) of mRNAs in the knocked down samples compared to the wildtype in both serotypes 51A and 51B. In addition, we also observed a statistically significant difference between the median mRNA expression among the knocked down samples. These observations assert why the control feeding samples clustered (Figure 4.10 A and B) closer with wildtype.

Subsequently, we performed differential gene expression analysis between the wildtype, and feeding samples. Figure 4.11 (C and D) shows set intersection plots of the statistically significantly differentially expressed genes for serotypes 51A and 51B, respectively. In both serotypes, DCR1 feeding has the highest and unique set of DE genes. However, approximately 30-40% of the DE genes

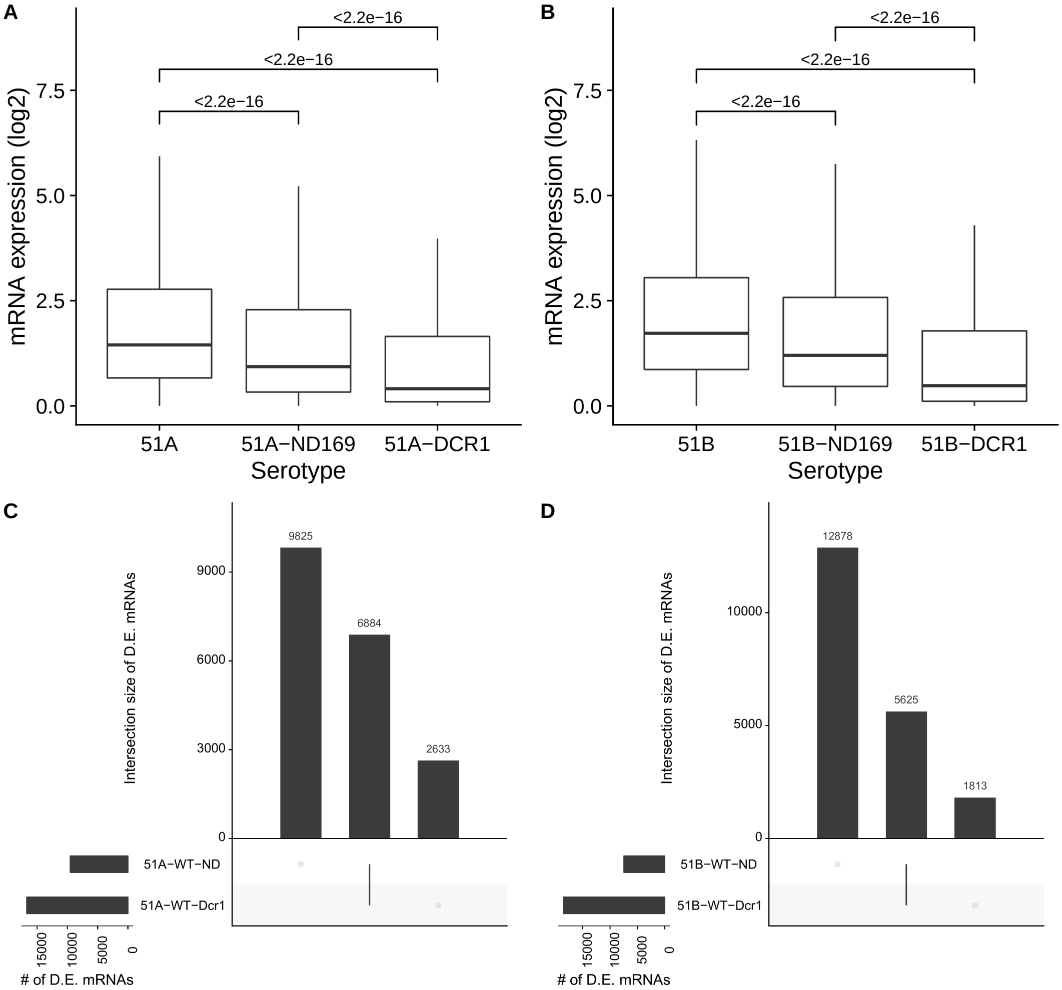


Figure 4.11: Quantification and differential expression of the knockdown transcriptome: (A and B) Boxplots of genome-wide mRNA expression (y-axis; TPM) of serotype (51A and 51B, respectively) and their knockdowns (ND169 and DCR1). Reads mapping to the feeding-associated regions was removed prior to expression quantification. The P-values indicated are from a two-tailed Wilcoxon test. (C and D) Set intersection plots of differentially expressed (D.E.) mRNA in the knockdown serotypes against each WT serotypes for 51A, and 51B, respectively.

in DCR1 are commonly found in the ND169 control feeding as well. It seems likely that genes, which are uniquely differentially expressed in DCR1, are due to direct effects of DCR1 being involved in endo-siRNA accumulation. In contrast, commonly differentially expressed genes are probably deregulated as a response to the feeding process rather than the causal effect of the knocked-down gene.

When we performed a GO enrichment analysis of these common DE genes, we observed diverse sets of biological processes like nucleoside phosphate metabolic process, gene expression, biosynthetic processes, ATPase activity, proteolysis, etc. These results indicate that feeding affects a diverse set of pathways, which seem to be involved in the general depletion of endo-siRNAs and mRNAs that we observe.

4.4.9 Off-target effects are unlikely to cause the observed drastic changes

The knocked down experiments are designed very carefully to avoid hitting unintended targets. Nevertheless, we wanted to verify if the drastic changes in endo-siRNAs and mRNAs we observe in knocked down (or feeding) samples are caused due to off-target effects (see 4.3.13). We found few exact matching off-target genes: five, and two for ND169, and DCR1 feeding regions, respectively. Of them, only a couple of off-targets are differentially expressed (see B.5). With this data we can cautiously conclude that a large number of DE genes are unlikely to be an off-target effect, but a general response to the massive dsRNA feeding.

4.5 Key conclusions

We reported the first genome wide endogenous small RNA profile of the vegetative *Paramecium.* While we did not discover any miRNAs, we identified that many endo-siRNAs are produced from the protein-coding genes. We confirmed the source of endo-siRNAs as mRNAs for a subset of endo-siRNAs, which are commonly expressed in different wildtype serotypes. In contrast to many organisms, we observed that endo-siRNAs do not strictly act in *cis* to silence their parent mRNA. We predicted phased small RNA loci and experimentally verified that they are products of the RNA interference machinery mediated by the enzymes RDR1 and RDR2. Future work is necessary to address the biogenesis, functional pathways of the subsets of endo-siRNAs whose source are not protein-coding genes, and the plausible *trans* acting pathways.

We also documented the aberrations of endogenous small RNA profile, and subsequently the respective transcriptomic alterations in different serotypes of *Paramecium* caused by feeding, a technique widely used to induce RNAi. Our study encourages RNAi researchers to cautiously select appropriate controls, and differentiate between the changes in expression caused by the feeding technique itself, and the actual regulatory changes. Our observation is crucial at a point when RNAi treated organisms are considered as free of Genetically Modified Organisms (GMOs), and being increasingly used to combat viral and pest resistance in bacteria and plants. Our results warrant a cautious systemic investigation, while studying the effects of RNAi using the feeding technique.

4.6 Contributions

The work presented in this chapter includes contribution from several colleagues and collaborators. Sequencing library preparation, knock down and mutant experiments were carried out by our wet-lab collaborators namely (in alphabetical order of last names) Miriam Cheaib, Franziska Drews, Gilles Gasparoni, Jasmin Kirch, Simone Marker, Marcello Pirritano, Angela M. Rodriguez-Viana, Martin Simon, and Raphael de Wijn. Sequenced data sets were trimmed by Karl Nordström. Vidya Oruganti generated the small RNA clusters (and their phasing prediction). All other data and analysis presented in this chapter were performed by me under the supervision of Prof. Dr. Martin C. Simon and Prof. Dr. Marcel H. Schulz.

Chapter 5

Epigenomic characteristics of the *Paramecium* macronuclear genome

This chapter summarises our work aimed at understanding the epigenome of *Paramecium tetraurelia.*

5.1 Background

The role of non-coding DNA, intergenic regions and introns, in gene regulation is widely studied. The intergenic regions host regulatory elements like promoters, enhancers and silencers. These regulatory elements and introns together control gene expression in several organisms (Nelson, Hersh, and Carroll, 2004; Shabalina and Spiridonov, 2004; Elkon and Agami, 2017). More information on non-coding DNA can be found in Chapter 2.

The macronuclear genome annotation of *Paramecium* revealed their coding density to be the highest among free living eukaryotes (Zagulski et al., 2004; Arnaiz et al., 2007). Table 5.1 shows the comparative statistics of different genomic features in *Paramecium* and humans. The highly condensed *Paramecium* genome (Figure 5.1) has short intergenic regions, with an average length of merely 352 bp, and small introns of mean size 25 bp. This raises the question on how the macronuclear genome expression is regulated.

In the previous chapter 4, we described the small RNA landscape of the vegetative MAC and their possible roles in regulating gene expression. Epigenetic inheritance of gene expression such as mating type or serotype determination has been phenotypically observed in *Paramecium* (Chalker, Meyer, and Mochizuki, 2013; Orias, Singh, and Meyer, 2017; Pilling et al., 2017). Extensive studies have also shown the epigenetic orchestration of programmed genome rearrangements in developing macronuclei (Betermier and Duharcourt, 2017). Molecular studies have also shown the role of chromatin assembly factors in programmed DNA elimination (Ignarski et al., 2014). Further, a study characterizing the transcriptomic landscape of the MAC genome alludes an epigenetic regulation of expression (Cheaib et al., 2015). However, the epigenome of the vegetative MAC remains uncharacterized.

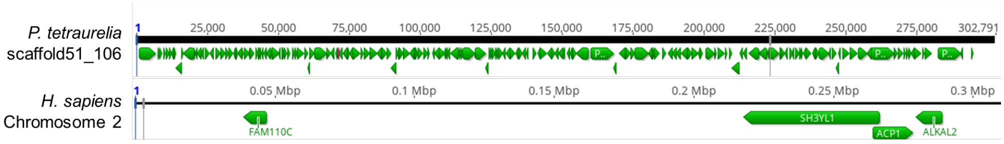


FIGURE 5.1: A genome browser view showing the comparison of intergenic regions in *P. tetraurelia* and *Homo sapiens*. Green arrow heads represent the direction of transcription in the annotate genes. This figure was created by Franziska Drews using the Geneious software `https://www.geneious.com/`.

Category	*P.tetraurelia*	*H.sapiens*
Genome size	72 Mb	3.1 Gb
Protein coding genes	40,460	22,802
Mean gene size	1,084 bp	62,825 bp
Mean intron size	25 bp	3,365 bp
Coding density	80%	3.3%
Mean intergenic size	352 bp	1500 bp

TABLE 5.1: Comparative statistics of different genomic features in *P.tetraurelia* and *H.sapiens*.

5.2 Research objectives

In this context, we defined the following objectives:

1. Does the *Paramecium*'s MAC genome have epigenetic features genome wide?

2. What is the relation of epigenomic changes with *Paramecium* gene expression?

5.3 Data and methodology

All the methods related to the work discussed in this chapter are described here. The biological steps involved in creating our data sets are not discussed, as it is beyond the scope of this thesis. However, the background chapter includes the idea behind these biological methods in brief. All data sets presented in this chapter are from the WT serotype 51A of *P. tetraurelia*.

5.3.1 Expression data

We quantified the mRNA expression of all the *P. tetraurelia* transcripts from the MAC annotation (version 2; stock 51), using Salmon (version 0.8.2; default parameters) (Patro et al., 2017). The mRNA expression data was obtained from ENA with accession number PRJEB9464 (Cheaib et al., 2015).

5.3.2 ChIP-Seq data and preprocessing

We sequenced four replicates each of MNase, H3K4me3, H3K27me3, and H3K9ac. We sequenced one replicate of Pol II ChIP. As a control for ChIP data, we created three replicates of Input. As a control for MNase, we created four replicates of naked DNA. All sequencing data sets were trimmed for adapter sequences using Trim Galore (`https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`), which employs cutadapt (Martin, 2011) using a stringency cutoff of 10. Using deeptools (Ramírez et al., 2016) we performed quality control of replicates (*multiBamSummary* and *plotCorrelation* tools). We downsampled one replicate each of H3K4me3, H3K9ac, and Input ChIP, which had rather high coverage.

5.3.3 Alignments

After preprocessing, all MNase, Pol II and histone ChIP-seq reads were aligned to the macronuclear genome of *P. tetraurelia* (stock 51). We aligned them using the local mode of bowtie2 (Langmead and Salzberg, 2012) with default parameters except the seed alignment mismatch parameter, which was set to 1 (-N 1). These alignments were used in all subsequent analysis except for the analysis mentioned in 5.3.6 and 5.3.7. These exceptions were performed by our collaborator Abdulrahman Salhab, who aligned all reads with GEM mapper (Marco-Sola et al., 2012) to the macronuclear genome of *P. tetraurelia* (stock 51). All replicates were merged in the downstream analysis, unless mentioned otherwise.

5.3.4 Occupancy quantification

For all MNase, Pol II and histone ChIP-seq samples, we used DANPOS2 (Chen et al., 2013) to quantify the respective control normalized occupancy values. We used the *dpos* functionality of DANPOS2 for MNase and Pol II occupancy and the *dpeak* functionality for histone ChIP-seq experiments. Default parameters were used for all functionalities of DANPOS2. Following the occupancy calculations we made use of the *profile* functionality of DANPOS2 to visualise occupancy distributions in a genomic annotation of interest.

Intron visualisation

For all introns, we created a 20 bp window centred on the first and last intron base of the 5'-exon-intron junction (EIJ) and the 3'-intron-exon junction (IEJ), respectively. We plotted the occupancy profile for 200 bp around this window with the centre of x-axis representing the junctions (see 5.4).

5.3.5 Pausing index

We defined a region starting at 30 bp upstream of the Transcription Start Site (TSS) till 300 bp downstream of the TSS as the Transcription Start Site Region (TSSR), and a region starting at 300 bp downstream of the TSS till the transcription end site (TES) as gene body (GB). We removed genes which had fewer than 10 reads in the TSSR. If R is the number of reads (TPM), then Pausing Index (PI) is calculated as the ratio of the reads in the TSSR over GB as shown in the below equation (Gates et al., 2017).

$$PI = \frac{R_{TSSR}}{R_{GB}} \tag{5.1}$$

5.3.6 Signal enrichment visualisation

Using the *plotProfile* and *plotHeatmap* functionality of deeptools2, we created the scaled enrichment plots of different chromatin features. In these plots, the gene body region was scaled to 500 bp unless mentioned otherwise. In Figure 5.5 only one replicate is shown.

5.3.7 Segmentation analysis of chromatin marks

We employed ChromHMM (Ernst and Kellis, 2017) to perform genome wide segmentation of the histone marks (H3K27me3, H3K4me3, H3K9ac), and MNase data. We binarized the genome into 200 bp bins based on a Poisson background model using *BinarizeBam* function. Further, the *LearnModel* function was used to learn a 5 state (-numstates) chromatin state model. Other parameters were set to default. We used the *OverlapEnrichment* function of ChromHMM with default parameters, to annotate the states with respect to the MAC genome annotation (stock 51; version 2) (Arnaiz et al., 2007). To identify genes overlapping with chromatin states at least by 80%, we used *intersectBed* (-f 0.8) function of BEDtools (version 2.23) (Quinlan and Hall, 2010).

5.4 Results and discussion

5.4.1 Well positioned nucleosomes are found near the TSS

To investigate the distribution of nucleosomes in the MAC genome of *Paramecium*, we performed MNase-seq analysis. Using the DANPOS2 software, we identified the positions where nucleosomes are enriched. We observed a well positioned +1 nucleosome located ≈ 100 bp downstream of the TSS (Figure 5.2A). As shown in Figure C.1, the shifted +1 is not uncommon in other organisms (Mavrich et al., 2008; Xiong et al., 2016). In contrary to many organisms, we found that the subsequent nucleosomes (+2, +3, etc.) in the gene body are less prominent. This does not seem to be an averaging effect on diverse gene length groups (Figure 5.2B). Interestingly, we also found a prominent peak before the +1 nucleosome. We investigated whether this is a -1 nucleosome or an effect of short intergenic regions.

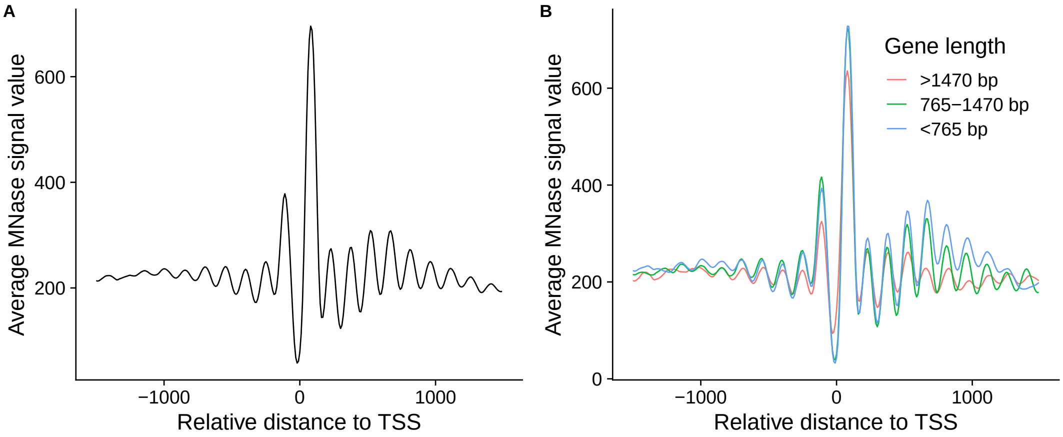


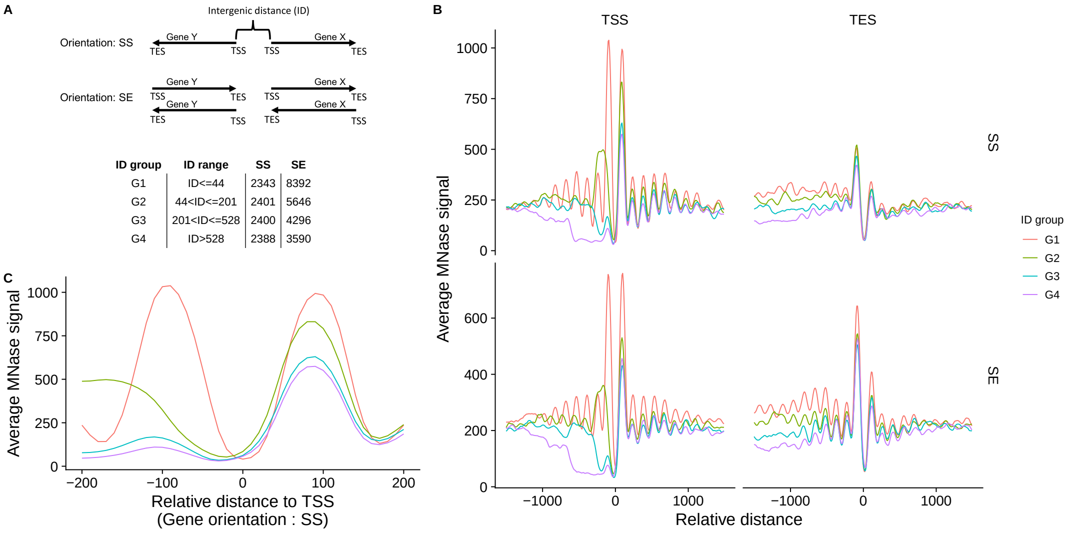
FIGURE 5.2: Nucleosome characteristics: A) The average MNase signal (y-axis) of all genes relative to their distance to TSS (x-axis) is shown for *Paramecium*. B) Same as A, but genes are split in different length groups (in bps).

5.4.2 Nucleosome positioning is influenced by intergenic distance

We categorized genes into different subsets based on their orientation and intergenic distance to their upstream gene. Figure 5.3A shows a schematic of the start-to-start (SS), and start-to-end (SE) orientation. Further, we also created four sub-groups of the SS and SE genes based on different intergenic distance cutoffs, which can be seen in the table in Figure 5.3A. We plotted the average MNase signal (Figure 5.3B) for the different gene subgroups in relative distance to TSS and TES position. Although the nucleosomes in the gene body were less prominent, we observed well positioned nucleosomes also at the TES. In Figure 5.3, we can observe that the abundance of the -1 nucleosome decreases with increasing intergenic distance in both SS and SE oriented genes. This shows that they are the +1 or the last nucleosome of the upstream gene, if they are in SS or SE orientation, respectively. We also observed that the +1 nucleosome shifts further downstream of the TSS with shorter intergenic distances (Figure 5.3C). These results show that intergenic distance influences the nucleosome positioning.

5.4.3 Expression levels are proportional to nucleosome abundance

Next, we investigated how the nucleosome abundance relates to gene expression. So, we categorized the genes into three gene expression groups as 19,090 high expressed ($TPM > 2$), 20,001 low expressed ($0 < TPM <= 2$) and 1,369 silent ($TPM = 0$) genes. Figure 5.4 shows that the abundance of nucleosomes is proportional to the gene expression groups in three gene structural elements namely TSS, introns and TES. The silent genes have very feeble amount of nucleosomes. The introns are mostly free of nucleosomes, but we observed



ID group	ID range	SS	SE
G1	ID<=44	2343	8392
G2	44<ID<=201	2401	5646
G3	201<ID<=528	2400	4296
G4	ID>528	2388	3590

FIGURE 5.3: Intergenic distance influences nucleosome positioning: A) Schematic representation of the gene sub-groups we used in the analysis. The table shows the counts of genes in the respective intergenic distance groups of each gene orientation. B) The average MNase signal (y-axis) of all genes relative to their distance to TSS or TES (x-axis) is shown for different gene orientation groups based on their intergenic distance. C) Same as B, but the x-axis showing relative distance to TSS is zoomed in for the SS orientation gene groups.

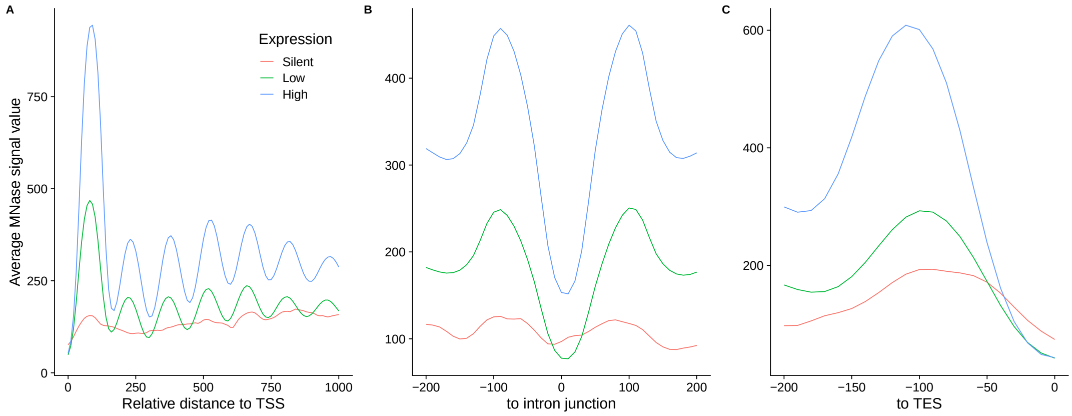


FIGURE 5.4: Nucleosome abundance correlates with gene expression: The average MNase signal (y-axis) of different gene expression groups are shown relative to their distance (x-axis) to TSS (A), intron junction (B) and TES (C).

two nucleosome peaks at the intron boundaries (Figure 5.4B). The *Paramecium* genome is known to be alternative splicing free. Nevertheless, it has been shown that efficient intron definition is crucial for splicing efficiency (Jaillon et al., 2008; Saudemont et al., 2017). Hence, the well positioned nucleosomes at the intron boundaries suggest that nucleosome positioning is likely important for efficient splicing. Additionally, it also suggests that during evolution the mean intron size of 25 bp is possibly influenced by the length of the linker DNA between the nucleosomes.

5.4.4 *Cis*-determination of epigenetic marks

In addition to the MNase, we sequenced histone-ChIP (H3K4me3, H3K9ac and H3K27me3) and Pol II-ChIP. Consequently, we wanted to investigate how these different epigenetic marks are distributed in a gene specific manner. We also investigated the GC content distribution, as they were shown to influence nucleosome positioning in other organisms (Fenouil et al., 2012). To this end, we created input normalized signals using deepTools. Figure 5.5 shows a heatmap of the signals along the gene body (scaled to 500 bp), and 1000 bp up/downstream of the gene body in decreasing order of gene length in each expression category. An occupancy profile of the epigenetic marks, MNase, Pol II, and GC content can be seen in Figure C.2. First, we found that there is a clear phasing of nucleosomes along the gene body except for the silent genes. Similarly, we observed clear phasing in all the epigenetic marks, Pol II and GC content as well. In all the heat maps, we observed a signal depletion at the gene boundaries (TSS and TES). Subsequently, we checked for the pairwise Pearson correlation of the normalized gene body signals (Figure C.3) of the epigenetic marks, MNase, Pol II and GC content. We found that the GC content has the highest positive correlation with Pol II ($r = 0.77$) and MNase ($r = 0.73$).

However, studies have reported that the combination of the size selection step in MNase-seq and the preference of MNase enzyme to cut at GC-rich regions can lead to false association of GC content and nucleosome placement (Chung et al., 2010). In order to verify that, we created 147 bp bins of the genome. Next, we calculated the Pearson correlations of the read counts of raw MNase, and naked DNA in these bins with their GC content (Figure C.4). We observed a significantly higher correlation of GC content with the raw MNase ($r = 0.6$) than the naked DNA ($r = 0.3$). These observations indicate that the GC content plays a dominant role in determining the positioning of epigenetic marks in an AT-rich *Paramecium* genome, similar to other organisms (Fenouil et al., 2012).

An intriguing aspect is that high expressed genes have high H3K27me3 (in other organisms associated with repressed genes) content spread throughout gene body (Figures 5.5 and C.2). Also, H3K27me3 shows a strong positive correlation with other epigenetic marks, GC content and Pol II as well (Figure C.3). This suggests that H3K27me3 possibly also plays a non-repressive role in *Paramecium*'s macronuclei.

5.4.5 Lack of measured epigenetic signals in predominant loci

To gain insights into combinatorial distribution of epigenetic signals genome wide, we used ChromHMM and identified chromatin states. Figure 5.6A shows a heatmap of the identified chromatin states with the signal contribution of each epigenetic mark to the respective state. Figure 5.6B shows the annotation of each state in different genomic elements, with the enrichments specific to each column. Further, we analysed the mRNA expression of genes (Figure 5.6C) which overlap with a chromatin state by at least 80%. We note that because of this overlap condition, we can analyse only ≈ 30% of the protein-coding genes.

Active state

State-1 comprises of genomic loci with increased coverage of activation-associated marks (H3K4me3, H3K9ac), decreased coverage of repression-associated mark (H3K27me3) and MNase (Figure 5.6A). The segment annotation of State-1 shows that they are enriched in TSS regions (Figure 5.6B), with a median expression (Figure 5.6C) of ≈ 0.8 TPM (log10). In other organisms, H3K4me3 and H3K9ac are often associated with active gene expression (Bannister and Kouzarides, 2011).

Repressive state

Genomic loci with high H3K27me3 signal comprise State-3, which is in contrast to State-1. State-3 is highly enriched in exonic regions. Although we only have 97 genes in this state, they show a large range of mRNA expression values. This further supports our observation that H3K27me3 likely plays a non-repressive role.

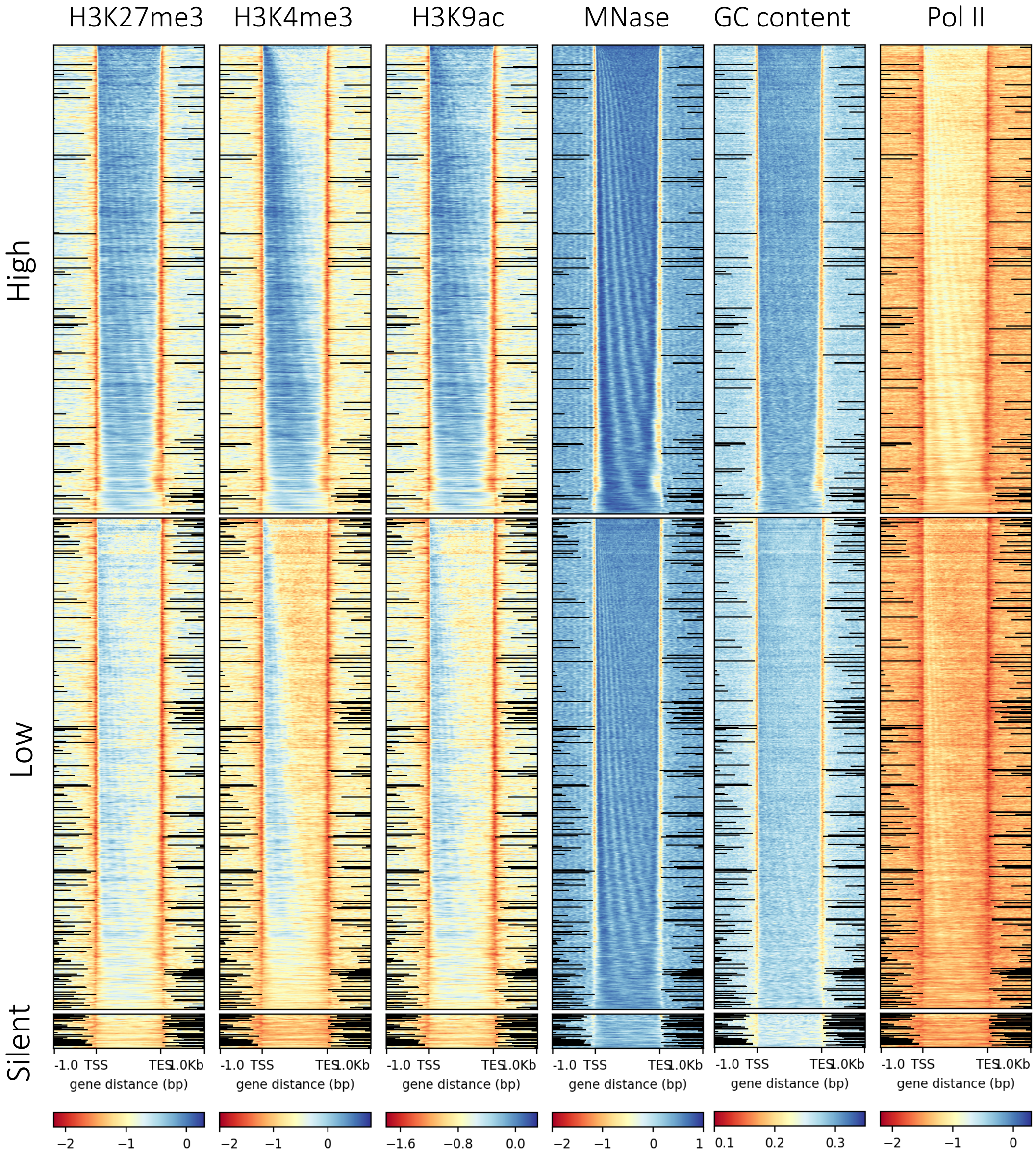


FIGURE 5.5: A heatmap showing the signal distributions of H3K27me3, H3K4me3, H3K9ac, MNase, GC content, and Pol II stratified according to gene expression groups. The signal shown is for the gene body (TSS to TES; scaled to 500 bp) and 1 Kbp upstream and downstream of gene body. This figure was created by Abdulrahman Salhab.

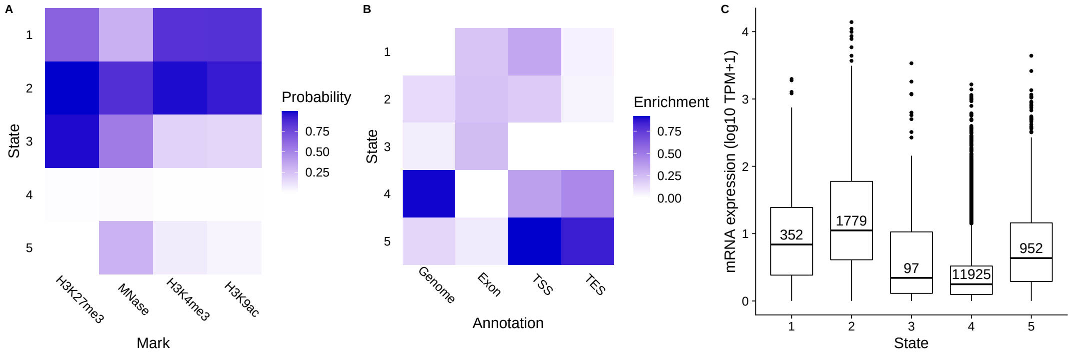


FIGURE 5.6: Identification of chromatin states: A) The chromatin states predicted by ChromHMM are shown as a heatmap. Each row corresponds to a chromatin state, and each column represents different epigenetic marks. Darker colour of an epigenetic mark in a particular state indicate that there is a higher probability of observing that epigenetic mark in that state. B) Heatmap showing the enrichment of a chromatin state (row) in different genomic annotations (columns). Enrichment values are obtained from overlap enrichment functionality of ChromHMM with a column specific colour scale (min-max scaled). C) Box plots showing the mRNA expression (y-axis; log10 TPM+1) of genes whose loci overlap at least by 80% with the respective chromatin states. This figure's elements A and B were recreated by me, by the ChromHMM segmentation performed by Abdulrahman Salhab.

Nucleosomal state

State-5 predominantly is composed of MNase and faint H3K4me3 signals. The State-5 annotation shows the highest enrichment among the TSS and TES regions (Figure 5.6B). The genes in State-5 shows a wide range of mRNA expression values (Figure 5.6C), although lower than States 1 and 2. This coincides with our earlier observations that MNase signals correlate with different gene expression groups (Figures 5.4 and 5.5).

Ambiguous states

The states 2 and 4 present rather peculiar patterns. State-2 has all the analysed epigenetic marks, while State-4 has almost no epigenetic marks (very faint MNase signal can be seen).

Interestingly, the entire genome is enriched for State-4 (Figure 5.6B) and it shows the lowest median expression of all states (Figure 5.6C). Subsequently, we investigated the GC content distribution of all the genes, whose loci overlap at least by 80% with the respective chromatin states (Figure C.5). We observed that the genes in State-4 has the lowest median GC content. In spite of the library optimisation strategies to minimise GC-bias, AT-rich genomic regions may still pose a challenge in NGS sequencing and alignment (Browne et al., 2020). In addition, all our ChIP- and MNase-seq experiments used in the segmentation analysis are normalised against control experiments. Nevertheless, there are likely AT-rich genomic regions where neither the ChIP/MNase nor the control experiment performed well, leading to sequencing and alignment artifacts. Hence, the AT richness of the genes in State-4 is likely a contributing factor to their observed low median expression (Figure 5.6C).

On the contrary, State-2 where both active and repressive marks are present shows the highest median expression. In multicellular organisms, State-2 is usually attributed to bivalent domains, which orchestrate cell differentiation during development through paused expression states (Voigt, Tee, and Reinberg, 2013; Sen et al., 2016; Blanco et al., 2020).

5.4.6 Paused genes are enriched in chromatin state with lowest expression

In order to investigate whether pausing occurs in *Paramecium*, we calculated a Pol II pausing index (PI) as shown in Equation 5.1, after removing genes which did not have at least 10 reads in the Transcription Start Site Region (TSSR). We labelled genes with a $PI > 1.5$ as Paused, and the rest as Not paused. The Pol II profile (Figure 5.7A) shows that the abundance of Pol II steadily decreases after the TSS. We also observed that there is a statistically significant difference in the median expression of 8,480 paused and 26,395 not paused genes (Figure 5.7B).

Further, we investigated the number of paused genes in each chromatin state (Figure 5.7C). Studies have shown that bivalent domains influence the pausing of gene expression (Blanco et al., 2020). In contrast, we did not find an enrichment of paused genes in our bivalent domains (State-2). However, we found

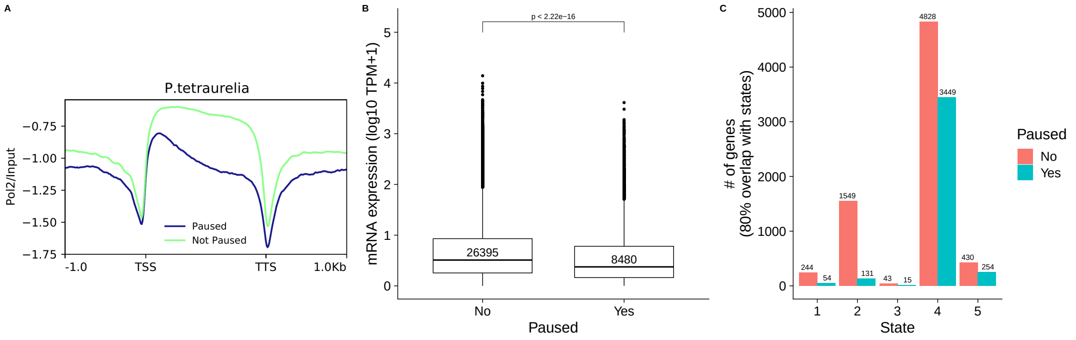


FIGURE 5.7: Pol II pausing occurs in *Paramecium*: A) Distribution profile showing the Pol II signal stratified according to the pausing status of genes. The signal shown is for the gene body (TSS to TES; scaled to 500 bp) and 1 Kbp upstream and downstream of the gene body. B) Box plots showing the mRNA expression (y-axis; log10 TPM+1) of genes which are paused and not paused. C) For the genes whose loci overlaps at least by 80% with a chromatin states we show the number of paused/unpaused genes and their The P-values indicated are from a two-tailed Wilcoxon test.

that chromatin State-4 to be enriched in paused genes. A chi-squared significance test for enrichment was found to be statistically significant $P < 2.2e-16$ (Figure C.6). With our analysis, the role of bivalent domains is unclear in *Paramecium*. Nevertheless, an enrichment of paused genes in the State 4 is interesting, as State-4 also has the lowest MNase signal (Figure 5.6A), and lowest median GC% (Figure 5.7D). This suggests that GC content is likely an influence in keeping the chromatin open for transcription similar to other organisms (Fenouil et al., 2012). However, as discussed in the previous Section 5.4.5, we cannot rule out that the AT richness of the genes in State-4 is confounding their low median expression, and low MNase signal.

5.5 Key conclusions

We report the first insights into the epigenomic characteristics of the macronuclei of *Paramecium*. We have shown that their nucleosomes are well positioned and shifted downstream of the TSS. The epigenetic marks (H3K4me3, H3K9ac, and H3K27me3), and Pol II are located along the gene structure and are less enriched in the non-coding regions around the gene boundaries. GC content may influence, in *cis*, the positioning of epigenetic marks, and Pol II. The introns are flanked by nucleosomes, which suggests a role in splicing regulation.

The highly condensed genome of *Paramecium* has several chromatin states. We characterized bivalent chromatin domains, whose role is unclear. We found the genome to be enriched in regions, which lack the measured epigenetic signals. We found that there are paused genes, which show lower median gene expression than the not paused genes. We acknowledge that our analysis raises novel questions on the epigenetic regulation in *Paramecium*'s MAC, which can be addressed only with future experiments.

5.6 Contributions

The work presented in this chapter includes contribution from several colleagues and collaborators. Sequencing library preparation, and sequencing were carried out by our wet-lab collaborators namely (in alphabetical order of last names) Miriam Cheaib, Franziska Drews, Gilles Gasparoni, Jasmin Kirch, Martin Jung, Simone Marker, Marcello Pirritano. Sequenced data sets were trimmed by Karl Nordström. Figures 5.5, C.2, and the chromatin state segmentations were done by Abdulrahman Salhab. All other data analysis presented in this chapter were performed by me under the supervision of Prof. Dr. Martin C. Simon and Prof. Dr. Marcel H. Schulz.

Chapter 6

Statistical analysis of macronuclear gene expression regulation in *Paramecium*

6.1 Background

In chapter 4, we characterised endo-siRNAs and showed that mRNAs are a predominant source of the small RNAs in the Genes associated with Small RNA Clusters (GSRCs). Correlation analysis of the endo-siRNAs and the mRNAs revealed that endo-siRNAs do not strictly act in *cis* to silence their parent mRNA. In chapter 5, we reported about the epigenomic characteristics of the vegetative MAC, and showed that abundance of the epigenetic marks are enriched in the genic regions. We also showed that the epigenetic marks are proportional with different gene expression groups (high/low/silent). Nevertheless, there are many aspects that are not clear from the simple analyses in the previous chapter. Hence, we wondered whether we can resort to more advanced statistical and machine learning methods to improve our understanding of the epigenetic regulation of macronuclear gene expression in *Paramecium*.

Researchers have employed machine learning methods to classify gene expression or predict gene expression levels quantitatively in other organisms. These methods can be broadly classified in two groups: methods, which use (i) DNA sequence features and (ii) epigenomic features. The methods which use the DNA sequence features, predominantly depend on capturing the sequence variations in large windows upstream of the TSS (Beer and Tavazoie, 2004; Vilar, 2010; Bessière et al., 2018). A recent study, using deep learning methods, was able to quantitatively predict gene expression merely from the DNA sequence in the promoter regions, in humans and mouse (Agarwal and Shendure, 2020). However, the window sizes around the TSS are usually several kilo base pairs. *Paramecium* has a mean gene size of only 1,084 bp. Hence, such large window sizes will create a lot of gene overlaps, and likely confound the prediction algorithm. If we remove overlapping genes from the analysis, a vast majority of the genes will be left out from our analysis, as the MAC genome is 80% protein coding with an average intergenic region of size 352 bp.

The latter group of methods, uses epigenomic features like histone marks and Pol II data. One such example is the study using support vector machines to classify protein coding gene expression in humans (Cheng et al., 2011). Another study constructed a random forests model on histone marks data from humans to classify gene expression as high or low (Dong et al., 2012). DeepChrome is a recent approach which applies deep learning methods on histone marks data to efficiently understand the complex interactions of histone marks and gene expression (Singh et al., 2016). Many more studies have also shown that epigenetic data is highly predictive of gene expression (Karlić et al., 2010; Huang et al., 2011).

6.2 Research objectives

In this context, we wanted to explore the following objectives:

1. How are epigenetic signals, small RNAs, and gene expression related to each other quantitatively?

2. Can we build an association network of epigenetic signals, small RNAs, and gene expression?

3. How accurately can we classify high and low expressed genes by applying machine learning on our data?

4. Can we identify general patterns of high/low expressed genes using model interpretation methods?

6.3 Data and methodology

6.3.1 RNA data

We used the WT serotype 51A data described in the Section 4.3.1. We quantifed the sRNA accumulation (in TPM) of all genes in this data set using our RAPID software (See Chapter 3). Similarly, we used the mRNA data of serotype 51A as described in Section 4.3.7.

6.3.2 Epigenetic data

We used all the ChIP data sets produced as part of the previous chapter described in Section 5.3.2. We used the occupancy values as calculated in Section 5.3.4. Further, we used bedtools to count the reads in different bins around the TSS and TES, as shown in Figure 6.1. All bin signals were normalised to the bin size and log2 transformed with a pseudocount of 1.

6.3.3 Sparse partial correlation network analysis

We performed the SPCN analysis using the scripts, with default parameters, made available by the authors `http://spcn.molgen.mpg.de/index.html`.

6.3.4 Classification of gene expression

The workflow of our machine learning based gene expression classification task is presented in Figure 6.4.

Data and feature definition

We removed the $1,369$ silent genes ($TPM = 0$), as it will lead to a heavily unbalanced data set. We used the remaining $19,090$ high expressed ($TPM > 2$), and $20,001$ low expressed ($0 < TPM < 2$) genes in our analysis. For these genes we defined different feature sets. We used the normalised gene body based signals of the following as *Epigenetic* features: H3K4me3, H3K27me3, H3K9ac, Pol II, MNase, and H3K4me3/H3K27me3 ratio. We used the following *Genic* features: gene length, intron frequency, intergenic length, and GC content of genes. We joined the *Epigenetic* features, *Genic* features and sRNA accumulation in genes (TPM) as a super set called *All* features. We reserved 10% of the high and low expressed genes to test our model, and used the rest for model training.

Model/classifier selection

We used the Experimenter application of WEKA machine learning workbench (version 3.8.4) (Frank, Hall, and Witten, 2016) to explore different machine learning algorithms: decision tree, neural networks (one layer with three sigmoid neurons), and random forest. The parameters of each algorithm were left to default. Each experiment was performed 10 times with different random seeds. Further, each experiment followed a 10-fold cross validation (CV) approach. Hence, we created 100 models for each algorithm. We evaluated the performance of the models using the area under the precision-recall curve (PR-AUC), and chose the best classification algorithm for subsequent steps. The steps described here are shown in green boxes in Figure 6.4.

Feature set evaluation and model interpretation

We trained three different models for the different feature sets (*All*, *Epigenetic*, *Genic*), using the best performing algorithm identified in previous step. We evaluated these models on the test data using the PR-AUC metric, to choose our final classifier. We implemented these steps using the scikit-learn module (Pedregosa et al., 2011) for python3. On the final classifier, we applied the TreeExplainer algorithm available from the SHAP python module (Lundberg et al., 2020), to explain the importance of each feature in our test data. The steps described here are shown in yellow boxes in Figure 6.4. We switched to python for these steps of our workflow, as the SHAP package was unavailable for WEKA.

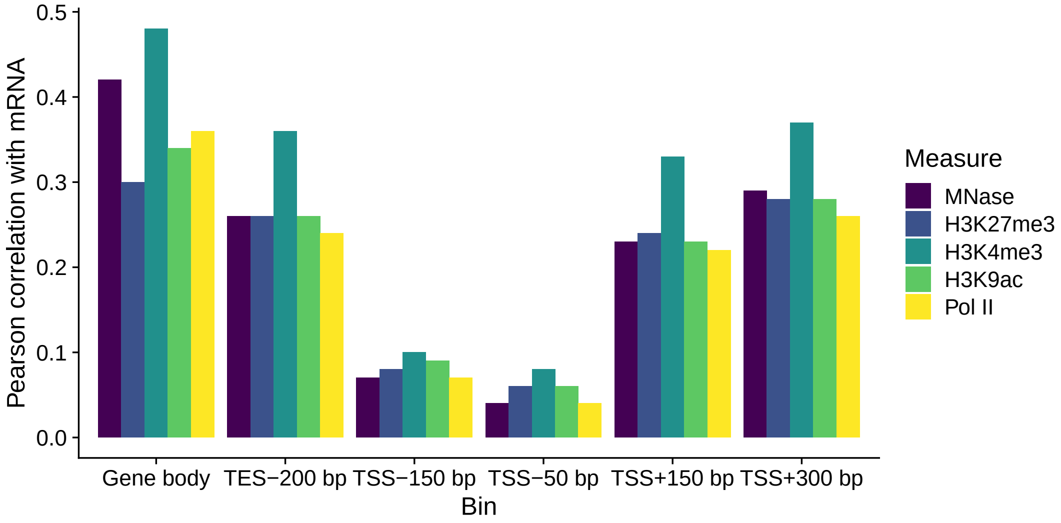


FIGURE 6.1: Epigenetic signals in the gene body correlate with gene expression: A bar chart showing the Pearson correlation coefficient (y-axis) of mRNA and different measurements (colors; see legend) in different bins (x-axis).

6.4 Results and discussion

6.4.1 Epigenetic signals in the gene body correlate with gene expression

Studies have shown the enrichment of various histone marks in different gene regions like promoters, and gene body and their correlation to active gene expression or silent genes (Cheng et al., 2011; Dong et al., 2012). We have seen earlier that our histone marks showed enrichment in different regions, around the TSS or gene body (Figure 5.5). Hence, we first set out to systematically investigate the correlation of gene expression with epigenetic signals in different bins of the annotated gene region. The bins were designed such that we can capture the variations in the enrichment of the different measurements (MNase, histone marks, and Pol II) in different genic loci. We considered the following bins: 50 bp or 150 bp upstream of the TSS (TSS$-X$ bp), 150 bp or 300 bp downstream of TSS (TSS$+X$ bp), gene body (TSS to TES), and 200 bp upstream (in the gene body) of TES (TES$-$200 bp). Figure 6.1 shows the Pearson correlation of the normalised epigenetic marks, MNase and Pol II in different bins with the mRNA expression. We observed a positive correlation in all the bins, with gene body signals showing the highest correlation ($0.28 < r < 0.46$) for all epigenetic marks, MNase and Pol II. This supplements our earlier observation (Figure C.2) that for most genes the epigenetic signals, MNase, and Pol II were spread through out the gene body. Hence, the subsequent analysis were performed using the gene body bin based signals.

6.4.2 Multicollinearity of epigenetic measurements

Since the gene body signals of all the epigenetic marks showed positive correlation with mRNA expression, we aimed to understand whether there is multicollinearity among the different epigenetic measurements. In other words, we wanted to investigate whether there is a linear relationship between some of the measured epigenetic marks. We included sRNA accumulation in the gene body of the respective genes as well in our analysis to gain further insights into the regulatory roles of sRNA.

Figure 6.2 shows a pair plot of all the variables we investigated. The diagonal of this plot shows a histogram of the signal distribution of individual variables. The scatter plots (below the diagonal), show the relationship of two variables and their respective Pearson correlation values are shown above the diagonal. We observed strong positive correlations between all histone marks, MNase, and Pol II ($0.65 < r < 0.85$), showing their multicollinearity. Alongside the positive correlation of mRNA with epigenetic marks ($0.28 < r < 0.46$), we also observed a positive correlation of sRNA accumulation in gene body ($r = 0.43$) with mRNA. This supports the idea that, genome-wide, small RNAs are involved in gene regulation, as we observed earlier in Chapter 4.

In addition, we found a positive correlation of GC content and mRNA ($r = 0.45$), alongside the other epigenetic marks ($0.69 < r < 0.77$). We have also seen earlier that GC content likely determines the placement of nucleosomes, and other epigenetic marks (Section 5.4.4). These results together suggests that the GC content plays a major role in orchestrating the regulatory landscape of the AT-rich *Paramecium* genome.

Interestingly, we found a strong positive correlation ($r = 0.74$) of H3K27me3 and H3K4me3. However, we observed that the positive correlation of H3K27me3 with mRNA ($r = 0.28$) is lower than that of the H3K4me3 with mRNA ($r = 0.46$). Such strong correlation between H3K27me3 and H3K4me3 posed a question, whether there is some cross-reactivity of histone antibodies used in the experiments. Our collaborators verified this and found no evidence of antibody cross-reactivity (Figure D.2).

The high correlation, alongside the bivalent marks (H3K4me3 and H3K27me3) we observed in Figure 5.6, can in part be due to the bulk measurements made in thousands of *Paramecium* cells with a $800n$ polyploid genome. In order to verify this, we calculated a log ratio of H3K4me3 and H3K27me3 (Figure D.1A), and found it to be positively correlated ($r = 0.24$) with mRNA expression. We refer to the ratio as K4K27ratio from here on. Figure D.1B shows the distribution of mRNA expression in different K4K27ratio groups. Only 7700 genes ($\approx 20\%$) showed equal amounts of H3K4me3 and H3K27me3 (log K4K27ratio $= 0$), whose median mRNA expression value is the lowest of all groups. Figure D.1B shows that 20438 genes ($\approx 51\%$ of the genome) have higher H3K27me3 signal than H3K4me3 (log K4K27ratio < 0), whose median expression is statistically significantly lower than the genes with higher H3K4me3 signal (log K4K27ratio > 0; 10920 genes). These results suggest that our bulk measurements on polyploid cells are likely contributing to the observed multicollinearity.

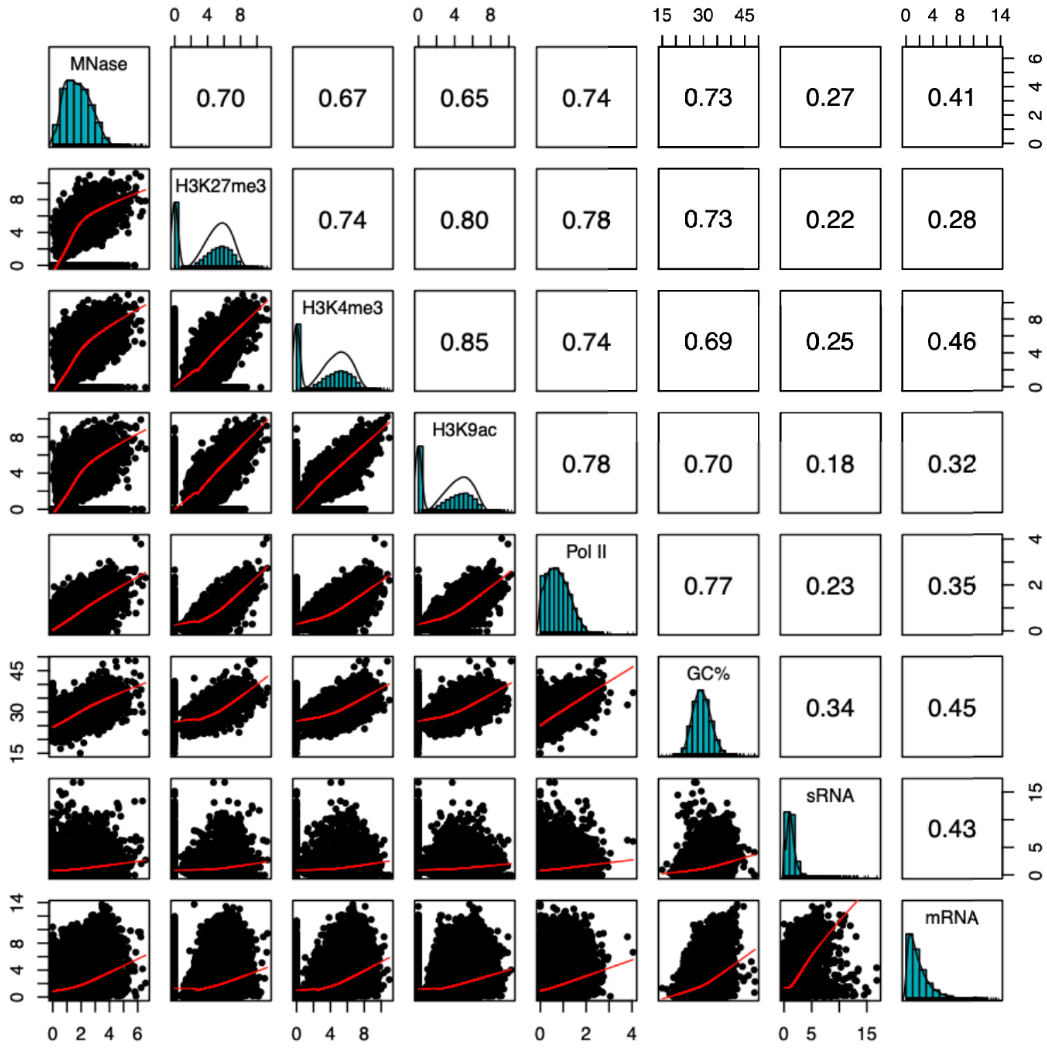


FIGURE 6.2: Multicollinearity of epigenetic marks: The distribution plot of the gene body signals of MNase, epigenetic marks, Pol II, GC content, sRNA expression, and mRNA expression are shown along the diagonal. The Pearson correlation coefficients (above the diagonal) are shown for the respective variables mentioned along the x- and y-axis of each box. The y-axis of scatter plots (below the diagonal) belongs to the variable mentioned along the horizontal line of that plot. mRNA and sRNA are measured in TPM units. All values shown are log2 transformed with a pseudocount of 1, except for the GC content.

6.4.3 Sparse partial correlation network of epigenetic marks, sRNA, and gene expression

Correlations can show the relationship between two variables, but they do not reveal causality and are susceptible to confounding factors. Collinear variables can be seen as an extreme case of confounding. We have seen our data is multicollinear, which makes the interpretation of our correlations of epigenetic marks with gene expression challenging. However, we can resort to partial correlation. Partial correlation removes the effect of an observed confounding variable from the two variables, whose partial correlation is calculated. We denote partial correlation coefficient as $r_{xy,z}$, where x and y are the variables of interest, and z represents one or more confounding variables. We aimed to construct a network of the MNase, histone marks, Pol II, mRNA and sRNA with their edges showing true regulatory connections, which cannot be explained by confounding factors. To this end, we used the sparse partial correlation network (SPCN) method, a special case of partial correlation, on our data (Lasserre, Chung, and Vingron, 2013). The SPCN method favours precision over completeness. Hence, the absence of edges should be interpreted with caution.

Commonly known association of epigenetic marks with gene expression are revealed

Figure 6.3 shows the network constructed by the SPCN method. The blue and red edges represent positive and negative sparse partial correlation values, which are shown as edge labels. We observed 13 out of 21 possible edges in our network, as shown in the Figure 6.3. The edges MNase-H3K4me3 ($r_{xy,z} = 0.12$), MNase-Pol II ($r_{xy,z} = 0.26$), MNase-mRNA ($r_{xy,z} = 0.14$), and H3K4me3-mRNA ($r_{xy,z} = 0.36$) are very much expected. The importance of open chromatin in gene expression is well known for over a decade (Li, Carey, and Workman, 2007). Similarly, the association of H3K4me3 with Pol II mediated active gene expression is widely studied (Bernstein et al., 2005; Liu et al., 2005; Zhang et al., 2009). Surprisingly, we did not observe an edge between Pol II and mRNA in our data, meaning their $r_{xy,z} = 0$. However, we observed earlier a Pearson correlation coefficient of 0.39 between Pol II and mRNA. Subsequently, we checked the contribution of confounding factors causing the difference between the Pearson correlation coefficient and the partial correlation coefficient (Figure D.3). We observed that H3K4me3 is the major confounding factor, not only between Pol II and mRNA, but for most of the associations.

We also noticed a commonly observed association, H3K9ac-H3K4me3, with a strong partial correlation ($r_{xy,z} = 0.58$). H3K9ac is known to be associated with active gene expression, similar to H3K4me3. The co-occurrence of the histone marks H3K9ac and H3K4me3 has also been shown in mammals using Co-ChIP measurements (Weiner et al., 2016). We also found an edge between H3K9ac and Pol II ($r_{xy,z} = 0.26$). In other organisms, H3K9ac is associated with the pause-release cycle of Pol II (Gates, Foulds, and O'Malley, 2017).

We found an association of H3K27me3 and MNase ($r_{xy,z} = 0.3$), alongside few unusual associations: H3K9ac with H3K27me3 ($r_{xy,z} = 0.32$), and

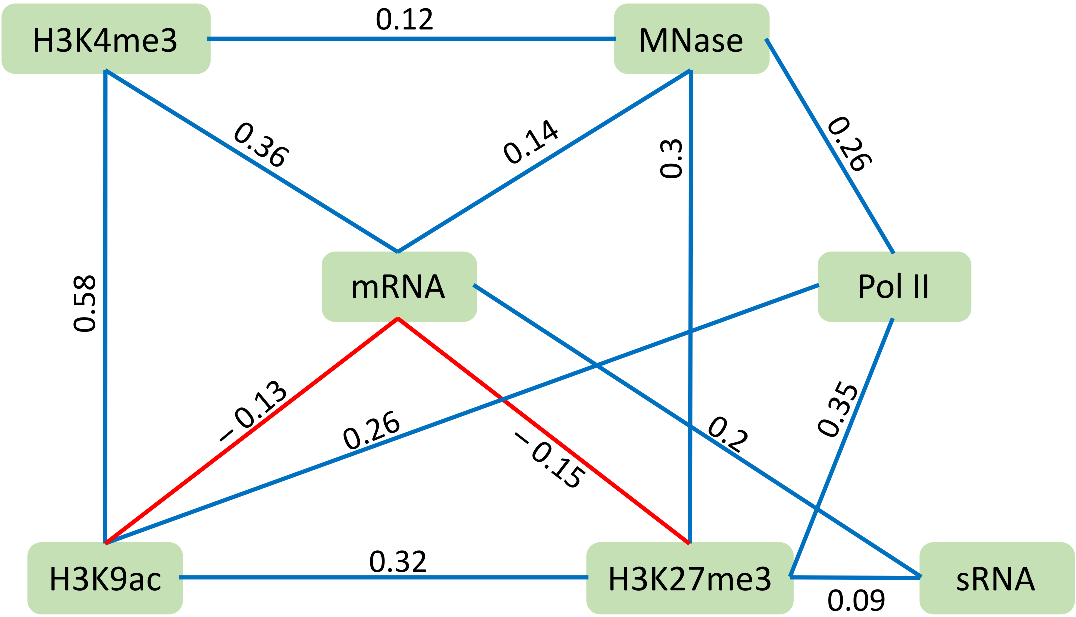


Figure 6.3: The sparse partial correlation network identified by the SPCN method is shown. The nodes show variable names (histone marks, Pol II, MNase, mRNA, and sRNA). The edges show a true association between the variables. The partial correlation coefficients are mentioned along side the edges. The edges with positive (or negative) partial correlation coefficients are shown in blue (or red).

H3K27me3 with Pol II ($r_{xy,z} = 0.35$). We questioned the repressive role of H3K27me3, based on two of our earlier observations: (i) H3K27me3 is spread over the whole gene body, and (ii) is directly proportional to different gene expression groups (please see Figures 5.5 and C.2). In that light, it makes sense to see a positive partial correlation coefficient between H3K9ac, H3K27me3, MNase, and Pol II.

Associations with small RNAs

We also observed two edges for sRNA, one with mRNA ($r_{xy,z} = 0.2$), and another with H3K27me3 ($r_{xy,z} = 0.09$). The sRNA-mRNA edge suggests that genome wide the sRNA accumulation is relevant for regulating gene expression. This coincides with our earlier observation that endo-siRNAs show a predominantly positive correlation with mRNA, suggesting that they do not strictly act in *cis* to silence their parent mRNA (see Section 4.4.4).

Similarly, the sRNA-H3K27me3 edge suggests an alternative pathway of our sRNAs to mediate gene expression by collaborating with H3K27me3. For instance, the developing zygotic MAC of *Paramecium* has been shown to depend on sRNAs, to deposit H3K27me3 marks in regions, which need to be eliminated during development (Frapporti et al., 2019). In other organisms, the RNA-induced transcriptional silencing (RITS) complex helps the siRNA production. Subsequently, these siRNAs influence the methylation enzymes to promote heterochromatin formation. This process effectively silences gene expression (Holoch and Moazed, 2015). While, the sRNA-H3K27me3 association we observed merely suggests such a silencing role, future experiments are necessary to confirm this hypothesis.

Curious case of repressive associations

The last two edges we discuss are H3K27me3-mRNA ($r_{xy,z} = -0.15$) and H3K9ac-mRNA ($r_{xy,z} = -0.13$), whose Pearson correlation coefficients were positive, 0.33 and 0.37, respectively. The negative partial correlation coefficient of H3K27me3-mRNA suggests a canonical repressive role of H3K27me3. However, we have seen through out this dissertation that H3K27me3 is likely not associated with repressed genes in *Paramecium*. Hence, we were curious whether there is a subset of genes whose expression show negative association to H3K27me3. To this end, we filtered genes, which have only H3K27me3 signal (*i.e.* no other histone mark signal is present for those genes). For this subset of genes, we found the H3K27me3-mRNA partial correlation coefficient to be 0.13. This suggests that the canonical repressive association of H3K27me3 with mRNA is still debatable in *Paramecium*.

In other organisms, histone marks are often associated with Pol II pausing (Gates et al., 2017). Hence, we investigated whether the paused genes (see Section 5.4.6) show negative association of mRNA expression with H3K27me3 and H3K9ac. For the paused genes, we found a negative partial correlation coefficient for the H3K27me3-mRNA ($r_{xy,z} = -0.16$), and H3K9ac-mRNA ($r_{xy,z} = -0.13$) association. While it is uncommon, several genes were reported to be repressed in the presence of H3K9ac in other organisms (Ha et al., 2011; Lai et al., 2017). These results show that the association of H3K27me3 with gene expression is less clear in *Paramecium*. One possibility is that each polyploid *Paramecium* chromosome undergoes histone modifications at different rates, and the bulk measurements are confounding the analysis. Future work, preferably at single-cell resolution, is necessary to investigate such a hypothesis.

6.4.4 Classification of gene expression

The results, we showed so far, demonstrate the complexity of the gene expression landscape in *Paramecium*. We wanted to decipher the general characteristics of genes which are highly expressed in *Paramecium* based on our data. To this end, we aimed to construct a supervised binary classifier to classify genes as high or low expressed. Figure 6.4 shows an overview of our machine learning analysis.

Classifier/model selection

We tested several binary classification algorithms on the training data set consisting of *All* features, using the WEKA software. We trained 100 models for each algorithm. We evaluated our models using area under the precision and recall curve (PR-AUC). A PR-AUC of 1.0 is a perfect model with 100% of genes

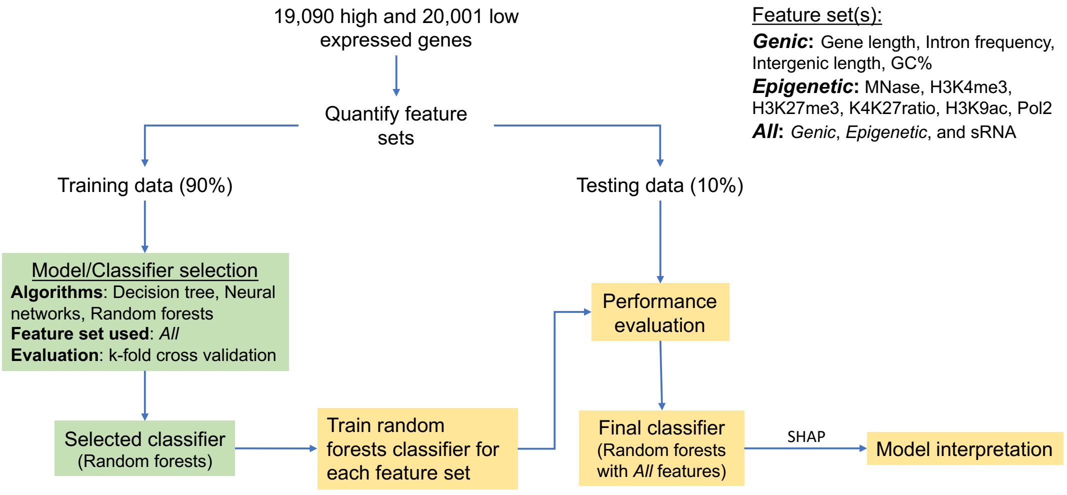

19,090 high and 20,001 low expressed genes
Quantify feature sets
Training data (90%)
Testing data (10%)
Feature set(s):
Genic: Gene length, Intron frequency, Intergenic length, GC%
Epigenetic: MNase, H3K4me3, H3K27me3, K4K27ratio, H3K9ac, Pol2
All: Genic, Epigenetic, and sRNA
Model/Classifier selection
Algorithms: Decision tree, Neural networks, Random forests
Feature set used: All
Evaluation: k-fold cross validation
Selected classifier (Random forests)
Train random forests classifier for each feature set
Performance evaluation
Final classifier (Random forests with All features)
SHAP
Model interpretation

Figure 6.4: Workflow of the machine learning analysis is shown. Green and yellow boxes show that they are implemented using WEKA machine learning workbench and Python scikit-learn package, respectively.

correctly classified as high or low. The mean PR-AUC value of the 100 models of each algorithm is shown as a bar plot in Figure 6.5. The jittered points are the PR-AUC of each of the 100 models evaluated. It shows that the models are robust, as their deviation from the mean PR-AUC is minimal. For our training data, the models with random forests algorithm showed the best performance with a PR-AUC of 0.82.

Feature set evaluation

As we have seen that the random forests algorithm performs best, we wanted to evaluate how different feature sets perform. To this end, we trained random forests model using the different feature sets: *All*, *Epigenetic* or *Genic*, and evaluated their performance on the test data. Figure 6.6 shows the precision and recall curve of the random forests model with different feature sets. The random forests model using the feature set *All* was found to be the best with a PR-AUC value of 0.83. The models with *Genic*, and *Epigenetic* features performed fairly well with a PR-AUC of 0.75, and 0.77, respectively.

6.4.5 Model interpretation

We wanted to know which of the features are used for classifying a given gene as high or low expressed in our best performing random forests model with *All* features. We employed the SHAP method to interpret our model.

Gene level feature effects

We can understand for any given gene, using the SHAP values, how each feature contributed to classify the gene as high or low expressed. Figure 6.7 shows

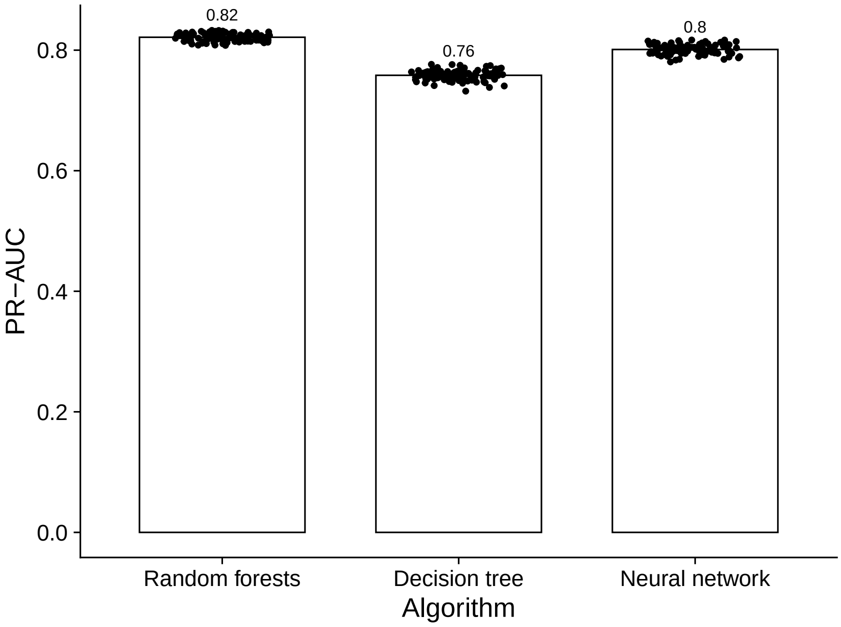


Figure 6.5: Model performance (y-axis; area under the precision-recall curve (PR-AUC)) is shown for the different classification algorithms (x-axis). Models are evaluated by WEKA in a 10-fold cross validation approach for 10 repetitions of each algorithm (100 models in total). The PR-AUC of the 100 models are jittered on the bar, with the mean PR-AUC written above the bars.

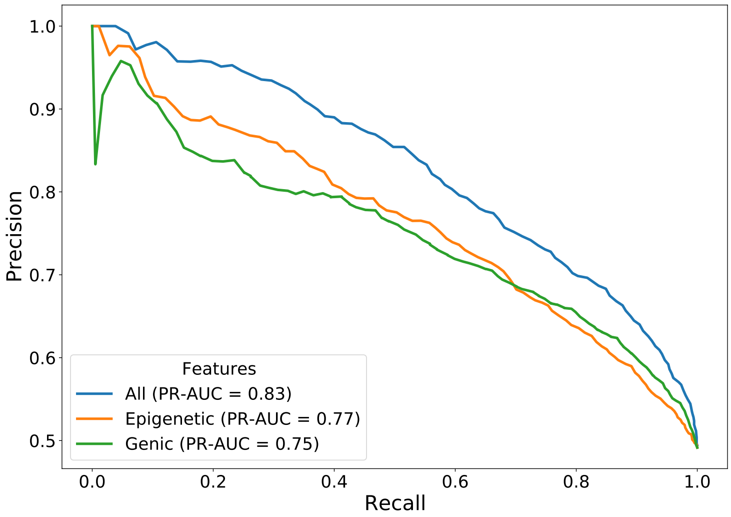


Figure 6.6: The precision (x-axis) and recall (y-axis) curve of the model based on random forest algorithm, with different feature sets (colors). The PR-AUCs of different feature sets are shown in the legend.

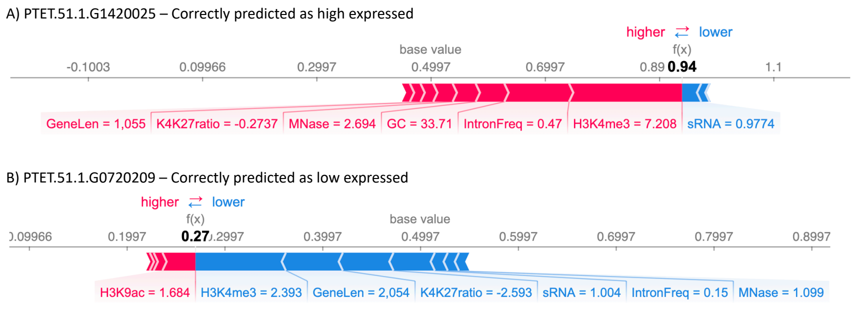


FIGURE 6.7: Force plots of SHAP values for two example genes (A and B) are shown. The base value shown (0.4997) is the average predicted probability of the model. The value at $f(x)$ shows the predicted probability of the gene of interest. The feature values of each gene are shown below the red or blue arrows. The width of the arrows are determined by the respective feature's SHAP value, *i.e.* the large width of an arrow shows a high importance of the feature. A feature is colored red or blue depending on whether that feature contributes to increase or decrease the $f(x)$ from the base value, respectively.

the force plots of two example genes, which were correctly predicted as high or low expressed by our classifier. In Figure 6.7A, we can see that the gene (PTET.51.1.G1420025) is classified as high expressed, because of the contributions from H3K4me3, intron frequeny, GC content, and so on. Figure 6.7B shows that the contributions of H3K4me3, gene length, and K4K27ratio are resulting in the prediction of the gene (PTET.51.1.G0720209) to be classified as low expressed. We can see that the order of features in Figure 6.7A and 6.7B are different from each other. Such plots are useful to know the feature effects of each gene's prediction.

Global feature importance

While it is good to know the feature importance of individual genes, by means of force plots, it would be beneficial to know a global importance of all features. Figure 6.8 shows a global feature importance plot for our test data set, which is measured by the mean absolute SHAP values. As we mentioned earlier, a high mean absolute SHAP value depicts a high importance of the respective feature in classifying a gene as high or low expressed. Figure 6.8 shows that, unsurprisingly, H3K4me3 is the most important feature globally. Further, we can see that intron frequency, gene length, sRNA, and GC content are in the top five globally important features. We found none of the other epigenetic features (H3K27me3, H3K9ac, MNase, K4K27ratio, Pol II) in the top five important features. The multicollinearity of epigenetic marks likely plays a role in these factors being treated as less important. Interestingly, we observed that H3K27me3 is the least important factor for our models. However, it is not clear from Figure 6.8, for instance, whether a high or low H3K27me3 signal is

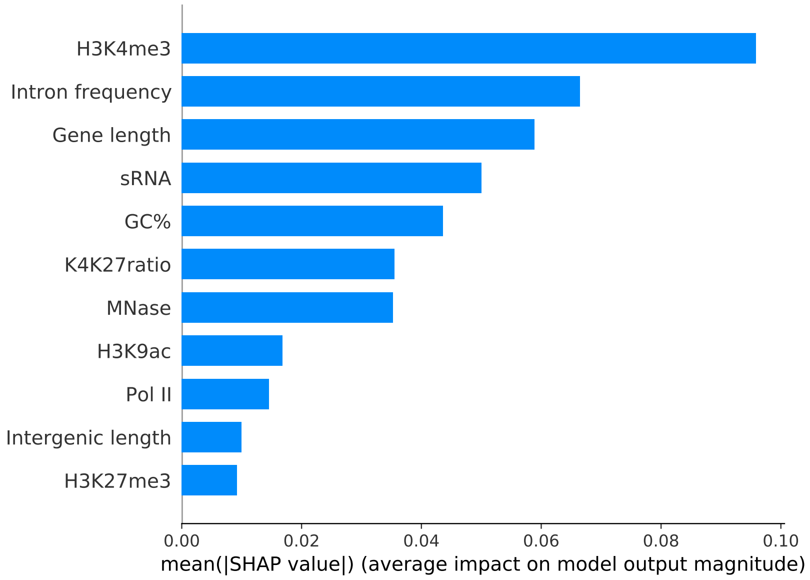


FIGURE 6.8: Global feature importance plot for our test data set is shown here. The mean absolute SHAP values (x-axis) shows the average impact of the feature (y-axis) on the model output. A high mean absolute SHAP value shows a high importance of the respective feature in classifying a gene as high or low expressed.

required to classify a gene as high or low expressed. To this end, we created a feature effect summary plot.

Global feature effects

The feature summary plot (Figure 6.9) combines the force plot and the global feature importance plot, which we discussed earlier. For instance, let us investigate the interpretation of the gene at the right most point (gene) of H3K4me3 (x-axis: 0.24). This gene has a high H3K4me3 value. A positive SHAP value of 0.24 shows that the high H3K4me3 value of this gene increases the prediction probability of this gene to be classified as high expressed, by a factor of 0.24. However, we will use this plot to discuss only the global patterns, and not the individual instances.

We observed that for a gene to be classified as high expressed the following are required: high H3K4me3, high MNase, low H3K9ac, high Pol II, and high GC content. We had discussed earlier, in this thesis, the importance of GC content in nucleosome placement, histone marks, and Pol II. Hence, it makes sense to see that genes with such characteristics are highly expressed. Gene expression prediction models on other organisms have shown the association of histone marks and GC content, as well (Singh et al., 2016; Agarwal and Shendure, 2020). Interestingly, owing to the debate on the role of H3K27me3

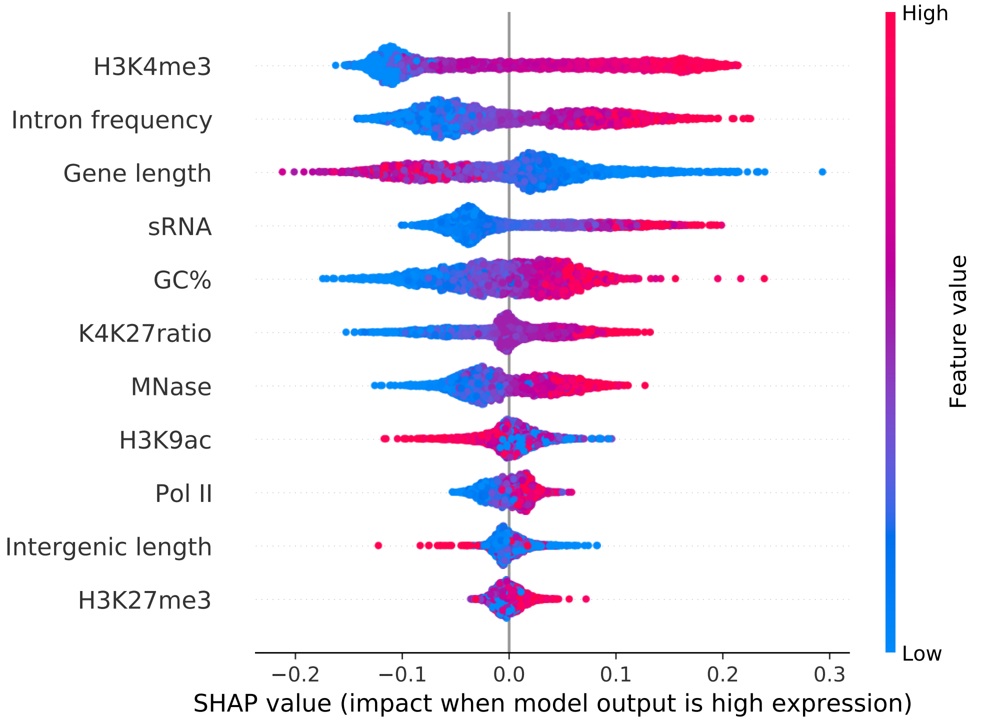


FIGURE 6.9: A summary plot showing the global feature effects is shown. The features (y-axis) are in the order of global feature importance. The SHAP value is shown on the x-axis. Each dot in the figure represents a gene, and the color gradient depicts the feature value in scale from low to high. The overlapping dots are jittered in the y-axis direction. A high feature value of H3K4me3 (located at x-axis: 0.24) of this gene can increase it's prediction probability to be classified as high expressed, by a factor of 0.24.

through out this thesis, we noticed that a high H3K27me3 content contributes to a prediction of high expression, although it is the least important feature.

Our classification model also suggests that a high intron frequency, and low intergenic length are characteristics of high expressed genes. This relation is, to our knowledge, not known in *Paramecium*. However, the role of intron frequency in regulating gene expression is widely studied in plants, and other eukaryotes (Deshmukh, Sonah, and Singh, 2016; Shaul, 2017). In organisms like *Caenorhabditis elegans* and *Arabidopsis thaliana*, studies have shown that in the absence of promoter regions, introns can increase gene expression (Rose, 2019). As we have discussed earlier (Table 5.1), the *Paramecium* genome is highly condensed and has very short intergenic regions, leaving very little room for promoter regions. Hence, the high intron frequency likely modulates the gene expression in *Paramecium*. We also found low gene length, and high sRNA to be characteristics of high gene expression in our model. The inverse relationship between gene expression and gene length has been shown in several organisms (Duret and Mouchiroud, 1999; Brown, 2021). The high sRNA content predicting high gene expression supplements our earlier observation that endo-siRNAs are likely not acting in *cis* to silence their parent mRNA (see Section 4.4.4).

6.5 Key conclusions

In this chapter, we showed the inherent multicollinearity in different epigenetics marks and their positive correlation with gene expression. We constructed an association network of the epigenetic marks, MNase, Pol II, sRNA and mRNA expression using sparse partial correlation network analysis. While we found several common associations, we found a positive association of H3K27me3 with mRNA, for genes which carry only H3K27me3 signals. However, paused genes showed negative partial correlation coefficient of H3K27me3 and mRNA. Our observations make the role of H3K27me3 in *Paramecium* less clear. We constructed a supervised learning based classifier which predicts high and low expressed genes with a test PR-AUC of 0.83. Not only did we build a classifier, we inferred the general patterns of high or low expressed genes in *Paramecium* using a model interpretation technique called SHAP. Consequently, we reported that a high expressed gene requires the following top five features: high H3K4me3, high intron frequency, low gene length, high sRNA, and high GC content. The SHAP analysis also revealed the plausible role of intron frequency in *Paramecium*'s gene expression machinery.

6.6 Contributions

The data used in this chapter was created as part of the works presented in Chapters 4 and 5. I merely re-purposed these data for the analysis presented in this chapter, all of which were performed by me under the supervision of Prof. Dr. Martin C. Simon and Prof. Dr. Marcel H. Schulz.

Chapter 7

Summary and outlook

Understanding of human gene regulation is paramount for improving healthcare. Several consortia projects like the encylopedia of DNA elements (ENCODE), and the International Human Epigenome Consortium (IHEC) have improved our understanding of transcriptomics and epigenomics tremendously. Researchers have developed numerous integrative data analysis tools to supplement the consortia projects. However, the existing tools are tailored to human data or few other model organisms, like mouse and drosophila. Often, these tools are not directly employable by researchers working on non-model organisms. Non-model organisms research is crucial for answering mechanistic biological and evolutionary questions like, how did multicellular organisms evolve?

Paramecium tetraurelia is a free-living unicellular ciliate, and a non-model organism to understand non-mendelian genetics, and epigenetics. *Paramecium* exhibits nuclear dimorphism with two germline micronuclei (MIC), and a somatic macronuclei (MAC). The MIC is diploid and transcriptionally inactive, while the MAC is transcriptionally active and exhibits polyploidy (800n) (Beale and Preer, 2008). During sexual reproduction, after fertilisation, a diploid MIC develops into a polyploid MAC by the fusion of several mitotic MIC and disintegration of the original (or parental) MAC. The mitotic MIC transitioning to a MAC is called the developmental MAC, which undergoes several genomic rearrangements. *Paramecium* also undergoes asexual (vegetative) reproduction, by separating into two daughter cells (Van Houten, 2019).

Decades of research on the developmental MAC have unraveled the crucial role of methylated histone marks, and small RNAs in controlling the genomic rearragements of a developmental MAC (Beisson et al., 2010). The vegetative MAC is known to express mutually exclusive surface antigens, a multigene family, ensuring expression of different serotypes (Wichterman, 1986). *Paramecium* switches their serotypes based on environmental changes, like temperature (Simon, Marker, and Schmidt, 2006). Small RNAs are known to play a role in controlling the expression of surface antigen genes in vegetative MAC (Marker et al., 2010). The prime diet of *Paramecium* is bacteria. Researchers fed genetically modified bacteria to *Paramecium*, and have shed light on the RNA interference (RNAi) pathways in vegetative MAC. The RNAi pathway depends on the RNA dependent RNA polymerase (RDR), the Dicer (DCR) enzyme to generate short interfering RNAs (siRNAs) (Carradec et al., 2015; Marker et al., 2014). However, the transcriptomic and epigenomic regulatory landscape of the MAC genome, which has 80% protein-coding genes and short intergenic regions, is poorly understood. In this thesis, we aimed to develop new tools, and adapt existing softwares for the analysis of transcriptomic and epigenomic data of *Paramecium.*

We developed an automated eukaryotic siRNA analysis tool, called RAPID. Our tool captures diverse siRNA characteristics from small RNA sequencing data, and provides easily navigable HTML visualisations. Some of the siRNA characteristics, RAPID captures include strand specific reads, and non-templated nucleotides. We also introduced a normalisation technique to facilitate the comparison of siRNA-based gene knockdown samples. We have also integrated differential expression analysis software, DESeq2. RAPID is made available for public use as a Conda package. Nevertheless, RAPID is not a one-stop solution for siRNA analysis. Some of the possible extensions of RAPID include capturing sequence-level attributes like, creating sequence logos of conserved siRNA sequences, or identifying siRNA target sites.

In addition to RAPID, we developed a pipeline to characterise the first genome wide small RNA profile of vegetative MAC of *Paramecium.* We identified many endogenous short interfering RNAs (endo-siRNAs) are produced from protein coding genes, which were confirmed through splice junction analysis. We did not discover any micro RNAs. In contrast to many organisms, we observed that the endo-siRNAs do not strictly act in *cis* to silence their parent mRNA. We also predicted phasing of siRNAs, which are regulated by the RNAi pathway. Further, using RAPID, we investigated the aberrations of endo-siRNAs, and their respective transcriptomic alterations caused by RNAi inducing technique, called feeding. Our findings alert RNAi researchers to select appropriate controls, and differentiate between the changes in expression caused by the feeding technique itself, and the actual regulatory changes. Our observation is crucial at a point when RNAi treated organisms are considered GMO-free (free from genetically modified organisms), and are being increasingly used to combat viral and pest resistance in bacteria and plants.

We also identified several sRNAs in the intergenic, and other non-coding regions. In another ciliate, *Oxytrichia trifallax*, sRNAs are shown to regulate the copy number of small chromosomes which carry only one gene (Khurana et al., 2018). We hypothesise that many of our sRNAs, especially the ones from non-coding regions, are likely to be involved in regulating the polyploid of the *Paramecium*'s MAC. However, future experiments are necessary to confirm such a hypothesis. While we documented the first genome-wide analysis of the sRNAs, using RAPID and other tools, their regulatory mechanisms are still unclear. Future analysis are required to identify the targets of the sRNAs. Small RNA target prediction is an active field of research. Predominant of existing tools primarily focus on bacterial small RNAs, and identify merely mRNA targets (King, Vanderpool, and Degnan, 2019; Wright et al., 2014). In *Paramecium*, the mRNAs are majorly positively correlated with sRNAs. Hence, it is likely that the sRNAs are not targeting mRNAs, but are involved in a complex landscape of sRNA-protein interaction to regulate the genome. In order to identify such interactions, the RNA and protein family domains of *Paramecium* needs to be queried. It is safe to say, that we have only laid the foundation for understanding the role of RNAi in *Paramecium.*

The vegetative MAC genome annotation of *Paramecium* revealed a high protein coding density of 80%, highest among the free living eukaryotes. This results in short intergenic regions of mere 352 bp, raising the question on how the *Paramecium*'s MAC is regulated. Hence, we shifted our focus to understand the epigenomic characteristics of MAC, which has never been characterised before. Using nucleosome positioning softwares, we identified well positioned nucleosomes, which are shifted downstream of the transcription start site (TSS). We found that the introns are flanked by nucleosomes, which suggests they play a role in regulating splicing efficiently.

The nucleosomes, epigenetic marks (H3K4me3, H3K9ac, and H3K27me3), and Pol II are located along the gene structure and are less enriched in the non-coding regions around the gene boundaries. They were also directly proportional to different gene expression groups. GC content seems to influence, in *cis*, the positioning of nucleosomes, epigenetic marks, and Pol II in the AT-rich *Paramecium* genome. We employed a chromatin state segmentation approach, on nucleosomes and histone marks, which revealed genes with active, repressive, and bivalent chromatin states. In multicellular organisms, genes in bivalent domains, carrying both H3K4me3 and H3K27me3 marks are associated with orchestration of cell differentiation during development, through pausing of gene expression (Voigt, Tee, and Reinberg, 2013; Sen et al., 2016; Blanco et al., 2020). However, among the genes with bivalent domains in *Paramecium*, we did not find an enrichment of paused genes. Hence, the role of bivalent domains in *Paramecium* is still unclear. Further, in order to confirm the bivalent domains, Re-ChIP experiments may need to be performed.

Further, we constructed a gene regulatory association network by applying the sparse partial correlation network method on different epigenetic marks, nucleosomes, small RNA and gene expression data. We observed common positive associations of H3K4me3 with mRNA and H3K9ac with H3K4me3. We also observed a negative association for H3K27me3 with mRNA, especially for the paused genes. However, the genes which carried only H3K27me3 marks show positive association for H3K7me3 and mRNA, which makes the role of H3K27me3 in *Paramecium* less clear.

It is important to note that, the *Paramecium* cells in a culture are not synchronised in their vegetative development cycle. This means different cells are under different epigenomic/transcriptomic states. The bulk epigenomic and transcriptomic measurements of the polyploid MAC are likely to confound our analysis. We would need to perform single cell data analysis of *Paramecium* to gain a better understanding of the different chromatin states. However, the experimental methods for obtaining single cell data in *Paramecium* are currently not available.

Next, we resorted to advanced machine learning methods to understand inherent patterns in the data. We constructed a Random Forests classifier, to classify gene expression as high or low, using *Genic* (gene length, intron frequency, etc.) and *Epigenetic* features (histone marks, Pol II, MNase, etc.). Our model has a test performance (PR-AUC) of 0.83. Upon evaluating different feature sets, we found that genic features are as predictive of gene expression, as the epigenetic features. Further, we applied the SHAP technique, on our Random Forests model, to infer general gene expression patterns. Consequently, we reported that the top five features of a high expressed gene are high H3K4me3, high intron frequency, low gene length, high sRNA, and high GC content. The role of intron frequency in gene expression regulation, revealed by SHAP analysis, has never been reported before in *Paramecium*.

The prediction model could be improved by incorporating DNA-sequence features, for example, using one-hot encoding of fixed windows around the TSS of genes, and applying deep convolutional neural networks. In mammals such methods have been successfully applied to predict gene expression both quantitatively and qualitatively (Beer and Tavazoie, 2004; Vilar, 2010; Bessière et al., 2018). However, the window sizes around the TSS are usually several kilo base pairs. *Paramecium* has a mean gene size of only 1,084 bp. Hence, such large window sizes will create a lot of overlaps, and likely confound the prediction algorithm. If we remove overlapping genes from the analysis, a vast majority of the genes will be left out from our analysis, as the MAC genome is 80% protein coding with short intergenic regions. Hence, we decided to use a prediction model, such that we can infer global gene expression patterns of all expressed genes.

There are several open questions, apart from the aforementioned, in relation to MAC genome regulation. For instance, in the *Paramecium*'s MAC, where are the transcription factor binding sites located in such short intergenic regions? Does *Paramecium* have distal regulatory elements, which are brought together through 3-D organisation of the genome?

In a nutshell, we developed novel tools, adapted existing bioinformatic methods, applied machine learning methods to shed light on the small RNA-omic and the epigenomic orchestration of gene expression in *Paramecium*'s macronuclei. We believe that our findings pave the way to better our communal understanding of regulatory omics.

Appendix A

Supplementary materials for Chapter 3

A.1 Supplementary tables

							Qualitative Plots			Quantitative Plots							
Tool/Supporting Feature	Contaminant removal	Supports other aligners	User-defined gene/region	Knockdown corrected normalisation	Offline	Hardcoded genomes	Heatmaps	PCA	MDS	Strand specificity	Soft-clipped bases	Coverage plots	Multi-sample comparison plots	Differential analysis	Enrichment analysis support	Interactive interface	miRNA or piRNA specific?
RAPID	✓	✓	✓	✓	✓	x	✓	✓	✓	✓	✓	✓	✓	✓	x	x	x
smallRNA toolkit(Moxon et al., 2008)	x	x	✓	NA	x	✓	x	x	x	x	x	x	x	x	x	x	✓
sRNA toolbox(Rueda et al., 2015)	x	x	✓	NA	✓	x	x	x	x	x	x	x	x	✓	✓	x	✓
Oasis(Capece et al., 2015)	x	x	x	x	x	x	✓	✓	x	x	x	x	✓	✓	✓	✓	✓
CPSS(Wan et al., 2017)	x	✓	x	NA	x	✓	x	x	x	x	x	x	x	x	✓	✓	✓
iSmart(Panero et al., 2017)	x	x	x	NA	✓	✓	✓	✓	x	x	x	x	x	✓	✓	✓	✓
iSRAP(Quek et al., 2015)	x	✓	x	NA	✓	x	✓	✓	✓	x	x	x	x	✓	x	x	x
PiPipes(Han et al., 2015)	✓	✓	x	NA	✓	x	x	x	x	✓	✓	x	x	✓	x	x	✓
ncPRO-Seq(Chen et al., 2012)	x	✓	✓	NA	✓	x	x	x	x	x	x	x	x	x	✓	x	✓
UEA sRNAworkbench (Mohorianu et al., 2017)	✓	✓	✓	x	✓	x	x	x	x	x	x	x	✓	✓	x	✓	x
NGSToolbox (Rosenkranz et al., 2015)	x	x	✓	NA	✓	x	x	x	x	x	x	x	x	x	x	x	✓
SePIA (Icay et al., 2016)	x	✓	x	NA	✓	x	x	x	x	x	x	x	x	✓	x	x	✓
SPAR (Kuksa et al., 2018)	x	✓	x	NA	x	✓	✓	x	x	x	x	x	x	x	x	✓	✓

Table A.1: Comparison of RAPID with other tools is shown. ✓- Feature supported, x - Feature not supported, NA - Feature is not in the scope of this tool. For instance, *Knockdown corrected normalisation* feature is *NA* for CPSS, because it does not support multiple sample comparison. The full description of the column headers are listed in Table A.2. *Note*: This list is not exhaustive. Tools whose primary focus is on identifying/annotating different classes of small RNAs are not included.

Supporting Feature	**Description**
Contaminant removal	Is there an option to remove set of contaminants (microbial, ribosomal, etc.) from the read files?
Supports other aligners	Does the tool support alignment files from other tools, instead of performing their own alignment?
User-defined gene/region	Could the user specify a list of regions to perform downstream analysis?
Knockdown corrected normalisation	Does the tool enable multiple-sample comparison by facilitating normalisation techniques specific to sRNA knockdown studies?
Offline	Can the tool be used offline?
Hardcoded genomes	Is the tool generic? i.e. Is the tool's ability somehow limited to a set of pre-defined genomes?
Quantitative, and Qualitative Plots	Does the tool support informative plots to gain understanding of the analyzed data (multi-dimensional scaling (MDS), principal component analysis (PCA))
Multi-sample comparison plots	Does the tool provide a comprehensive view of multiple samples (not just differential analysis)? For instance, how does the read distribution vary across multiple samples in different genes of interest?
Differential analysis	Is the tool equipped with modules to perform pairwise differential analysis?
Enrichment analysis support	Is there any support to perform functional enrichment within the tool
Interactive interface	Does the tool have an interactive interface, or plots?
miRNA or piRNA specific?	Is the tool specific to analyze miRNA or piRNA only?

Table A.2: Table describing the supporting features of RAPID mentioned in the column headers of Table A.1.

Appendix B

Supplementary materials for Chapter 4

B.1 Supplementary Methods

B.1.1 SRC Boundary Modifications

Before correlating sRNA accumulation with mRNA expression, we did a quick scan of SRCs in IGV Browser (version 2.3.91). This investigation of SRCs showed occurrences of non-specific boundaries.

What are non-specific boundaries?

Consider example in Supplementary Figure B.1. When performing an annotation overlap (intersectBed; bedtools v2.23), of a SRC region (C1732; shown in the above figure) overlaps with three genes. Whereas, only one gene is responsible for the SRC in the example, two neighboring genes got included in the overlap as well. This is due to the expansion of the SRC localization, which could have been due to one or more of the following reasons: (i) Alignment artifacts, (ii) Padding parameter in Shortstack, and (iii) Unifying the identified clusters from different serotypes to obtain the SRC. In the example in Supplementary Figure B.1, one could see the interplay of all three reasons for the expansion of SRC localization. We decided to explore the effect of unification of clusters in introducing non-specific overlaps, as the other two reasons are not entirely with in our scope.

How many genes are introduced because of unification?

From the Supplementary Figure B.2, we can observe that, we are introducing a lot of new distinct genes. There is a small effect on other annotation categories as well, but it is meagre.

How many clusters are introduced because of unification?

From the Supplementary Figure B.3, we can see that because of the cluster merging we are introducing approximately 50% extra clusters in individual serotypes.

Boundary modification criteria

In order to correct for these non-specific partial overlap problems, we tried filtering SRCs with different overlap percentage to a gene.

Category-1: Should overlap at least 70% of a gene

This rather strict condition, removed all SRCs which overlap multiple genes and naturally, retained genes which overlap multiple SRCs. The table below shows the summary of SRCs, Genes and its overlaps with each other.

<table>
<tr><th colspan="2">Number of SRCs overlapping Multiple Genes</th></tr>
<tr><td>No. of SRCs</td><td>No. of overlapping genes</td></tr>
<tr><td>1687</td><td>1</td></tr>
<tr><th colspan="2">Number of Genes overlapping Multiple SRCs</th></tr>
<tr><td>No. of Genes</td><td>No. of overlapping SRCs</td></tr>
<tr><td>1243</td><td>1</td></tr>
<tr><td>120</td><td>2</td></tr>
<tr><td>38</td><td>3</td></tr>
<tr><td>12</td><td>4</td></tr>
<tr><td>6</td><td>5</td></tr>
<tr><td>2</td><td>6</td></tr>
</table>

Category-2: Should overlap at least 10% of a gene

This rather relaxed condition, retained most SRCs which overlap with multiple genes and a well as, retained genes which overlap multiple clusters. The table below shows the summary of SRCs, Genes and its overlaps with each other.

<table>
<tr><th colspan="2">Number of SRCs overlapping Multiple Genes</th></tr>
<tr><td>No. of SRCs</td><td>No. of overlapping genes</td></tr>
<tr><td>1643</td><td>1</td></tr>
<tr><td>189</td><td>2</td></tr>
<tr><td>11</td><td>3</td></tr>
<tr><td>1</td><td>5</td></tr>
<tr><th colspan="2">Number of Genes overlapping Multiple SRCs</th></tr>
<tr><td>No. of Genes</td><td>No. of overlapping SRCs</td></tr>
<tr><td>1645</td><td>1</td></tr>
<tr><td>123</td><td>2</td></tr>
<tr><td>36</td><td>3</td></tr>
<tr><td>11</td><td>4</td></tr>
<tr><td>3</td><td>5</td></tr>
</table>

Category-3:

As a trade-off, we tried altering the SRC boundaries for non-specific overlaps. i.e. In a gene-cluster overlap, if the gene is covered $> 80\%$ and the SRC is covered at least by 20%, then the SRC's boundary is limited to the gene's boundary. Doing so, limits the non-specific overlaps. But, this condition missed genes with multiple clusters being overlapped. So, we added another criteria. If the gene is covered $> 10\%$ and the cluster is covered more than 80%, those

SRCs are retained with out any boundary changes. The table below shows the summary of SRCs, Genes and its overlaps with each other.

<table>
<tr><th colspan="2">Number of SRCs overlapping Multiple Genes</th></tr>
<tr><td>No. of SRCs</td><td>No. of overlapping genes</td></tr>
<tr><td>1335</td><td>1</td></tr>
<tr><td>32</td><td>2</td></tr>
<tr><td>1</td><td>3</td></tr>
<tr><td>1</td><td>4</td></tr>
<tr><th colspan="2">Number of Genes overlapping Multiple SRCs</th></tr>
<tr><td>No. of Genes</td><td>No. of overlapping SRCs</td></tr>
<tr><td>1283</td><td>1</td></tr>
<tr><td>87</td><td>2</td></tr>
<tr><td>15</td><td>3</td></tr>
<tr><td>2</td><td>4</td></tr>
</table>

As a results of these two conditions, we arrived at 1618 SRCs in total, which can be used for comparing with mRNA expression. These set filtered and boundary modified 1618 SRC loci are termed endo-siRNAs, short for endogenous small interfering RNAs. They will be utilised in all downstream knockdown and mutant analysis, unless mentioned otherwise.

B.2 Supplementary Figures

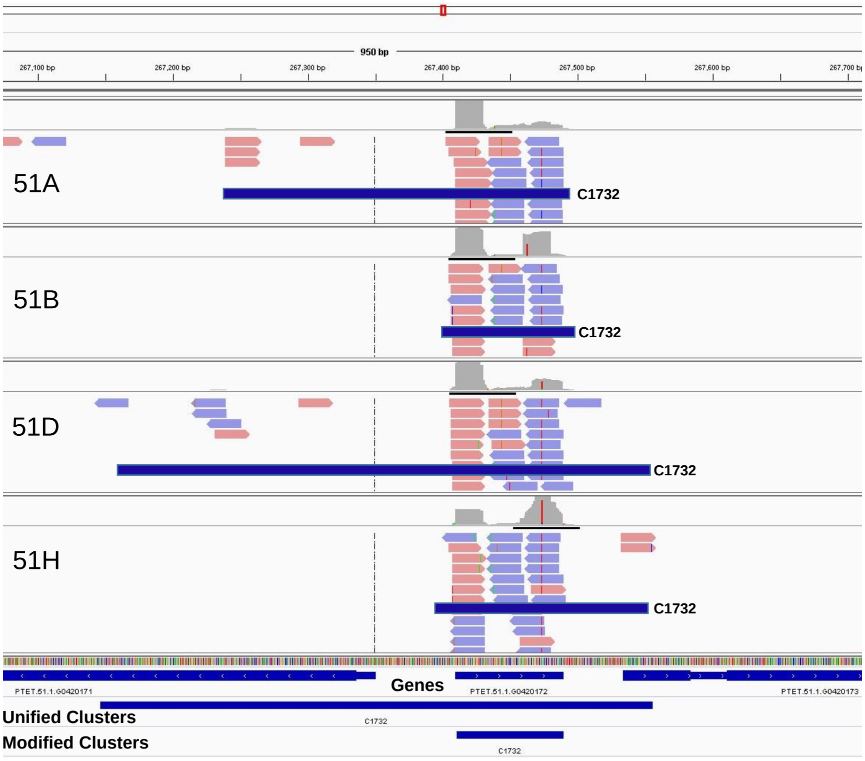


FIGURE B.1: The localization of SRC C1732 in different serotypes is shown. In the bottom panel the SRC boundaries before and after modification along with the gene annotation is shown.

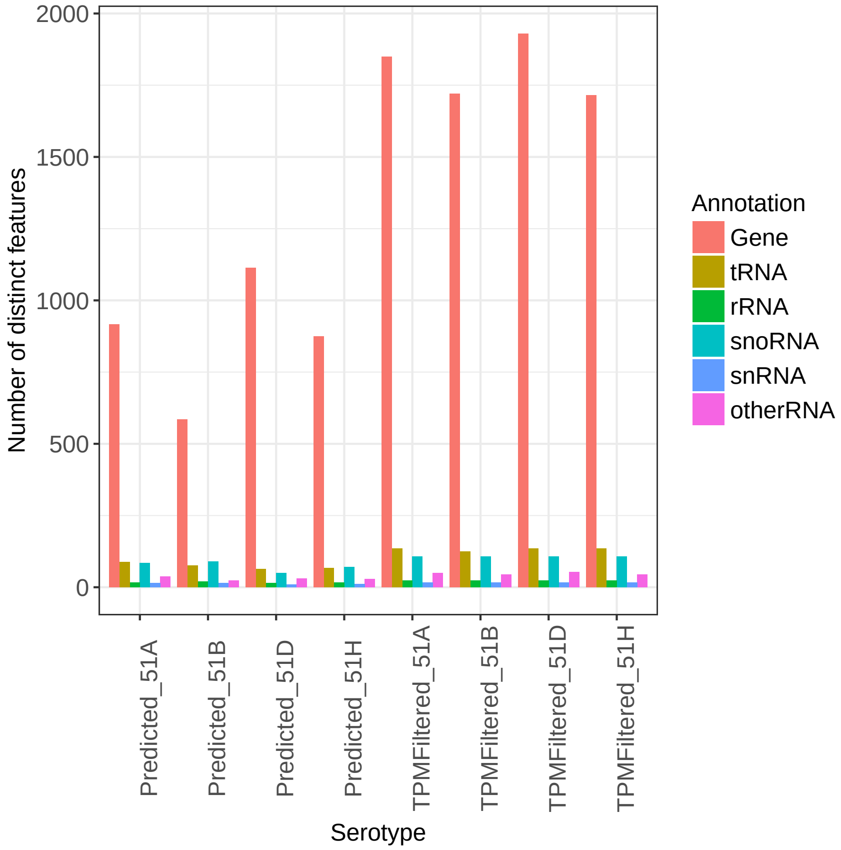


FIGURE B.2: The effect of unification in the number of distinct features identified is shown. SRCs expressed in the wildtype serotype (51A, 51B, 51D, 51H; replicates were merged) samples were overlapped with annotated regions. Each annotated element is counted only once (distinct counting) and the number of elements of the different types (colors) is shown on the y-axis for all 4 serotypes. Predicted represents the distinct feature annotations while using the Shortstack's prediction results directly (i.e. Before unifying them to form SRCs). TPMFiltered represents what we call serotype specific SRC (i.e. SRC with a $TPM > 1$ in each serotype).

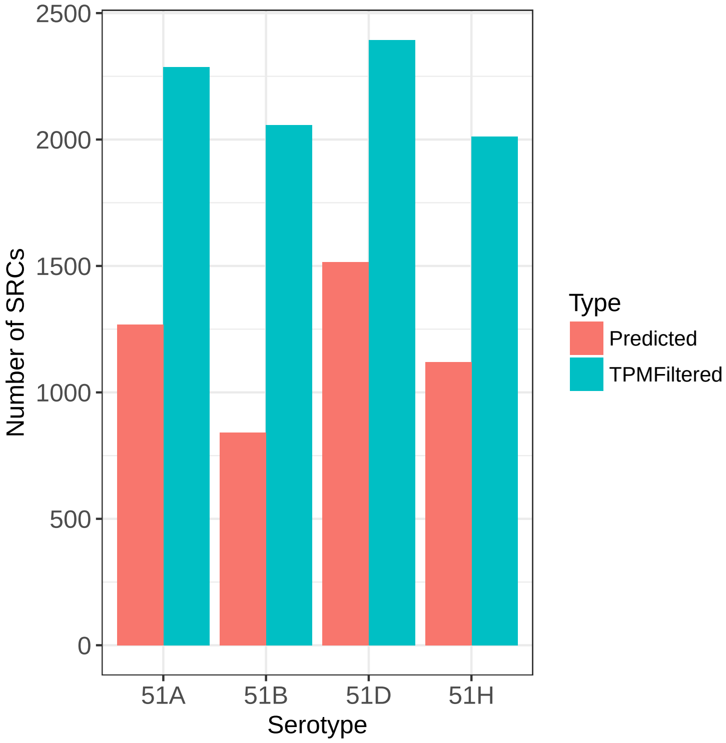


Figure B.3: The number of SRCs identified in each wildtype serotype is shown. The type Predicted represents the distinct feature annotations while using the Shortstack's prediction results directly (i.e. Before unifying them to form SRCs). TPM-Filtered represents what we call serotype specific SRC (i.e. SRC with a $TPM > 1$ in each serotype).

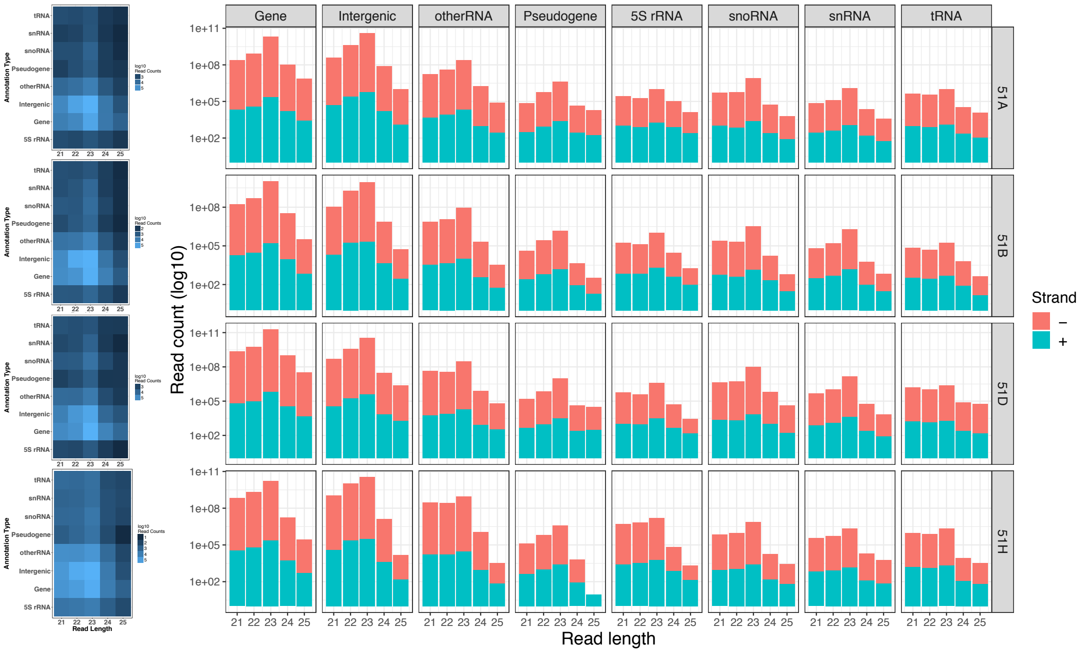
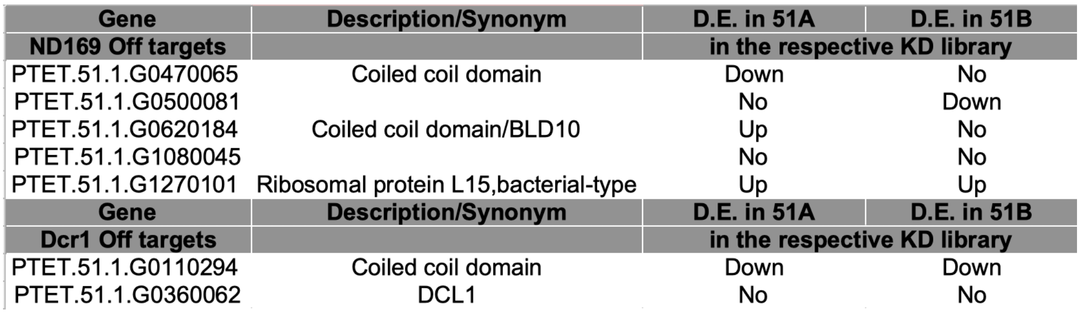


FIGURE B.4: Supplementary figure for Figure 4.4. Left to right: A heatmap of sRNA accumulation (log10, color scale) in SRCs overlapping different genomic annotations and restricted to small RNA length (x-axis) for serotypes 51A, 51B, 51D and 51H is shown. Barplots showing the length distribution of sense (green) and antisense (red) sRNAs mapping to different genomic annotations in serotypes 51A, 51B, 51D and 51H is shown.



Gene	Description/Synonym	D.E. in 51A	D.E. in 51B
ND169 Off targets		in the respective KD library	
PTET.51.1.G0470065	Coiled coil domain	Down	No
PTET.51.1.G0500081		No	Down
PTET.51.1.G0620184	Coiled coil domain/BLD10	Up	No
PTET.51.1.G1080045		No	No
PTET.51.1.G1270101	Ribosomal protein L15,bacterial-type	Up	Up
Gene	Description/Synonym	D.E. in 51A	D.E. in 51B
Dcr1 Off targets		in the respective KD library	
PTET.51.1.G0110294	Coiled coil domain	Down	Down
PTET.51.1.G0360062	DCL1	No	No

FIGURE B.5: Differential expression status of different off-target genes

Appendix C

Supplementary materials for Chapter 5

C.1 Supplementary Figures

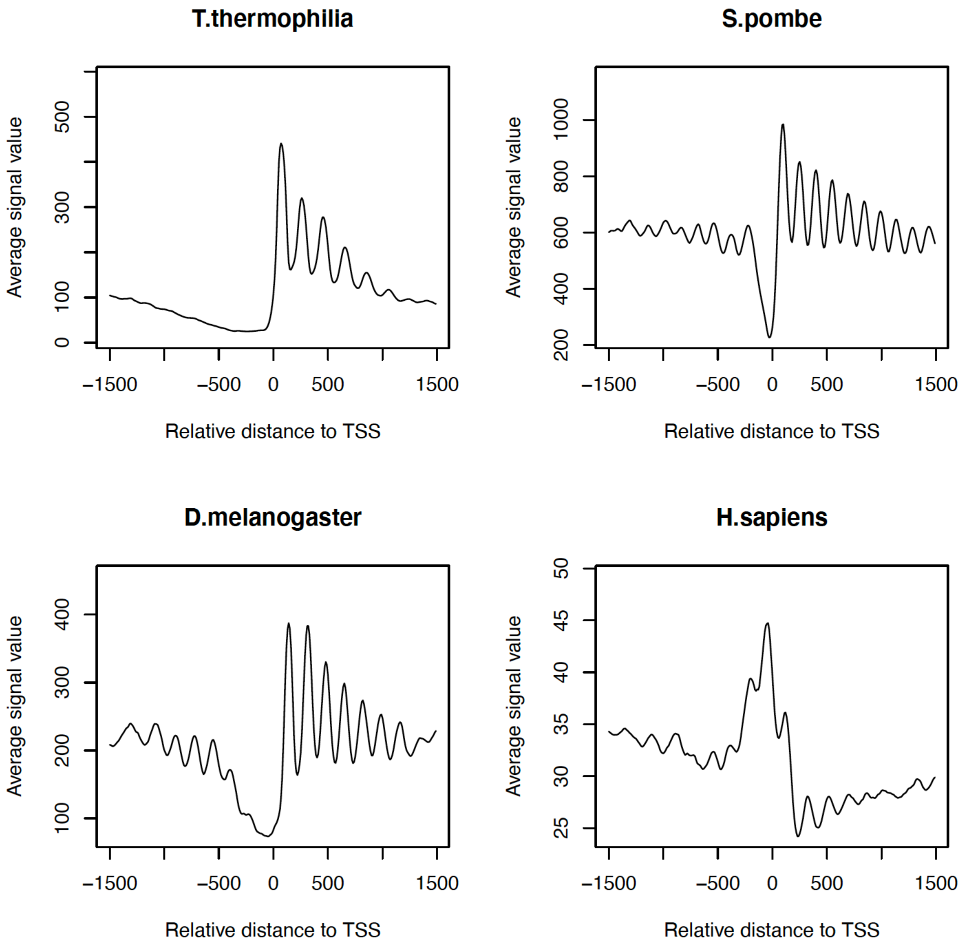


Figure C.1: Nucleosome occupancy profiles at the TSS is shown for all genes of different organisms.

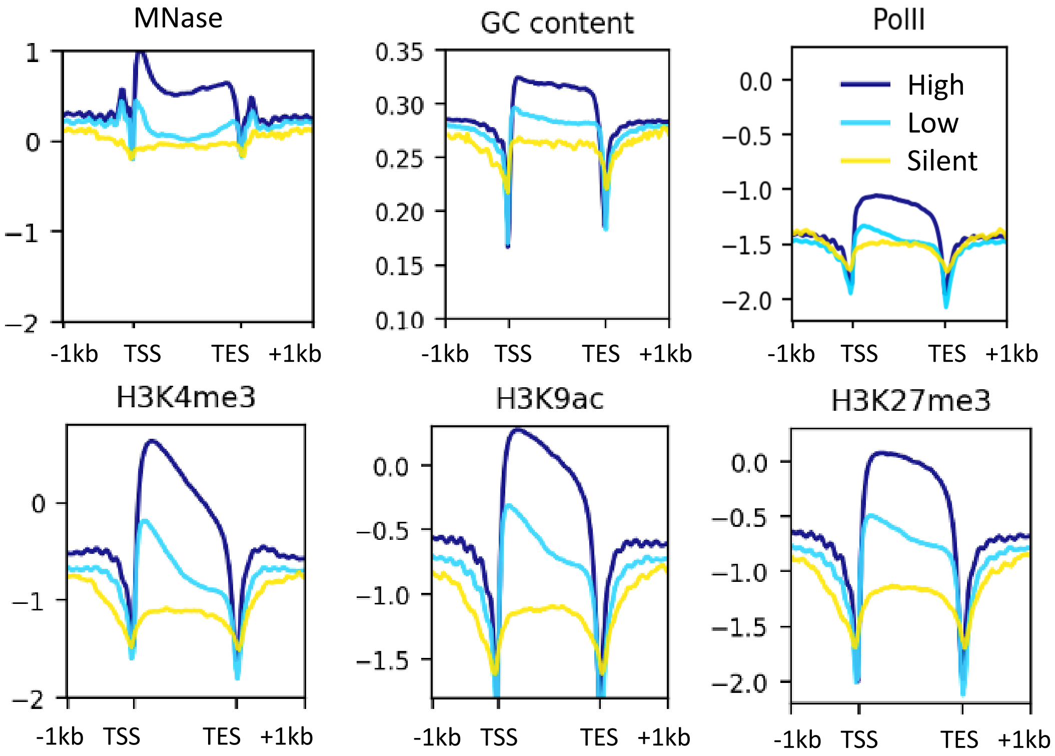


FIGURE C.2: Distribution profile showing the signals of H3K27me3, H3K4me3, H3K9ac, MNase, GC content, and Pol II stratified according to the gene expression groups. The signal shown is for the gene body (TSS to TES; scaled to 500 bp) and 1 Kbp upstream and downstream of gene body.

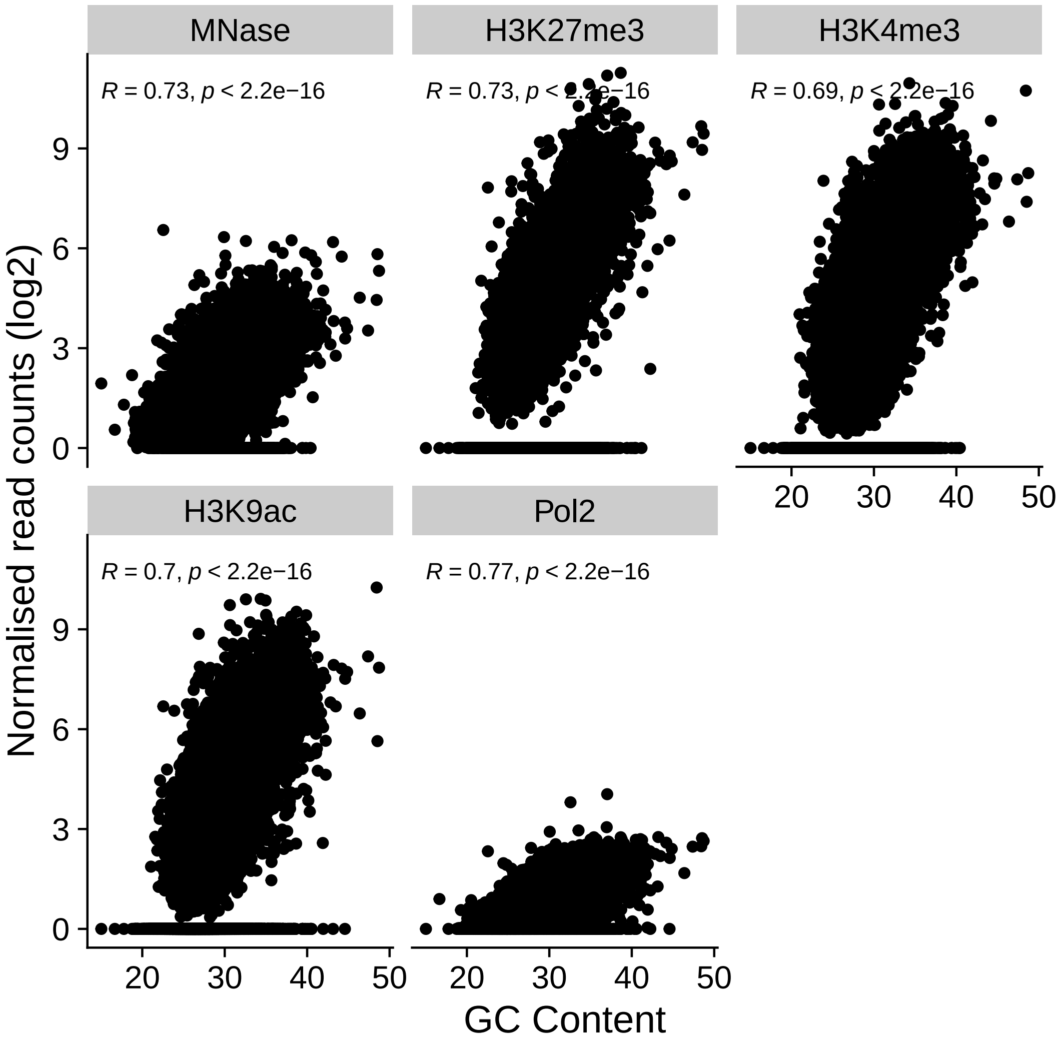


FIGURE C.3: Scatter plots of the GC content (x-axis) and the normalised read counts in the gene body for MNase, H3K27me3, H3K4me3, H3K9ac, and Pol2. The counts are normalised to the gene body length, and log2 transformed with a pseudocount of 1. The GC percentage is calculated for each gene. The Pearson correlation values are mentioned on the top left corner of the plots.

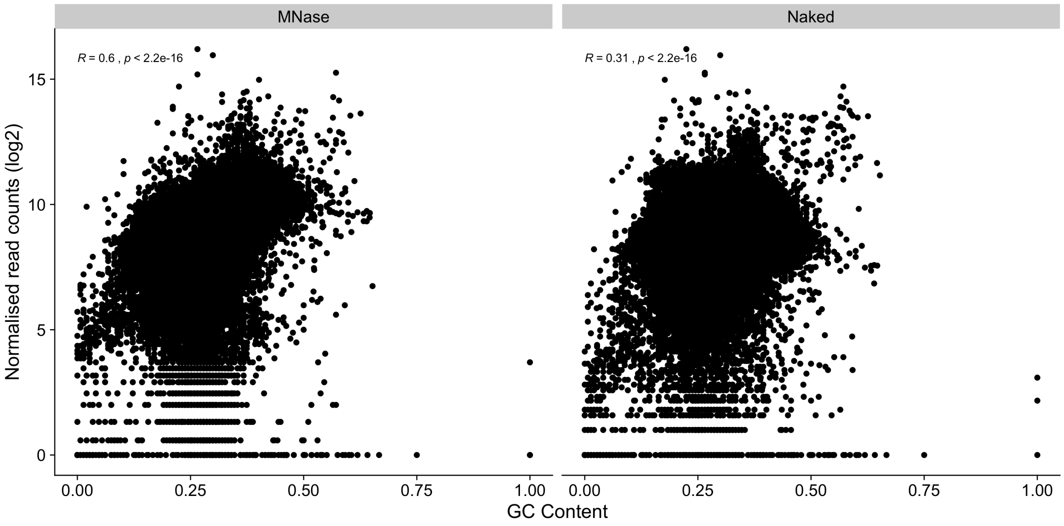


FIGURE C.4: Scatter plots of the GC content (x-axis) vs bin-length normalised read counts (y-axis; log2) of raw MNase and naked DNA measured in 147 bp bins of the genome. The Pearson correlation values are mentioned on the top left corner of the plots.

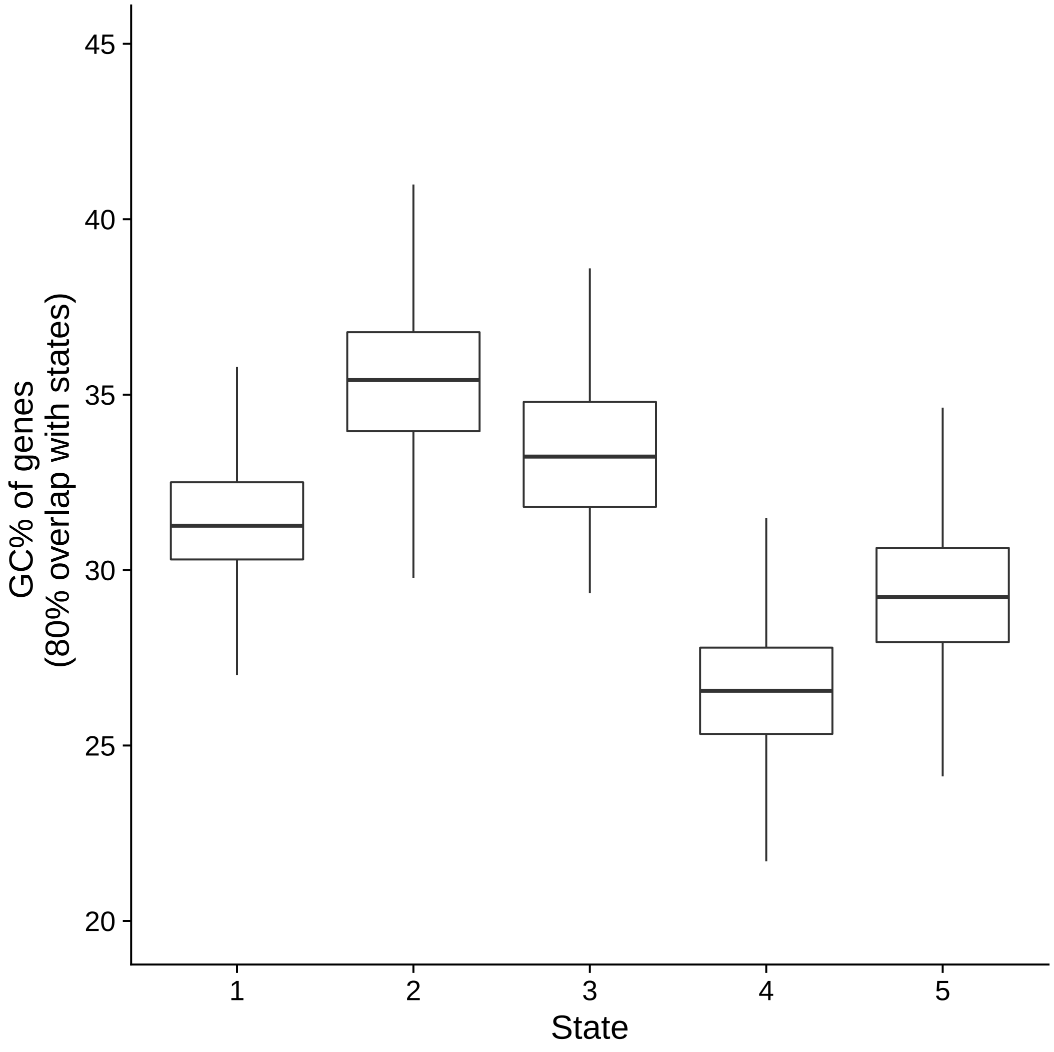


FIGURE C.5: GC content of the genes (y-axis) overlapping at least 80% with a chromatin state (x-axis) is shown.

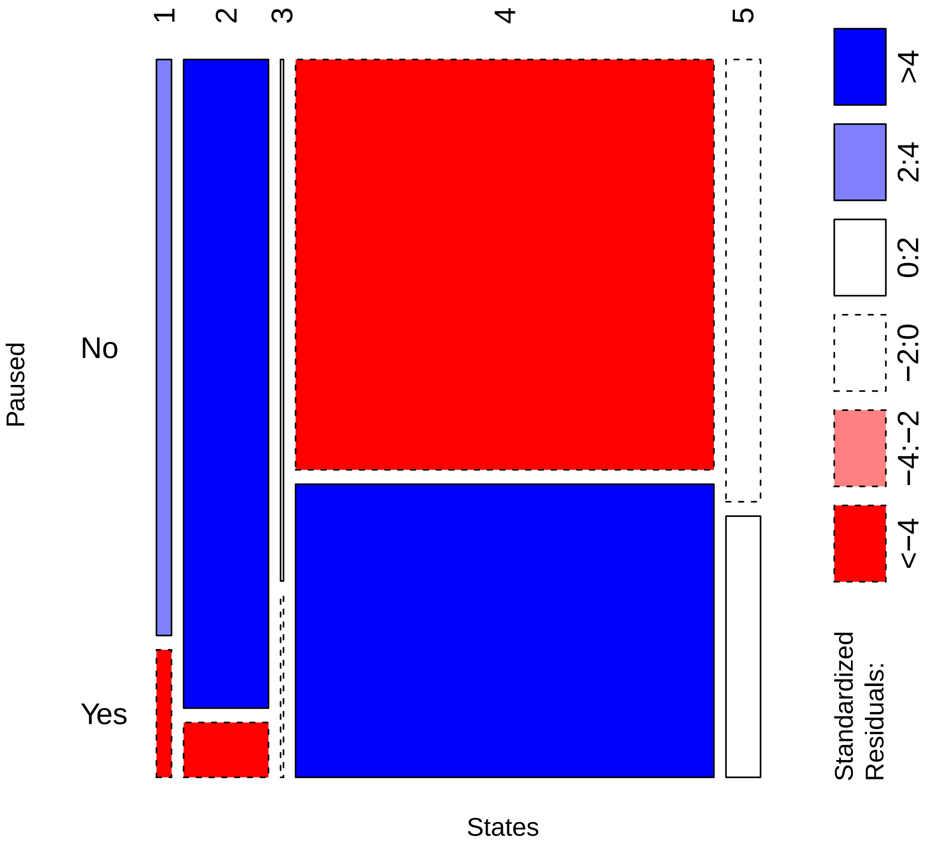


FIGURE C.6: A mosaic plot is shown for the enrichment of paused or unpaused genes (y-axis) in each chromatin state (x-axis). This mosaic plot shows the calculated chi-squared residuals, and to be interpreted as follows. The bar height and width represent the paused genes counts (scaled to 100%), and the total number of genes in each state (scaled to 100%). The blue and red colour indicate that the observed value is higher or lower than the expected value if the data were random, respectively.

Appendix D

Supplementary materials for Chapter 6

D.1 Supplementary Figures

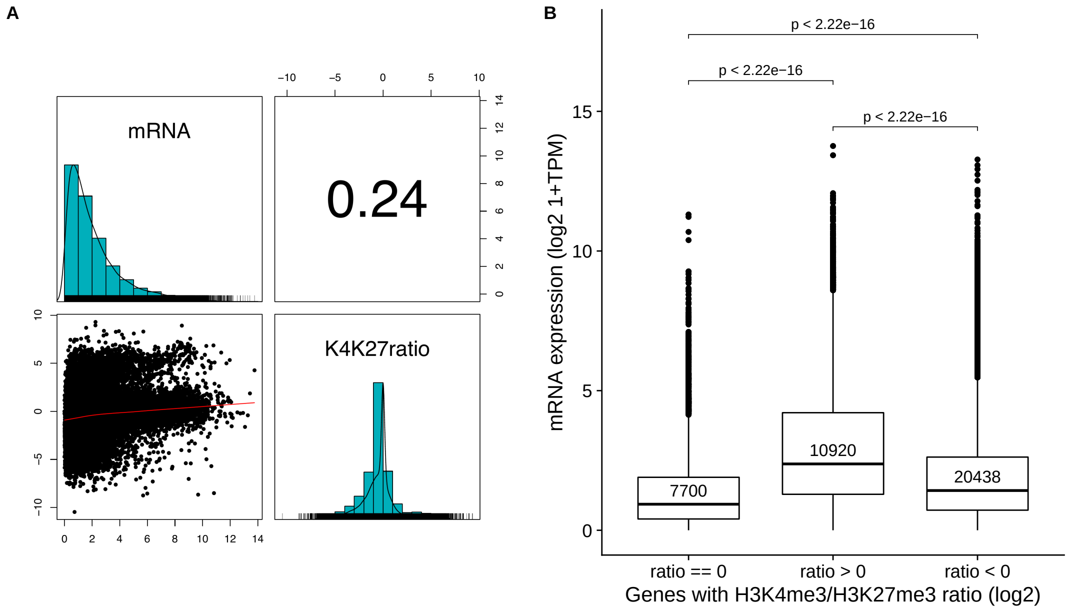


FIGURE D.1: A) The distribution plot of K4K27ratio (*i.e.* H3K4me3/H3K27me3) and mRNA are shown along the diagonal. The Pearson correlation coefficients (above the diagonal) are shown for the K4K27ratio and mRNA, and their scatter plots are shown below the diagonal. belongs to the variable mentioned along the horizontal line of that plot. The y-axis of scatter plot belongs to K4K27ratio, and x-axis belongs to the mRNA. Note: mRNA is shown in TPM units, and log2 transformed with a pseudocount of 1. B) The mRNA expression (y-axis; log2 1+TPM) is shown for genes with different H3K4me3/H3K27me3 ratio (x-axis; log2).

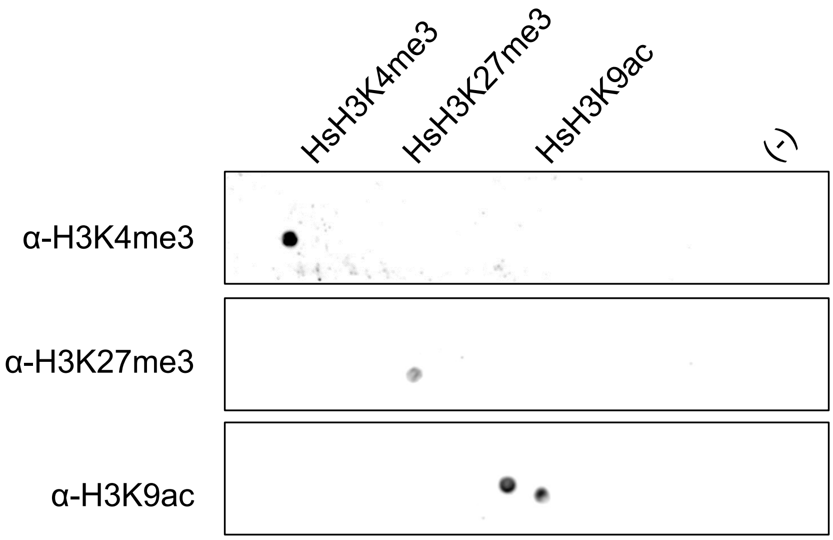
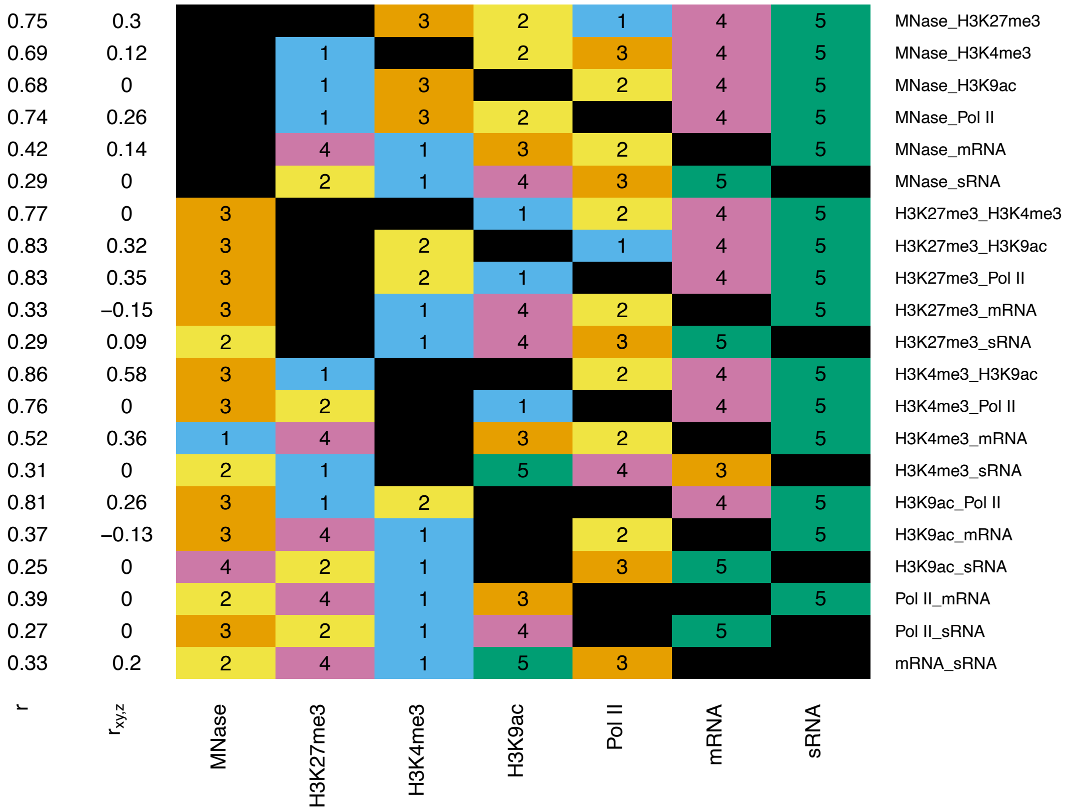


FIGURE D.2: A western (dot) blot showing that the histone mark antibodies used, are not cross-reacting with other histone marks. The rows represent the $\alpha-$ chain of the antibody used. The columns show the histone marks peptide, which the antibodies are supposed to recognise. A dot shows that the antibody recognises the respective histone mark peptide. The last column shows a negative control, which is not expected to show a dot. Please note that, we used the histone marks peptides designed for *Homo sapiens*(Hs), although the *Paramecium* histones (H3) have few mismatches in the H3 peptide tail.



r	$r_{xy,z}$	MNase	H3K27me3	H3K4me3	H3K9ac	Pol II	mRNA	sRNA	
0.75	0.3			3	2	1	4	5	MNase_H3K27me3
0.69	0.12		1		2	3	4	5	MNase_H3K4me3
0.68	0		1	3		2	4	5	MNase_H3K9ac
0.74	0.26		1	3	2		4	5	MNase_Pol II
0.42	0.14		4	1	3	2		5	MNase_mRNA
0.29	0		2	1	4	3	5		MNase_sRNA
0.77	0	3			1	2	4	5	H3K27me3_H3K4me3
0.83	0.32	3		2		1	4	5	H3K27me3_H3K9ac
0.83	0.35	3		2	1		4	5	H3K27me3_Pol II
0.33	−0.15	3		1	4	2		5	H3K27me3_mRNA
0.29	0.09	2		1	4	3	5		H3K27me3_sRNA
0.86	0.58	3	1			2	4	5	H3K4me3_H3K9ac
0.76	0	3	2		1		4	5	H3K4me3_Pol II
0.52	0.36	1	4		3	2		5	H3K4me3_mRNA
0.31	0	2	1		5	4	3		H3K4me3_sRNA
0.81	0.26	3	1	2			4	5	H3K9ac_Pol II
0.37	−0.13	3	4	1		2		5	H3K9ac_mRNA
0.25	0	4	2	1		3	5		H3K9ac_sRNA
0.39	0	2	4	1	3			5	Pol II_mRNA
0.27	0	3	2	1	4		5		Pol II_sRNA
0.33	0.2	2	4	1	5	3			mRNA_sRNA

FIGURE D.3: A heat map showing the ranks of the contribution of confounding factors (columns) to the difference they cause between the Pearson correlation coefficient (r) and the partial correlation coefficient ($r_{xy,z}$) of each correlation pair (rows). The confounding factors are colored as per their rank (mentioned in the rectangles; 1-5). The highest confounding factor is ranked 1, and the least confounding factor is given the rank 5. The black rectangles denote the correlation pair.

Appendix E

Scientific contributions

E.1 Peer-reviewed publications

First authorships

1. Exogenous RNAi mechanisms contribute to transcriptome adaptation by phased siRNA clusters in *Paramecium*, **2020**, Nucleic acids research 47 (15), 8036-8049. doi: `https://doi.org/10.1093/nar/gkz553`

2. Automated analysis of small RNA datasets with RAPID, **2019**, PeerJ 7, e6719. doi: `https://doi.org/10.7717/peerj.6710`

Shared first authorships

1. Epiregio: analysis and retrieval of regulatory elements linked to genes, **2020**, Nucleic acids research 48 (W1), W193-W199. doi: `https://doi.org/10.1093/nar/gkaa382`

2. Feeding exogenous dsRNA interferes with endogenous sRNA accumulation in *Paramecium*, **2020**, DNA Research 27 (1). doi: `https://doi.org/10.1093/dnares/dsaa005`

Co-authorships

1. Broad domains of histone marks in the highly condensed *Paramecium* macronucleus, **2021**, *manuscript in communication.* bioRxiv doi: `https://doi.org/10.1101/2021.08.05.454756`

2. Two Piwis with Ago-like functions silence somatic genes at the chromatin level, **2021**, RNA Biology. `https://doi.org/10.1080/15476286.2021.1991114`

3. Environmental temperature controls accumulation of transacting siRNAs involved in heterochromatin formation, **2018**, MDPI Genes 9 (2), 117. doi: `https://doi.org/10.3390/genes9020117`

E.2 Conference talks

1. Genome-wide small RNA profiling in *Paramecium tetraurelia*, Ciliate Molecular Biology Conference, Washington, DC, USA, July 17-23, **2018**.

2. Dynamic transcriptome adaptation by endogenous small RNA diversity in *Paramecium tetraurelia* GDRE 2017, Conference on Paramecium Epigenome Organization, Dynamics and Evolution, Nohfelden, Germany, October 3-6, **2017**.

E.3 Poster presentations

1. Epiregio: analysis and retrieval of regulatory elements linked to genes, ISCB Recomb Systems Genetics conference 2020, Virtual event, November 16-19, 2020.

2. Genome-wide small RNA profiling in *Paramecium tetraurelia*, 27th Conference on Intelligent Systems for Molecular Biology (ISMB) and 18th European Conference on Computational Biology (ECCB) 2019, Basel, Switzerland, July 21-25, 2019.

3. Cell type specific prediction of monoallelically expressed genes from human epigenomes, 17th European Conference on Computational Biology, Athens, Greece, Sep 8-12, 2018

4. Genome wide prediction of monoallelic gene expression from human epigenetic data, 22nd International Conference on Research in Computational Molecular Biology (RECOMB) 2018, Paris, France, April 21-24, 2018

5. Automated analysis and comparison of multiple small RNA datasets with RAPID, GDRE 2017: Conference on Paramecium Epigenome Organization, Dynamics and Evolution, Nohfelden, Germany, October 3-6, 2017.

6. Automated analysis and comparison of multiple small RNA datasets with RAPID, Joint 25th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) and 16th European Conference on Computational Biology (ECCB) 2017, Prague, Czech Republic, July 21-25, 2017.

E.4 Teaching assistant

1. Computational epigenomics course (summer semester 2021, summer semester 2020)

2. Workshops on analysis of RNA- and ChIP-seq data (winter semester 2020/2021, summer semester 2019)

E.5 Academic volunteering

1. Young researchers session coordinator, German conference on Bioinformatics (2020)

2. Equal opportunity working group specialist, MaxPlanck PhDnet

E.6 Reviewer activities

1. BMC Genomics - 2021, 2020, 2017

2. ECCB Conference - 2020

3. PLoS Computational Biology - 2019

4. ISMB ECCB Conference - 2019

5. F1000 research - 2018

6. ISMB Conference - 2018

Bibliography

Agarwal, Vikram and Jay Shendure (May 2020). “Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks”. en. In: *Cell Reports* 31.7, p. 107663. ISSN: 22111247. DOI: 10.1016/j.celrep.2020.107663. URL: https://linkinghub.elsevier.com/retrieve/pii/S2211124720306161 (visited on 05/25/2020).

Allen, Edwards et al. (Apr. 2005). “microRNA-Directed Phasing during Trans-Acting siRNA Biogenesis in Plants”. English. In: *Cell* 121.2. Publisher: Elsevier, pp. 207–221. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2005.04.004. URL: https://www.cell.com/cell/abstract/S0092-8674(05)00345-4 (visited on 01/01/2021).

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber (Sept. 2014). “HT-Seq—a Python framework to work with high-throughput sequencing data”. In: *Bioinformatics* 31.2, pp. 166–169. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu638. URL: https://doi.org/10.1093/bioinformatics/btu638.

Andrews, Simon et al. (Jan. 2015). *FastQC*. Babraham Institute. Babraham, UK.

Anzalone, A. V. et al. (2019). “Search-and-replace genome editing without double-strand breaks or donor DNA”. In: *Nature* 576.7785, pp. 149–157.

Arnaiz, O. et al. (2007). “ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data”. In: *Nucleic Acids Research* 35.Database issue, pp. D439–444. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkl777. URL: https://www.ncbi.nlm.nih.gov/pubmed/17142227https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1669747/pdf/gkl777.pdf.

ATDBio (2016). *Sequencing-forensic-analysis-and-genetic-analysis*. URL: http://www.atdbio.com/content/20/Sequencing-forensic-analysis-and-genetic-analysis.

Aufderheide, K. J. (1986). “Clonal aging in Paramecium tetraurelia. II. Evidence of functional changes in the macronucleus with age”. In: *Mech Ageing Dev* 37.3, pp. 265–279.

Aury, J. M. et al. (Nov. 2006). “Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia”. In: *Nature* 444.7116, pp. 171–178.

Avery, O. T., C. M. Macleod, and M. McCarty (1944). “ Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III”. In: *J Exp Med* 79.2, pp. 137–158.

Bannister, A. J. and T. Kouzarides (Mar. 2011). “Regulation of chromatin by histone modifications”. In: *Cell Res* 21.3, pp. 381–395.

Bauer, Sebastian et al. (2008). "Ontologizer 2.0:a multifunctional tool for GO term enrichment analysis and data exploration". In: *Bioinformatics* 24.14, pp. 1650–1651. DOI: `10.1093/bioinformatics/btn250`. eprint: `/oup/backfile/content_public/journal/bioinformatics/24/14/10.1093/bioinformatics/btn250/2/btn250.pdf`. URL: `+http://dx.doi.org/10.1093/bioinformatics/btn250`.

Baulcombe, David C. (Jan. 2007). "Amplified Silencing". en. In: *Science* 315.5809. Publisher: American Association for the Advancement of Science Section: Perspective, pp. 199–200. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.1138030`. URL: `https://science.sciencemag.org/content/315/5809/199` (visited on 01/01/2021).

Beale, G. and J.R. Preer (2008). *Paramecium: Genetics and Epigenetics*. NetLibrary, Inc. CRC Press. ISBN: 9780203491904. URL: `https://books.google.de/books?id=Qt\_WVjigjFIC`.

Beer, M. A. and S. Tavazoie (Apr. 2004). "Predicting gene expression from sequence". In: *Cell* 117.2, pp. 185–198.

Beisson, J. and T. M. Sonneborn (1965). "Cytoplasmic inheritance of the organization of the cell cortex in Paramecium aurelia". In: *Proc Natl Acad Sci U S A* 53, pp. 275–282.

Beisson, J. et al. (2010). "Paramecium tetraurelia: the renaissance of an early unicellular model". In: *Cold Spring Harb Protoc* 2010.1, pdb.emo140.

Bernstein, Bradley E. et al. (Jan. 2005). "Genomic maps and comparative analysis of histone modifications in human and mouse". eng. In: *Cell* 120.2, pp. 169–181. ISSN: 0092-8674. DOI: `10.1016/j.cell.2005.01.001`.

Bessière, C. et al. (Jan. 2018). "Probing instructions for expression regulation in gene nucleotide compositions". In: *PLoS Comput Biol* 14.1, e1005921.

Betermier, Mireille and Sandra Duharcourt (2017). "Programmed Rearrangement in Ciliates: Paramecium". en. In: p. 20.

Blanco, Enrique et al. (Feb. 2020). "The Bivalent Genome: Characterization, Structure, and Regulation". eng. In: *Trends in genetics: TIG* 36.2, pp. 118–131. ISSN: 0168-9525. DOI: `10.1016/j.tig.2019.11.004`.

Brown, Jay C. (2021). *Role of Gene Length in Control of Human Gene Expression: Chromosome-Specific and Tissue-Specific Effects*. en. Research Article. DOI: `https://doi.org/10.1155/2021/8902428`. URL: `https://www.hindawi.com/journals/ijg/2021/8902428/` (visited on 03/07/2021).

Browne, P. D. et al. (Feb. 2020). "GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms". In: *Gigascience* 9.2.

Buffenstein, Rochelle and J. Graham Ruby (2021). "Opportunities for new insight into aging from the naked mole-rat and other non-traditional models". en. In: *Nature Aging* 1.1, pp. 3–4. ISSN: 2662-8465. DOI: `10.1038/s43587-020-00012-4`. URL: `https://www.nature.com/articles/s43587-020-00012-4` (visited on 03/12/2021).

Capece, V. et al. (2015). "Oasis: online analysis of small RNA deep sequencing data". In: *Bioinformatics* 31.13, pp. 2205–2207. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btv113`.

Carradec, Q. et al. (2015). "Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate Paramecium tetraurelia". In: *Nucleic*

Acids Research 43.3, pp. 1818–33. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gku1331. URL: https://www.ncbi.nlm.nih.gov/pubmed/25593325https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4330347/pdf/gku1331.pdf.

Carter, Shelli and Lumen Learning (2021). *Introduction to Groups of Protists.* [Online; accessed 04-April-2021]. URL: https://courses.lumenlearning.com/ivytech-bio1-1/chapter/groups-of-protists/.

Carthew, R. W. and E. J. Sontheimer (2009). "Origins and Mechanisms of miRNAs and siRNAs". In: *Cell* 136.4, pp. 642–655. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking). DOI: 10.1016/j.cell.2009.01.035. URL: https://www.ncbi.nlm.nih.gov/pubmed/19239886https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675692/pdf/nihms-107442.pdf.

Chalker, Douglas L., Eric Meyer, and Kazufumi Mochizuki (Dec. 2013). "Epigenetics of Ciliates". In: *Cold Spring Harbor Perspectives in Biology* 5.12. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a017764. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3839606/ (visited on 10/22/2020).

Charpentier, E. and J. A. Doudna (2013). "Biotechnology: Rewriting a genome". In: *Nature* 495.7439, pp. 50–51.

Cheaib, M. et al. (2015). "Epigenetic regulation of serotype expression antagonizes transcriptome dynamics in Paramecium tetraurelia". In: *DNA Research* 22.4, pp. 293–305. ISSN: 1756-1663 (Electronic) 1340-2838 (Linking). DOI: 10.1093/dnares/dsv014. URL: https://www.ncbi.nlm.nih.gov/pubmed/26231545https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4535620/pdf/dsv014.pdf.

Chen, Chong Jian et al. (2012). "NcPRO-seq: A tool for annotation and profiling of ncRNAs in sRNA-seq data". In: *Bioinformatics* 28.23, pp. 3147–3149. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts587.

Chen, Kaifu et al. (Jan. 2013). "DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing". en. In: *Genome Research* 23.2, pp. 341–351. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.142067.112. URL: http://genome.cshlp.org/content/23/2/341 (visited on 08/01/2019).

CHEN, T. T. (June 1946). "Varieties and mating types in Paramecium bursaria; new variety and types, from England, Ireland, and Czechoslovakia". In: *Proc Natl Acad Sci U S A* 32, pp. 173–181.

Cheng, Chao et al. (2011). "A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets". In: *Genome Biology* 12.2, R15. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-2-r15.

Chung, Ho-Ryun et al. (Dec. 2010). "The Effect of Micrococcal Nuclease Digestion on Nucleosome Positioning Data". In: *PLoS ONE* 5.12. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0015754. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3012088/ (visited on 02/18/2021).

Cobb, M. (2017). "60 years ago, Francis Crick changed the logic of biology". In: *PLoS Biol* 15.9, e2003243.

Collins, F. S., M. Morgan, and A. Patrinos (2003). "The Human Genome Project: lessons from large-scale biology". In: *Science* 300.5617, pp. 286–290.

Cusick, M. E. et al. (Nov. 1981). "Structure of chromatin at deoxyribonucleic acid replication forks: prenucleosomal deoxyribonucleic acid is rapidly excised from replicating simian virus 40 chromosomes by micrococcal nuclease". In: *Biochemistry* 20.23, pp. 6648–6658.

Davis, Jesse and Mark Goadrich (2006). "The Relationship between Precision-Recall and ROC Curves". In: *Proceedings of the 23rd International Conference on Machine Learning.* ICML '06. Association for Computing Machinery, 233–240. ISBN: 1595933832. DOI: 10.1145/1143844.1143874. URL: https://doi.org/10.1145/1143844.1143874.

Deshmukh, Rupesh K., Humira Sonah, and Nagendra K. Singh (2016). "Intron gain, a dominant evolutionary process supporting high levels of gene expression in rice". en. In: *Journal of Plant Biochemistry and Biotechnology* 25.2, pp. 142–146. ISSN: 0974-1275. DOI: 10.1007/s13562-015-0319-5. URL: https://doi.org/10.1007/s13562-015-0319-5 (visited on 03/07/2021).

Di Resta, C. et al. (2018). "Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities". In: *EJIFCC* 29.1, pp. 4–14.

Dillies, M A et al. (2013). "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". In: *Briefings in bioinformatics.* ISSN: 1477-4054; 1467-5463. DOI: 10.1093/bib/bbs046[doi].

Dong, Xianjun et al. (June 2012). "Modeling gene expression using chromatin features in various cellular contexts". In: *Genome Biology* 13.9, R53. ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-9-r53.

Doshi-Velez, Finale and Been Kim (2017). *Towards A Rigorous Science of Interpretable Machine Learning.* arXiv: 1702.08608 [stat.ML].

Duret, Laurent and Dominique Mouchiroud (1999). "Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis". en. In: *Proceedings of the National Academy of Sciences* 96.8, pp. 4482–4487. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.96.8.4482. URL: https://www.pnas.org/content/96/8/4482 (visited on 03/07/2021).

Elkon, Ran and Reuven Agami (Aug. 2017). "Characterization of noncoding regulatory DNA in the human genome". en. In: *Nature Biotechnology* 35.8. Number: 8 Publisher: Nature Publishing Group, pp. 732–746. ISSN: 1546-1696. DOI: 10.1038/nbt.3863. URL: https://www.nature.com/articles/nbt.3863 (visited on 02/19/2021).

Epstein, L. M. and J. D. Forney (1984). "Mendelian and non-mendelian mutations affecting surface antigen expression in Paramecium tetraurelia". In: *Mol Cell Biol* 4.8, pp. 1583–1590.

Ernst, Jason and Manolis Kellis (Dec. 2017). "Chromatin-state discovery and genome annotation with ChromHMM". en. In: *Nature Protocols* 12.12. Number: 12 Publisher: Nature Publishing Group, pp. 2478–2492. ISSN: 1750-2799. DOI: 10.1038/nprot.2017.124. URL: https://www.nature.com/articles/nprot.2017.124 (visited on 10/15/2020).

Felsenfeld, G. (2014). "A brief history of epigenetics". In: *Cold Spring Harb Perspect Biol* 6.1.

Fenouil, Romain et al. (Dec. 2012). “CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters”. eng. In: *Genome Research* 22.12, pp. 2399–2408. ISSN: 1549-5469. DOI: 10.1101/gr.138776.112.

Fields, S. and M. Johnston (2005). “Cell biology. Whither model organism research?” In: *Science* 307.5717, pp. 1885–1886.

Fire, A. et al. (Feb. 1998). “Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans”. eng. In: *Nature* 391.6669, pp. 806–811. ISSN: 0028-0836. DOI: 10.1038/35888.

Frank, E., M. A. Hall, and I. H. Witten (2016). “The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"”. In: *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Ed. by O. Maimon and L. Rokach. Berlin: Morgan Kaufmann. URL: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.

Frapporti, Andrea et al. (June 2019). “The Polycomb protein Ezl1 mediates H3K9 and H3K27 methylation to repress transposable elements in Paramecium”. en. In: *Nature Communications* 10.1. Number: 1 Publisher: Nature Publishing Group, p. 2710. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10648-5. URL: https://www.nature.com/articles/s41467-019-10648-5 (visited on 10/22/2020).

French, J. D. and S. L. Edwards (2020). “The Role of Noncoding Variants in Heritable Disease”. In: *Trends Genet* 36.11, pp. 880–891.

Friedländer, Marc R. et al. (2012). “MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades”. In: *Nucleic Acids Research* 40.1, pp. 37–52. ISSN: 03051048. DOI: 10.1093/nar/gkr688.

Galvani, Angélique and Linda Sperling (2002). “RNA interference by feeding in Paramecium”. In: *Trends in Genetics* 18.1, pp. 11–12. DOI: 10.1016/s0168-9525(01)02548-3.

Gates, Leah A., Charles E. Foulds, and Bert W. O’Malley (Dec. 2017). “Histone marks in the ‘drivers seat’: functional roles in steering the transcription cycle”. In: *Trends in biochemical sciences* 42.12, pp. 977–989. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2017.10.004. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5701853/ (visited on 03/07/2021).

Gates, Leah A. et al. (Sept. 2017). “Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation”. In: *The Journal of Biological Chemistry* 292.35, pp. 14456–14472. ISSN: 0021-9258. DOI: 10.1074/jbc.M117.802074. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5582839/ (visited on 03/07/2021).

Gebert, Luca F. R. and Ian J. MacRae (Jan. 2019). “Regulation of microRNA function in animals”. en. In: *Nature Reviews Molecular Cell Biology* 20.1. Number: 1 Publisher: Nature Publishing Group, pp. 21–37. ISSN: 1471-0080. DOI: 10.1038/s41580-018-0045-7. URL: https://www.nature.com/articles/s41580-018-0045-7 (visited on 01/01/2021).

Giurato, Giorgio et al. (2013). “IMir: An integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq”. In:

BMC Bioinformatics 14, p. 362. ISSN: 14712105. DOI: 10.1186/1471-2105-14-362.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning.* http://www.deeplearningbook.org. MIT Press.

Götz, Ulrike et al. (2016). "Two sets of RNAi components are required for heterochromatin formation in trans triggered by truncated transgenes". In: *Nucleic Acids Research* 44 (12), pp. 5908–5923. ISSN: 13624962. DOI: 10.1093/nar/gkw267.

Griffith, F. (1928). "The Significance of Pneumococcal Types". In: *J Hyg (Lond)* 27.2, pp. 113–159.

Grüning, Björn et al. (July 2018). "Bioconda: sustainable and comprehensive software distribution for the life sciences". en. In: *Nature Methods* 15.7. Number: 7 Publisher: Nature Publishing Group, pp. 475–476. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0046-7. URL: https://www.nature.com/articles/s41592-018-0046-7 (visited on 02/24/2021).

Ha, Misook et al. (Apr. 2011). "Coordinated histone modifications are associated with gene expression variation within and between species". In: *Genome Research* 21.4, pp. 590–598. ISSN: 1088-9051. DOI: 10.1101/gr.116467.110. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3065706/ (visited on 03/07/2021).

Han, Bo W. et al. (2015). "PiPipes: A set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome-and CAGE-seq, ChIP-seq and genomic DNA sequencing". In: *Bioinformatics* 31.4, pp. 593–595. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu647.

Han, P. and C. P. Chang (2015). "Long non-coding RNA and chromatin remodeling". In: *RNA Biol* 12.10, pp. 1094–1098.

Heather, J. M. and B. Chain (Jan. 2016). "The sequence of sequencers: The history of sequencing DNA". In: *Genomics* 107.1, pp. 1–8.

Herron, M. D. (2016). "Origins of multicellular complexity: Volvox and the volvocine algae". In: *Mol Ecol* 25.6, pp. 1213–1223.

Holoch, Daniel and Danesh Moazed (Feb. 2015). "RNA-mediated epigenetic regulation of gene expression". en. In: *Nature Reviews Genetics* 16.2. Number: 2 Publisher: Nature Publishing Group, pp. 71–84. ISSN: 1471-0064. DOI: 10.1038/nrg3863. URL: https://www.nature.com/articles/nrg3863 (visited on 03/07/2021).

Howell, Miya D et al. (Mar. 2007). "Genome-Wide Analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 Pathway in Arabidopsis Reveals Dependency on miRNA- and tasiRNA-Directed Targeting". In: *The Plant Cell* 19.3, pp. 926–942.

Huang, Wen et al. (Jan. 2011). "MTML-msBayes: approximate Bayesian comparative phylogeographic inference from multiple taxa and multiple loci with rate heterogeneity". eng. In: *BMC bioinformatics* 12, p. 1. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-1.

Hunter, P. (2008). "The paradox of model organisms. The use of model organisms in research will continue despite their shortcomings". In: *EMBO Rep* 9.8, pp. 717–720.

Hwang, W. Y. et al. (2013). "Efficient genome editing in zebrafish using a CRISPR-Cas system". In: *Nat Biotechnol* 31.3, pp. 227–229.

Icay, Katherine et al. (2016). "SePIA: RNA and small RNA sequence processing, integration, and analysis". In: *BioData Mining* 9, p. 20. ISSN: 17560381. DOI: 10.1186/s13040-016-0099-z.

Ignarski, Michael et al. (Oct. 2014). "Paramecium tetraurelia chromatin assembly factor-1-like protein PtCAF-1 is involved in RNA-mediated control of DNA elimination". In: *Nucleic Acids Research* 42.19, pp. 11952–11964. ISSN: 0305-1048. DOI: 10.1093/nar/gku874. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4231744/ (visited on 10/21/2020).

Ishino, Y. et al. (1987). "Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product". In: *J Bacteriol* 169.12, pp. 5429–5433.

Jaillon, Olivier et al. (Jan. 2008). "Translational control of intron splicing in eukaryotes". en. In: *Nature* 451.7176. Number: 7176 Publisher: Nature Publishing Group, pp. 359–362. ISSN: 1476-4687. DOI: 10.1038/nature06495. URL: https://www.nature.com/articles/nature06495 (visited on 10/21/2020).

James, Gareth et al. (June 2013). *An Introduction to Statistical Learning*. 2013th ed. Vol. 103. Springer Series in Statistics. Springer New York. ISBN: 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7. URL: http://dx.doi.org/10.1007/978-1-4614-7138-7.

Jiang, W. et al. (2013). "RNA-guided editing of bacterial genomes using CRISPR-Cas systems". In: *Nat Biotechnol* 31.3, pp. 233–239.

Jo, B. S. and S. S. Choi (Dec. 2015). "Introns: The Functional Benefits of Introns in Genomes". In: *Genomics Inform* 13.4, pp. 112–118.

Johnson, Nathan R. et al. (2016). "Improved Placement of Multi-mapping Small RNAs". In: *G3 & Genes-Genomes-Genetics* 6.7, pp. 2103–2111. ISSN: 2160-1836. DOI: 10.1534/g3.116.030452.

Karlić, Rosa et al. (Feb. 2010). "Histone modification levels are predictive for gene expression". eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.7, pp. 2926–2931. ISSN: 1091-6490. DOI: 10.1073/pnas.0909344107.

Karunanithi, Sivarajan, Martin Simon, and Marcel H. Schulz (Apr. 2019). "Automated analysis of small RNA datasets with RAPID". In: *PeerJ* 7, e6710. ISSN: 2167-8359. DOI: 10.7717/peerj.6710. URL: https://doi.org/10.7717/peerj.6710.

Karunanithi, Sivarajan et al. (June 2019). "Exogenous RNAi mechanisms contribute to transcriptome adaptation by phased siRNA clusters in Paramecium". In: *Nucleic Acids Research* 47.15, pp. 8036–8049. ISSN: 0305-1048. DOI: 10.1093/nar/gkz553. eprint: https://academic.oup.com/nar/article-pdf/47/15/8036/29213734/gkz553.pdf. URL: https://doi.org/10.1093/nar/gkz553.

Karunanithi, Sivarajan et al. (Apr. 2020). "Feeding exogenous dsRNA interferes with endogenous sRNA accumulation in Paramecium". In: *DNA Research* 27.1. dsaa005. ISSN: 1756-1663. DOI: 10.1093/dnares/dsaa005. eprint: https://academic.oup.com/dnaresearch/article-pdf/27/1/dsaa005/

33426925/dsaa005.pdf. URL: https://doi.org/10.1093/dnares/dsaa005.

Khurana, J. S. et al. (Jan. 2018). "Small RNA-mediated regulation of DNA dosage in the ciliate Oxytricha". In: *RNA* 24.1, pp. 18–29.

Kim, Y., H. G. Nam, and D. R. Valenzano (2016). "The short-lived African turquoise killifish: an emerging experimental model for ageing". In: *Dis Model Mech* 9.2, pp. 115–129.

King, A. M., C. K. Vanderpool, and P. H. Degnan (Jan. 2019). "sRNA Target Prediction Organizing Tool (SPOT) Integrates Computational and Experimental Data To Facilitate Functional Characterization of Bacterial Small RNAs". In: *mSphere* 4.1.

Krzyszczyk, P. et al. (2018). "The growing role of precision and personalized medicine for cancer treatment". In: *Technology (Singap World Sci)* 6.3-4, pp. 79–100.

Kuksa, Pavel P. et al. (2018). "SPAR: Small RNA-seq portal for analysis of sequencing experiments". In: *Nucleic Acids Research* 46.W1, W36–W42. ISSN: 13624962. DOI: 10.1093/nar/gky330.

Lahortiga, Idoya and Luk Cox. *Somersault1824*. URL: https://www.somersault1824.com/.

Lai, Qingwei et al. (2017). "H3K9ac and HDAC2 Activity Are Involved in the Expression of Monocarboxylate Transporter 1 in Oligodendrocyte". eng. In: *Frontiers in Molecular Neuroscience* 10, p. 376. ISSN: 1662-5099. DOI: 10.3389/fnmol.2017.00376.

Langmead, Ben and Steven L Salzberg (Mar. 2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4, pp. 357–359. ISSN: 1548-7091. DOI: 10.1038/nmeth.1923. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/.

Lasserre, Julia, Ho-Ryun Chung, and Martin Vingron (Sept. 2013). "Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks". en. In: *PLOS Computational Biology* 9.9. Publisher: Public Library of Science, e1003168. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003168. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003168 (visited on 07/08/2020).

Ledford, H. (2020). "Quest to use CRISPR against disease gains ground". In: *Nature* 577.7789, p. 156.

Lengauer, T., N. Pfeifer, and R. Kaiser (2014). "Personalized HIV therapy to control drug resistance". In: *Drug Discov Today Technol* 11, pp. 57–64.

Li, Bing, Michael Carey, and Jerry L. Workman (Feb. 2007). "The Role of Chromatin during Transcription". English. In: *Cell* 128.4. Publisher: Elsevier, pp. 707–719. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2007.01.015. URL: https://www.cell.com/cell/abstract/S0092-8674(07)00109-2 (visited on 03/07/2021).

Li, Heng et al. (2009). "The Sequence Alignment / Map format and SAMtools". In: *Bioinformatics* 25 (16), pp. 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

Li, Sisi and Dinshaw J Patel (May 2016). "Drosha and Dicer: Slicers cut from the same cloth". In: *Cell Research* 26.5, pp. 511–512. ISSN: 1001-0602. DOI:

10.1038/cr.2016.19. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4856758/ (visited on 01/01/2021).

Li, Yang et al. (Oct. 2013). “RNA interference functions as an antiviral immunity mechanism in mammals”. eng. In: *Science (New York, N.Y.)* 342.6155, pp. 231–234. ISSN: 1095-9203. DOI: 10.1126/science.1241911.

Liu, Chih Long et al. (Oct. 2005). “Single-nucleosome mapping of histone modifications in S. cerevisiae”. eng. In: *PLoS biology* 3.10, e328. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0030328.

Love, Michael I., Wolfgang Huber, and Simon Anders (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15.12, p. 550. DOI: 10.1186/s13059-014-0550-8.

Lundberg, Scott M. et al. (2020). “From local explanations to global understanding with explainable AI for trees”. In: *Nature Machine Intelligence* 2.1, pp. 2522–5839.

Maillard, P. V. et al. (Oct. 2013). “Antiviral RNA interference in mammalian cells”. eng. In: *Science (New York, N.Y.)* 342.6155, pp. 235–238. ISSN: 1095-9203. DOI: 10.1126/science.1241930.

Marco-Sola, Santiago et al. (2012). “The GEM mapper: fast, accurate and versatile alignment by filtration”. en. In: p. 7.

Marker, S. et al. (2010). “Distinct RNA-dependent RNA polymerases are required for RNAi triggered by double-stranded RNA versus truncated transgenes in Paramecium tetraurelia”. In: *Nucleic Acids Research* 38.12, pp. 4092–107. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/nar/gkq131. URL: https://www.ncbi.nlm.nih.gov/pubmed/20200046https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896523/pdf/gkq131.pdf.

Marker, Simone et al. (2014). “A forward genetic screen reveals essential and non-essential RNAi factors in Paramecium tetraurelia”. In: *Nucleic Acids Research* 42.11, pp. 7268–7280.

Martin, Marcel (2011). “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1, pp. 10–12. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: https://journal.embnet.org/index.php/embnetjournal/article/view/200.

Maston, G. A., S. K. Evans, and M. R. Green (2006). “Transcriptional regulatory elements in the human genome”. In: *Annu Rev Genomics Hum Genet* 7, pp. 29–59.

Matsuda, A. and J. D. Forney (2005). “Analysis of Paramecium tetraurelia A-51 surface antigen gene mutants reveals positive-feedback mechanisms for maintenance of expression and temperature-induced activation”. In: *Eukaryot Cell* 4.10, pp. 1613–1619.

Mavrich, Travis N. et al. (May 2008). “Nucleosome organization in the Drosophila genome”. en. In: *Nature* 453.7193. Number: 7193 Publisher: Nature Publishing Group, pp. 358–362. ISSN: 1476-4687. DOI: 10.1038/nature06929. URL: https://www.nature.com/articles/nature06929 (visited on 02/21/2021).

McCarty, M. (2003). “Discovering genes are made of DNA”. In: *Nature* 421.6921, p. 406.

Meister, Gunter and Thomas Tuschl (Sept. 2004). "Mechanisms of gene silencing by double-stranded RNA". en. In: *Nature* 431.7006, pp. 343–349. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature02873. URL: http://www.nature.com/articles/nature02873 (visited on 12/31/2020).

Mignone, F. et al. (2002). "Untranslated regions of mRNAs". In: *Genome Biol* 3.3, REVIEWS0004.

Moazed, Danesh (Jan. 2009). "Small RNAs in transcriptional gene silencing and genome defence". eng. In: *Nature* 457.7228, pp. 413–420. ISSN: 1476-4687. DOI: 10.1038/nature07756. URL: https://pubmed.ncbi.nlm.nih.gov/19158787.

Mohorianu, Irina et al. (2017). "The UEA small RNA workbench: A suite of computational tools for small RNA analysis". In: *Methods in Molecular Biology*. Vol. 1580. Springer, pp. 193–224. ISBN: 9781493968664. DOI: 10.1007/978-1-4939-6866-4{_}14.

Molnar, Christoph (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/.

Momozawa, Y. and K. Mizukami (2021). "Unique roles of rare variants in the genetics of complex diseases in humans". In: *J Hum Genet* 66.1, pp. 11–23.

Moran, V. A., R. J. Perera, and A. M. Khalil (Aug. 2012). "Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs". In: *Nucleic Acids Res* 40.14, pp. 6391–6400.

Moxon, Simon et al. (2008). "A toolkit for analysing large-scale plant small RNA datasets". In: *Bioinformatics* 24.9, pp. 2252–2253. ISSN: 13674803. DOI: 10.1093/bioinformatics/btn428.

Müller, Fabian (Oct. 2017). *Epigenomic marks associated with euchromatin and heterochromatin.* DOI: 10.6084/m9.figshare.5057566.v1. URL: https://figshare.com/articles/figure/Epigenomic_marks_associated_with_euchromatin_and_heterochromatin/5057566/1.

Nelson, Craig E, Bradley M Hersh, and Sean B Carroll (2004). "The regulatory content of intergenic DNA shapes genome architecture". In: *Genome Biology* 5.4, R25. ISSN: 1465-6906. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC395784/ (visited on 10/22/2020).

O'Connor, C. (2008). "Discovery of DNA as the hereditary material using Streptococcus pneumoniae". In: *Nature Education* 1.1, p. 104. URL: https://www.nature.com/scitable/topicpage/discovery-of-dna-as-the-hereditary-material-340/.

Orias, Eduardo, Deepankar Pratap Singh, and Eric Meyer (2017). "Genetics and Epigenetics of Mating Type Determination in Paramecium and Tetrahymena". In: *Annual Review of Microbiology* 71.1. _eprint: https://doi.org/10.1146/annurev-micro-090816-093342, pp. 133–156. DOI: 10.1146/annurev-micro-090816-093342. URL: https://doi.org/10.1146/annurev-micro-090816-093342 (visited on 10/22/2020).

Panero, Riccardo et al. (2017). "ISmaRT: A toolkit for a comprehensive analysis of small RNA-Seq data". In: *Bioinformatics* 33.16, pp. 938–940. ISSN: 14602059. DOI: 10.1093/bioinformatics/btw734.

Patro, Rob et al. (Apr. 2017). "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4, pp. 417–419. ISSN: 1548-7091. DOI: 10.1038/nmeth.4197. URL: http:https://doi.org/10.1038/nmeth.4197.

Patwardhan, Bhushan et al. (2017). "Chapter 8 - Transcriptomics and Epigenomics". In: *Innovative Approaches in Drug Discovery*. Elsevier, p. 235. ISBN: 9780128018149. DOI: 10.1016/B978-0-12-801814-9.00008-8.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830. URL: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Pennisi, E. (2012). "Genomics. ENCODE project writes eulogy for junk DNA". In: *Science* 337.6099, pp. 1159–1161.

Pilling, Olivia A. et al. (Oct. 2017). "Insights into transgenerational epigenetics from studies of ciliates". eng. In: *European Journal of Protistology* 61.Pt B, pp. 366–375. ISSN: 1618-0429. DOI: 10.1016/j.ejop.2017.05.004.

Quek, Camelia et al. (2015). "iSRAP - A one-touch research tool for rapid profiling of small RNA-seq data". In: *Journal of Extracellular Vesicles* 4, p. 29454. ISSN: 20013078. DOI: 10.3402/jev.v4.29454.

Quinlan, Aaron R and Ira M Hall (Mar. 2010). "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842. DOI: 10.1093/bioinformatics/btq033.

Ramírez, Fidel et al. (Apr. 2016). "deepTools2: a next generation web server for deep-sequencing data analysis". In: *Nucleic Acids Research* 44.W1, W160–W165. ISSN: 0305-1048. DOI: 10.1093/nar/gkw257. eprint: https://academic.oup.com/nar/article-pdf/44/W1/W160/7631613/gkw257.pdf. URL: https://doi.org/10.1093/nar/gkw257.

Ratti, M. et al. (June 2020). "MicroRNAs (miRNAs) and Long Non-Coding RNAs (lncRNAs) as New Tools for Cancer Therapy: First Steps from Bench to Bedside". In: *Target Oncol* 15.3, pp. 261–278.

Robinson, James T et al. (Jan. 2011). "Integrative genomics viewer". In: *Nature Biotechnology* 29, p. 24.

Rose, Alan B. (2019). "Introns as Gene Regulators: A Brick on the Accelerator". English. In: *Frontiers in Genetics* 9. ISSN: 1664-8021. DOI: 10.3389/fgene.2018.00672. URL: https://www.frontiersin.org/articles/10.3389/fgene.2018.00672/full (visited on 03/07/2021).

Rosenkranz, David et al. (2015). "Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence". In: *Genomics Data* 5, pp. 309–313. ISSN: 22135960. DOI: 10.1016/j.gdata.2015.06.026.

Rueda, Antonio et al. (2015). "SRNAtoolbox: An integrated collection of small RNA research tools". In: *Nucleic Acids Research* 43.W1, W467–W473. ISSN: 13624962. DOI: 10.1093/nar/gkv555.

Russell, J. J. et al. (2017). "Non-model model organisms". In: *BMC Biol* 15.1, p. 55.

Sallam, H. N. (2010). "Aristotle, godfather of evidence-based medicine". In: *Facts Views Vis Obgyn* 2.1, pp. 11–19.

Saudemont, Baptiste et al. (Oct. 2017). "The fitness cost of mis-splicing is the main determinant of alternative splicing patterns". In: *Genome Biology* 18.1,

p. 208. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1344-6. URL: https://doi.org/10.1186/s13059-017-1344-6 (visited on 02/21/2021).

Schmidt, Florian. *Florian Schmidt.* URL: https://scholar.google.com/citations?user=EIMLugsAAAAJ&hl=en.

Sen, Subhojit et al. (Sept. 2016). "Genome-wide positioning of bivalent mononucleosomes". en. In: *BMC Medical Genomics* 9.1, p. 60. ISSN: 1755-8794. DOI: 10.1186/s12920-016-0221-6. URL: https://doi.org/10.1186/s12920-016-0221-6 (visited on 10/22/2020).

Setten, Ryan L., John J. Rossi, and Si-ping Han (June 2019). "The current state and future directions of RNAi-based therapeutics". en. In: *Nature Reviews Drug Discovery* 18.6, pp. 421–446. ISSN: 1474-1776, 1474-1784. DOI: 10.1038/s41573-019-0017-4. URL: http://www.nature.com/articles/s41573-019-0017-4 (visited on 12/30/2020).

Shabalina, Svetlana A. and Nikolay A. Spiridonov (Mar. 2004). "The mammalian transcriptome and the function of non-coding DNA sequences". In: *Genome Biology* 5.4, p. 105. ISSN: 1474-760X. DOI: 10.1186/gb-2004-5-4-105. URL: https://doi.org/10.1186/gb-2004-5-4-105 (visited on 02/19/2021).

Shapley, Lloyd (1953). *Contributions to the Theory of Games (AM-28), Volume II. A value for n-person games.*

Shaul, Orit (Oct. 2017). "How introns enhance gene expression". eng. In: *The International Journal of Biochemistry & Cell Biology* 91.Pt B. ISSN: 1878-5875. DOI: 10.1016/j.biocel.2017.06.016.

Shendure, J. et al. (Oct. 2017). "DNA sequencing at 40: past, present and future". In: *Nature* 550.7676. [DOI:10.1038/nature24286] [PubMed:18243105], pp. 345–353.

Simon, M. C., S. Marker, and H. J. Schmidt (2006). "Posttranscriptional control is a strong factor enabling exclusive expression of surface antigens in Paramecium tetraurelia". In: *Gene Expr* 13.3, pp. 167–178.

Simon, Martin. *Martin Simon.* URL: https://scholar.google.de/citations?user=14dOBuMAAAAJ&hl=de.

Simon, Martin and Helmut Plattner (2014). "Unicellular Eukaryotes as Models in Cell and Molecular Biology". en. In: *International Review of Cell and Molecular Biology.* Vol. 309. Elsevier, pp. 141–198. ISBN: 978-0-12-800255-1. DOI: 10.1016/B978-0-12-800255-1.00003-X. URL: https://linkinghub.elsevier.com/retrieve/pii/B978012800255100003X (visited on 10/15/2020).

Singh, Ritambhara et al. (Sept. 2016). "DeepChrome: deep-learning for predicting gene expression from histone modifications". In: *Bioinformatics* 32.17, pp. i639–i648. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw427. URL: https://doi.org/10.1093/bioinformatics/btw427 (visited on 03/07/2021).

Sonneborn, T. M. (1947). "Recent advances in the genetics of Paramecium and Euplotes". In: *Adv Genet* 1, pp. 263–358.

Stark, R., M. Grzelak, and J. Hadfield (Nov. 2019). "RNA sequencing: the teenage years". In: *Nat Rev Genet* 20.11. [DOI:10.1038/s41576-019-0150-2] [PubMed:31341269], pp. 631–656.

Sultan, Marc et al. (2008). “A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome”. In: *Science* 321.5891, pp. 956–960. ISSN: 0036-8075. DOI: 10.1126/science.1160342. eprint: http://science.sciencemag.org/content/321/5891/956.full.pdf. URL: http://science.sciencemag.org/content/321/5891/956.

Van Houten, Judith (2019). “Paramecium Biology”. In: *Evo-Devo: Non-model Species in Cell and Developmental Biology*. Ed. by Waclaw Tworzydlo and Szczepan M. Bilinski. Cham: Springer International Publishing, pp. 291–318. ISBN: 978-3-030-23459-1. DOI: 10.1007/978-3-030-23459-1_13. URL: https://doi.org/10.1007/978-3-030-23459-1_13.

Vilar, J. M. (Oct. 2010). “Accurate prediction of gene expression by integration of DNA sequence statistics with detailed modeling of transcription regulation”. In: *Biophys J* 99.8, pp. 2408–2413.

Voigt, Philipp, Wee-Wei Tee, and Danny Reinberg (June 2013). “A double take on bivalent promoters”. eng. In: *Genes & Development* 27.12, pp. 1318–1338. ISSN: 1549-5477. DOI: 10.1101/gad.219626.113.

Wan, Changlin et al. (2017). “CPSS 2.0: A computational platform update for the analysis of small RNA sequencing data”. In: *Bioinformatics* 33.20, pp. 3289–3291. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx066.

Wang, Kai et al. (2014). “Prediction of piRNAs using transposon interaction and a support vector machine”. In: *BMC Bioinformatics* 15, p. 419. ISSN: 14712105. DOI: 10.1186/s12859-014-0419-6.

Watson, James D. et al. (Dec. 2003). *Molecular Biology of the Gene, Fifth Edition*. 5th ed. Benjamin Cummings. ISBN: 080534635X. URL: http://www.worldcat.org/isbn/080534635X.

Weiner, Assaf et al. (Sept. 2016). “Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution”. eng. In: *Nature Biotechnology* 34.9, pp. 953–961. ISSN: 1546-1696. DOI: 10.1038/nbt.3652.

Whangbo, Jennifer S. and Craig P. Hunter (June 2008). “Environmental RNA interference”. eng. In: *Trends in genetics: TIG* 24.6, pp. 297–305. ISSN: 0168-9525. DOI: 10.1016/j.tig.2008.03.007.

Wichterman, Ralph (1986). “The Antigens of Paramecium”. en. In: *The Biology of Paramecium*. Boston, MA: Springer US, pp. 357–373. ISBN: 978-1-4757-0374-0 978-1-4757-0372-6. DOI: 10.1007/978-1-4757-0372-6_10. URL: http://link.springer.com/10.1007/978-1-4757-0372-6_10 (visited on 04/03/2021).

Wikimedia, Commons (2017). *File:Paramecium diagram.png — Wikimedia Commons, the free media repository*. [Online; accessed 3-April-2021]. URL: https://upload.wikimedia.org/wikipedia/commons/f/f7/Paramecium_diagram.png.

— (2021). *File:Aminoacids table.svg — Wikimedia Commons, the free media repository*. [Online; accessed 23-March-2021]. URL: https://commons.wikimedia.org/w/index.php?title=File:Aminoacids_table.svg&oldid=526774118.

Wilusz, J. E., H. Sunwoo, and D. L. Spector (July 2009). “Long noncoding RNAs: functional surprises from the RNA world”. In: *Genes Dev* 23.13, pp. 1494–1504.

Wright, P. R. et al. (July 2014). “CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains”. In: *Nucleic Acids Res* 42.Web Server issue, W119–123.

Wu, Jinyu et al. (2013). “MirTools 2.0 for non-coding RNA discovery, profiling and functional annotation based on high-throughput sequencing”. In: *RNA Biology* 10.7, pp. 1087–1092. ISSN: 15558584. DOI: 10.4161/rna.25193.

Xiong, Jie et al. (Aug. 2016). “Dissecting relative contributions of cis- and trans-determinants to nucleosome distribution by comparing Tetrahymena macronuclear and micronuclear chromatin”. In: *Nucleic Acids Research* 44.21, pp. 10091–10105. ISSN: 0305-1048. DOI: 10.1093/nar/gkw684. eprint: https://academic.oup.com/nar/article-pdf/44/21/10091/9598539/gkw684.pdf. URL: https://doi.org/10.1093/nar/gkw684.

Zagulski, Marek et al. (Aug. 2004). “High Coding Density on the Largest Paramecium tetraurelia Somatic Chromosome”. English. In: *Current Biology* 14.15. Publisher: Elsevier, pp. 1397–1404. ISSN: 0960-9822. DOI: 10.1016/j.cub.2004.07.029. URL: https://www.cell.com/current-biology/abstract/S0960-9822(04)00533-0 (visited on 10/09/2020).

Zhang, Chunxiang (Dec. 2009). “Novel functions for small RNA molecules”. eng. In: *Current Opinion in Molecular Therapeutics* 11.6, pp. 641–651. ISSN: 2040-3445.

Zhang, F. and J. R. Lupski (2015). “Non-coding genetic variants in human disease”. In: *Hum Mol Genet* 24.R1, R102–110.

Zhang, Xiaoyu et al. (2009). “Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana”. eng. In: *Genome Biology* 10.6, R62. ISSN: 1474-760X. DOI: 10.1186/gb-2009-10-6-r62.

Zhang, Zhaolei et al. (Mar. 2006). “PseudoPipe: an automated pseudogene identification pipeline”. In: *Bioinformatics* 22.12, pp. 1437–1439. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl116. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/22/12/1437/543080/btl116.pdf. URL: https://doi.org/10.1093/bioinformatics/btl116.