Modelling the structure, function and evolution of Polycomb/Trithorax Response Elements

Bjørn André Bredesen

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2020

UNIVERSITY OF BERGEN

Modelling the structure, function and evolution of Polycomb/Trithorax Response Elements

Bjørn André Bredesen

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 20.11.2020

Year: 2020

Title: Modelling the structure, function and evolution of Polycomb/Trithorax Response Elements

Name: Bjørn André Bredesen

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

The work contained within this thesis was funded by a PhD fellowship granted by the University of Bergen, Norway, and conducted at the Computational Biology Unit (CBU) at the Institute of Informatics (II), Department of Mathematics and Natural Sciences, of the University of Bergen (UiB), Norway. My PhD work was supervised by Prof. Dr. Marc Rehmsmeier—professor at the II of UiB during the first year of my PhD, and, during the remainder, employed at the Integrated Research Institute (IRI) for the Life Sciences and Department of Biology, Humboldt-Universität zu Berlin—, Prof. Dr. Leonie Ringrose—professor at the Integrated Research Institute (IRI) for the Life Sciences and Department of Biology, Humboldt-Universität zu Berlin—and Prof. Dr. Inge Jonassen—professor at the II of UiB. Over the course of my PhD, I have been a member of the Molecular and Computational Biology (MCB) research school, and the NORBIS research school.

Acknowledgements

The work contained within this thesis has only been possible to complete with the support from a great number of people, whom I owe my deep gratitude for their direct and indirect contributions.

First of all, I thank my main supervisor, Marc Rehmsmeier, who introduced me to the exciting field of Polycomb/Trithorax epigenetics, for providing me with his excellent supervision and collaboration throughout both my master's and my PhD, with inspiring and informative conversations, and with the freedom to pursue new ideas. During my master's and the first year of my PhD, Marc was a group leader and professor at the Institute of Informatics at the University of Bergen, where he created an inspiring environment. A year into my PhD, Marc left the University of Bergen, but continued to grant me his supervision remotely, until I could finish my thesis—even well beyond the four years of my PhD fellowship. Marc, it has been a pleasure to work with you, and you have my deepest gratitude.

After Marc left his position at the UiB, Inge Jonassen has been my local supervisor. Thank you, Inge, for providing me with your supervision and meetings with helpful advice towards finishing my PhD. Additionally, Inge included me in the meetings of his

own research group, which has been inspiring. I also thank Leonie Ringrose for being my co-supervisor. I thank both Marc and Leonie for organizing and letting me take part in their highly inspirational summer school in Berlin—Mathematics meets epigenetics. I also thank Marc, Leonie and Inge for their invaluable feedback on my thesis.

I thank the Institute of Informatics at the University of Bergen for hiring me for the PhD fellowship, providing me with the chance to dedicate all of my time to working on exciting research for four years. I would also like to thank the Computational Biology Unit of the UiB for providing an inspiring and inclusive environment, with regular seminars and board game nights.

For the past year-and-a-half, I have worked at Knowit Reaktor Solutions. I thank Knowit for providing me with flexibility and a friendly working environment as I was finishing my thesis in my spare time.

I thank María Josefina Puig Ferreira for her love, support and encouragement over the years. I also thank the family of María Josefina for letting me stay at their ranch in Uruguay while I wrote the introduction of my thesis to the inspiring soundtrack of squawking parrots in the trees.

Thank you also to Takaya Saito of the former Rehmsmeier group for inspiring and friendly conversations, and for our time sharing a flat.

Finally, I extend my deep thanks to my mother, father, brothers, uncles, aunts, my late grandparents and remaining friends for their support and encouragement.

Summary

The correct development of animals and plants depends on carefully coordinated gene regulation. Polycomb/Trithorax Group (PcG/TrxG) proteins are conserved epigenetic regulators that are recruited to Polycomb/Trithorax Response Elements (PREs), a class of DNA *cis*-regulatory elements (CREs) originally discovered in the fruit fly. The structure and function of PREs has been progressively unravelled over the past three decades, with the identification of sequence motifs and the subsequent motif-based modelling and prediction of PREs, and with the genome-wide experimental mapping of PcG/TrxG binding. Whereas binding patterns vary for different cells, computational prediction holds the potential to predict PREs comprehensively. In this thesis, we exploit the recent explosion of data to conduct new investigations into the structure, function and evolution of PREs, presenting two papers with scientific investigations and two for tools.

Previous studies for computationally predicting fruit fly PREs have used a small training set and selections of known PRE motifs, leaving open the question of how training with genome-wide data might affect generalization. To address this, we trained PRE-predictors using genome-wide PcG binding sites, which we found improves gen-

eralization to independent PREs. We also trained models using different motif sets, where the addition of the GTGT motif further improved generalization. We were interested in how well a more advanced model would generalize, and we developed the Support Vector Machine Motif Occurrence Combinatorics Classification Algorithm (SVM-MOCCA), a hierarchical method that trains one Support Vector Machine (SVM) for each motif in a set and combines motif predictions. SVM-MOCCA significantly improved generalization to independent PREs. We predict large new sets of candidate PREs in the fruit fly genome that are enriched in experimental PcG/TrxG signals.

The low number of verified vertebrate PREs and a limited knowledge of relevant motifs has hampered the application of motif-based PRE predictors to vertebrate genomes. Methods such as k-spectrum SVMs can learn motifs from sequences, but the resulting models are high-dimensional and the specification of negative training sets is complicated. Previous computational studies for vertebrate PcG target prediction have focused exclusively on either predicting PcG target genes or on modelling genome-wide clusters of a small set of PcG markers. We developed a reinforcement learning regimen that exploits larger arsenals of genome-wide experimental data for the training of non-linear k-spectrum SVMs, yielding iteratively more precise models. We applied our methods to the fruit fly, mouse and human genomes. The final fruit fly model is competitive with models that incorporate prior motif knowledge. For all three species, we predict candidate PcG target sites genome-wide. We performed model analysis, which revealed a variety of motifs, subsets of which are conserved between models.

The success of SVM-MOCCA with predicting PREs prompted me to develop a polished and configurable implementation that can be useful for the broader community of CRE researchers—the Motif Occurrence Combinatorics Classification Algorithms

(MOCCA) suite. MOCCA provides polished implementations of SVM-MOCCA and baseline methods, and also the ability to combine feature set formulations with machine learning methods. Additionally, MOCCA presents RF-MOCCA, a derivative of SVM-MOCCA using the method of Random Forests (RFs). For ease of use, MOCCA implements functionality for generating negative training data and performing genome-wide prediction. We applied our methods for modelling fruit fly PREs and boundary elements. Our MOCCA-based methods improved generalization to both classes of CREs compared with previous methods. MOCCA is open source and extensible.

A Python package that streamlines the specification and application of CRE sequence models has been lacking. I developed Gnocis, a feature-rich package for Python 3 that provides tools for data preparation and analysis and a flexible vocabulary for feature set and model specification, and with implementations of functionality for model evaluation and genome-wide prediction. Gnocis integrates with Scikit-learn and TensorFlow for state-of-the-art machine learning. We demonstrated the use of Gnocis by modelling fruit fly PREs using a selection of methods, including a 5-spectrum mismatch kernel SVM and a Convolutional Neural Network. Gnocis is open source and extensible, and can be installed using the PyPI package manager.

List of scientific contributions

Article I: DNA sequence models of genome-wide *Drosophila melanogaster* Polycomb binding sites improve generalization to independent Polycomb Response Elements, Bredesen B. A., Rehmsmeier M., Nucleic Acids Research, 2019. BB conceived and designed the work, devised the methods and implemented them, ran the analyses, prepared the figures and wrote the manuscript; MR conceived and designed the work and wrote the manuscript.

Article II: Biomarker reinforcement learning with k-spectra enables precise Polycomb target site prediction without prior motif knowledge, Bredesen B. A., Rehmsmeier M., In submission. BB conceived and designed the work, devised the methods and implemented them, ran the analyses, prepared the figures and wrote the manuscript; MR made conceptual contributions to the analyses and wrote the manuscript.

Article III: MOCCA: A flexible suite for modelling DNA sequence motif occurrence combinatorics, Bredesen B. A., Rehmsmeier M., In revision. BB conceived and designed the work, devised the methods and implemented them, ran the analyses, prepared the figures and wrote the manuscript; MR revised the manuscript.

Article IV: Gnocis: An integrated system for interactive and reproducible analysis

and modelling of *cis*-regulatory elements in Python 3, Bredesen B. A., Rehmsmeier M., In preparation. BB conceived and designed the work, devised the methods and implemented them, ran the analyses, prepared the figures and wrote the manuscript; MR revised the manuscript.

Abbreviations

Abbreviation	Meaning
ANN	Artificial Neural Network
BE	Boundary Element
bp	Basepairs (unit)
CGI	CpG island
ChIP	Chromatin immunoprecipitation
ChIP-chip	Chromatin immunoprecipitation combined with microarrays
ChIP-seq	Chromatin immunoprecipitation combined with high-throughput sequencing
CNN	Convolutional Neural Network
CRE	*Cis*-regulatory element
DNA	Deoxyribonucleic acid
DT	Decision Tree
GFF	General Feature Format
GUI	Graphical User Interface
H3K27ac	Histone tail modification: Histone 3 lysine 27 acetylation
H3K27me3	Histone tail modification: Histone 3 lysine 27 trimethylation

Abbreviation	Meaning
H3K4me1	Histone tail modification: Histone 3 lysine 4 monomethylation
H3K4me2	Histone tail modification: Histone 3 lysine 4 dimethylation
H3K4me3	Histone tail modification: Histone 3 lysine 4 trimethylation
HBME	Higly BioMarker-Enriched locus
I.i.d.	Identically and independently distributed
IUPAC	International Union of Pure and Applied Chemistry
kb	Kilobasepairs (unit; 1000 bp)
LBME	Lowly BioMarker-Enriched locus
lncRNA	Long non-conding RNA
PcG	Polycomb Group
PhoRC	Pleiohomeotic Repressive Complex
PRC1	Polycomb Repressive Complex 1
PRC2	Polycomb Repressive Complex 2
PRE	Polycomb/Trithorax Response Element
PWM	Position Weight Matrix
RF	Random Forest
RNA	Ribonucleic acid
SVM	Support Vector Machine
TFBS	Transcription Factor Binding Site
TIP	Transcribed Intergenic Polycomb target site
TrxG	Trithorax Group
TSS	Transcription Start Site
XML	Extensible Markup Language

Contents

Part I

Introduction

Chapter 1

Biological background—genetics and epigenetics

The focus of this thesis is on the modelling of the sequences of Polycomb/Trithorax Response Elements—a class of *cis*-regulatory DNA sequence elements. I begin this thesis with a brief introduction to the biology of the Polycomb system.

1.1 Gene regulation in multicellular organisms

Contained within almost every cell of the body of every living organism is a copy of its genome (a notable exception being mature red blood cells in mammals, for which

the nuclei are removed [1]). The genomic sequence—a string of deoxyribonucleotides (DNA), divided among chromosomes—forms a blueprint for every constituent of each cell, and for the organism at large. Different cell types of a multicellular organism can display a wide variety of specific traits, yet, in terms of genomic sequence composition, they are largely identical. This diversity is enabled by means of epigenetic mechanisms—"*epi*", from Greek, meaning "above"—, imparting additional levels of information onto the otherwise static genomic sequence [2].

1.1.1 Evolution of phenotypic diversity

The central dogma of molecular biology states that the genome contains genes, which are transcribed from DNA to ribonucleic acid (RNA) sequences, where in turn RNA transcripts are translated to amino acids sequences, termed proteins [3]. The numbers of protein-coding genes—termed the G-number—can be similar in complex vertebrates, such as *Homo sapiens*, and comparatively simpler organisms, such as the nematode *Caenorhabditis elegans* [4]. In addition to protein-coding genes, genomes contain non-coding regions. For large genomes, non-coding regions generally make up more of the genomic sequence than do protein-coding genes [4]. These non-coding sequences were previously termed "junk DNA", due to their functions being unknown [5]. The high conservation of protein-coding sequences among species with clear phenotypic differences, such as humans and apes [6, 7], begs the question of what underlies the observed phenotypic diversity.

As whole-genome sequences became available, focus for the study of the genetic determinants of phenotypic diversity shifted from being mainly on protein-coding genes to non-coding sequences [8]. Within non-coding sequences, discrete elements with critical

regulatory functions have been discovered [9, 10]. The interactions among regulatory sequences and genes give rise to complex gene regulatory networks, whose function and evolution can explain the vast diversity in nature that cannot be attributed to protein sequence evolution.

1.1.2 *Cis*-regulatory elements (CREs)

Cis-regulatory elements (CREs) are non-coding stretches of DNA that regulate transcriptional levels of specific sets of genes by recruiting *trans*-acting factors [9, 10]. Multiple categories of CREs have been identified, distinguished by functions and mechanisms. Promoters recruit the transcriptional machinery (RNA polymerase) to target gene Transcription Start Sites (TSSes) [9, 10], where as the name suggests, transcription starts. Basal levels of transcription from genes and their promoters alone are often low [9]. Enhancers positively stimulate transcriptional levels of their target genes [9], and silencers silence genes [10]. The factors that bind to these elements can deteriorate over development, leaving the job of maintaining the established transcriptional states for another class of CREs [2]. Polycomb/Trithorax Response Elements (PREs)—the focus of this thesis—maintain epigenetic transcription state memories for their target genes across DNA replication and mitosis, for many cell generations [11, 12, 2]. Further enriching the repertoire of CREs, Boundary Elements (BEs) or insulators delimit the domains within which enhancers, repressors and PREs act [13, 14, 10].

CREs are enriched in Transcription Factor Binding Sites (TFBSs), which can be characterized by sequence motifs [10]. Sequence motifs are short (4-20 basepair), degenerate sequence patterns [15]. Different classes of CREs can be divided into subclasses based on the factors that bind them [16, 17].

1.1.3 Post-translational histone tail modifications

CREs can be distant from their target genes—in some cases over 10 kilobases [10]. CREs have been found to recruit factors that deposit a variety of post-translational histone tail modifications, some of which can be spread across large domains [18]. Specific marks have been associated with specific CRE-classes, such as histone 3 lysine 27 acetylation (H3K27ac) with active enhancers [19], histone 3 lysine 4 trimethylation (H3K4me3) at active promoters [20], and histone 3 lysine 27 trimethylation (H3K27me3) with repressing PREs [21]. Several histone tail modifications can spread within broad chromatin domains, including H3K27me3 [22].

1.1.4 Chromatin–chromatin interactions

Recent advances in experimental methods for the mapping of chromatin–chromatin interactions—including Chromatin Conformation Capture (3C), and the derivative methods 4C and Hi-C—, have resulted in a number of studies yielding a vast number of new insights into how CREs interact with their targets and surrounding chromatin [22, 13, 23]. Experimental mapping of chromatin–chromatin interactions has revealed that chromosomes are divided into Topologically Associating Domains (TADs) [13]—stretches of DNA within which interactions are more frequent than with surrounding regions. TADs have been found to correlate with histone modification marks [22, 13], and the boundaries of TADs are often enriched in insulator binding sites [13]. Intra-TAD gene expression is often correlated [13]. In the nucleus, PcG proteins have been found to cluster together, forming what has been termed “Polycomb bodies” [24]. Several lines of evidence support that TADs form regulatory units that delimit the actions

of CREs, where inter-TAD regulatory interactions are restricted to TADs of similar regulatory states [13]. In addition to TADs, for *Drosophila*, a Hi-C study identified loops between PRC1-bound regions and gene promoters [23].

1.2 The Polycomb/Trithorax system

The first PcG gene was discovered over 70 years ago, in the fruit fly *Drosophila melanogaster*, as a repressor of *Hox* genes [25]. *Hox* genes determine the body plans of bilaterians [26, 27]. Reduction of Polycomb group proteins in fruit flies yields developmental phenotypes such as the growth of additional sex combs [28]. Polycomb group (PcG) and Trithorax group (TrxG) proteins are recruited to DNA, where PcG proteins maintain target gene repression, and TrxG proteins antagonize PcG-mediated repression [29, 30]. Due to the historical significance of fruit flies for research into the Polycomb/Trithorax system, I start by introducing this system in fruit flies.

1.2.1 In flies

Today, a variety of PcG and TrxG proteins have been identified in *D. melanogaster* (Table 1.1). PcG and TrxG proteins form distinct complexes, including Polycomb Repressive Complexes 1 and 2 (PRC1 and PRC2), Polycomb repressive deubiquitinase (PR-DUB), dRING-associated factors (dRAF), and Pleiohomeotic Repressive Complex (PhoRC) [28]. For simplicity, I will refer to both genes and their protein products using established gene names (italicized), where all noted functions are executed by the corresponding protein products.

The core of PRC2 in flies consists of the proteins encoded by the genes *E(z)* (en-

hancer of zeste), *Esc* (extra sex combs), *Su(z)12* (suppressor of zeste 12) and *Nurf-55* [28]. *E(z)* encodes a histone methyltransferase that trimethylates histone 3 lysine 27 (H3K27me3). As a result, H3K27me3 is a defining histone mark of Polycomb repression [31]. PRC1 contains products of *Pc* (Polycomb), *dRING/Sce* (Sex combs extra), *Ph* (Polyhomeotic), *Psc* (Posterior sex combs) and *Scm* (Sex combs on midleg) [31]. *Pc* contains a chromodomain, which binds to H3K27me3 [28]. *dRING* monoubiquitinates H2A lysine 118 [28] (119 in vertebrates [25]). No identified constituents of PRC1/2 bind DNA with sequence specificity. PhoRC, however, contains *Pho* (Pleiohomeotic), which has been found to bind specific motifs, in addition to *dSFMBT* (*Drosophila* Scm-related gene containing four malignant brain tumour (MBT) domains) [31]. No known DNA sequence motifs are—by themselves—sufficient to recruit PcG [25], and PcG recruitment by *Pho* has been found to depend on combinatorial interactions [32].

The TrxG proteins also form several complexes, including the COMPASS family complexes and the SWI/SNF complexes [25]. At the core of COMPASS complexes are the four proteins *Wds*, *Ash2*, *Rbbp5* and *Dpy30* (abbreviated as *WARD* [25]). H3K4me1 is a mark of Trithorax activation, mediated by *Trx* and *Trr* [33]. H3K4me2 has also been proposed to play a role [34].

1.2.2 From plants to vertebrates

Since the discovery of PcG genes in the fruit fly, orthologs of PcG and TrxG genes have been discovered across the animalia and plantae kingdoms [25]. Orthologs of *Drosophila* PcG/TrxG genes are given in Table 1.1. With animal complexity rises the level of complexity of the Polycomb system. For instance, five orthologs of the *D. melanogaster* PcG gene *Pc* have been identified in the human genome: *CBX2/4/6–8*

Complex	*Drosophila*	Vertebrate	Function (*Drosophila*)
PcG—PRC1			
	Pc [28, 35, 31]	*CBX2,4,6–8* [35, 36]	Chromodomain binds H3K27me3 [25, 28, 36]
	Ph [28, 35, 31]	HPH1-3 [35, 31]	Oligomerization [25]
	Psc [28, 35, 31]	BMI1 [31]	H2AK119ub [25]; Oligomerization [25]
	Sce (*dRING*) [28, 35, 31]	*RING1A/B* [35, 31]	H2AK118ub [28] / H2AK119ub [25]
	Sxc * [28]		
PcG—PRC2			
	E(z) [28, 35, 31]	*EZH1/2* [35, 31]	Deposits H3K27me3 [25, 28]
	Su(z)12 [28, 35, 31]	*SUZ12* [35, 31]	RNA/DNA-binding [25]; H3K27me3 [28]
	Esc/Escl [28, 35, 31]	*EED* [35, 31]	H3K27me-binding [25]
	Caf1-55 [28, 35, 31]	*RpAp46/48* [35, 31]	H3K36me3-binding [25]
	Pcl * [28, 35]	*PCL1-3* * [35]	H3K36me3-binding [25]
	(*Jing*?) * [28, 35]	*AEBP2* * [28]	H2Aub-binding [25]; DNA-binding [25]
	(*Jarid2*?) * [28, 35]	*Jarid2* * [28]	H2Aub-binding [25]; RNA-binding [25]
PcG—PhoRC			
	Sfmbt [28, 35]		
	Pho/*Phol* [28, 35, 36]	*YY1* [28, 36]	DNA-binding [28]
PcG—PR-DUB			
	Calypso [28, 25]	*BAP1* [25]	Deubiquitinates H2Aub1 at K118/K119 [28]
	Asx [28]	*ASXL1/2* [25]	Chromatin-binding [25]
PcG—dRAF			
	Psc/*Su(z)2* [28]		
	Sce (*dRING*) [28]		Required for H2AK118ub-deposition [28]
	Kdm2 [28]	*KDM2B* [25]	Demethyl. H3K36me2 [28, 25]; DNA-binding [25]
PcG—Unassigned			
	Scm [28]	*SCMH1-2* [31]	Potential link between PRC1/2/PhoRC [28]
	Mxc [28]		
	Crm [28]		
TrxG—COMPASS-core (*WARD*)	*Wds* [25]	*WDR5* [25]	Histone-binding [25]
	Ash2 [28, 25]	*ASH2L* [37, 25]	DNA-binding [25]
	Rbbp5 [25]	*RBBP5* [25]	Histone-binding [25]
	Dpy30 [25]	*DPY30* [25]	
TrxG—MLL1/2 COMPASS-like	*WARD* [25]	*WARD* [25]	
	Trx [28]	*MLL1/2* [37, 36]	Deposits H3K4me [28, 25]
TrxG—MLL3/4 COMPASS-like	*WARD* [25]	*WARD* [25]	
	Trr [25]	*MLL3/4* [25]	Deposits H3K4me [28, 25]
TrxG—SET1/COMPASS			
	WARD [25]	*WARD* [25]	
	dSet1 [25]	*SET1A/B* [25]	Deposits H3K4me [28, 25]
TrxG—SWI/SNF (BAF/PBAF)	*Brm* [25]	*BRM/BRG1* [25]	Chromatin-remodelling [25]

Table 1.1: Polycomb/Trithorax group proteins in *D. melanogaster*, and vertebrate homologs. Not a comprehensive list. * Associated with the complex, and modulates its function.

(chromobox protein homologs 2/4/6–8) [25]. The histone modification H3K27me3 is catalysed by the PRC2 member *E(z)/EZH2* [38, 39, 40], which is largely conserved across metazoans [25]. Like the PcG proteins and complexes, the TrxG is also conserved, with the COMPASS-core (*WARD*) conserved across metazoans [25]. H3K4me1 is also a mark of Trithorax activation in vertebrates, mediated by MLL1–4 (vertebrate orthologs of *Trx* and *Trr*) [33]. As in *Drosophila*, H3K4me2 has also been proposed to play a role in vertebrates [34].

1.3 Polycomb/Trithorax recruitment

In *D. melanogaster*, PcG and TrxG proteins are recruited to a class of CREs termed Polycomb/Trithorax Response Elements (PREs) [11, 12], through which they maintain epigenetic memories of transcriptional states of their target genes over DNA replication and mitosis [41]. Over the last decade, PREs have also been identified in vertebrates. However, other genomic features have also been associated with PcG/TrxG recruitment and regulation.

1.3.1 Polycomb/Trithorax Response Elements

The first PREs were discovered in the fruit fly [11], and recently, a number of PREs have also been discovered in vertebrate genomes, including frog [42], mouse [43, 44] and human [45, 46, 44].

Fruit fly PREs are a few hundred to thousands of basepairs long [52], and are enriched in a selection of characteristic sequence motifs [2, 44, 47]. Known fruit fly PRE-motifs are listed in Table 1.2. The majority of the motifs identified are binding

Factor	IUPAC-motif
GAF/Psq	GAGAG [47, 2, 48]
GAF	GAGAGAGAGA [47]
Pho	GCCAT [47, 2, 48]
Pho	CNGCCATNDNND [47]
Pho	GCCATHWY [47]
Zeste	YGAGYG [47, 48]
Zeste	BGAGTGV [48]
Dsp1	GAAAA [2, 48]
Grh	TGTTTTT [2, 48]
Grh	WCHGGTT [48]
Sp1/KLF	RRGGYGY [2, 48]
Combgap	GTGT [47, 49]
? (EN 1)	GSNMACGCCCC [47]
? (Site A)	GAACNG [48]

Table 1.2: Identified motifs of *D. melanogaster* Polycomb/Trithorax Response Elements.

sequences for DNA-binding factors. The motifs show no obvious patterns in their occurrence, and PREs have little or no homology [47]. Efforts for modelling fruit fly PRE sequences have revealed that the pairing of motif occurrences better distinguishes PREs from background than singular motifs, suggesting that motif pairing is a defining sequence feature of PREs [47, 54]. PREs at the fruit fly *invected/engrailed* locus, together with experimentally mapped PcG proteins and trimethylated H3K27, is shown in Figure 1.1.

Vertebrate PREs have also been analysed in terms of motif composition [43, 44], and a selection of DNA-binding factors have been identified as candidate recruiters, including *YY1* [44]—an ortholog of the fly PcG gene *pho*—, *REST* [44], *RUNX1* [44] and *E2F6* [44]. Although *YY1* is a homolog of the fly PcG gene *Pho*, the involvement of *YY1* in PcG recruitment is disputed [44].

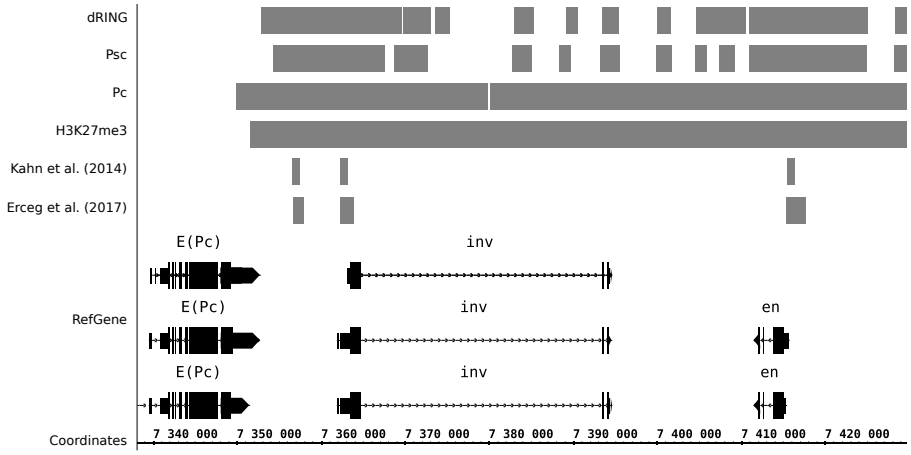


Figure 1.1: PREs at the *engrailed/invected* locus. The figure was adapted from a screenshot from the Integrated Genome Browser [50]. The peaks for *dRING*, *Psc*, *Pc* and *H3K27me3* are taken from modENCODE [51] (IDs: 5071_1819; 3955_1820; 3957_1816; 3960_1817). The Kahn *et al.* [32] regions (taken from their Supplementary Table S1) are computationally defined PREs, based on ChIP-chip data. The Erceg *et al.* [52] regions (taken from their Supplementary Table S1) are functionally verified PREs. The R5 *D. melanogaster* genome assembly [53] was used.

1.3.2 CpG islands

CpG dinucleotides in mammalian genomes are prone to the methylation of the cytosine, and methylated cytosines spontaneously mutate to thymines, resulting in low genome-wide occurrence frequencies of CpG dinucleotides in mammalian genomes [55]. However, regions of clusters of unmethylated CpG dinucleotides and elevated GC-content have been identified, termed CpG islands (CGIs) [55]. The precise mechanisms that protect CGIs from methylation are not fully understood, but multiple models have been proposed, including one of steric hindrance of factors that methylate CpG dinucleotides, such as by the binding of transcription factors [55]. Additionally, subsets of CGIs can

acquire or lose methylation status over development, where methylation yields target gene repression [55]. Around half of identified mouse and human CGIs have been associated with TSSes, and the remainder have been proposed to correspond to unannotated promoters [44].

A substantial fraction of CGIs can recruit PcG/TrxG proteins [44]. Accordingly, it has been debated whether or not a subset of CGIs may be vertebrate PREs [44]. Computational identification of CGIs in terms of CpG dinucleotide frequencies and GC-content identifies hypomethylated CGIs with highly variable precision and recall, and there is experimental evidence that motifs of DNA-binding factors are important for maintaining hypomethylation [55]. A recent study revealed a role for the PcG in the maintenance of CpG hypomethylation [56], which could be a sign that causality goes in the reverse direction: rather than that CpG islands are vertebrate PREs, perhaps vertebrate PREs often give rise to CpG islands over evolution, such as through the protection of CpG dinucleotides from spontaneous mutation. Further worth noting, many intergenic PcG/TrxG target sites do not overlap with CpG islands, and not all CpG islands have been found to recruit PcG/TrxG [44].

1.3.3 Non-coding RNAs

Many fruit fly PREs have been found to be transcribed into long non-coding RNAs (lncRNAs)—RNAs with lengths ranging from several hundred basepairs to several kilobases—[57]. A number of murine transcribed promoter-proximal (TSS -/+ 5kb) intergenic loci have been found to coincide with H3K27me3- and SUZ12-enrichment in the underlying DNA sequences, and were termed Transcribed Intergenic Polycomb target sites (TIPs) by the authors [58]. The association of PREs and PcG/TrxG target

sites with non-coding transcription has raised the question of whether PRE-transcription plays a functional role for the recruitment of the Polycomb/Trithorax machinery, or for the switching of transcriptional states [57, 44, 59].

Hekimoglu *et al.* [58] experimentally tested the function of three of the TIPs they had identified. When inserting each TIP upstream of a reporter gene, the authors found above endogenous levels of transcription of the TIPs, and that two out of the three TIPs tested yielded substantial repression of the reporter gene.

Glazko *et al.* [60] published a machine learning study that demonstrated that lncRNAs associated with PRC2 can be distinguished from other lncRNAs in terms of sequence features. For the features, the authors employed RNA sequence structure patterns (RSSPs), k-mers and PWM motifs, filtered for significant difference in enrichment in PRC2-associated lncRNAs and non-associated lncRNAs. Importantly, enrichment of RSSPs were not significantly different in PRC2-associated lncRNAs than in other lncRNAs, and the discriminative features could correspond to DNA sequence motifs (as a result of the lncRNA originating from a transcribed PRE), rather than features involved in PRC2 recruitment to lncRNAs.

1.3.4 The identification of Polycomb/Trithorax Response Elements

A number of methods have been employed in the endeavour of identifying PREs. The first PREs were discovered as genomic elements that maintain *Hox* gene expression states, by means of laborious testing of chromosomal segments [11, 61, 12]. The most commonly used assay for testing PRE-function in *Drosophila* is the *miniwhite* assay—a transgenic assay where a candidate regulatory element is linked to a derivative of the *white* gene (with a large part of the first intron removed, and the regulatory region

shortened) and inserted into the fly genome, with a white mutant background [44, 62]. Importantly, the *white* gene gives rise to red eye pigment, is non-essential, and viable homozygous flies lacking expression of the gene can be generated [62]. For mutant flies with the reporter construct inserted, gene expression is measured in terms of eye colour, with changes arising solely from the transgenic reporter gene, which ranges between white and red when the gene is fully silenced or expressed, respectively, with yellow intermediates.

In the 2000s, the advent of methods for the genome-wide mapping of DNA-binding factors, such as chromatin immunoprecipitation combined with microarrays (ChIP-chip) [63] or with high-throughput sequencing (ChIP-seq) [64], and DNA adenine methyltransferase identification (DamID) [65], yielded a number of studies and public data repositories that mapped the genome-wide binding of PcG/TrxG proteins and modified histones in fruit flies [66, 65, 67, 68, 69, 70, 32, 51] and vertebrates [71, 72, 73, 42, 74]. Genome-wide sets of clusters of binding sites for multiple PcG/TrxG proteins and H3K27me3/H3K4me1 are likely to contain many PREs. Nonetheless, these methods are not perfect for the identification of PREs. Experiments that map DNA-binding only map binding in the cells in which the experiments are performed. Given the dynamic nature of gene regulation, these experiments are unlikely to comprehensively map all PcG/TrxG-binding sites of an organism. Thus, creating a map of PREs based solely on clusters of PcG/TrxG peaks from ChIP-chip or ChIP-seq experiments with homogeneous cells is likely to result in Type II errors (false negatives). On the other hand, PREs have been found to make long-range physical contacts with their targets [24, 22, 23]. A chromatin conformation study further found that ChIP-signals for PcG proteins can be observed both at a PRE and its target promoter [24]. Thus, Type

I errors (false positives) may arise if one loop end corresponds to a PRE, and the other is merely a shadow of the interaction. Experimental parameters, such as the choice of antibodies, can further influence data accuracy.

The observation that *Drosophila* PREs are enriched in a variety of DNA sequence motifs has made way for an additional approach to the identification of PREs: genome-wide bioinformatic prediction. The first study to predict PREs was that of the PREdictor [47], in which PRE sequences were modelled based on the occurrence frequencies of 7 PRE-motifs, weighted based on a set of 12 PREs and 16 non-PREs. Importantly, the authors [47] found that motif pairs are predictive of *Drosophila* PREs—whereas singular motifs are not—, and a substantial selection of candidate PRE predictions have been experimentally verified. Other studies have built upon this work [75, 76, 54]. Notably, these studies all relied on similar, small training sets of PREs and non-PREs.

As experimentally determined binding sites of PcG/TrxG proteins may not all constitute PREs, I make a distinction in this thesis between a PRE and a PcG target site.

Definition 1 (Polycomb/Trithorax target site (PcG target site)) *A genomic region that recruits or interacts with Polycomb/Trithorax group proteins.*

Definition 2 (Polycomb/Trithorax Response Element (PRE)) *A PcG target site that maintains transcription state memories of target genes over DNA replication and mitosis.*

Accordingly, PcG target sites are candidate PREs with experimental support, but additional experiments are required in order to prove that a PcG target site is a PRE.

1.3.5 Implications for medical research

Although understanding the epigenetics of development is interesting in its own right, the value of studying the Polycomb/Trithorax system extends beyond that of providing insights into developmental biology, and holds the potential to improve the quality of human lives, by virtue of its medical implications. The Polycomb/Trithorax system is highly conserved in bilaterians, all the way up to humans, and is central to the maintenance of cell identity and stem cell differentiation [25]. A number of diseases are associated with genetic dysregulation, including cancers [77] and arthritis [78]. As PcG/TrxG proteins are important epigenetic regulators, they may be involved in multiple genetic diseases.

Both PcG and TrxG proteins have been associated with human cancers [77]. Targets of the Polycomb system in the human genome include the tumour suppressor genes *CDKN2A* and *CDKN2B*, whose silencing can yield uncontrolled proliferation, which in turn can be induced by the overexpression of PcG proteins [77]. *EZH2* has also been found to prohibit tumour necrosis factor-mediated programmed cell death [77]. Accordingly, an increased understanding of the Polycomb/Trithorax system and its involvement in cancer may enable the development of new therapies.

Chapter 2

Machine learning with biological sequences

A number of tasks are difficult to formulate as specific algorithms. For example, a photograph of a face may contain all the necessary data for facial recognition, but hard-cording the recognition of distinguishing features is a daunting task. The field of Machine Learning is concerned with the development and application of algorithms that can learn from data. Examples of problem domains where Machine Learning has been successfully applied include facial recognition [79], the playing of video games by a computer [80], and the generation of photo realistic imagery from text [81].

I begin this chapter with a brief introduction to Machine Learning methods, and then proceed to narrow the focus to Machine Learning with biological sequences.

2.1 A brief introduction to Machine Learning

Within this thesis, I define any method that can approximate a target function/behaviour based on input data as a machine learning method.

Definition 3 (Machine Learning method) *Any computational method that approximates a desired target function/behaviour based on input data.*

In this thesis, we restrict our focus to Machine Learning methods that construct models based on observations in an n-dimensional vector space, termed a *feature space*.

Definition 4 (Feature space) *A feature space, $\mathbb{F} : \mathbb{O} \rightarrow \mathbb{R}^n$ is a mapping from observations $o \in \mathbb{O}$ to n-dimensional real-valued vectors, describing properties of the observations.*

Generally, Machine Learning methods can be divided into three categories: supervised, semi-supervised and unsupervised methods.

2.1.1 Supervised learning

Supervised machine learning methods are *function approximation* methods, that construct models based on a number of examples of inputs and corresponding, desired outputs.

Definition 5 (Function approximation) *A function approximation is a function $\hat{f}(x) \approx f(x)$, which approximates the output of a target function for the same input. A hat over the function name is used to denote function approximations.*

Outputs can be one or more continuous values, discrete classes, or a combination thereof. In this thesis, I restrict the focus to models with singular outputs. Models that

predict a continuous output are called *regression models*, and models that predict class labels are called *classifiers*.

Definition 6 (Regression model) *Given a space of observations, $\mathbb{O}$, a regression model, $\hat{r} : \mathbb{O} \to \mathbb{R}$, is a function that maps every observation $o \in \mathbb{O}$ to a real value $v \in \mathbb{R}$.*

Definition 7 (Classifier) *Given a space of objects to classify, $\mathbb{O}$, and a space of labels, $\mathbb{L}$, a classifier, $\hat{c} : \mathbb{O} \to \mathbb{L}$, is a function that maps every observation $o \in \mathbb{O}$ to a label $l \in \mathbb{L}$.*

If there are two output classes, a classifier is called binary. For binary classifiers, the labels are typically designated the titles *positive* and *negative*, with *positive* denoting the class of primary interest.

Definition 8 (Binary class set) *The set of binary labels is denoted as $\mathbb{B} = \{\oplus, \ominus\}$, representing positive and negative classes, respectively.*

Classifiers can be constructed from regression models by thresholding, treating one side of the number line as negative, and the other as positive.

Definition 9 (Binary classifier by thresholding) *A binary classifier can be obtained from a regression model, $\hat{r} : \mathbb{O} \to \mathbb{R}$, by applying a threshold τ.*

$$\hat{c}(o) = \begin{cases} \oplus & \textit{if } \hat{r}(o) \geq \tau \\ \ominus & \textit{otherwise} \end{cases}$$

Examples of supervised machine learning methods include Support Vector Machines (SVMs) [82], Random Forests (RFs) [83] and Convolutional Neural Networks (CNNs) [84, 85].

2.1.2 Unsupervised learning

For many tasks, assigning labels or output values to training data may be laborious or impossible. For instance, the possible states in a video game can be prohibitively difficult or impossible to enumerate in advance. For classification problems, examples of a target class may be easy to define, but negatives difficult. For example, if training an algorithm to identify dogs in photographs, how does one enumerate everything that is not a dog, and everything that looks *almost* like a dog, but not quite? Unsupervised methods can discover structure within supplied data, without the need for pre-defined labels. Example methods include clustering algorithms, such as k-means clustering [86], and variational autoencoders [87].

2.1.3 Semi-supervised learning

Semi-supervised machine learning methods form intermediates between supervised and unsupervised methods, by making use of unlabelled training data and progressively constructing sets for supervised learning. The method of Mapping-Convergence is a semi-supervised method for training Support Vector Machines (introduced later in this chapter). Reinforcement learning methods are semi-supervised methods that model a machine learning problem in terms of an *agent* that learns by trial and error within an *environment*, iteratively improving its performance on the task at hand by a given measure [88].

2.1.4 Linear models

A wide array of Machine Learning methods have been developed. Linear models map values from an input feature space to predicted values by means of linear weights and an offset term.

Definition 10 (Linear model) *A linear model maps feature vectors $\vec{x} \in \mathbb{R}^n$ to predicted values $\hat{y} \in \mathbb{R}$ by means of linear weights $\vec{\beta}$ and an offset term α.*

$$\hat{f}(\vec{x}) = \vec{\beta}^T \vec{x} + \alpha$$

A variety of linear Machine Learning methods have been developed that can assign values to the weight and offset terms of linear models, either by direct calculation, or by optimization methods, including linear regression, log-odds modelling [47] and linear Support Vector Machines [82].

The simplest linear model that can be formulated is an unweighted sum. As this corresponds to setting all weights equal to 1, there is no learning involved. Within the context of this thesis, we refer to this model as a *dummy model*. In spite of the simplicity of a dummy model, if the features used have been determined based on insights relevant to the problem at hand, this can yield a successfully predictive model (later discussed in **Article I**).

Definition 11 (Dummy model) *A dummy model maps feature vectors $\vec{x} \in \mathbb{R}^n$ to sums of their components.*

$$\hat{f}(\vec{x}) = \sum_i x_i$$

When the features are frequencies, an alternative simple model formulation, which

does not forgo training, is the *log-odds model*. The log-odds model independently assigns weights to all features based on the logarithm of the ratio of average values in the positive and negative training examples.

Definition 12 (Log-odds model) *A log-odds model maps feature vectors $\vec{x} \in \mathbb{R}^n$ to a log-odds weighted sum.*

$$\hat{f}(\vec{x}) = \sum_i \beta_i x_i,$$

where weights are assigned as

$$\beta_i = \log \frac{\sum_{p \in P} p_i / |P|}{\sum_{n \in N} n_i / |N|},$$

for sets P and N of positive and negative training example feature vectors.

A key advantage of linear models is that model analysis is straightforward: the features with the largest absolute weights have the largest influence on the predicted value.

Real-world problems are not always linearly solvable, and a variety of non-linear methods of varying complexity have been developed. It is worth noting that with increased model complexity come additional ways of solving the optimization problem at hand, many of which may generalize poorly beyond the training data. The problem of models that adapt well to the training data but generalize poorly to independent data is known in the field of Machine Learning as the problem of *overfitting* [89].

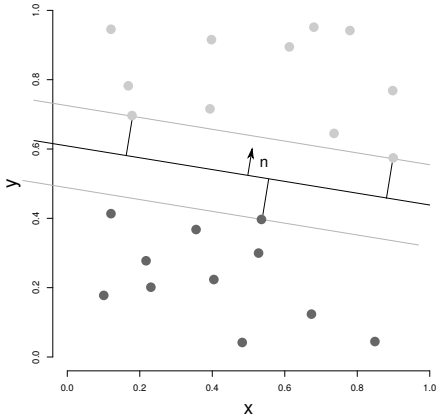


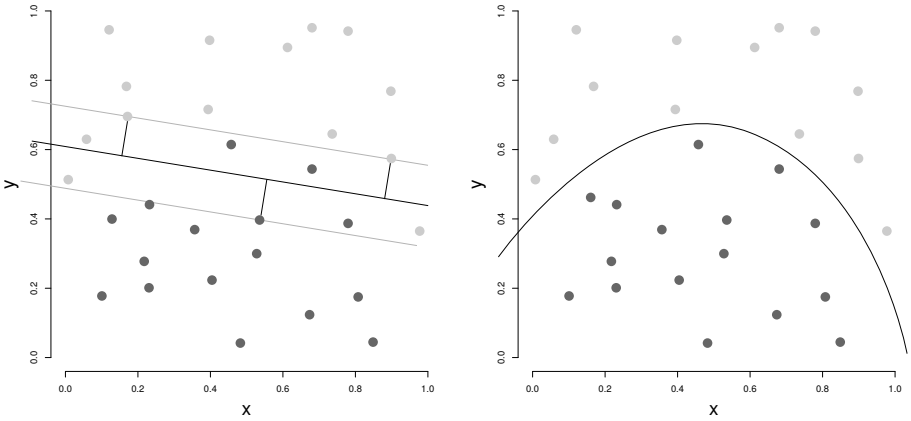
Figure 2.1: A linear Support Vector Machine constructs a decision surface with maximal margins to training examples of opposite classes.

2.1.5 Support Vector Machines

The method of Support Vector Machines (SVMs) solves a binary classification problem by placing a decision surface between training examples of two classes such that the margin to opposing classes is maximized [82]. The vectors closest to the decision surface define its normal vector, and are called the *support vectors*. The decision surface of a linear SVM is illustrated in Figure 2.1.

When classes are not linearly separable, the method of SVMs supports two strategies to rectify this problem: 1) subsets of training examples can be identified as noise during the search for the optimal decision surface (Figure 2.2a), and 2) SVMs can map the feature space into a higher-dimensional space, by means of *kernel functions*, in which the training examples may be separated (Figure 2.2b).

Definition 13 (Support Vector Machine kernel function) *A Support Vector Machine*



(a) Non-linearly separable set and soft margin (b) Non-linearly separable set and non-linear kernel

Figure 2.2: Examples of Support Vector Machines with a training set in two dimensions that is not linearly separable. (a) The soft margin property of a linear SVM enables the training algorithm to treat a subset of training examples as noise, and thus to train a model on data that is not linearly separable. (b) Non-linear kernels map the feature space to a higher-dimensional space, in which the data may be linearly separable.

kernel function is defined as

$$k(\vec{x}, \vec{y}) = \Phi(\vec{x})^T \Phi(\vec{y}),$$

where $\Phi : \mathbb{R}^a \rightarrow \mathbb{R}^b, b > a$ *is a mapping to a higher-dimensional space.*

Definition 14 (Support Vector Machine decision function)

$$\hat{c}(\vec{x}) = \sum_{y, \vec{s} \in SV} yk(\vec{s}, \vec{x}) - \rho,$$

where $\vec{s}$ *and* y *are support vectors and their weights, respectively, and* ρ *is the offset of*

the decision surface.

A Support Vector Machine is trained by solving the Lagrangian dual of a quadratic programming problem that optimizes decision surface class margin size (the details are outside the scope of this thesis) [82]. The solution to the optimization problem is unique. Each weight y is the product of a Lagrange multiplier α_i and a class label value $c_i \in \{+1, -1\}$. Only for the support vectors are the Lagrange multipliers non-zero.

Common kernel functions include the linear kernel, the N-th degree polynomial kernel, and the Radial Basis Function (RBF) kernel [90].

Definition 15 (Linear kernel) *The linear kernel is a linear combination of a feature vector and a support vector.*

$$k(\vec{x}, \vec{y}) = \vec{x}^T \vec{y}.$$

Definition 16 (Polynomial kernel) *The N-th degree polynomial kernel is a linear combination of a feature vector and a support vector, raised to the N-th power.*

$$k(\vec{x}, \vec{y}) = (\gamma \vec{x}^T \vec{y} + c_0)^n,$$

where $\gamma > 0$ and c_0 are kernel parameters.

Definition 17 (Radial Basis Function kernel) *The Radial Basis Function kernel is defined by feature vector and support vector distances, by*

$$k(\vec{x}, \vec{y}) = e^{-\gamma ||\vec{x} - \vec{y}||^2},$$

where $||\vec{x} - \vec{y}||$ is the distance between vectors $\vec{x}$ and $\vec{y}$, and γ is a kernel parameter.

The formulation of the model in terms of linear combinations of support vectors and a maximal margin yields an optimization problem with a unique solution [82]. The optimization problem of an SVM is outside of the scope of this thesis, and will not be discussed in further detail.

When more than two classes of data are available, multiclass classification with Support Vector Machines can be performed by training one binary SVM per class-class boundary, and predicting the class label by means of voting among the binary classifiers, as is implemented in libsvm [90].

2.1.6 Random Forests

A decision tree (DT) is a machine learning model with a branching tree structure [91]. Each node of the tree has either a decision criterion and two or more branches leading to subsequent nodes, or is a leaf node yielding an output. The prediction of a DT is computed by traversing the tree from the root to a leaf based on the decisions made at each node.

The method of Random Forests (RFs) [83] is an ensemble machine learning method, whereby a population of DTs is constructed with random variation, and their predictions are combined upon model application by averaging. Both classification and regression formulations of RFs have been made. Importantly, the randomization reduces overfitting [83]. Random variations employed by RFs include the random selection of training examples per tree, random subsampling of features, and random selection of tree splits [83]. The details of DT and RF training are outside the scope of this thesis.

2.1.7 Convolutional Neural Networks

An Artificial Neural Network (ANN) is a machine learning model that propagates input signals through a graph of computational nodes, commonly organized in layers [91]. Each node sums and weights the output values of its inputs, and applies a non-linear transformation, such as the sigmoid function [91]. The weights of an ANN are typically trained using gradient-based methods, such as the backpropagation algorithm [91].

A Convolutional Neural Network (CNN) [84, 85] is an ANN with weight-sharing such that the weights between layers form convolutions. This reduces the number of weights to train, and enforces the use of localized structure in the input. CNNs are able to learn predictive features from spatially organized input, and have been successfully applied to complex machine learning problems such as image recognition [84, 85]. The details of the training of ANNs/CNNs are beyond the scope of this thesis.

2.2 Features of biological sequences

In this thesis, the focus is on the construction and application of models of DNA sequences. Biological sequences can be presented to machine learning models in a number of ways. Regulatory DNA sequences are commonly enriched in motifs—short, recurring DNA sequence patterns—[15], corresponding to the binding sites for DNA-binding factors [9]. A common approach is to define sets of motifs, and to model the sequences based on numerical measures derived from motif occurrences, such as their occurrence frequencies [47, 54, 92], or paired occurrence frequencies [47, 75].

IUPAC nucleotide code	Nucleotides
A	A
T/U	T/U (U for RNA)
G	G
C	C
R	A, G
Y	C, T/U
S	G, C
W	A, T/U
K	G, T/U
M	A, C
B	C, G, T/U
D	A, G, T/U
H	A, C, T/U
V	A, C, G
N	A, C, G, T/U

Table 2.1: IUPAC nucleotide codes [93].

2.2.1 Motif formulations

Motifs can be defined with different schemes, such as precise DNA sequence strings, DNA sequence strings with ambiguous positions, regular expressions, and Position Weight Matrices (PWMs). The IUPAC [93] nucleotide codes are a popular scheme for noting nucleic acid sequences with ambiguous positions, and are listed in Table 2.1. For a PWM, every position of the motif has a weight per nucleotide. Upon application, the PWM is slid across a DNA sequence, and weights for subsequent matched positions are summed together, yielding a single score per position. Setting a threshold yields a number of discrete matched occurrences across the sequence. Yet more advanced motif formulations exist, such as based on Hidden Markov Models (HMMs) [94], and taking structural DNA features into account [95]. This thesis focuses on motifs defined in IUPAC nucleotide codes and PWMs.

2.2.2 Motif-based features

A variety of feature sets can be formulated based on motifs, the perhaps simplest of which is the motif occurrence frequency spectrum.

Definition 18 (Motif occurrence frequency) *For a motif, $m \in \mathbb{M}$, the occurrence frequency of m in a sequence $s \in \mathbb{S}$ is defined as the number of occurrences of m in s divided by the length of s.*

Definition 19 (Motif occurrence frequency spectrum) *Given a set of motifs, $M \subset \mathbb{M}$, the motif occurrence frequency spectrum, $\mathcal{S}_M : \mathbb{S} \to \mathbb{R}^{|M|}$, maps any sequence $s \in \mathbb{S}$ to a vector $\vec{y} \in \mathbb{R}^{|M|}$, with one occurrence frequency per motif as components. Motif occurrences in both the forward and the reverse complement direction are counted.*

Definition 20 (Motif pair occurrence frequency spectrum) *Given a set of motifs, $M \subset \mathbb{M}$, the motif pair occurrence frequency spectrum, $\mathcal{S}^2_{M(d)} : \mathbb{S} \to \mathbb{R}^n$, maps any sequence $s \in \mathbb{S}$ to a vector $\vec{y} \in \mathbb{R}^{|M|}$, with one occurrence frequency per unique motif pair, for a pairing cutoff distance d. Motif occurrences in both the forward and the reverse complement direction are counted.*

2.2.3 Motif discovery

Various motif discovery algorithms have been developed that enable the automated identification of sequence motifs [96, 97, 98, 99], which can be employed in order to identify enriched sequence motifs *de novo* in a set of input sequences. However, the process of *de novo* motif discovery is computationally complex, and biologically relevant motifs may go undiscovered unless prior biological knowledge is used to narrow the search [100].

2.2.4 Motif databases

For a number of organisms, including the fruit fly [101], mouse [102] and human [102], large databases of DNA-binding factor consensus motifs have been determined. Accordingly, DNA sequence models can be constructed by use of entire motif databases. Two key benefits of this approach are that 1) predictive motifs need not be identified in advance for the sequence class of interest, and 2) the model may identify new motifs as predictive of the target sequence class, yielding new biological insights.

It is important to note that the PWMs in motif databases are themselves models of sequences bound by factors, and that the motifs are unlikely to perfectly reflect the preferred binding sequences for the noted factors. Furthermore, cooperative binding of factors has been found to in some cases substantially alter binding sequence preference [103]. Additionally, most motifs within the database may be irrelevant to the modelling problem at hand, adding a potentially large number of irrelevant features to the model, for which overfitting can occur.

2.2.5 *k*-spectrum kernels

An unbiased, comprehensive approach for selecting motifs is to use a k-spectrum [104], which is the set of occurrence frequencies of all possible, unambiguous motifs of length k, also known as k-mers.

Definition 21 (*k*-spectrum kernel) *The k-spectrum kernel is a mapping $\mathcal{S}_k : \mathbb{S} \rightarrow \mathbb{R}^{4^k}$, of any sequence $s \in \mathbb{S}$ to a feature vector $\vec{v} \in \mathbb{R}^{4^k}$, where the components are the occurrence frequencies of all k-mers—motifs of length k.*

Combining k-spectra with machine learning models, such as SVMs, enables models

to learn predictive motifs *de novo*, based only on training sequences. It is worth noting that dimensionality increases exponentially with the choice of k. However, due to k-spectra being comprehensive sets, overlaps of multiple k-mers can enable the modelling of motifs that are longer than k nucleotides, and a small value of k may be sufficient. k-spectra may contain many k-mers that are irrelevant to the modelling problem at hand, leading to a potential risk of overfitting randomly enriched signals in the training sequences, and complicating the definition of strict negative training sets.

Motifs of DNA-binding factors commonly contain degenerate positions (positions where more than one nucleotide can be allowed). Several alternative formulations of k-spectra have been published, allowing mismatches or gaps [105, 106, 107]. Allowing one mismatch per k-mer in an arbitrary position yields a k-spectrum mismatch kernel. In the context of this thesis, I focus on allowing only a single mismatch.

Definition 22 (k-spectrum mismatch kernel) *The k-spectrum (single) mismatch kernel is a mapping $\mathcal{S}_k^* : \mathbb{S} \rightarrow \mathbb{R}^{4^k}$, of any sequence $s \in \mathbb{S}$ to a feature vector $\vec{v} \in \mathbb{R}^{4^k}$, where the components are the occurrence frequencies of all k-mers—motifs of length k—, with one mismatch allowed in an arbitrary position.*

2.2.6 Structural DNA sequence features

The structural properties of DNA sequences can be influenced by their sequence composition. A variety of sequence features have been experimentally determined that influence structural properties of DNA, which in turn have been found to influence nucleosome occupancy [108]. In the work of [108], structural features are provided in the form of weighted dimer and trimer spectra. DNA structural features thus provide an additional feature set that can be used when modelling regulatory DNA sequences. In

the context of this thesis, for the sake of flexibility, instead of using published weighted dimer and trimer spectra, dimer frequencies are used directly with machine learning methods in order to model structural properties of sequences.

2.3 Generative sequence models

It is often useful to be able to randomly generate DNA sequences with desirable properties. The simplest generative DNA sequence model is one that outputs nucleotides with given probabilities (in the simplest case with all probabilities equal to $p(n) = \frac{1}{4}, n \in \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$). This is called an independent and identically distributed (i.i.d.) sequence model.

Definition 23 (Independent and identically distributed (i.i.d.) sequence model) *An independent and identically distributed (i.i.d.) sequence model randomly outputs nucleotides based on per-nucleotide probabilities, $p(n)$, for $n \in \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$.*

An i.i.d. model can be trained by counting the number of observations of each nucleotide in training sequences, and dividing by the total length. An i.i.d. sequence model does not conserve motifs. When conservation of motifs is desirable, an N-th order Markov chain can be employed.

Definition 24 (Sub-sequence) *For a sequence $s \in \mathbb{S}$, $s_{a\ldots b}$ will denote the sub-sequence of s from nucleotide index a up to, and including, nucleotide index b.*

Definition 25 (N-th order Markov chain sequence model) *An N-th order Markov chain sequence model randomly outputs the next nucleotide based on the probability of observing each nucleotide after the preceding N-mer, $p(n|o_{i\ldots i+N-1})$, for $n \in$*

$\{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$, *where* $o_{i\ldots i+N-1}$ *is the preceding* N*-mer.*

An N-th order Markov chain can be trained by calculating the fraction of times each nucleotide succeeds each N-mer out of the total number of occurrences of the N-mer. This yields one categorical distribution over nucleotides per possible N-mer.

2.4 Training nucleic acid sequence models

The aforementioned methods, with the exception of dummy models, all require model training. A variety of approaches can be used when training sequence models, with the optimal approach depending on the machine learning method used, as well as on the problem at hand. In the context of this thesis, all predictive machine learning models are binary or multiclass, and require a positive training set and one or more negative training sets.

2.4.1 Positive training set construction

Positive training sequences can be sourced from the literature for high-confidence functional examples. Alternatively, positives can be determined by an automated scheme with genome-wide experimental data. For Polycomb target sites, candidate PREs have previously been defined based on clusters of ChIP-chip and ChIP-seq peaks for relevant biological signatures [69, 70, 32]. Positives may have variable sequence lengths, necessitating a scheme for training with variable-length sequences.

One strategy that can be employed when features are based on frequencies (which all of the aforementioned features are) is to derive the frequencies from the entire sequences, yielding one feature vector per sequence. A potential complication with this

approach is that predictive features can be "watered out" in longer sequences, if the longer sequences contain irrelevant sequence material, depleted of the predictive features.

An alternative strategy is to use a fixed training sequence length, by either sampling all sliding windows from the positive sequences, or alternatively, sampling one or more sliding windows based on a selection criterion, expected to identify the true positive sub-regions. Zeng *et al.* [54] sampled the windows from the positives with the highest number of relevant motif occurrences. An alternative to this, which we explore in **Article II**, is that of prediction within the positives in order to home in on the core predictive regions.

2.4.2 Negative training set construction

There are many approaches for constructing negative training sets for the training of a discriminative, predictive model. A simple strategy is to use a generative sequence model to generate a set of negatives. For generated sequences, it is desirable to avoid having the negatives be "too null", such that the distance between the positives and negatives in the feature space is large, and the trained model decision surface may be far away from the true positives. For i.i.d. models, where nucleotides are generated with fixed probabilities, the resulting sequences will generally be highly null. For N-th order Markov chains trained genome-wide, the probability of generating true positives will in general also be low, unless the target class occurs more frequently in the genome than other sequence classes. The resulting generated sequences are, however, likely to be too null. Alternatively, if a Markov chain trained on positives is used to generate negatives, the generated negatives preserve motif composition of the positives, but not

higher-order structure, such as motif pairing.

An alternative strategy is to construct a biologically informed set of negatives. Ringrose *et al.* [47] constructed a negative set based on promoters that were found not to be regulated by Polycomb, but nonetheless regulated by GAF and ZESTE, which are involved in Polycomb recruitment to chromatin. A key advantage with this approach is that it pushes the decision surface towards the positives in a biologically informed way. A disadvantage is that their approach was manual, and does not readily scale for the construction of a larger training set. Alternatively, if a class of genomic sequences is known not to coincide with the positives, this class can be used as negatives. In **Article I**, we use annotated coding sequences as negatives. Negatives can also be determined based on experimental data by taking genomic regions depleted of relevant signals. When using signal depletion for designating negatives, absence of evidence is not evidence of absence. However, if true positives are relatively rare, randomly selected windows depleted of relevant signals can be expected to only rarely contain positives. Alternatively, randomly selected genomic windows can be used. Notably, randomly selected genomic windows may be too null if the relevant features are rare.

Yet other alternatives arise from iterative training set refinement. One such approach is that of Mapping-Convergence (M-C) with Support Vector Machines [109]. The idea behind the M-C algorithm exploits the maximum-margin property of SVMs. The algorithm starts with a positive set and an unlabelled set, and the model is trained with an initial set. Iteratively, the unlabelled set is classified, and negatively classified instances are appended to the negative training set, forcing the decision surface towards the positives. An alternative approach to iterative refinement that we explore in **Article II** is in terms of reinforcement learning, where both the positive and negative training sets are

refined based on the scrutinizing of predictions.

As with positives, sampling sliding windows, either all or by a selection criterion, may be beneficial in order to avoid warping of the training set feature space.

Chapter 3

Model deployment and the quantification of generalization ability

Once a model has been trained, it is important to measure its generalization ability. Such measures can elucidate how well the model will fare at predicting new instances, and aid in selecting the best from a set of models. Various approaches can be used in order to measure generalization and to deploy a sequence model for prediction. In this chapter, I cover the background of how these tasks can be accomplished.

3.1 Scoring sequences

In order to measure the generalization ability of classifiers, the classifiers should be applied to independent test data. The application of classifiers to sequences can be performed in several ways. For the sake of flexibility, I first focus on the scoring of sequences. Binary classification can later be achieved by means of thresholding.

For feature sets that are applicable to sequences of variable lengths, such as motif occurrence frequencies, a model can be applied to entire sequences that are to be scored. When sequences are of variable lengths, this can bias scoring by warping the feature space and averaging out predictive features.

Alternatively, a model can be applied to sequences using a sliding window. This avoids feature space warping by keeping observed sequence lengths constant. The resulting window scores can be handled in one of two ways: 1) all scores per class can be gathered (across all sequences), in order to calculate validation metrics, or 2) the set of window scores per sequence can be condensed into a sequence score, such as by taking the mean, median or maximum window score for a sequence as a representative sequence score.

For regulatory element sequences, known positives have variable lengths [47, 52, 110], as do experimentally determined PcG/TrxG peak clusters [69, 70, 32], and we can generally not be confident that every window in a positive sequence corresponds to a true positive. Accordingly, using all window scores for the calculation of generalization performance metrics is likely to yield misleading results. Taking the average window score over a sequence as a representative sequence score is also problematic, as the scores of longer sequences will be watered out. Using the maximum sequence

window score as a representative sequence score has the benefits of both keeping observed sequence lengths (as presented to the model) constant and preserving positive prediction. Accordingly, we focus all validation on this score condensation approach, throughout this thesis.

3.2 Quantifying model generalization ability

The most commonly used measures of classifier generalization performance are those derived from the *confusion matrix*.

3.2.1 The confusion matrix and associated measures

For a set of scored sequences and a classification threshold value, binary class predictions can be assigned to the sequences and the numbers of *True Positives* (TP; correctly classified positives), *False Positives* (FP; negatives classified as positives), *True Negatives* (TN; correctly classified negatives) and *False Negatives* (FN; positives classified as negatives) can be calculated. These values form the constituents of the *confusion matrix* [111].

Definition 26 (Confusion Matrix) *The confusion matrix is a 2x2 contingency table, of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN).*

	Actual positive	*Actual negative*
Predicted positive	TP	FP
Predicted Negative	FN	TN

A variety of measures of model generalization have been developed based on the

confusion matrix, including the *True Positive Rate* (TPR), also known as *Recall* and *Sensitivity*, *False Positive Rate* (FPR), and *Positive Predictive Value* (PPV), also known as *Precision* [112].

Definition 27 (True Positive Rate) *The True Positive Rate (TPR), also known as Recall and Sensitivity, is given by*

$$TPR = Recall = Sensitivity = \frac{TP}{TP + FN}.$$

Definition 28 (False Positive Rate) *The False Positive Rate (FPR), is given by*

$$FPR = \frac{FP}{FP + TN}.$$

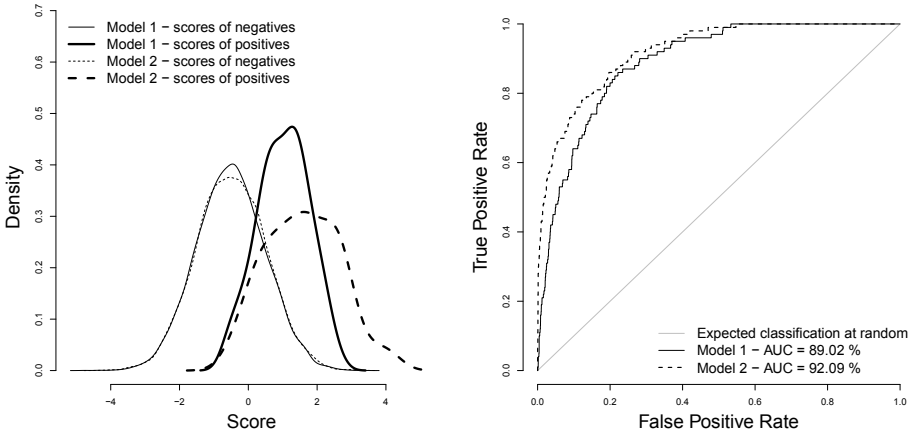
Definition 29 (Positive Predictive Value) *The Positive Predictive Value (PPV), also known as Precision, is given by*

$$PPV = Precision = \frac{TP}{TP + FP}.$$

For ease of comparison, it can be useful to condense multiple measures into a single measure of generalization ability. This can be achieved either for a single classification threshold, or for multiple thresholds. For a given classification threshold, measures that achieve this include the F_1 score—the harmonic mean of precision and recall—[113].

Definition 30 (F_1 score) *The F_1 score is given by*

$$F_1 = 2\frac{Recall * Precision}{Recall + Precision}.$$



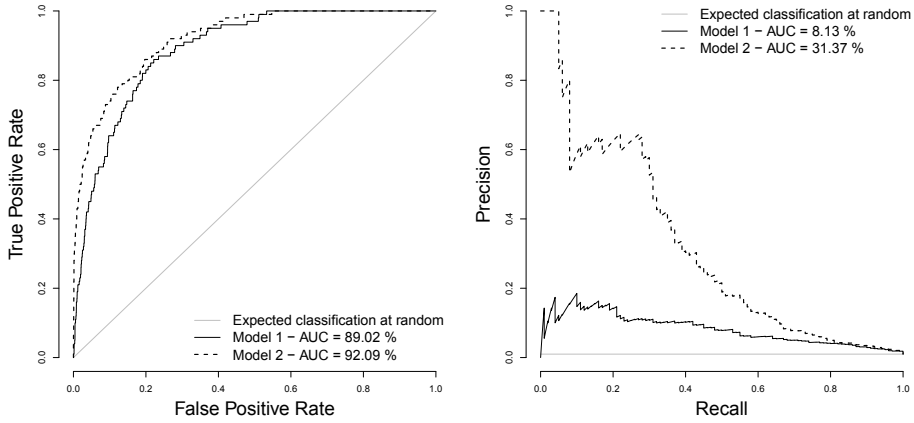
(a) Score densities for two simulated models

(b) Receiver Operating Characteristic curves

Figure 3.1: Simulated scores for two models, and the corresponding ROC curves. The ROC curves were generated with R [114] and the `precrec` library [115]. Each model is composed of normally distributed, randomly generated positive class scores and negative class scores, with 100 positives and 10000 negatives, yielding a ratio of positives to negatives of $1/100$. For model 1, $\mu_{\oplus} = 1.0, \sigma_{\oplus} = 0.8, \mu_{\ominus} = -0.5, \sigma_{\ominus} = 1.0$, and for model 2, $\mu_{\oplus} = 1.5, \sigma_{\oplus} = 1.0, \mu_{\ominus} = -0.5, \sigma_{\ominus} = 1.0$, where $\oplus$ and $\ominus$ correspond to the positive and negative classes, respectively.

3.2.2 Thresholdless measures of generalization

A common approach for threshold-independent characterization of generalization is to generate a curve by plotting two confusion matrix-based measures against one another by varying the prediction threshold from minimal to maximal. An instance of this is the *Receiver Operating Characteristic curve* (ROC curve), which plots the *True Positive Rate* in the y-axis against the *False Positive Rate* in the x-axis [116], and further to calculate the Area Under the Curve (*ROC AUC*) [116]. Figure 3.1 shows an example of two simulated models, with score density curves and a corresponding Receiver



(a) Receiver Operating Characteristic curves

(b) Precision/Recall curves

Figure 3.2: Comparison of simulated ROCs and PRCs. Whereas the ROC curves are similar for models 1 and 2, the difference between the PRCs is substantially larger. The simulated scores are the same as in Figure 3.1. The curves were generated with R [114] and the `precrec` library [115].

Operating Characteristic curve.

Definition 31 (Receiver Operating Characteristic curve) *The Receiver Operating Characteristic (ROC) curve is the curve generated when varying the classification threshold from minimal to maximal (or vice versa) and plotting the True Positive Rate in the y-axis against the False Positive Rate in the x-axis.*

A better model will have higher a TPR for a lower FPR, such that the curve tends towards the upper left. A model that predicts both classes with equal probability (a fair coin-toss) will have a ROC curve tending towards the diagonal from $TPR = FPR = 0$ to $TPR = FPR = 1$.

The ROC curve does not reflect any imbalance that may exist in class prominence,

which can yield skewed measures of generalization in genome-wide prediction tasks, where positives are often rare compared with negatives [117]. The Precision/Recall Curve (PRC), and the area beneath it (PRC AUC), addresses this issue by taking model precision into account.

Definition 32 (Precision/Recall Curve) *The Precision/Recall Curve (PRC) is the curve generated when varying the classification threshold from minimal to maximal (or vice versa) and plotting the Precision in the y-axis against the Recall in the x-axis.*

Precision/Recall curves for the simulated models in Figure 3.1 are given in Figure 3.2. The expected classification performance at random is equal to the ratio of positives to the total number of instances tested, $\frac{n_{\oplus}}{n_{\oplus}+n_{\ominus}}$, for $n_{\oplus}$ positives and $n_{\ominus}$ negatives. The area under the PRC is equal to the average precision. A better PRC will tend towards the upper right. Where the ROC curves are similar for the two models, the PRC for *Model 2* is vastly superior, achieving much higher precision over a large range of recall values.

3.2.3 Cross-validation

The measurement of classification performance on the training set generally is not a good measure for how well a classifier will perform on new data. In fact, the models that show the best performance on the training set are often the models that overfit it, and in turn perform poorly on new data [89]. Several methods have been developed in order to avoid this issue, collectively referred to as *cross-validation* techniques.

Common techniques for cross-validation include unstratified or stratified n-fold cross-validation, and leave-one-out (LOO) cross-validation [118]. For unstratified n-

fold cross-validation, the training set is divided into n independent portions (folds) of equal size, and iteratively, each portion is left out from training and independently used for validation. For stratified cross-validation, an equal number of examples for each training set class is used for each fold. Leave-one-out (LOO) cross-validation, as the name suggests, iteratively leaves exactly one training example out, until the entire training set has been independently classified. Larger values of n ($n \geq 10$) for n-fold cross-validation (and in turn LOO cross-validation) have been associated with increased bias, and a lower n ($n \leq 5$) has been associated with increased variance of the measured generalization [118].

3.3 Classifier threshold calibration

In order to predict candidate CREs using a scoring model, a prediction threshold must be set. Several methods have been developed and employed to this end. One approach for setting a prediction threshold is to control for the expected number of false positive predictions. Ringrose *et al.* [47] calibrated the prediction threshold of their PREdictor algorithm by randomly generating 100 genome-sized sequences using an i.i.d. model trained genome-wide (the background model), applying the PREdictor to the randomly generated sequences, and based on the resulting score profiles setting a threshold that yielded an average of one prediction per sequence, corresponding to an expectation of one false positive prediction genome-wide (an E-value of 1). For increased stringency, the same approach can be adapted to more complex background models (such as N-th order Markov chains). A pro of the E-value approach is that it does not require having an independent set of positives for calibration. A con is that it does not take into account

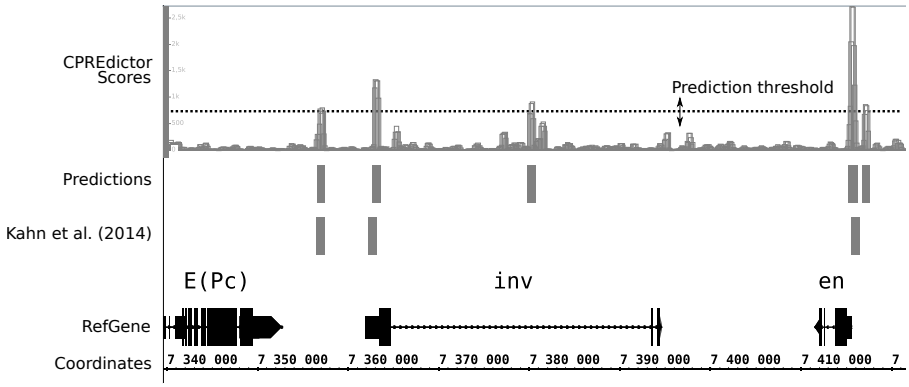


Figure 3.3: Window scores and predictions by the CPREdictor at the *engrailed/invected* locus. CPREdictor was here trained as in the MOCCA tutorial (see **Article III**). The figure was adapted from a screenshot from the Integrated Genome Browser [50]. The Kahn *et al.* [32] regions (taken from their Supplementary Table S1) are computationally defined PREs, based on ChIP-chip data. The R5 *D. melanogaster* genome assembly [53] was used.

the ability of a model to positively classify independent positives.

An alternative approach is to set a prediction threshold based on sets of both positives and negatives. This can for instance be achieved by selecting a threshold for a desired $Precision$ in Precision/Recall space. Precision/Recall curves commonly have spikes, with up to multiple thresholds yielding the same or higher Precision. A solution is to set the threshold for maximal $Recall$ that achieves the desired $Precision$ or higher (discussed further in **Article I**).

3.4 Genome-wide prediction

A sequence model can be applied for genome-wide prediction. This is generally achieved by applying the model genome-wide using a sliding window, with fixed window and

step sizes [47, 75, 54]. This results in a genome-wide score profile curve. Applying a prediction threshold yields predicted windows, and overlapping predicted windows are merged into non-overlapping predicted regions [47, 75]. This is illustrated in Figure 3.3.

3.5 Region overlap evaluation

Given sets of genome-wide CRE-predictions and of experimentally determined CREs, the overlaps can give an indication of the precision and recall that the predictors have achieved. Similarly, overlaps can be used to quantify the agreement between sets of experimentally determined CREs or of predictions. In this thesis, we use two measures of overlaps between region sets.

Definition 33 (Overlap recall) *Given two sets of regions, A and B, the overlap recall of A to B is defined as*

$$Overlap\ Recall(A, B) = \frac{|B \to A|}{|B|},$$

where $B \to A$ is the set of regions from B that overlap with at least one nucleotide of at least one region in A.

Definition 34 (Overlap precision) *Given two sets of regions, A and B, the overlap precision of A to B is defined as*

$$Overlap\ Precision(A, B) = \frac{|A \to B|}{|A|} = Overlap\ Recall(B, A).$$

The definition of overlaps based on regions of one set that overlap with at least one nucleotide of a region in the other set avoids biasing measurements of overlaps when region lengths are variable.

3.6 Target gene prediction

Based on predicted regions, candidate target genes can be predicted. The simplest approach to this end is to predict the closest gene to each predicted region, as was done by Ringrose *et al.* [47]. The closest gene to a prediction need not necessarily be the only target of a regulatory region, or even a target at all. For example, the presence of boundary elements can block the action of regulatory elements on nearby promoters in one direction [14], or enable skipping across domains [119]. However, in the absence of additional information, the closest gene is the most probable candidate. As an alternative to predicting the closest gene, the gene with the closest promoter can be predicted. In order to reduce the risk of false negatives, target gene prediction can be performed bidirectionally. Zeng *et al.* [54] predicted the closest gene in each direction of a prediction if the second closest gene in the other direction was within four kilobases.

Chapter 4

A brief history of Polycomb/Trithorax target sequence models

Over the past two decades, a variety of studies have modelled and predicted candidate PREs and PcG/TrxG target sites in the fruit fly and vertebrate genomes. I here review the history of genome-wide prediction of PREs and PcG/TrxG targeting, focusing first on the fruit fly, in which the most work has been conducted, and then moving on to vertebrate Polycomb target prediction.

4.1 Fruit fly PRE prediction

4.1.1 The PREdictor—the pioneering work

In 2003, Ringrose, Rehmsmeier *et al.* [47] published the PREdictor—the first method developed with the goal of predicting Polycomb/Trithorax Response Elements genome-wide in the *D. melanogaster* genome. The PREdictor is a log-odds model with motif pair occurrence frequencies as features, with a pairing cutoff distance of 220 basepairs. The cutoff distance enables the model to learn locally clustered pair occurrence frequencies. The PREdictor was trained on a small set of 12 PREs and 16 non-PREs, with 7 motifs. The motif set focused exclusively on IUPAC nucleotide code motifs, and consisted of two motifs for the GAGA-binding factor (GAF), three for *Pleiohomeotic* (*Pho*), one for *Zeste*, and one motif termed EN1, found important for the function of the *engrailed* PRE. The PREs used for training were previously published, experimentally identified PREs. The non-PREs were promoter sequences that were determined to not be regulated by Polycomb, in spite of being regulated by GAF or Z. In addition to the PREdictor, Ringrose *et al.* [47] had tested a model of singular motif occurrence frequencies, and found that only the paired motif model could distinguish PREs from non-PREs, suggesting an importance of motif pairing. Ringrose *et al.* [47] then set a prediction threshold for the PREdictor for an E-value of 1 (based on 100 genomes randomly generated by an i.i.d. background model trained genome-wide) and applied the PREdictor for genome-wide prediction using a sliding window with a size of 500bp. Windows with scores above the threshold were predicted, and overlapping windows were merged into non-overlapping candidate PRE predictions. With this experimental set-up, the PREdictor predicted 167 candidate PREs. Ringrose *et al.* [47]

experimentally determined *Pc*-enrichment at PRE predictions, and found over half to be *Pc*-enriched. Furthermore, Ringrose *et al.* [47] experimentally tested four PRE predictions using the *miniwhite* assay, and found that all four predictions exhibit hallmark signs of being functional PREs: all four silenced the *miniwhite* reporter gene, three of four showed pairing-sensitive silencing (repression was stronger for homozygotes than for heterozygotes), and several lines showed variegation and a dependence on PcG for repression (repression was lost in PcG mutant flies).

4.1.2 The jPREdictor

In 2006, Fiedler *et al.* [75] published the jPREdictor, which is a Java-implementation of the PREdictor method, sporting a Graphical User Interface (GUI). The jPREdictor is a flexible tool, allowing the user to define their own sets of motifs and their pairing. In addition to supporting IUPAC nucleotide code motifs, the jPREdictor supports the use of PWM motifs. Fiedler *et al.* [75] trained the jPREdictor with a similar motif set to that used by Ringrose *et al.* [47], but with the addition of a *Pho* PWM motif and minor pairing modifications. This model yielded a larger set of predictions for the same E-value of 1, of 378 candidate PREs genome-wide.

4.1.3 Evolutionary plasticity of PREs across drosophilids

The genomes of close evolutionary relatives of *D. melanogaster* have been sequenced, including those of *D. simulans*, *D. yakuba* and *D. pseudoobscura*. In 2008, Hauenschild *et al.* [76] published an article which addressed the question of how PREs have evolved among these close evolutionary relatives (within 50 million years since the last common ancestor). To this end, Hauenschild *et al.* [76] applied the jPREdictor [75] to

the genomes of all four species, using the same configuration as was used in 2003 by Ringrose *et al.* [47], with the modification of using a window step size of 10 basepairs. The authors found that the numbers of predictions made varied between the species, with *D. pseudoobscura* having over twice as many predictions as *D. melanogaster*. The authors tested predictions from all four species using chromatin immunoprecipitation, and found that predictions in the Bithorax complex of each species were PcG-enriched in embryos. Additionally, the authors tested—and verified—the orthologous predictions of the *bxd*-PRE in all three of the relatives of *D. melanogaster*. From this, the authors concluded that the sequence criteria of PREs are highly conserved among these species.

In addition to using the classic PREdictor [47] algorithm, Hauenschild *et al.* [76] introduced a derivative method, which boosts the score of predictions that are conserved among the different fly species, henceforth referred to as the Dynamic PREdictor. The Dynamic PREdictor performs a BLAST search for each prediction in one genome, in order to identify orthologs in the other genomes, and, if an orthologous region is found, prediction is performed within a window centred at the ortholog. The closest predicted PRE in another species, if any, is predicted as a candidate orthologous PRE. The authors found that many predicted PREs are not conserved in position, and they experimentally verified with ChIP that selected predictions recruit PcG, both with conserved and non-conserved positions. Predictions with conserved positions were found to have motif turnover. In order to explain the evolutionary mobility of PREs, the authors proposed a model where motif turnover enables PREs to gradually shift in position.

4.1.4 The EpiPredictor

In 2012, Zeng *et al.* [54] published the EpiPredictor, a method that models singular motif occurrence frequencies in PREs using a Support Vector Machine (SVM) for binary classification, and scores positively classified windows based on the number of motif occurrences. The EpiPredictor additionally filters out positively predicted windows whose GC-content is below 44%. In addition to the SVM, Zeng *et al.* tested using a Bayesian Additive Regression Trees (BART) model, but found that the SVM yielded superior generalization.

When training the EpiPredictor, Zeng *et al.* [54] used the same set of 12 PREs, 16 non-PREs, and 7 motifs as were used by Ringrose *et al.* [47], with the addition of seven more non-PREs. Zeng *et al.* [54] maintained a constant window size of 200bp throughout training, and sampled from the PREs and non-PREs with a sliding window, using a step size of 20bp. For the PREs, Zeng *et al.* [54] sampled the window with the highest number of motif occurrences, and for the non-PREs, every window. Zeng *et al.* [54] observed improved generalization using their SVM model.

4.2 Vertebrate PcG target models

4.2.1 PcG target gene prediction in mouse embryonic stem cells

The first published study to train models of vertebrate PcG targeting was that of Liu *et al.* in 2010 [120]. Liu *et al.* [120] trained a Bayesian Additive Regression Tree (BART) model to discriminate murine PcG-enriched promoters from PcG-depleted promoters (TSS -8kb/+2kb). The feature space of the BART model consisted of occurrence fre-

quencies of all motifs from a vertebrate transcription factor consensus motif database, and seven transcription factor enrichment signals from previously published genome-wide ChIP-chip and ChIP-seq experiments.

The authors found that their BART model and CpG-density could distinguish PcG-enriched promoters from PcG-depleted ones, whereas conservation scores could not, with their BART model yielding the best distinction performance, in terms of cross-validation ROC AUC. The removal of ChIP-based transcription factor features (e.g. using only motifs) resulted in a minor decrease in the cross-validation AUC (their Figure 2a). Different choices of machine learning methods (among BART, Group LASSO and SVM) also had a minor impact on the cross-validation ROC AUC (their Supplementary Figure 1), whereas different choices of motif databases (among TRANSFAC, JASPAR and UniProbe) had a larger impact (their Supplementary Figure 2). The authors attempted pan-species prediction for fruit fly and mouse genomes, which yielded poor generalization. Notably, the authors found the motif for *Yy1*—a mouse ortholog of the fruit fly PcG protein *Pho*, which has a highly conserved DNA-binding consensus motif—to be significantly depleted at mouse PcG-enriched promoters.

4.2.2 PcG target gene prediction in human embryonic stem cells

In 2013, Xiao *et al.* [121] published a study in which the authors trained a Support Vector Machine using the method of Mapping-Convergence (M-C). This enabled an iterative convergence of the classifier decision surface towards the positive training set, based on training with a set of positives and an unlabelled set. The positive training set consisted of gene promoters (TSS -1/+1 kb) enriched in EZH2, and the initial negative training set of randomly selected unlabelled instances. For the features of the model,

Xiao *et al.* [121] used 10 histone modifications and 63 transcription factor consensus binding motifs.

The authors achieved high $Precision$ and $Recall$ to PcG-targeted promoters over 3-fold cross-validation, and observed reduced generalization when removing either histone modifications or motifs as features. Notably, the initial training set used corresponded to a training set of positives and randomly selected negatives, and the authors were able to converge towards a more stringent negative set by means of the Mapping-Convergence procedure, yielding higher $Precision$ and $Recall$.

4.2.3 Genome-wide prediction of H3K27me3 nucleation sites in *Xenopus tropicalis*

In 2014, van Heeringen *et al.* [42] published a study where an 8-spectrum SVM was trained to distinguish PcG-repressed domains from background in the Western clawed frog, *Xenopus tropicalis*. Their model was trained on 1kb fragments of H3K27me3 domains as positives, and randomly selected 1kb H3K27me3-depleted genomic sequences as negatives.

Van Heeringen *et al.* found that the SVM could accurately distinguish H3K27me3-domains from depleted regions, and identified motifs within the predictions. Additionally, the authors trained SVMs on comparable zebrafish and human data, and applied all three models for pan-species prediction, which yielded high generalization in the same genome as the model was trained on, and low to moderate in other genomes. The authors additionally found that H3K27me3 nucleation sites in *X. tropicalis* do not correlate with GC-content, but that they do correlate with Non-Methylated Islands (loci depleted of DNA methylation) and a variety of sequence signatures.

4.3 A bird's eye perspective on past efforts

Over the past 17 years, a variety of published studies have trained sequence models in order to predict PcG/TrxG targeting. These studies have yielded novel insights into PRE and PcG target site architecture and evolution, and have also provided predictions as resources for further study.

The majority of PcG target site model studies have focused on fruit fly PREs. All of the aforementioned fruit fly PRE-models have been trained with small sets of PREs and non-PREs, and small sets of known PRE-motifs. This has left several questions open, including: 1) How well do models trained on genome-wide PcG target sites generalize to PREs, and how stable is the generalization for different sets? 2) How well do PRE-predictors generalize when taking the expected genomic imbalance of PREs to non-PREs into account? 3) Are there additional motifs important for fruit fly PRE-function? 4) How well would a more advanced machine learning algorithm trained with larger data sets generalize to PREs?

The majority of vertebrate PcG target site models have focused on the prediction of PcG target genes, with no attempt made at predicting intergenic PREs. The study by van Heeringen *et al.* [42] modelled genome-wide H3K27me3-enriched loci, and experimentally confirmed several predictions as repressive elements, none of which were CpG island. Several questions are left open by the vertebrate PcG target site studies: 1) How well can models of vertebrate genome-wide PcG/TrxG signal clusters generalize? 2) How can we best construct a stringent negative training set for vertebrate PcG/TrxG cluster models? 3) How well can vertebrate PcG target site models predict experimentally verified vertebrate PREs? 4) Are there conserved sequence features

among intergenic vertebrate PcG target sites, and if so, what are they?

In the remainder of the thesis, I aim to address the aforementioned, previously open questions.

Part II

Present investigation

Chapter 5

Aims of the thesis

I set out on my PhD project with two primary goals: 1) to further the state-of-the-art in the modelling and prediction of Polycomb/Trithorax Response Elements, and 2) to further our biological understanding of i) the structure, ii) the function and iii) the evolution of PREs, via the exploitation of the vast amounts of relevant experimental data that have become available, in connection with DNA sequence. This thesis presents four scientific contributions: two full-length scientific articles with novel investigations (**Article I** and **Article II**), and two shorter papers for tools that I developed over the course of my thesis work (**Article III** and **Article IV**). In accordance with goal 1), all four articles included in my thesis focus on sequence-based machine learning and Polycomb/Trithorax Response Elements or PcG/TrxG target sites. **Article I** focuses on prediction in the fruit fly, *D. melanogaster*, and **Article II** on prediction in both fly and vertebrate genomes, thus addressing goal 2).

Chapter 6

Contribution summaries

In this chapter, I outline the motivations and results of each of the four articles of this thesis. The articles summarize the results of the thesis in more detail.

6.1 DNA sequence models of genome-wide *Drosophila melanogaster* Polycomb binding sites improve generalization to independent Polycomb Response Elements

The past one-and-a-half decades have seen the publication of a number of genome-wide maps of PcG and TrxG signals in the fruit fly genome [66, 67, 68, 69, 70, 32], which have not previously been used in order to train fruit fly PRE-sequence models. Past studies have also not compared expected generalization of fruit fly PRE-predictors

while taking the effects of imbalanced data into account. Furthermore, the GTGT-motif [47, 49] has not previously been used for training PRE-predictors.

In **Article I**, we trained a re-implementation of the PREdictor algorithm using experimentally determined genome-wide PcG/TrxG clusters and generated negatives. We additionally developed a new hybrid method, the Support Vector Machine Motif Occurrence Combinatorics Classification Algorithm (SVM-MOCCA). SVM-MOCCA constructs one SVM per motif, in order to predict whether occurrences are similar to those in PREs, with a feature space consisting of local dinucleotide and motif occurrence frequencies, and combines positively classified motif occurrence frequencies using a log-odds model.

We found that models trained on genome-wide PcG/TrxG clusters improve generalization to independent PREs—with the addition of the GTGT-motif further improving generalization, and SVM-MOCCA outperforming the other methods—, and we precisely predicted several PREs that had been left out during training. The improvement observed with GTGT could not be attributed only to an increased model complexity. We additionally tested models that included published motifs that were not in the Ringrose *et al.* [47] motif set, and found that these did not improve generalization. When training with only 12 PcG/TrxG clusters, we also observed improvement over when training with 12 *Hox* PREs, suggesting a qualitative difference between the sets. Our models predict orders of magnitude more candidate PREs genome-wide than previous methods, with large subsets enriched in biologically relevant signals.

6.2 Biomarker reinforcement learning with k-spectra enables precise Polycomb target site prediction without prior motif knowledge

Genome-wide maps of PcG/TrxG binding and DNA accessibility are available for a number of species, including the fruit fly [51], mouse [122, 123] and human [74, 124]. PcG/TrxG target sites in *D. melanogaster* have previously been defined based on genome-wide experimental data by the clustering of signals [69, 70, 32], which we exploited for the training of our models in **Article I**. Although motif knowledge for vertebrate PcG/TrxG target sites is limited, the use of k-spectra [104] with SVMs enables models to identify predictive motifs based on the training data.

In **Article II**, we trained 5-spectrum mismatch quadratic kernel SVMs with iterative training set refinement by means of a reinforcement learning regimen that we developed—Positive Convergence, Additive Negative Biomarker Reinforcement Learning (PeCAN-BioRL). PeCAN-BioRL uses genome-wide experimental data to scrutinize predictions made by the previous iteration, with the goal of iteratively improving accuracy. We applied PeCAN-BioRL to the fruit fly, mouse and human genomes.

The initial models are unable to distinguish biomarker-enriched and accessible biomarker-depleted sequences, and make unrealistic numbers of predictions. Generalization improves over iterations of PeCAN-BioRL. The final fruit fly model is competitive with SVM-MOCCA, in spite of not having been trained with known PRE-motifs. The false positives rates decrease for all three species. In conclusion, our PeCAN-BioRL procedure coupled with the use of 5-spectrum mismatch SVMs enabled us to train successful

classifiers without prior motif knowledge. Detailed analysis of the fruit fly model identified a number of motifs that can be mapped to factors with support for PRE-related function in the literature. Mammalian model analysis also identified factors that may be involved in PRE function. Pan-species prediction and model analysis identified conserved predictive features. Substantial subsets of our predictions are enriched in relevant experimental signals, and are provided as a resource for further study.

6.3 MOCCA: A flexible suite for modelling DNA sequence motif occurrence combinatorics

Given a CRE-class and a set of known predictive motifs, a number of motif-based predictive sequence models can be formulated based on different feature space formulations and machine learning algorithms. In **Article I**, we developed the SVM-MOCCA method, which proved highly successful at predicting PcG/TrxG target sites in *Drosophila melanogaster*. However, we did not publish a polished implementation, which can be useful for the wider community of CRE researchers. We also had not tested the use of PWM motifs with SVM-MOCCA.

In **Article III**, we present MOCCA—Motif Occurrence Combinatorics Classification Algorithms—, a flexible suite for the motif-based modelling of CRE-sequences that I developed over the course of working on **Article I**, and which is based on preliminary work that was done while working on my master's thesis. MOCCA is an optimized suite of tools, implemented in C++. MOCCA contains an optimized implementation of SVM-MOCCA, as well as a reimplementation of the PREdictor, and supports both IUPAC and PWM motifs. In addition, MOCCA presents a derivative of SVM-MOCCA

using the method of Random Forests, called RF-MOCCA. Finally, MOCCA includes a variety of feature space formulations that can be combined with different modelling methods.

We applied SVM-MOCCA and RF-MOCCA to the tasks of modelling PREs and Boundary Elements (BEs), making this study the first to model PREs using Random Forests, and BEs using MOCCA-based methods. We found that SVM-MOCCA and RF-MOCCA improve generalization to both PREs and BEs over that of the PREdictor, and over that of cdBEST [92] for BEs, with RF-MOCCA yielding the best generalization in both cases.

MOCCA provides a flexible suite of tools for researchers interested in experimenting with motif-based modelling of CRE-sequences. In addition to presenting a new method that further improves generalization to PREs, we have demonstrated the broader applicability of MOCCA by modelling BEs. MOCCA is open source and extensible, with the full source code available on Github: `https://github.com/bjornbredesen/MOCCA`.

6.4 Gnocis: An integrated system for interactive and reproducible analysis and modelling of *cis*-regulatory elements in Python 3

The Python programming language has recently risen to become the most widely used programming language in the world [125]. A wide arsenal of packages are available for Python, such as Pandas [126] and Scikit-learn [127], which enable Python as a

powerful data analysis platform. Nonetheless, a Python package has been lacking that streamlines the modelling of CREs by enabling the combination of feature sets with machine learning methods and by implementing functionality for model comparison and genome-wide prediction.

In **Article IV**, we present Gnocis, a feature-rich package for Python 3 that streamlines the analysis and the modelling of CREs. Gnocis implements functionality for data preparation and flexible vocabularies for feature set and model specification. Feature sets in Gnocis are composed of graphs that can efficiently extract, combine and transform features. The modelling API in Gnocis implements common functionality for model application, such as scoring and prediction using sliding windows, with multiprocessing support. Multiple feature space formulations are included, such as motif occurrence frequencies, paired occurrence frequencies and k-spectra, and furthermore, feature spaces can be combined. Gnocis implements support for four modelling methods that can be combined with feature spaces: dummy models (unweighted sums), log-odds models, Support Vector Machines (via Scikit-learn [127]) and Random Forests (via Scikit-learn [127]). Gnocis also implements support for neural networks via TensorFlow [128]. Additional models can be implemented by extending the Gnocis model base class. For model comparison, Gnocis contains a cross-validation engine that supports imbalanced, multiclass data. In order to facilitate interactive use and visualization, Gnocis integrates with IPython [129] and Matplotlib [130]. I implemented Gnocis in Cython, yielding a high-performance, compiled library.

We applied a 5-spectrum mismatch quadratic kernel SVM and a Convolutional Neural Network (CNN) to the problem of modelling fruit fly PREs, making this the first study to use CNNs for modelling PREs. We additionally applied a Gnocis im-

plementation of the PREdictor, and also SVM-MOCCA using a wrapper for MOCCA (included with Gnocis). The 5-spectrum mismatch SVM yielded the best generalization to independent PREs versus the same class of negatives as was used for training (dummy PREs). However, the SVM generalized poorly to PREs versus coding sequences, whereas the CNN achieved generalization comparable to that of the PREdictor and SVM-MOCCA. All steps in the experiments are provided in a Jupyter Notebook for easy reproduction of our results.

The broad suite of tools included in Gnocis not only reduces the need for resorting to using external libraries but also enables additional optimizations, further aiding in efficient model application, and enabling model specification and application with a minimal code footprint. Gnocis is open source, with the full source code available on Github: `https://github.com/bjornbredesen/gnocis`

Chapter 7

Discussion

In Chapter 4, I reviewed past work on the modelling and prediction of PREs and PcG/TrxG target sites, and I discussed questions left open. In this thesis, I present four articles, two of which address these questions through the development and application of methods, and analysis of results—**Article I** and **Article II**—, and the other two of which are for tools that I developed in connection with experimental work—MOCCA (presented in **Article III**) and Gnocis (presented in **Article IV**).

In this chapter, I discuss the contributions I have made in this thesis, in terms of methodological innovation and tools for the broader scientific community studying *cis*-regulatory elements, and in order to expand our understanding of the structure, function and evolution of Polycomb/Trithorax Response Elements and PcG/TrxG target sites. Finally, I wrap up the thesis with a discussion of future work.

7.1 Pushing the methodological boundaries of the field

Over the course of working on my master's and PhD theses, I and my main supervisor, Marc Rehmsmeier, identified several opportunities for improving the state-of-the-art in methodology within the field of modelling PREs and PcG/TrxG target sites in terms of DNA sequence composition.

Genome-wide experimental data for the training of *Drosophila* PRE-predictors

All previously published work for the modelling of fruit fly PRE-sequences has focused on using a small set of 12 experimentally verified PREs for training [47, 75, 76, 54]. The use of a small, high-confidence PRE training set is accompanied by several problems, including the inability to train methods that require larger amounts of data, issues with cross-validation, and difficulties with scaling up the approach through the experimental verification of PREs on a large scale. Furthermore, the majority of these 12 PREs are from *Hox* complexes, potentially biasing models.

Advances in experimental methods have yielded an abundance of genome-wide binding data. In **Article I**, we exploited these advances by training our models with experimentally determined genome-wide PcG/TrxG target sites. We also needed a comparably large negative training set. Another innovation that we present in **Article I** is the use of Markov chains to generate non-PRE training sequences, which retain PRE motif composition but lose pairing. Our results in **Article I** show a substantial improvement in generalization to independent PcG/TrxG target sites and PREs—which cannot be attributed solely to the increased training set size—, and make way for the training of PRE-predictors that require larger training sets.

In **Article II**, we furthermore define PcG/TrxG target sites in terms of clustered relevant genome-wide biomarkers obtained from public data repositories, for the fruit fly, mouse and human genomes. This way, we were able to define PcG/TrxG target sites by the clustering of higher counts of signals than have previously been used for the identification of PcG/TrxG target sites. We then used these newly defined PcG target site sequences to train our models.

An important limitation to note is that we cannot be confident that all PcG/TrxG clusters determined using genome-wide binding assays, such as ChIP-chip and ChIP-seq, are PREs. Binding signals can be detected at non-PREs for multiple reasons, including physical interactions between regulatory elements and experimental noise. We also cannot assume that a set of PcG/TrxG clusters is comprehensive, as different epigenetic states may yield different recruitment patterns. Nonetheless, we have found the sequence signals of PREs to be sufficiently strong in the sequences to be useful for training PRE-predictors.

The quantification of model generalization by means of Precision/Recall curves with imbalanced data

A variety of approaches can be employed for comparing the generalization of models. For fair and informative comparisons, all models should be cross-validated using the same data. The ratio of PREs to non-PREs in a genome is highly imbalanced, and accordingly, Precision/Recall curves are more informative than Receiver Operating Characteristic curves [117]. So as to not overestimate generalization, the test set employed should preserve the imbalance of positives to negatives. Studies published prior to **Article I** for the prediction of PREs in *D. melanogaster* have included only limited model

comparisons, such as in terms of sensitivity and specificity for cross-validation with a small, balanced training set [54], overlaps of PRE predictions to ChIP-chip peaks [76], and ROC curves for genes [54]. The question of how the models performed in terms of Precision/Recall curves when using larger, imbalanced test sets—which can be very different from the generalization indicated by ROC curves and balanced test sets [117]—remained unaddressed.

In **Article I** and **Article II**, we compared model generalization using Precision/Recall curves and imbalanced test sets, with numerous variations. Our work in **Article I** and **Article II** raises the bar for how PRE-predictors should be compared.

It should be noted that even when preserving the expected genomic ratio of positives to negatives in the test set, the types of positives and negatives used can strongly influence generalization measures. Evaluation with dummy genomic sequences can paint an overly optimistic picture, as the sequences will not retain the complexity of real genomic sequence. On the other hand, if using real genomic sequences, the presence of positives in the negative set may negatively impact measured precision.

SVM-MOCCA and RF-MOCCA: new contenders to the throne of fruit fly PRE-predictors

We presented a new CRE-predictor in **Article I**, which models CREs by means of training one SVM per motif, and combining motif-predictions by means of a log-odds model. SVM-MOCCA yielded the best generalization of the methods that we compared. In **Article III**, we presented a derivative method called the Random Forest Motif Occurrence Combinatorics Classification Algorithm (RF-MOCCA), where the SVMs are replaced by Random Forests (RFs). RF-MOCCA further improves generalization

to PREs.

MOCCA: a flexible suite for CRE-prediction

The potential use cases of SVM-MOCCA extend beyond PREs, to other classes of CREs with known motifs, such as boundary elements [92]. We have written a polished tool, MOCCA (presented in **Article III**), which contains efficient implementations of SVM-MOCCA and RF-MOCCA, as well as support for training models using combinations of different classifiers and feature space formulations. To demonstrate the broader applicability of SVM-MOCCA and RF-MOCCA beyond that of PREs, we trained SVM-MOCCA and RF-MOCCA with Boundary Elements (BEs), and found that both models improve generalization over that of classic SVMs and RFs using motifs or a 4-spectrum as features.

PeCAN Biomarker Reinforcement Learning

The reinforcement learning method that we introduced in **Article II**, PeCAN-BioRL, enables the training of PcG/TrxG target site models without prior motif knowledge. This enabled us to train PcG target site models for both fruit fly and vertebrate genomes. This method can be generalized to new genomes where sufficient data is available.

It should be noted that the successful application of the PeCAN-BioRL procedure depends on the quality and variety of relevant data available for the target organism. In our experiments, using only four experimental signals yielded poor generalization. Training with data from only one type of cell may also bias models.

Gnocis: a feature-rich package for CRE-bioinformatics in Python 3

The ability to interactively inspect data and explore analytical and modelling approaches can greatly improve the efficiency with which data science–based research can be conducted, as is evidenced by the popularity of tools such as Jupyter Notebooks [131] and libraries such as Pandas [126]. In **Article IV**, we present Gnocis, a feature-rich package for Python 3, with the aim of streamlining the analysis and the modelling of CREs, and in turn, making the life of the CRE-researcher easier. In addition to providing a set of tools that can readily be employed, and a tutorial that demonstrates all steps necessary for training and deploying multiple PRE-predictors, Gnocis provides classes that can be extended with user-defined models, and is open source.

7.2 Advancing our understanding of Polycomb/Trithorax Response Elements

In line with the title of this thesis, a central aim has been to expand our knowledge of i) the structure, ii) the function and iii) the evolution of PREs. I here address each of these three aspects.

The structure of Polycomb/Trithorax Response Elements

In **Article I**, we trained models using the motifs used by previous *Drosophila* PRE-models [47, 76, 54], and a variety of additional motifs from the literature [47, 2, 48, 49]. Of the additional motifs, only the GTGT-motif improved generalization. This adds to the evidence of the importance of the GTGT-motif to the structure of *Drosophila* PREs.

When we constructed the negative training set in **Article I**, we generated negatives (dummy PREs) with a 4th order Markov chain trained on PREs, preserving motif composition in the generated sequences, but not motif-pairing, as was previously identified as predictive of PREs [47]. The fact that we observed improved generalization when training models on genome-wide experimental data and dummy PREs further illustrates that combinatorial motif occurrence is a central feature of *Drosophila* PRE architecture, as otherwise, models would have failed to distinguish the two classes.

In **Article II**, our move beyond training with known PRE-motifs, and in-depth model analysis, enabled us to identify a variety of motifs for both fly and vertebrate PcG/TrxG target sites. For the fly model, a number of motifs can be mapped to factors whose involvement in PREs is supported by the literature. Elucidating whether the remaining motifs and putative binding factors contribute to PRE function will require further investigation.

The function of Polycomb/Trithorax Response Elements

The genome-wide prediction of PREs can potentially uncover PcG target sites that are not easily discovered using genome-wide binding assays, such as ChIP-seq, due to their dependence on cellular and experimental conditions. This is exemplified by predictions made by our models that land outside of *Drosophila* chromatin accessible early in development. We have performed target gene prediction in *Drosophila* in **Article I**, and performed gene ontology analysis. As our models make orders of magnitude more predictions than were made by previous studies, and our models generalize better to independent PcG target sites, the ontologies of our target gene predictions may be more representative of the genome-wide picture for PcG-regulation than those of previous

sequence model predictions. However, it is likely that target gene predictions that take chromatin domains into account will yield a more precise picture.

A central functional question: How likely is an experimentally determined PcG/TrxG target site or a predicted PRE to be a functional PRE? As noted earlier, there are multiple ways by which a non-PRE can be experimentally enriched in PcG/TrxG. As such, it is likely that not all experimentally determined PcG/TrxG target sites are PREs. Accordingly, I have made an effort to distinguish between the two in this thesis. First and foremost, peaks may arise as shadows caused by chromatin interactions [24]. Furthermore, even if we can extract the subset of true positive binding sites for PcG/TrxG proteins with complete precision, a PRE is defined by its ability to maintain target gene repression or activation, previously established by a separate regulatory element. The question remains of whether or not all *bonafide* PcG binding sites correspond to functional PREs, or whether the presence or absence of certain sequence features can yield a PcG-bound but dysfunctional PcG target site. The results of experimental testing of PRE predictions in previous studies [47, 76] demonstrate that sequence models trained on PREs with known PRE-motifs do predict PREs with high precision—as measured by means of the *miniwhite* assay—, and that the main issue has been a lack of sensitivity. It should be noted that although the *miniwhite* assay can test that a regulatory element is functional, it does not test the memory function of PREs, as it only gives a readout in the adult (Leonie Ringrose, personal correspondence). Our models in **Article I** are able to separate the training set used by Ringrose *et al.* [47] to a moderate degree, in spite of not having been trained with the set, and our models in **Article II** to a high degree. Although the above question cannot readily be answered without functional testing of candidate PREs, the results of past prediction validation do suggest that

high-confidence predictions enriched in experimental signals are likely to be *bonafide* PREs.

The evolution of Polycomb/Trithorax Response Elements

The evolution of PREs among drosophilids has previously been investigated in terms of motif turnover and mobility over evolution [76]. For vertebrates, the generalization of sequence models trained on PcG target gene promoters [120] or H3K27me3-domains [42] has been measured between multiple species.

The models presented in **Article II** move beyond the confines of gene promoters, and incorporate more comprehensive experimental data than have previously been used for PcG/TrxG target site models, in multiple species, using a reinforcement learning regimen. This has enabled us to investigate genome-wide PcG target gene conservation based on PcG target site predictions, which, to our knowledge, has not previously been done. Our choice of model formulation further enabled us to probe the predictive features of a non-linear model in detail, and in turn gave us access to information about the conservation of the features that define PcG/TrxG target sites (according to the models), and potentially define PREs, across evolutionarily distant species.

Further investigation will be required in order to shed light on the biological significance of the predicted target gene conservation and conserved predictive sequence features. Few vertebrate PREs have yet been verified, and although our predictions provide a resource that can aid in the discovery of new vertebrate PREs, experimental functional testing will be required. As sufficient data becomes available to apply our or similar methods to more species, it will become possible to paint an ever clearer picture of the evolution of PcG target sites and PREs.

7.3 Future work

Time did not permit me to investigate all of the relevant ideas that I conceived of, and many of them, at the time of this writing, still remain open for future work.

First of all, there are a number of features that I would like to implement for MOCCA and Gnocis. Work during my master's thesis—following the work of [132]—showed promising results for the use of Particle Swarm Optimization (PSO) for hyper-parameter tuning, and I have been interested in the possibility of a derivative of SVM-MOCCA that identifies predictive motifs within a motif database by means of PSO. For Gnocis, implementing support for Bayesian modelling, such as Bayesian Convolutional Neural Networks, would be interesting.

I have also had ideas for a number of experiments. The combination of modelling and predicting PREs and boundary elements could enable more precise PcG/TrxG target gene predictions, and in turn yield a more accurate picture of the targets of the Polycomb/Trithorax system for the organism as a whole. Modelling the endpoints of chromatin loops containing PREs, such as from [23], might enable the prediction of both PREs and any non-PRE sequences that PREs physically interact with. Given the success of the methods developed in **Article II**, it would be very interesting to train a Convolutional Neural Network using a similar approach to PeCAN-BioRL. As sufficient experimental data becomes available in additional species, it would be very interesting to analyse the evolution of PcG/TrxG targeting in more detail.

List of Figures

List of Tables

List of Definitions

Notation overview

- $|S|$: Cardinality of a set S.
- $\oplus$: Positive label.
- $\ominus$: Negative label.
- $\mathbb{B} = \{\oplus, \ominus\}$: Binary label set.
- $f : A \to B$: Function mapping elements of A to elements of B.
- $\hat{c}(x)$: Approximation of function $c(x)$.
- TP, TN, FP, FN: True Positives, True Negatives, False Positives and False Negatives, respectively.
- $S_{a...b}$: Sub-sequence of a sequence S, from index a up to, and including, b.

Bibliography

[1] A. R. Migliaccio, "Erythroblast enucleation," *Haematologica*, vol. 95, no. 12, pp. 1985–1988, 2010.

[2] L. Ringrose and R. Paro, "Polycomb/Trithorax response elements and epigenetic memory of cell identity," *Development*, vol. 134, no. 2, pp. 223–232, 2007.

[3] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[4] R. R. Copley, "The animal in the genome: comparative genomics and evolution," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1496, pp. 1453–1461, 2008.

[5] M. Hulten, M. Stacey, and S. Armstrong, "Does junk DNA regulate gene expression in humans?," *Clinical molecular pathology*, vol. 48, no. 3, pp. M118–M123, 1995.

[6] M.-C. King and A. C. Wilson, "Evolution at two levels in humans and chimpanzees," *Science*, vol. 188, no. 4184, pp. 107–116, 1975.

[7] L. F. Franchini and K. S. Pollard, "Human evolution: the non-coding revolution," *BMC Biology*, vol. 15, no. 1, p. 89, 2017.

[8] S. A. Shabalina and N. A. Spiridonov, "The mammalian transcriptome and the function of non-coding DNA sequences," *Genome Biology*, vol. 5, no. 4, p. 105, 2004.

[9] P. J. Wittkopp and G. Kalay, "*Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence," *Nature Reviews Genetics*, vol. 13, no. 1, pp. 59–69, 2012.

[10] R. C. Hardison and J. Taylor, "Genomic approaches towards finding *cis*-regulatory modules in animals," *Nature Reviews Genetics*, vol. 13, no. 7, pp. 469–483, 2012.

[11] J. Simon, A. Chiang, W. Bender, M. J. Shimell, and M. O'Connor, "Elements of the *Drosophila* bithorax complex that mediate repression by Polycomb group products," *Developmental Biology*, vol. 158, no. 1, pp. 131–144, 1993.

[12] C.-S. Chan, L. Rastelli, and V. Pirrotta, "A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression.," *The EMBO Journal*, vol. 13, no. 11, pp. 2553–2564, 1994.

[13] D. Chetverina, T. Aoki, M. Erokhin, P. Georgiev, and P. Schedl, "Making connections: Insulators organize eukaryotic chromosomes into independent *cis*-regulatory networks," *Bioessays*, vol. 36, no. 2, pp. 163–172, 2014.

[14] T. Ali, R. Renkawitz, and M. Bartkuhn, "Insulators and domains of gene expression," *Current opinion in genetics & development*, vol. 37, pp. 17–26, 2016.

[15] P. D'haeseleer, "What are DNA sequence motifs?," *Nature Biotechnology*, vol. 24, no. 4, pp. 423–425, 2006.

[16] U. Ohler, "Identification of core promoter modules in *Drosophila* and their application in accurate transcription start site prediction," *Nucleic Acids Research*, vol. 34, no. 20, pp. 5943–5950, 2006.

[17] A. M. Bushey, E. Ramos, and V. G. Corces, "Three subclasses of a *Drosophila* insulator show distinct and cell type-specific genomic distributions," *Genes & Development*, vol. 23, no. 11, pp. 1338–1350, 2009.

[18] A. J. Bannister and T. Kouzarides, "Regulation of chromatin by histone modifications," *Cell research*, vol. 21, no. 3, pp. 381–395, 2011.

[19] M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, *et al.*, "Histone H3K27ac separates active from poised enhancers and predicts developmental state," *Proceedings of the National Academy of Sciences*, vol. 107, no. 50, pp. 21931–21936, 2010.

[20] H. Santos-Rosa, R. Schneider, A. J. Bannister, J. Sherriff, B. E. Bernstein, N. T. Emre, S. L. Schreiber, J. Mellor, and T. Kouzarides, "Active genes are tri-methylated at K4 of histone H3," *Nature*, vol. 419, no. 6905, pp. 407–411, 2002.

[21] R. Cao, L. Wang, H. Wang, L. Xia, H. Erdjument-Bromage, P. Tempst, R. S. Jones, and Y. Zhang, "Role of histone H3 lysine 27 methylation in Polycomb-group silencing," *Science*, vol. 298, no. 5595, pp. 1039–1043, 2002.

[22] T. Cheutin and G. Cavalli, "Polycomb silencing: from linear chromatin domains to 3D chromosome folding," *Current opinion in genetics & development*, vol. 25, pp. 30–37, 2014.

[23] K. P. Eagen, E. L. Aiden, and R. D. Kornberg, "Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map," *Proceedings of the National Academy of Sciences*, vol. 114, no. 33, pp. 8764–8769, 2017.

[24] F. Bantignies and G. Cavalli, "Polycomb group proteins: repression in 3D," *Trends in Genetics*, vol. 27, no. 11, pp. 454–464, 2011.

[25] B. Schuettengruber, H.-M. Bourbon, L. Di Croce, and G. Cavalli, "Genome regulation by Polycomb and Trithorax: 70 years and counting," *Cell*, vol. 171, no. 1, pp. 34–57, 2017.

[26] E. B. Lewis, "A gene complex controlling segmentation in *Drosophila*," in *Genes, Development and Cancer*, pp. 205–217, Springer, 1978.

[27] M. Mallo, D. M. Wellik, and J. Deschamps, "*Hox* genes and regional patterning of the vertebrate body plan," *Developmental Biology*, vol. 344, no. 1, pp. 7–15, 2010.

[28] J. A. Kassis, J. A. Kennison, and J. W. Tamkun, "Polycomb and Trithorax group genes in *Drosophila*," *Genetics*, vol. 206, no. 4, pp. 1699–1725, 2017.

[29] V. Chinwalla, E. P. Jane, and P. Harte, "The *Drosophila* trithorax protein binds to specific chromosomal sites and is co-localized with Polycomb at many sites.," *The EMBO Journal*, vol. 14, no. 9, pp. 2056–2065, 1995.

[30] T. Klymenko and J. Müller, "The histone methyltransferases Trithorax and Ash1 prevent transcriptional silencing by Polycomb group proteins," *EMBO Reports*, vol. 5, no. 4, pp. 373–377, 2004.

[31] B. Schuettengruber, D. Chourrout, M. Vervoort, B. Leblanc, and G. Cavalli, "Genome regulation by Polycomb and Trithorax proteins," *Cell*, vol. 128, no. 4, pp. 735–745, 2007.

[32] T. G. Kahn, P. Stenberg, V. Pirrotta, and Y. B. Schwartz, "Combinatorial interactions are required for the efficient recruitment of pho repressive complex (PhoRC) to Polycomb Response Elements," *PLoS Genetics*, vol. 10, no. 7, p. e1004495, 2014.

[33] F. Tie, R. Banerjee, A. R. Saiakhova, B. Howard, K. E. Monteith, P. C. Scacheri, M. S. Cosgrove, and P. J. Harte, "Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing," *Development*, vol. 141, no. 5, pp. 1129–1139, 2014.

[34] R. Rickels, D. Hu, C. K. Collings, A. R. Woodfin, A. Piunti, M. Mohan, H.-M. Herz, E. Kvon, and A. Shilatifard, "An evolutionary conserved epigenetic mark of Polycomb Response Elements implemented by Trx/MLL/COMPASS," *Molecular Cell*, vol. 63, no. 2, pp. 318–328, 2016.

[35] M. Entrevan, B. Schuettengruber, and G. Cavalli, "Regulation of genome architecture and function by Polycomb proteins," *Trends in Cell Biology*, vol. 26, no. 7, pp. 511–525, 2016.

[36] S. J. Geisler and R. Paro, "Trithorax and Polycomb group-dependent regulation: a tale of opposing activities," *Development*, vol. 142, no. 17, pp. 2876–2887, 2015.

[37] A. Piunti and A. Shilatifard, "Epigenetic balance of gene expression by Polycomb and COMPASS families," *Science*, vol. 352, no. 6290, p. aad9780, 2016.

[38] B. Czermin, R. Melfi, D. McCabe, V. Seitz, A. Imhof, and V. Pirrotta, "*Drosophila* enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites," *Cell*, vol. 111, no. 2, pp. 185–196, 2002.

[39] R. Margueron, N. Justin, K. Ohno, M. L. Sharpe, J. Son, W. J. Drury Iii, P. Voigt, S. R. Martin, W. R. Taylor, V. De Marco, *et al.*, "Role of the polycomb protein EED in the propagation of repressive histone marks," *Nature*, vol. 461, no. 7265, pp. 762–767, 2009.

[40] J. A. Simon and R. E. Kingston, "Mechanisms of Polycomb gene silencing: knowns and unknowns," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 10, pp. 697–708, 2009.

[41] P. A. Steffen and L. Ringrose, "What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 5, pp. 340–356, 2014.

[42] S. J. van Heeringen, R. C. Akkers, I. van Kruijsbergen, M. A. Arif, L. L. Hanssen, N. Sharifi, and G. J. C. Veenstra, "Principles of nucleation of H3K27 methylation

during embryonic development," *Genome Research*, vol. 24, no. 3, pp. 401–410, 2014.

[43] A. Sing, D. Pannell, A. Karaiskakis, K. Sturgeon, M. Djabali, J. Ellis, H. D. Lipshitz, and S. P. Cordes, "A vertebrate Polycomb Response Element governs segmentation of the posterior hindbrain," *Cell*, vol. 138, no. 5, pp. 885–897, 2009.

[44] M. Bauer, J. Trupke, and L. Ringrose, "The quest for mammalian Polycomb response elements: are we there yet?," *Chromosoma*, vol. 125, no. 3, pp. 471–496, 2016.

[45] C. J. Woo, P. V. Kharchenko, L. Daheron, P. J. Park, and R. E. Kingston, "A region of the human HOXD cluster that confers Polycomb-group responsiveness," *Cell*, vol. 140, no. 1, pp. 99–110, 2010.

[46] C. J. Woo, P. V. Kharchenko, L. Daheron, P. J. Park, and R. E. Kingston, "Variable requirements for DNA-binding proteins at Polycomb-dependent repressive regions in human HOX clusters," *Molecular and Cellular Biology*, vol. 33, no. 16, pp. 3274–3285, 2013.

[47] L. Ringrose, M. Rehmsmeier, J.-M. Dura, and R. Paro, "Genome-wide prediction of Polycomb/Trithorax Response Elements in *Drosophila melanogaster*," *Developmental Cell*, vol. 5, no. 5, pp. 759–771, 2003.

[48] J. L. Brown and J. A. Kassis, "Architectural and functional diversity of Polycomb group response elements in *Drosophila*," *Genetics*, vol. 195, no. 2, pp. 407–419, 2013.

[49] P. Ray, S. De, A. Mitra, K. Bezstarosti, J. A. Demmers, K. Pfeifer, and J. A. Kassis, "Combgap contributes to recruitment of Polycomb group proteins in *Drosophila*," *Proceedings of the National Academy of Sciences*, vol. 113, no. 14, pp. 3826–3831, 2016.

[50] J. W. Nicol, G. A. Helt, S. G. Blanchard Jr, A. Raja, and A. E. Loraine, "The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets," *Bioinformatics*, vol. 25, no. 20, pp. 2730–2731, 2009.

[51] S. E. Celniker, L. A. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, *et al.*, "Unlocking the secrets of the genome," *Nature*, vol. 459, no. 7249, pp. 927–930, 2009.

[52] J. Erceg, T. Pakozdi, R. Marco-Ferreres, Y. Ghavi-Helm, C. Girardot, A. P. Bracken, and E. E. Furlong, "Dual functionality of *cis*-regulatory elements as developmental enhancers and Polycomb response elements," *Genes & Development*, vol. 31, no. 6, pp. 590–602, 2017.

[53] R. A. Hoskins, J. W. Carlson, C. Kennedy, D. Acevedo, M. Evans-Holm, E. Frise, K. H. Wan, S. Park, M. Mendez-Lago, F. Rossi, *et al.*, "Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin," *Science*, vol. 316, no. 5831, pp. 1625–1628, 2007.

[54] J. Zeng, B. D. Kirk, Y. Gou, Q. Wang, and J. Ma, "Genome-wide Polycomb target gene prediction in *Drosophila melanogaster*," *Nucleic Acids Research*, vol. 40, no. 13, pp. 5848–5863, 2012.

[55] R. S. Illingworth and A. P. Bird, "CpG islands–'A rough guide'," *FEBS letters*, vol. 583, no. 11, pp. 1713–1720, 2009.

[56] Y. Li, H. Zheng, Q. Wang, C. Zhou, L. Wei, X. Liu, W. Zhang, Y. Zhang, Z. Du, X. Wang, *et al.*, "Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys," *Genome Biology*, vol. 19, no. 1, p. 18, 2018.

[57] B. Hekimoglu and L. Ringrose, "Non-coding RNAs in Polycomb/Trithorax regulation," *RNA Biology*, vol. 6, no. 2, pp. 129–137, 2009.

[58] B. Hekimoglu-Balkan, A. Aszodi, R. Heinen, M. Jaritz, and L. Ringrose, "Intergenic Polycomb target sites are dynamically marked by non-coding transcription during lineage commitment," *RNA Biology*, vol. 9, no. 3, pp. 314–325, 2012.

[59] L. Ringrose, "Noncoding RNAs in Polycomb and Trithorax regulation: a quantitative perspective," *Annual Review of Genetics*, vol. 51, pp. 385–411, 2017.

[60] G. V. Glazko, B. L. Zybailov, and I. B. Rogozin, "Computational prediction of Polycomb-associated long non-coding RNAs," *PLoS One*, vol. 7, no. 9, p. e44878, 2012.

[61] J. Müller and M. Bienz, "Long range repression conferring boundaries of Ultrabithorax expression in the *Drosophila* embryo.," *The EMBO Journal*, vol. 10, no. 11, pp. 3147–3155, 1991.

[62] V. Pirrotta, "Vectors for P-mediated transformation in *Drosophila*," in *Vectors*, pp. 437–456, Elsevier, 1988.

[63] C. E. Horak and M. Snyder, "ChIP-chip: A genomic approach for identifying transcription factor binding sites," *Methods in enzymology*, vol. 350, pp. 469–483, 2002.

[64] E. R. Mardis, "ChIP-seq: welcome to the new frontier," *Nature Methods*, vol. 4, no. 8, pp. 613–614, 2007.

[65] B. Tolhuis, I. Muijrers, E. de Wit, H. Teunissen, W. Talhout, B. van Steensel, and M. van Lohuizen, "Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in *Drosophila melanogaster*," *Nature Genetics*, vol. 38, no. 6, pp. 694–699, 2006.

[66] Y. B. Schwartz, T. G. Kahn, D. A. Nix, X.-Y. Li, R. Bourgon, M. Biggin, and V. Pirrotta, "Genome-wide analysis of Polycomb targets in *Drosophila melanogaster*," *Nature Genetics*, vol. 38, no. 6, pp. 700–705, 2006.

[67] K. Oktaba, L. Guitiérrez, J. Gagneur, C. Girardot, A. K. Sengupta, E. E. Furlong, and M. Jürg, "Dynamic regulation by Polycomb group protein complexes controls pattern formation and the cell cycle in *Drosophila*," *Developmental Cell*, vol. 15, no. 6, pp. 877–889, 2008.

[68] B. Schuettengruber, M. Ganapathi, B. Leblanc, M. Portoso, R. Jaschek, B. Tolhuis, M. v. Lohuizen, A. Tanay, and G. Cavalli, "Functional anatomy of Polycomb and Trithorax chromatin landscapes in *Drosophila* embryos," *PLoS Biology*, vol. 7, no. 1, p. e1000013, 2009.

[69] Y. B. Schwartz, T. G. Kahn, P. Stenberg, K. Ohno, R. Bourgon, and V. Pirrotta, "Alternative epigenetic chromatin states of Polycomb target genes," *PLoS Genetics*, vol. 6, no. 1, p. e1000805, 2010.

[70] D. Enderle, C. Beisel, M. B. Stadler, M. Gerstund, P. Athri, and R. Paro, "Polycomb preferentially targets stalled promoters of coding and noncoding transcripts," *Genome Research*, vol. 21, no. 2, pp. 216–226, 2011.

[71] L. A. Boyer, K. Plath, J. Zeitlinger, T. Brambrink, L. A. Medeiros, T. I. Lee, S. S. Levine, M. Wernig, A. Tajonar, M. K. Ray, *et al.*, "Polycomb complexes repress developmental regulators in murine embryonic stem cells," *Nature*, vol. 441, no. 7091, pp. 349–353, 2006.

[72] X. Shen, Y. Liu, Y.-J. Hsu, Y. Fujiwara, J. Kim, X. Mao, G.-C. Yuan, and S. H. Orkin, "EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency," *Molecular Cell*, vol. 32, no. 4, pp. 491–502, 2008.

[73] M. Ku, R. P. Koche, E. Rheinbay, E. M. Mendenhall, M. Endoh, T. S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie, A. S. Chi, *et al.*, "Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains," *PLoS Genetics*, vol. 4, no. 10, p. e1000242, 2008.

[74] E. P. Consortium *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[75] T. Fiedler and M. Rehmsmeier, "jPREdictor: a versatile tool for the prediction of *cis*-regulatory elements," *Nucleic Acids Research*, vol. 34, no. suppl_2, pp. W546–W550, 2006.

[76] A. Hauenschild, L. Ringrose, C. Altmutter, R. Paro, and M. Rehmsmeier, "Evolutionary plasticity of Polycomb/Trithorax Response Elements in *Drosophila* species," *PLoS Biology*, vol. 6, no. 10, p. e261, 2008.

[77] H. Richly, L. Aloia, and L. Di Croce, "Roles of the Polycomb group proteins in stem cells and cancer," *Cell Death & Disease*, vol. 2, no. 9, p. e204, 2011.

[78] K. Klein and S. Gay, "Epigenetic modifications in rheumatoid arthritis, a review," *Current Opinion in Pharmacology*, vol. 13, no. 3, pp. 420–425, 2013.

[79] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997.

[80] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[81] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5907–5915, 2017.

[82] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[83] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[84] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, pp. 396–404, 1990.

[85] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[86] G. Kerr, H. J. Ruskin, M. Crane, and P. Doolan, "Techniques for clustering gene expression data," *Computers in Biology and Medicine*, vol. 38, no. 3, pp. 283–293, 2008.

[87] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[88] S. D. Whitehead and D. H. Ballard, "Learning to perceive and act by trial and error," *Machine Learning*, vol. 7, no. 1, pp. 45–83, 1991.

[89] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM Computing Surveys*, vol. 27, no. 3, pp. 326–327, 1995.

[90] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[91] T. M. Mitchell, *Machine Learning*. McGraw-Hill Book Co, 1997.

[92] A. Srinivasan and R. K. Mishra, "Chromatin domain boundary element search tool for *Drosophila*," *Nucleic Acids Research*, vol. 40, no. 10, pp. 4385–4395, 2012.

[93] CBN, "IUPAC-IUB commission on biochemical nomenclature (CBN). abbreviations and symbols for nucleic acids, polynucleotides and their constituents. recommendations 1970.," *The Biochemical journal*, vol. 120, no. 3, pp. 449–454, 1970.

[94] A. Mathelier and W. W. Wasserman, "The next generation of transcription factor binding site prediction," *PLoS Computational Biology*, vol. 9, no. 9, p. e1003214, 2013.

[95] L. Yang, T. Zhou, I. Dror, A. Mathelier, W. W. Wasserman, R. Gordân, and R. Rohs, "TFBSshape: a motif database for DNA shape features of transcription factor binding sites," *Nucleic Acids Research*, vol. 42, no. D1, pp. D148–D155, 2013.

[96] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, no. suppl_2, pp. W202–W208, 2009.

[97] T. L. Bailey, "DREME: motif discovery in transcription factor ChIP-seq data," *Bioinformatics*, vol. 27, no. 12, pp. 1653–1659, 2011.

[98] P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," *Bioinformatics*, vol. 27, no. 12, pp. 1696–1697, 2011.

[99] S. J. van Heeringen and G. J. C. Veenstra, "GimmeMotifs: a *de novo* motif prediction pipeline for ChIP-sequencing experiments," *Bioinformatics*, vol. 27, no. 2, pp. 270–271, 2010.

[100] D. Simcha, N. D. Price, and D. Geman, "The limits of *de novo* DNA motif discovery," *PLoS One*, vol. 7, no. 11, p. e47836, 2012.

[101] L. J. Zhu, R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. A. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe, *et al.*, "FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system," *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D111–D117, 2010.

[102] I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, A. D. Fedorova, E. I. Rumynskiy, Y. A. Medvedeva, A. Magana-Mora, V. B. Bajic, D. A. Papatsenko, *et al.*, "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis," *Nucleic Acids Research*, vol. 46, no. D1, pp. D252–D259, 2017.

[103] A. Jolma, Y. Yin, K. R. Nitta, K. Dave, A. Popov, M. Taipale, M. Enge, T. Kivioja, E. Morgunova, and J. Taipale, "DNA-dependent formation of transcription factor pairs alters their binding specificity," *Nature*, vol. 527, no. 7578, pp. 384–388, 2015.

[104] C. Leslie, E. Eskin, and W. S. Noble, "The spectrum kernel: A string kernel for SVM protein classification," in *Biocomputing 2002*, pp. 564–575, World Scientific, 2001.

[105] E. Eskin, J. Weston, W. S. Noble, and C. S. Leslie, "Mismatch string kernels for SVM protein classification," in *Advances in Neural Information Processing Systems*, pp. 1441–1448, 2003.

[106] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, vol. 20, no. 4, pp. 467–476, 2004.

[107] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support vector machines and kernels for computational biology," *PLoS Computational Biology*, vol. 4, no. 10, p. e1000173, 2008.

[108] Y. Gan, J. Guan, S. Zhou, and W. Zhang, "Structural features based genome-wide characterization and prediction of nucleosome organization," *BMC Bioinformatics*, vol. 13, no. 1, p. 49, 2012.

[109] H. Yu, "Single-class classification with mapping convergence," *Machine Learning*, vol. 61, no. 1-3, pp. 49–69, 2005.

[110] B. A. Bredesen and M. Rehmsmeier, "DNA sequence models of genome-wide *Drosophila* melanogaster Polycomb binding sites improve generalization to independent Polycomb Response Elements," *Nucleic Acids Res.*, vol. 47, no. 15, pp. 7781–7797, 2019.

[111] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[112] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

[113] G. Hripcsak and A. S. Rothschild, "Agreement, the F-measure and reliability in information retrieval," *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296–298, 2005.

[114] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[115] T. Saito and M. Rehmsmeier, "Precrec: fast and accurate precision–recall and ROC curve calculations in R," *Bioinformatics*, vol. 33, no. 1, pp. 145–147, 2017.

[116] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[117] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS One*, vol. 10, no. 3, p. e0118432, 2015.

[118] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th international joint conference on Artificial intelligence*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

[119] N. Postika, M. Metzler, M. Affolter, M. Müller, P. Schedl, P. Georgiev, and O. Kyrchanova, "Boundaries mediate long-distance interactions between en-

hancers and promoters in the *Drosophila* Bithorax complex," *PLoS Genetics*, vol. 14, no. 12, p. e1007702, 2018.

[120] Y. Liu, Z. Shao, and G.-C. Yuan, "Prediction of Polycomb target genes in mouse embryonic stem cells," *Genomics*, vol. 96, no. 1, pp. 17–26, 2010.

[121] X. Xiao, Z. Li, H. Liu, J. Su, F. Want, X. Wu, H. Liu, Q. Wu, and Y. Zhang, "Genome-wide identification of Polycomb target genes in human embryonic stem cells," *Gene*, vol. 518, no. 2, pp. 425–430, 2013.

[122] I. Yevshin, R. Sharipov, T. Valeev, A. Kel, and F. Kolpakov, "GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments," *Nucleic Acids Research*, vol. 45, no. D1, pp. D61–D67, 2016.

[123] I. Yevshin, R. Sharipov, S. Kolmykov, Y. Kondrakhin, and F. Kolpakov, "GTRD: a database on gene transcription regulation — 2019 update," *Nucleic Acids Research*, vol. 47, no. D1, pp. D100–D105, 2018.

[124] C. A. Davis, B. C. Hitz, C. A. Sloan, E. T. Chan, J. M. Davidson, I. Gabdank, J. A. Hilton, K. Jain, U. K. Baymuradov, A. K. Narayanan, *et al.*, "The encyclopedia of DNA elements (ENCODE): data portal update," *Nucleic Acids Research*, vol. 46, no. D1, pp. D794–D801, 2017.

[125] S. Cass, "The 2017 top programming languages," *IEEE Spectrum*, vol. 18, 2017.

[126] W. McKinney *et al.*, "pandas: a foundational Python library for data analysis and statistics," *Python for High Performance and Scientific Computing*, vol. 14, no. 9, 2011.

[127] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[128] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[129] F. Pérez and B. E. Granger, "IPython: a system for interactive scientific computing," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 21–29, 2007.

[130] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in science & engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[131] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, *et al.*, "Jupyter Notebooks—a publishing format for reproducible computational workflows.," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87–90, 2016.

[132] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines,"

Expert Systems with Applications, vol. 35, no. 4, pp. 1817–1824, 2008.

Part III

Scientific contributions

Published online 24 July 2019

Nucleic Acids Research, 2019, Vol. 47, No. 15 ***7781–7797***
doi: 10.1093/nar/gkz617

DNA sequence models of genome-wide *Drosophila melanogaster* Polycomb binding sites improve generalization to independent Polycomb Response Elements

Bjørn André Bredesen[1] and Marc Rehmsmeier[1,2,*]

[1]Computational Biology Unit, Department of Informatics, University of Bergen, P.O. Box 7803, N-5020 Bergen, Norway and [2]Integrated Research Institute (IRI) for the Life Sciences and Department of Biology, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Received January 25, 2019; Revised July 01, 2019; Editorial Decision July 07, 2019; Accepted July 11, 2019

ABSTRACT

Polycomb Response Elements (PREs) are *cis*-regulatory DNA elements that maintain gene transcription states through DNA replication and mitosis. PREs have little sequence similarity, but are enriched in a number of sequence motifs. Previous methods for modelling *Drosophila melanogaster* PRE sequences (PREdictor and EpiPredictor) have used a set of 7 motifs and a training set of 12 PREs and 16-23 non-PREs. Advances in experimental methods for mapping chromatin binding factors and modifications has led to the publication of several genome-wide sets of Polycomb targets. In addition to the seven motifs previously used, PREs are enriched in the GTGT motif, recently associated with the sequence-specific DNA binding protein Combgap. We investigated whether models trained on genome-wide Polycomb sites generalize to independent PREs when trained with control sequences generated by naive PRE models and including the GTGT motif. We also developed a new PRE predictor: SVM-MOCCA. Training PRE predictors with genome-wide experimental data improves generalization to independent data, and SVM-MOCCA predicts the majority of PREs in three independent experimental sets. We present 2908 candidate PREs enriched in sequence and chromatin signatures. 2412 of these are also enriched in H3K4me1, a mark of Trithorax activated chromatin, suggesting that PREs/TREs have a common sequence code.

INTRODUCTION

The body plan of the fruit fly, *Drosophila melanogaster*, is genetically determined by transcription factors whose expression patterns are carefully coordinated and localized (1). Some transcription factors are produced early in development, where they gather at initiation elements in DNA that in turn establish the expression states of developmentally important genes (1). Later in development, these initiating factors deteriorate, and a memory of gene transcription states must be maintained (2,3).

Polycomb Response Elements (PREs) are cellular memory elements in DNA that maintain a memory of transcription states of their target genes over cell division (4,5). To accomplish this, PREs recruit Polycomb group (PcG) proteins, which maintain repression, and Trithorax group (TrxG) proteins, which antagonize PcG repression (6,7) (see Materials and Methods for a discussion of response elements nomenclature). PcG proteins were first identified as *Hox* gene regulators in *Drosophila melanogaster*, where PcG mutant flies exhibit ectopic *Hox* gene expression along the anterior-posterior axis (3,8). It has since been discovered that PcG proteins target a much wider range of genes (9–12) and that PcG proteins have mammalian homologs, with important roles in development and with implications in cancer (13,14).

The Polycomb system is best characterized in *Drosophila melanogaster*, where tens of PREs have been experimentally verified (1,15–17) and tens of PcG/TrxG proteins have been identified (13,14). *Drosophila* PREs are several hundred base pairs long, with little sequence homology between them (1). Nonetheless, they are enriched in the binding motifs for several DNA binding factors. PcG proteins in *D. melanogaster* include Pc (Polycomb), Psc (Posterior sex

*To whom correspondence should be addressed. Tel: +49 30 2093 49767; Fax: +49 30 2093 49771; Email: marc.rehmsmeier@hu-berlin.de

combs), Pho (Pleiohomeotic) and Sfmbt (Scm-related gene containing four mbt domains) (13). Pho is the only PcG protein known to bind DNA with sequence specificity (18). PcG proteins form three major complexes on chromatin: Polycomb Repressive Complex 1 (PRC1) (19), Polycomb Repressive Complex 2 (PRC2) (20–23) and Pleiohomeotic Repressive Complex (PhoRC) (24). Polycomb repressed chromatin is marked by histone 3 lysine 27 trimethylation (H3K27me3) (20–23). Trithorax activated chromatin is marked by histone 3 lysine 4 monomethylation (H3K4me1) (25) or dimethylation (H3K4me2) (26).

Drosophila PREs were originally discovered by testing segments of DNA for their ability to maintain previously established transcription states when taken out of their endogenous context (4,5). In 2003, Ringrose *et al.* published a computational method to model PRE sequences, named the PREdictor, which predicted 167 candidate PREs genome-wide in *D. melanogaster* for one expected false-positive prediction (9). The PREdictor scores sequence windows by a linear combination of motif pair occurrence frequencies, weighted by log-odds of occurrence frequencies in a training set of PREs and non-PREs. Ringrose *et al.* trained the PREdictor on a set of 12 PREs (11 PREs from *D. melanogaster* and 1 from *D. virilis*) and 16 non-PREs (promoters that are enriched in PRE sequence motifs but that do not recruit Polycomb), together with a set of seven motifs. Six of these motifs correspond to DNA binding factors (two for GAGA binding factor, three for Pleiohomeotic, one for Zeste), and one is a motif that was identified by conservation between *D. melanogaster* and *D. virilis* in the engrailed PRE and whose deletion abrogates silencing function (EN1) (27). The authors found that paired motif occurrence frequencies can distinguish PREs from non-PREs, whereas single motif occurrence frequencies cannot. This suggests that the sequence criteria for recruiting Polycomb are of a combinatorial nature and that DNA binding factors cooperate on PREs to recruit Polycomb regulatory complexes. Furthermore, Ringrose *et al.* identified several new candidate PRE sequence motifs, including the GTGT motif. Since then, the GTGT motif has been shown to be essential for silencing in the *vg* PRE (28), and it has been shown to be bound by the sequence-specific DNA binding protein Combgap, which is involved in PcG recruitment (29). The GTGT motif has also been rediscovered as the CACA motif in a ChIP-on-chip study of genome-wide binding profiles of PcG and other proteins (30). The PREdictor (9) method was later extended to the jPREdictor (31), a re-implementation in Java, providing a graphical user interface and offering the ability to flexibly define motifs and their combinations.

In 2012, Zeng *et al.* published the EpiPredictor (32), a PRE predictor that uses the machine learning method of Support Vector Machines (SVMs). Support Vector Machines model feature space class boundaries by placing a decision surface between the points of two classes such that the margin to the closest points is maximized, with room for treating data points as noise by use of a soft margin, and with the possibility of non-linear modelling by use of kernel functions (33). The EpiPredictor filters sequence windows using the SVM and a GC-content filter and scores them based on the total number of motif occurrences they contain. The SVM feature space consists of single motif occurrence frequencies. The EpiPredictor was trained on the same set of PREs and with the same motifs as used by Ringrose *et al.* (9). Zeng *et al.* (32) found that non-linear kernels distinguish PREs from non-PREs better than linear kernels, adding further evidence of the importance of motif occurrence combinatorics for PRE sequences.

Recent advances in experimental methods have led to the publication of several sets of candidate PREs genome-wide in *Drosophila* (10,11,30,34–38). These methods include chromatin immunoprecipitation (ChIP) combined with microarray (ChIP-chip) (39), ChIP combined with high-throughput sequencing (ChIP-seq) (40) and DNA adenine methyltransferase identification (DamID) (37).

The published candidate PRE sets vary in the number and identity of candidate PREs they contain (1,12). Several factors may underlie these discrepancies, such as differences in experimental methods (ChIP-chip versus ChIP-seq) or differences in antibodies used. The results of experimental mapping methods also depend on the cells being studied and on their genetic states. Furthermore, PREs physically interact with other genomic loci, forming higher-order structures (41). Experimental mapping methods do not discriminate between recruiting and interacting sites and can as a result capture regions that PREs interact with, in addition to the PREs themselves (1). *In silico* PRE prediction methods have no such limitations and can help us to understand the sequence criteria for what constitutes a PRE.

Sequences that recruit PcG proteins in other organisms are also being studied, though few mammalian PREs have so far been identified (15). PcG recruitment has been modelled in human embryonic stem cells using Support Vector Machines (42). In the frog *Xenopus tropicalis*, Support Vector Machines were able to identify a *k*-mer spectrum that characterizes H3K27me3 nucleation sites that are not CpG islands and that work as repressive elements when taken out of their endogenous context (43). Du *et al.* (44) reported three classes of response elements in human: Polycomb Response Elements (PREs), Trithorax Response Elements (TREs) and Polycomb/Trithorax Response Elements (P/TREs).

Previous publications on modelling *Drosophila* PREs have used small sets of experimentally tested PREs for training the models. The resulting genome-wide predictions have limited overlaps with genome-wide experimentally determined PcG-recruiting chromatin regions. Furthermore, the GTGT motif has not previously been included in *Drosophila* PRE sequence models. We here seek to refine the state of the art in DNA sequence models of *Drosophila* Polycomb Response Elements by investigating whether the training of sequence models on genome-wide experimentally determined PcG-recruiting DNA and including the GTGT motif increases the agreement between *in silico* PRE predictions and independent experimentally determined genome-wide sets of PcG target regions. We further address the question whether a more advanced modelling approach can additionally improve model generalization and present a new method for modelling *cis*-regulatory elements, SVM-MOCCA.

MATERIALS AND METHODS

Nomenclature of response elements

The nomenclature of response elements is evolving. Chang *et al.* (45) identified a 440-bp fragment in the *postbithorax/bithoraxoid* region of *Ultrabithorax* that contains both a PRE (Polycomb Response Element) and a TRE (Trithorax Response Element). Tillib *et al.* (46) distinguish TREs and PREs as discrete sequences in a TRE-PRE module. The closeness of PREs and TREs is described by (47) as an 'intermingling of elements', and the authors propose that PREs/TREs acquire the new name 'maintenance elements', to reflect their dual function. Bloyer *et al.* (48) conclude that (then) recent data strongly suggests that 'each PRE/TRE is composed of multiple different *cis*-DNA modules, which can be bound by different subsets of PC-G and TRX-G at defined spatial and temporal positions in the embryo'. While some authors consistently use the term PRE/TRE (1), emphasizing the dual nature of these maintenance elements, others primarily use the term PRE and conclude from experimental data that 'PREs are also TREs' (34). Enderle *et al.* (35) present a set of 'PcG binding sites' that is not only defined on the basis of proteins from the Polycomb group, but also on TRX-C, and also use the term PRE. Kahn *et al.* (36) also use the term PRE for regions defined from overlapping peaks of E(Z), TRX and PC and coinciding with H3K27me3 domains. It thus appears that more recently, the term PRE is universally used for PcG target sites that can also be TrxG target sites and have potential to be both Polycomb and Trithorax Response Elements (with the caveat that the response function of these sites has not been tested). In accordance with this, we primarily use the term PRE (Polycomb Response Element), but mean it to encompass such elements' potential function as TREs.

Genome assembly

We used the *D. melanogaster* genome assembly release 6 (2014) (49,50). All published genomic coordinates that we considered that were for a previous genome assembly were converted to release 6 using the FlyBase (51) coordinate converter.

DNA sequence motifs

We used motifs defined in IUPAC notation (52), as used or reported in Ringrose *et al.* (9): EN 1: `GSNMACGCCCC` (one mismatch allowed), G10: `GAGAGAGAGA` (one mismatch allowed), GAF: `GAGAG`, PF: `GCCATHWY`, PM: `CNGCCATNDN ND`, PS: `GCCAT`, Z: `YGAGYG`, GTGT: `GTGT`. Throughout the manuscript, when comparing classifiers with and without the GTGT motif, those with have been marked 'w. GTGT'. SVM-MOCCA always makes use of this motif and has not been marked explicitly.

For comparison experiments, we also used the following motifs, reported in (53): one additional motif for Zeste: `BGAGTGV`, one for Sp1/KLF: `RRGGYG`, one for Dsp1: `GAAAA`, two for Grainyhead: `TGTTTTTT` and `WCHGGTT`, and one for 'site A': `GAACNG`.

To investigate how the addition of GTGT to a PRE model compares to adding a random 4-mer, we randomly generated 19 unique 4-mers (unique also when considering reverse complements).

Sequence-generating nth-order Markov chains

For every n-mer s (a DNA sequence of length n), we obtained the probability of observing each nucleotide $q \in \{\texttt{A}, \texttt{T}, \texttt{G}, \texttt{C}\}$ next as the fraction of times we observe q after s versus the total number of observations of s. To account for double-strandedness, we also obtained n-mer frequencies on the reverse complement of each sequence. We added a pseudocount of 1 for each nucleotide for each n-mer to ensure none had zero observations. To generate a sequence, we randomly picked an n-mer with the probability of observing this n-mer, and generated each subsequent nucleotide based on the nucleotide probability distribution for the last generated n-mer.

Training and validation sequences

We acquired the training set used by Ringrose *et al.* (9), consisting of 12 PREs and 16 non-PREs, henceforth referred to as the T2003 training set.

Additionally, we acquired genome-wide candidate PcG target sites determined by Schwartz *et al.* (34), Enderle *et al.* (35) and Kahn *et al.* (36). We considered including data from Schuettengruber *et al.* (30), but as they did not publish candidate PRE coordinates and we already consider three more recently published PRE sets, we opted not to include their data in our analysis. For the Schwartz *et al.* (34) set, computationally defined PREs were downloaded from the article's Supplementary Table S6, and coordinates were converted from *D. melanogaster* genome assembly 4 to assembly 6. PcG target regions from the Enderle *et al.* (35) set were acquired from the article's Supplementary Table 3 and converted from *D. melanogaster* genome assembly 5 to assembly 6. The Kahn *et al.* (36) set of computationally defined PREs was extracted from the article's Supplementary Table S1 and converted from genome assembly 5 to assembly 6. All coordinate conversions between genome assemblies were performed using the FlyBase (51) coordinate converter. Only regions localized on chromosomes 2L, 2R, 3L, 3R, 4 and X were considered. Heterochromatic regions ('Het' chromosomes in the FlyBase annotation) were discarded. After coordinate conversions, in order to account for any distancing between recruited factors and recruiting sequences, all regions were resized to a length of 3 kb each (1.5 kb bidirectionally from each region center), and corresponding sequences were extracted from the assembly 6 genome.

We generated three sets of negative control sequences for training and testing: (a) For each PcG target region set, we generated a set of one hundred times as many 3 kb-long random sequences, using a fourth-order Markov chain trained on the respective set, henceforth referred to as dummy PREs. (b) A fourth-order Markov chain was trained on the *D. melanogaster* genome and used to generate a set of a hundred times as many 3 kb-long random sequences as in the largest Polycomb target set (20 100 sequences in total), henceforth referred to as dummy genomic sequences. Dummy sequences mirror average 5-mer distributions of

their set of origin, but are unlikely to retain any higher-order structure such as motif pairing or clustering. (c) Finally, we acquired coding sequences from the FlyBase (51) r6.04 annotation. In order to get a set of uniformly sized coding sequences for training and testing, we concatenated the coding sequences and split the resulting sequence into non-overlapping 3 kb-long fragments, henceforth referred to as coding sequences. Additionally, in order to have a coding sequence region set to check for genomic overlaps with predictions, unlikely to contain gene-proximal PREs, we defined core coding sequences as annotated coding sequences shrunk bi-directionally by 250 bp, with regions too small to shrink omitted.

We refer to training sets consisting of PREs from a genome-wide experimental set and corresponding dummy PREs by the name T2017. For the main figures, T2017 refers to the Schwartz *et al.* (34) set of PREs and of corresponding dummy PREs as controls. For supplementary figures where we train models on the Enderle *et al.* (35) and Kahn *et al.* (36) sets, the meaning of T2017 is modified to refer to the specified PRE set and corresponding dummy PREs.

Cross-validation

To account for random variation in generalization performance, we cross-validated with 50 repetitions, resulting in 50 sets of independent training and test sequences. Over cross-validation, each sequence set was randomly shuffled, and the first 110 sequences were reserved for training. Of the remainder, the first 50 PRE sequences and 5000 non-PRE sequences of each set were used for testing. This 100:1 ratio of controls to PREs reflects the expected genome-wide context, based on the assumption that the 140 Mb-long *D. melanogaster* genome contains 1400 1 kb-long PREs. Note that the precise number is neither known nor necessary to be known for this analysis, since any number between a few hundred and a few thousand PREs in the *Drosophila* genome will be reflected accurately enough in the performance evaluations.

Classifier performance evaluation

When testing model generalization, we applied our models using a sliding window across all test sequences, where the maximum window score was taken as the final test sequence score. When visualizing model generalization, we focused on Precision/Recall curves (PRCs), which plot Precision = TP/(TP + FP) in the Y-axis and Recall = TP/(TP + FN) in the X-axis. TP denotes the number of true positives, FP the number of false positives and FN the number of false negatives. PRCs, unlike ROC (Receiver Operating Characteristics) curves, are informative of generalization performance on highly imbalanced datasets, such as genome-wide predictions, where the number of positives is small compared to the number of negatives (54). The area under the Precision/Recall curve (PRC AUC) gives a threshold-independent measure of expected classifier generalization. Note that, as a consequence, PRC AUC does not refer to any particular number of predictions nor to any particular number of true and false positives. Rather, such numbers correspond to a point on the Precision/Recall curve. Depending on requirements, e.g. with respect to an expected precision, a score cutoff can be chosen which will then determine specific numbers such as the number of predictions and true and false positives. We use the mean PRC AUC over cross-validation, with 95% confidence intervals calculated based on normally distributed means.

CPREdictor

We have reimplemented the PREdictor (9) algorithm in C++, following the formulation given in (9) and in the jPREdictor (31) source code. We henceforth refer to our implementation as the CPREdictor. The CPREdictor has been tested for functional equivalence with PREdictor and jPREdictor, in order to ensure comparability.

SVM-MOCCA

The Support Vector Machine Motif Occurrence Combinatorics Classification Algorithm (SVM-MOCCA) constructs one Support Vector Machine (SVM) per motif in order to model local sequence composition around motif occurrences in a target class versus one or more negative classes. Given a DNA sequence, a feature vector is constructed for each occurrence of each motif, consisting of occurrence frequencies of motifs and dinucleotides within 250 bp of the occurrence, giving a feature space in $|M| + 4^2$ dimensions for a set of $|M|$ motifs. For a given set of training sequences, each motif SVM is trained on all occurrences of its respective motif in the training sequences, with the view of predicting the sequence class (positive or negative) of a motif occurrence.

Once each SVM has been trained, occurrences of all motifs in the training set are classified by the corresponding SVMs. Let M denote a set of motifs, P and N sets of positive and negative training sequences, respectively, and $f(m, s)$ the frequency of positively classified occurrences of motif m in sequence s. For each motif $m \in M$, a weight is calculated as

$$w_m = \log \frac{\sum_{p \in P} f(m, p)/|P|}{\sum_{n \in N} f(m, n)/|N|}.$$

Given a sequence to classify, feature vectors are constructed for all motif occurrences in the sequence, which are in turn classified by their corresponding SVM. Frequencies of positively classified motif occurrences, $f(m)$ for a motif m, are weighted and summed, giving a score for the sequence:

$$S = \sum_m w_m f(m).$$

We used LibSVM (55) for the Support Vector Machine implementation. SVMs were trained with linear kernels and also with polynomial kernels with degrees 2 and 3 (henceforth referred to as quadratic and cubic kernels, respectively). As SVMs support the use of more than two classes, we used PREs together with all three control classes for training (dummy PREs, dummy genomic sequences and coding sequences).

When more than two classes are used, each SVM models all class boundaries using binary SVM classifiers, and

the class of each motif occurrence is predicted by majority vote, as implemented in LibSVM (55). One of the classes is designated as positive and the remaining classes as negative, giving a binary classification.

Prediction threshold calibration

We considered the model trained for cross-validation fold 1. The test set PREs were taken as positives. For the calibration negatives, we trained a fourth-order Markov chain on the *D. melanogaster* genome, and we generated 44 626 sequences, each 3 kb long, adding up to approximately the size of the *D. melanogaster* genome, at a total of 133.9 Mb. We searched the precision/recall space for the threshold with highest recall for the desired precision, with linear interpolation if necessary. For reasons of stability, we took the mean threshold over 10 repetitions of random-genome construction.

Genome-wide prediction

We applied each classifier across chromosomes 2L, 2R, 3L, 3R, 4 and X using a sliding window, with a step size of 10 bp, and a window size determined by the classifier. Windows with a score above the classifier threshold were noted as predictions, and overlapping predictions were merged into non-overlapping predicted candidate PREs.

Chromatin accessibility

We acquired DNaseI-seq data from the Berkeley Drosophila Transcription Network Project (BDTNP) (http://bdtnp.lbl.gov:8080/Fly-Net/access.jsp) for five different developmental stages (embryonic stages 5, 9, 10, 11 and 14). For a given set of regions, we defined accessible regions of the set as the subset of regions that overlap with regions in at least one of the five DNaseI-seq sets.

Genomic region overlaps

To measure genomic region overlaps between two sets A and B, we took the subset of regions in A that overlap with at least one region in B by at least one base pair. When comparing predictions to published genome-wide data sets, in order to account for potential distancing of recruited factors from recruitment sites, we extended regions in the published sets bi-directionally by 1 kb before checking overlaps (with the exception of modENCODE histone marks).

ModENCODE data sets

We acquired GFF/GFF3 genomic coordinate files from modENCODE (56) for *D. melanogaster*: H3K27me3 (13 sets); H3K4me1 (10 sets); H3K4me3 (14 sets); Pc (Polycomb) (6 sets); Psc (Posterior sex combs) (3 sets); dSFMBT (2 sets). The full paths from the modENCODE FTP archive are given in Supplementary Table S1. The datasets were downloaded in April 2016, and later datasets were not considered. The sets include data from animals (Adult-Female, Adult-Male, Embryos-0-12-hr, Embryos-0-4-hr, Embryos-12-16-hr, Embryos-14-16-hr-OR, Embryos-16-20-hr, Embryos-2-4-hr-OR, Embryos-20-24-hr, Embryos-4-8-hr, Embryos-8-12-hr, Larvae-3rd-instar, Larvae-L1-stage, Larvae-L2-stage, Larvae-L3-stage, Late-Embryonic-stage), as well as cell-lines (ML-DmBG3-c2, S2-DRSC).

Extraction of PRE predictions with biologically relevant signals

For each set of predictions by CPREdictor T2017 w. GTGT and SVM-MOCCA (Supplementary Files 1 and 3), we extracted the subsets of predictions that overlapped both with at least one H3K27me3 peak and with at least one peak of Pc, Psc or Sfmbt. For the H3K27me3, Pc, Psc and Sfmbt signals, we used merged sets of peaks from modENCODE as noted above. The resulting sets of candidate PREs are henceforth referred to as CPREdictor T2017 w. GTGT HC (1036 candidate PREs; Supplementary File 2) and SVM-MOCCA HC (2908 candidate PREs; Supplementary File 4), respectively, with 'HC' standing for 'high-confidence'. In addition, we extracted predictions enriched in H3K4me1 as candidate TREs (Supplementary Files 10 and 11).

Core sequence fragment prediction

From the 3 kb-long (or longer when merged) SVM-MOCCA predictions, we identified the most predictive subregions, henceforth referred to as SVM-MOCCA HC Core (Supplementary File 5). We applied SVM-MOCCA to its genome-wide predictions, with an iteratively larger window size from the following sequence of sizes: 500 bp, 600 bp, 750 bp, 1 kb, 1.5 kb, 2 kb, 2.5 kb, 3 kb, and with a step size of 50 bp. The highest-scoring window for each window size was collected, and the overall maximally scoring window (with the score normalized by window length), was defined as the core sequence.

Target gene prediction

We acquired the FlyBase genome annotation release R6.04. For a given region, any gene overlapping with a region was defined as a candidate target gene. For each region that did not overlap with any gene, the gene closest to the region (as determined by the closest region and gene endpoints) was defined as a candidate target gene.

Candidate PcG target genes were predicted for the complete PRE prediction sets from CPREdictor T2003, CPREdictor T2017, CPREdictor T2017 w. GTGT (Supplementary File 6) and SVM-MOCCA (Supplementary File 7).

Target genes from other publications

We downloaded published sets of predicted PcG target genes for PREdictor (9) and EpiPredictor (32), and from Schwartz *et al.* (34) and Enderle *et al.* (35).

The Schwartz *et al.* (34) PcG target genes were extracted from Supplementary Tables S2 and S4 from their article (class I and class II high-confidence PcG target genes, respectively), and these two sets were merged. For the Enderle *et al.* (35) set, target genes were extracted from the article's Supplementary Table 4 (first column). Genes that could not

be found in the FlyBase (51) r6.04 annotation were omitted. No further validation was performed on sets, except for the predictions from (9), which we validated using FlyBase, giving higher numbers of genes recognized in the annotation we used. Since no target genes were published in (36), we predicted target genes for that study by proximity, following the same procedure as for our own PRE predictions.

Gene ontology analysis

A list of all gene names was extracted from the FlyBase (51) r6.04 annotation. For each set of candidate PcG target genes, gene ontology analysis was performed using GOrilla (57) with two unranked lists of genes, where the first was the list of candidate PcG target genes and the second was the list of all annotated genes.

Software and packages

All figures except for Figure 2D were generated using R (58). The Precrec (59) library was used for generating average Precision/Recall curves and corresponding confidence intervals (Figure 1A and C). The Plotrix (60) library was used when generating the pie charts in Figure 2C. For generating the Venn diagrams in Figures 3D and Supplementary Figure S11, the VennDiagram (61) library was used. Tomtom (62) was used to search for factors that bind a *k*-mer. Gene ontology analysis was performed using GOrilla (57). The *vestigial*, *invected* and *engrailed* loci in Figure 2D were visualized using the Integrated Genome Browser (63).

RESULTS

Training sequence models on genome-wide PcG target sites improves PRE sequence model generalization

We wanted to see how models trained on genome-wide experimentally determined PcG-enriched regions compare to models trained on the Ringrose *et al.* (9) set of PREs, in terms of their ability to distinguish independent experimentally determined PcG-enriched regions from different classes of background sequences. To this end, we extracted genomic sequences for PcG-enriched regions from three publications (34–36), as described in Materials and Methods. We focus on the (34) set for training. Our models are discriminative, necessitating a set of non-PREs for training. We used three classes of non-PRE sequences for training and testing: (a) dummy PREs, (b) dummy genomic sequences and (c) coding sequences, as described in Materials and Methods. Dummy PREs, due to their motif composition being similar to that of PREs, form the strictest of our control sets, but are also unlikely to retain the characteristic motif occurrence clustering that has been found to be predictive of PREs (9). We thus assume that dummy PREs are unlikely to model functional PREs, and we include this set in the training of all of our models. Core coding sequences have zero or close to zero overlaps with experimentally determined PRE sets when promoter-overlapping PREs are omitted (data not shown). We speculate that any overlaps of PREs with coding sequences are due to promoter-promixal PREs, lack of positional precision for ChIP data, and factor mobility, rather than that PREs occur in coding sequences. With this assumption, coding sequences constitute a set of real genomic sequences that are unlikely to contain PREs. Both dummy genomic sequences and coding sequences share only minimal resemblence with PREs, making them more null than dummy PREs. We thus focused most of our attention on training with dummy PREs, but we include dummy genomic and coding sequences in our model evaluation and when training multi-class models, as independent control sets. This enabled us to investigate any over-fitting to dummy PREs that may occur and to train multi-class models.

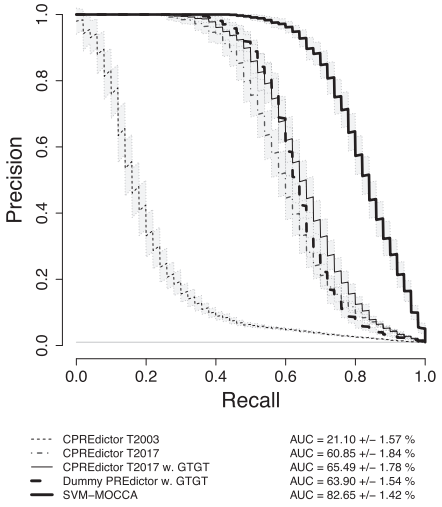
In order to test model generalization, we split the PRE and control sets into independent training and test sets, with 50-fold cross-validation to account for random variation, and a 1:100 ratio of PREs to non-PREs to reflect the expected genome-wide context, as described in Materials and Methods.

When training the CPREdictor algorithm on the T2017 set, using the same motifs as Ringrose *et al.* (9), and evaluating the trained classifiers on independent cross-validation PREs versus dummy genomic sequence controls, we observed a 2.9-fold increase in the mean Area Under the Precision Recall Curve (PRC AUC) compared to training with the training set used by Ringrose *et al.* (9) (T2003) (Figure 1A). This increase in PRC AUC is robust over cross-validation (Figure 1A and B), with non-overlapping 95% confidence intervals of the mean PRC AUCs (Figure 1A). We also observed increased PRC AUC for T2017 when evaluating with dummy PRE controls (Figure 1C and D) and coding sequence controls (Supplementary Figure S1).

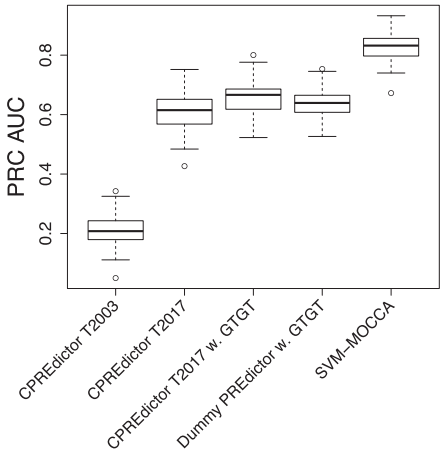
These results demonstrate that training models on genome-wide experimentally determined PcG target sites, and with controls generated by a fourth-order Markov chain trained on those sites, results in models that better distinguish independent PcG target sites (from the same set) from genomic background and PRE-like non-PRE sequences than models trained on the set compiled by Ringrose *et al.* (9). Training and evaluating PRE sequence models using other published sets of PcG-recruiting regions shows the same trend, where models trained on PcG-recruiting regions generalize better to independent PcG target regions than models trained on the T2003 set (Supplementary Figure S2).

The improvement in model generalization is independent of training set size

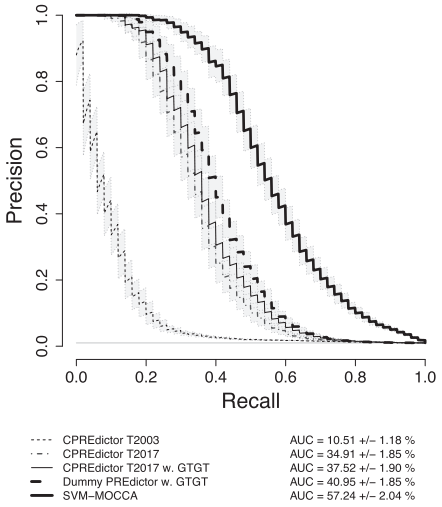
To determine the influence of training set size on generalization performance, we additionally trained the CPREdictor using sets of 12 and sets of 50 PRE and control sequences each. We observed only negligible differences in generalization performance across the sets of 12, 50 and 110 training sequences (mean PRC AUCs, with 95% confidence intervals, were $34.62 \pm 1.84\%$, $34.79 \pm 1.82\%$ and $34.91 \pm 1.85\%$, respectively; all values from an evaluation against dummy PREs; compare Figure 1C), demonstrating that training set size does not play a role in generalization performance and



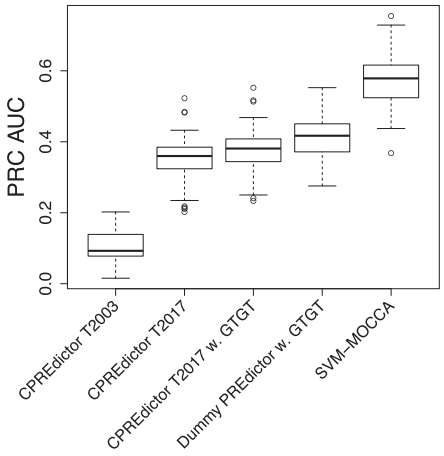
A Training PRE classifiers on ChIP-data improves generalization, with SVM-MOCCA giving the highest PRC AUC.



B Increases in cross-validation PRC AUC values associated with training PRE classifiers on ChIP-based data are robust to random variation.

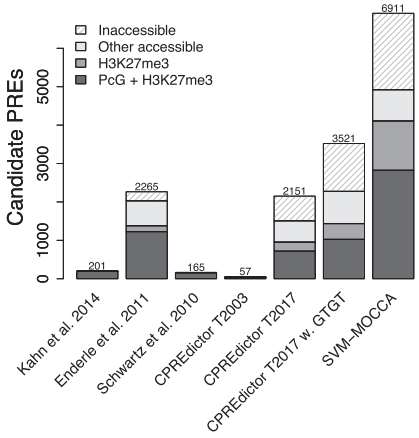


C Sequences generated by a naive PRE model (dummy PREs) are more difficult to separate from PREs, but models still manage to do so.

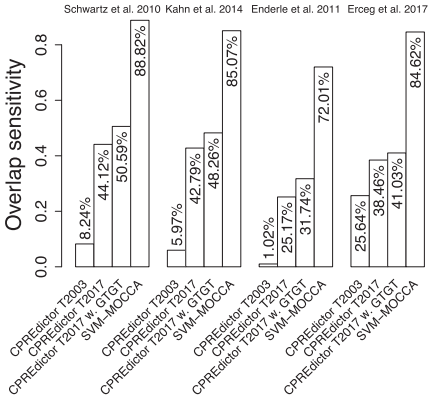


D Distinction of naively generated PRE sequences from experimentally determined PREs is robust over cross-validation.

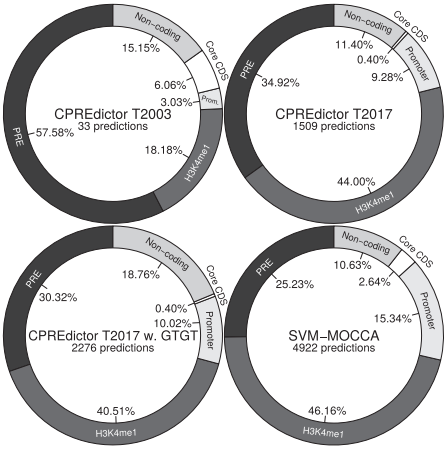
Figure 1. Classifier generalization when trained on genome-wide experimental data for PRE prediction. (**A**) Average Precision/Recall plot for classifiers applied to PREs determined by Schwartz *et al.* (34) (independent from training set PREs) versus 100 times as many control sequences generated by a fourth-order Markov Chain trained genome-wide, as according to the plot legend. Average curves over all 50 folds are shown, together with 95% confidence intervals for the mean precision. AUC values are percentages rounded to two digits. (**B**) PRC AUC box plot for multiple classifiers over all 50 folds. (**C**) Average Precision/Recall plot for PREs determined by Schwartz *et al.* (34) (independent from training set PREs) versus 100 times as many sequences generated randomly using a fourth-order Markov Chain trained on PREs, constituting a naive PRE model (dummy PREs). Average curves over all 50 folds are shown, together with 95% confidence intervals for the mean precision. (**D**) PRC AUC box plot for multiple classifiers over all 50 folds.



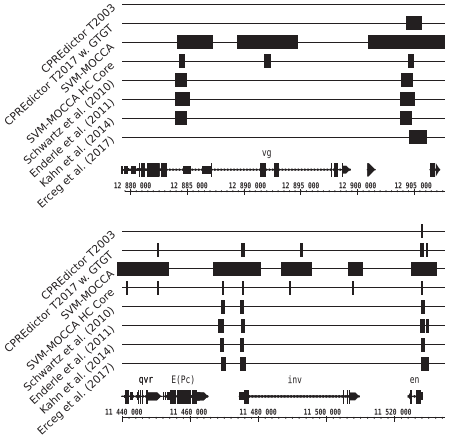
A Training PRE classifiers on ChIP-based data yields over 10-fold more candidate PRE/TRE predictions genome-wide, and the majority of those in H3K27me3 domains also recruit Polycomb group proteins.



B The larger set of candidate PRE/TRE predictions has a comparable increase in overlaps with independent genome-wide datasets. Note that CPREdictor T2017, CPREdictor T2017 w. GTGT and SVM-MOCCA were trained on a cross-validation subset of the Schwartz *et al.* (34) dataset.

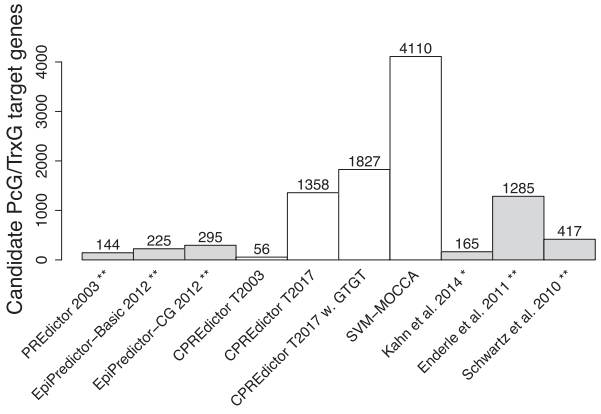


C Our new predictions are less exclusive to the merged set of experimentally determined PREs, but difference in precision is smaller when considering H3K4me1 as a signature of TREs.

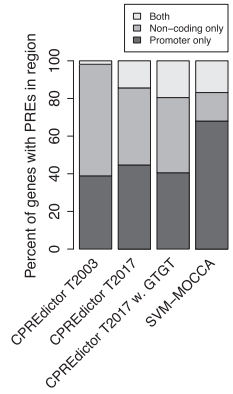


D Our classifiers predict verified PREs that were left out during training.

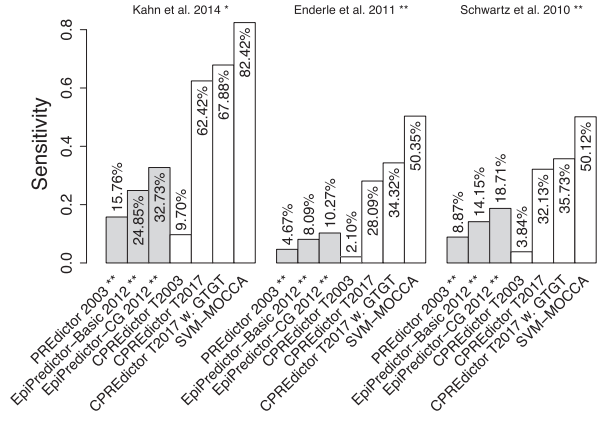
Figure 2. Results of genome-wide candidate PRE/TRE prediction for an expected precision of 80%. **(A)** Numbers of experimentally determined and computationally predicted candidate PREs. Accessible portions in Polycomb repressed domains (H3K27me3) have been marked, as well as the portions of those regions that are enriched in Polycomb. Chromatin accessibility was derived from DNaseI-seq data; see Materials and Methods, also for H3K27me3 and Polycomb datasets. **(B)** Overlap sensitivity of each classifier's predictions to two genome-wide, experimentally determined candidate PRE sets (35,36) and a set of functionally validated PREs (69) (see Materials and Methods for the definition of these three sets). Overlap sensitivity is defined as the fraction of regions in an experimental set that are overlapped by at least one prediction. **(C)** Proportions of the sets of predictions that overlap with different genomic loci. Only predictions in accessible chromatin are considered. The merged set of experimentally determined PREs by Kahn *et al.* (36), Enderle *et al.* (35) and Schwartz *et al.* (34) are considered first, and from the leftover, H3K4me1, then promoters, then core CDS; the final leftover set of predictions is marked as non-coding. See Materials and Methods for H3K27me3 datasets. Promoters are predicted as 3 kb upstream to 0.5 kb downstream from annotated gene transcription start sites. Core CDS is annotated coding sequence (CDS) shrunk bi-directionally by 250 bp (see Materials and Methods). **(D)** *invected*/*engrailed* and *vestigial* loci, visualized with the Integrated Genome Browser (63).



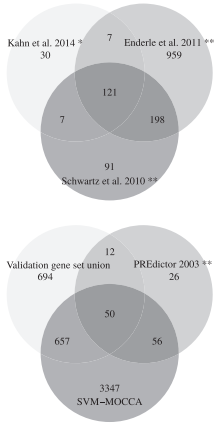
A Target gene prediction by proximity to our predicted candidate PREs gives over 10-fold more PcG/TrxG target gene candidates.



B Predicted target genes primarily have PREs either at their promoters or in non-coding regions.



C The increase in numbers of target gene predictions for our predicted candidate PREs is accompanied by increased sensitivity to PcG/TrxG target genes from genome-wide datasets.



D Gene set overlaps are incomplete.

Figure 3. PcG/TrxG target gene prediction results. (**A**) Numbers of target genes predicted by each algorithm, as well as in each experimentally published set. (**B**) Fractions of predicted target genes that have predicted PREs in either promoter regions (TSS −3 kb/+0.5 kb), in non-coding regions (not on promoters or core coding regions) or both. (**C**) Sensitivity of each classifier target gene prediction set to experimentally determined sets. Sensitivity is defined as the fraction of experimentally determined genes that are also predicted. * Kahn *et al.* (36) did not publish a set of genes, so we predicted target genes by proximity. ** Genes that were not found in the current annotation were omitted. (**D**) Venn diagrams of gene set overlaps for validation gene sets and target gene predictions.

suggesting that the T2017 training set is qualitatively different from the T2003 training set.

The choice of negative training sequences is instrumental in PRE prediction performance

It is interesting to ask how models trained with the training set used by Ringrose *et al.* (9) fare compared with models trained using their set of PREs and randomly generated non-PRE sequences (dummy PREs). To test this, we trained a fourth-order Markov chain on the Ringrose *et al.* (9) training set PREs, and we randomly generated 12 dummy PREs, each with length equal to the mean PRE length (2914 bp). After training CPREdictor on this training set, we observed increased generalization to the Schwartz *et al.* (34) set versus dummy PREs compared to when using the negative training set from Ringrose *et al.*, with PRC AUCs close to those

obtained when training with ChIP-based data (Supplementary Figure S3). In summary, combining the 2003 positive training sequences with dummy PREs derived from these as negatives generalizes better to independent PcG targets than models trained with the original 2003 positive and negative training sets, demonstrating that the choice of negative training sequences is instrumental in PRE prediction performance.

Including the GTGT motif improves PRE sequence model generalization

The accumulating evidence for the GTGT motif being a component of Polycomb regulation (9,28–30) prompted us to investigate whether the inclusion of the GTGT motif in our PRE sequence models improves generalization to independent PREs. When we added the GTGT motif to our CPREdictor T2017 model, we observed an additional 1.1-fold increase in the mean PRC AUC on independent PREs and dummy genomic sequences in comparison to the T2017 model without the GTGT motif (Figure 1A). This increase is robust over cross-validation, different PRE sets and different control classes (Figure 1A–D, and Supplementary Figures S1 and S2). In summary, the inclusion of the GTGT motif in PRE models improves generalization across different training and test sets, providing additional evidence that this motif plays an important role in Polycomb regulation.

The improvement in model generalization cannot be attributed to increased model complexity, and GTGT performs better than other reported motifs

To assess the degree to which the improvement in generalization performance upon adding the GTGT motif might be explained by the increased model complexity (owing to the inclusion of a motif and the associated parameters), we added random 4-mers to our CPREdictor T2017 model. The inclusion of GTGT resulted in a 1.04-fold to 1.16-fold increase in the mean PRC AUC over the inclusion of 18 out of 19 other unique, randomly generated 4-mers, for Schwartz PREs versus dummy genomic controls (Supplementary Figure S4), demonstrating that the GTGT motif contributes to a performance improvement beyond that expected from increased model complexity. The only 4-mer that gave higher PRC AUC was GGCG. Searching for *D. melanogaster* factors that bind GGCG using Tomtom (62) gave Brinker (Brk) as a match (P-value = 7.78e–04), a transcriptional repressor of Dpp target genes (64–66). We also tested six other published motifs that have been associated with PcG recruitment: one additional motif for Zeste, one for Sp1/KLF, one for Dsp1, two for Grainyhead and one for 'site A' ((53) and references therein; see also Materials and Methods). The GTGT motif gives the largest improvement in model generalization (1.06-fold to 1.10-fold higher PRC AUC compared with the inclusion of the other motifs, for Schwartz PREs versus dummy genomic sequences), while the other motifs affect model generalization to only a smaller extent and similarly to each other (PRC AUCs range from 59.73% to 61.96% for Schwartz PREs versus dummy genomic controls, and the majority of the confidence intervals overlap with one another) (Supplementary Figure S5), suggesting that the GTGT motif plays a more decisive role in PcG recruitment.

Genome-wide PcG target sites and Ringrose *et al.* training PREs have different sequence properties

Given that the models trained with the T2017 set and the GTGT motif and those trained with the T2003 set showed highly different generalization abilities to independent PcG target sites, we were interested in how the models differ and what might cause the difference in generalization ability. We thus investigated the weights of CPREdictor models trained with the T2003 set and T2017 set, and also with PREs from the T2003 set and generated non-PREs (Supplementary Figure S6).

We found a moderate negative correspondence between motif pair weights assigned using T2003 versus T2017 (Pearson's correlation coefficient < -0.5). Weight correlation when using T2003 PREs and generated non-PREs versus when using the T2017 set is low (Pearson's correlation coefficient < 0.4). For T2003 versus when using T2003 PREs and generated non-PREs, correlation is similarly low (Pearson's correlation coefficient < 0.4). Whereas the T2003 model has three negatively weighted motif pairs (G10:G10, G10:GAF and GAF:GAF), with all three weights being substantial, the T2017 model has two (PM:PM and PS:GTGT), both with weights close to zero. In fact, the most negatively predictive T2003 motif pair, G10:G10, is the most highly weighted motif pair for the T2017 model. The discrepancy might be due to clusters of GAF motifs in the negative training set in (9) which includes promoters of genes that are regulated by GAF and Z (9). The small size of the T2003 set can result in one or a few more pair occcurrences in the negative training set compared to the positive training set which would have a large influence on the final model weights. The seven highest weighted motif pairs in the T2003 model all include Pho binding site variants (PF:PM, GAF:PF, PM:PS, G10:PM, G10:PF, PF:PF and PM:Z). These weights have approximately been reduced by half or more for the T2017 model, and the top four highest weighted motif pairs for the T2017 model do not include any Pho binding site variants and are instead enriched for G10 (G10:G10, G10:GAF, G10:Z, G10:GTGT). The dominance of G10 in the top T2017 motif pair weights may in part be attributed to properties of control sequences generated by Markov chains of fixed order and the long length of G10. Models trained using the T2003 versus T2017 sets are thus dissimilar, meaning that motif composition is different in the training sets.

Models trained with genome-wide PcG targets can distinguish Ringrose *et al.* training PREs from background

As the models trained on T2003 and T2017 are so different, we wanted to see how our models score the training set used by Ringrose *et al.* (9). The models that we trained on ChIP data versus dummy PREs have lower PRC AUC to the Ringrose *et al.* (9) training set than does CPREdictor trained on this set, but PRC AUCs are still above random (Supplementary Figure S7). The best generalization to the Ringrose *et al.* (9) training set that we observe for

models not trained on this set is for CPREdictor including the GTGT motif, with a mean PRC AUC of 70.20 ± 0.99%. The lowest is for SVM-MOCCA, with a mean PRC AUC of 61.75 ± 1.14%. We also investigated the degree to which our models can distinguish the Ringrose *et al.* (9) PREs from dummy PREs. For this case, CPREdictor with GTGT and SVM-MOCCA obtain the highest PRC AUCs, at 81.23 ± 0.26% and 98.45 ± 0.32%, respectively. In conclusion, our models are still able to distinguish the set of PREs and non-PREs used by Ringrose *et al.* (9), though to a lower degree than CPREdictor trained on this set, and our models are better at distinguishing the Ringrose *et al.* (9) PREs from randomly generated controls.

Uniformly weighted motif pair clustering distinguishes PREs from background

Considering the large differences in model weights obtained when using T2003, T2017 and a set consisting of PREs from T2003 and generated non-PREs, we wanted to see how well a uniformly weighted PREdictor model would distinguish PREs from non-PREs. We thus constructed a PREdictor model with all weights set equal to 1, henceforth referred to as the Dummy PREdictor. We found that the Dummy PREdictor generalizes comparably to CPREdictor trained with T2017 when including the GTGT motif and testing with Schwartz PREs as positives and dummy genomic sequences as negatives (Figure 1A). When we evaluate our models using Schwartz PREs as positives and dummy PREs as negatives, where the CPREdictor has been trained with this set, the Dummy PREdictor outperforms the CPREdictor (Figure 1C). This was a surprise to us, as we expected a trained model would have an advantage, with weights fitted both to PREs and a randomly generated non-PRE distribution. The Dummy PREdictor corresponds to a uniformly weighted motif pair clustering.

A more advanced PcG target site sequence model improves generalization

We have developed SVM-MOCCA (see Materials and Methods), a new method for modelling *cis*-regulatory elements, and we wanted to test how such a more advanced modelling method would fare in modelling PcG target sites in comparison to the CPREdictor.

We trained SVM-MOCCA using the T2017 set with all three control classes and with the motifs used by Ringrose *et al.* (9), with the addition of the GTGT motif. The training sequences are 3 kb long. Ringrose *et al.* (9) used a 500 bp window. We thus tested how CPREdictor and SVM-MOCCA models generalize when using windows that are 500 bp or 3 kb long. We found that for SVM-MOCCA, using a 3 kb sequence window gave similar generalization performance to a 500 bp window, and we focus on a 3 kb window due to it potentially capturing more diffuse PREs. For the CPREdictor, a 500 bp sequence window gives the best generalization, so we focus on using this window size (Supplementary Figure S8).

The method of Support Vector Machines supports non-linear classification, which prompted us to test SVM-MOCCA with linear, quadratic and cubic kernels (see Materials and Methods). The best generalization performance was achieved with the quadratic kernel (Supplementary Figure S9). We thus focus on the quadratic kernel in subsequent analyses, referring to the corresponding run as SVM-MOCCA.

When testing with Schwartz PREs versus dummy genomic sequences, we observed a 1.3-fold increase in PRC AUC when using SVM-MOCCA (with a quadratic kernel, trained with T2017 with three control classes, and including the GTGT motif) compared to the best CPREdictor result (trained with T2017 and including GTGT) (Figure 1A). This increase is robust over cross-validation, different PRE sets and different control classes (Figure 1A–D, and Supplementary Figures S1 and S2), and the 95% confidence intervals of the mean PRC AUCs are non-overlapping (Figure 1A and C). SVM-MOCCA is particularly good at distinguishing PREs from dummy PREs, giving a 1.5-fold increase in the mean PRC AUC over CPREdictor (Figure 1C). These results demonstrate that a more advanced modelling approach can substantially contribute to an improved generalization performance.

Models trained on genome-wide PcG target sites predict more candidate PREs for the same expected precision

Having trained our models, we can predict candidate PREs genome-wide. Previous efforts of modelling PREs (9) have yielded candidate PRE predictions of high reliability, but with only moderate overlap with sets of genome-wide PcG target sites (67). We wanted to see whether training models on genome-wide PcG target sites would result in predictions with higher agreement with independent genome-wide PcG target sites.

We set a score threshold for each model for an expected precision of 80% genome-wide. Having trained CPREdictor with the T2017 set, we predicted over 37 times more candidate PREs genome-wide compared to having trained CPREdictor with the T2003 set (Figure 2A). Including the GTGT motif led to another 1.6-fold increase in predictions (Supplementary File 1). Using SVM-MOCCA gave a further 2-fold increase in predictions over CPREdictor (Supplementary File 3).

CPREdictor trained with T2003 predicts less than half as many PREs as the PREdictor predicted genome-wide (9). This can be explained by differences in the threshold calibration procedure. Ringrose *et al.* (9) calibrated the PREdictor threshold for one expected false positive prediction genome-wide, based on 100 genome-size sequences generated by an i.i.d. genome model. Our method differs in that we find a threshold for which we obtain a desired precision for a set of independent PREs and controls generated by a fourth-order Markov chain trained genome-wide, where the total control sequence length adds up to the size of the genome. Sequences generated by a fourth-order Markov chain are more difficult for our models to distinguish from PREs than are sequences generated by an i.i.d. model (data not shown). As a result, we can expect a reduction in numbers of predictions made using our control sequences for calibration. Also, the ability of a model to positively classify PREs is taken into account by our method, which can affect the numbers of predictions made if precision is only high for low recall, which is the case for CPREdictor trained on

T2003. We can expect some further difference in numbers of predictions for these calibration methods on the basis that Ringrose *et al.* (9) use genome-length random sequences, whereas we use sets of PRE-length sequences with total set length equal to that of the genome. The calibration methods are thus not comparable. However, we use our method for calibrating all the classifiers that we consider, where possible.

SVM-MOCCA motif model weights are heterogeneous and enriched for interacting dinucleotide patterns

Given the improved generalization of SVM-MOCCA with a quadratic kernel, we were interested in what the sequence criteria encoded in the model are. In order to investigate this, we transformed the SVM quadratic kernel into a sum of weighted feature pairs (Supplementary Text 1). Our SVM-MOCCA models are multi-class, giving a large number of weights. We wanted to condense the weights involved in distinguishing PREs from non-PREs into one weight per feature pair. We thus summed up all feature pair weights across all PRE versus non-PRE class boundaries. Duplicate features, due to reverse complements and reversed pairing order were added together, giving a set of 171 unique feature pair weights.

Strikingly, each SVM has different motif pair weighting, even though all of the SVMs have been trained on the same sets of PREs and non-PREs. The only difference lies in the motifs for which each SVM is trained to classify its local sequence landscape. This suggests that PRE sequence criteria may vary per motif, with different local sequence landscapes for different PRE motifs.

For all motifs except the En motif, all weights involving motif pairs are negatively weighted, and positively weighted feature pairs are with dinucleotide pairs. Positively weighted dinucleotides generally include 'GA'/'AG', which likely correspond with GAGA site enrichment, as well as 'AC'/'CA', which may correspond with GTGT sites. 'AA' self-pairing is generally positively weighted, as is 'CC' self-pairing, but interestingly, 'AA' paired with 'CC' is negatively weighted.

In conclusion, SVM-MOCCA classifier weights are enriched for patterns in agreement with previous work, such as GAGA, GTGT and poly-A, but also in 'CC'-dinucleotide self-pairing, and there are weight interactions for the 'AA' and 'CC' dinucleotides.

A quarter to half of genome-wide PRE predictions are in chromatin that is inaccessible early in development

ChIP-chip and ChIP-seq can only detect the PcG target regions that are accessible for binding in the cells that are being studied. We were thus interested in how many of our predictions fall in chromatin that is accessible over development. We acquired DNaseI-seq peaks for cells in five different embryonic stages (Materials and Methods). We refer to regions that overlap with peaks in at least one of the DNaseI-seq sets as being in accessible chromatin. The experimentally determined PcG target sets that we consider (34–36) were determined by ChIP-chip and ChIP-seq on ML-DmBG3-c2, ML-DmD23-c4, S2 and Sg4 cell lines, derived from embryonic cells and the developing nervous system. As expected, all regions in these sets overlap with accessible chromatin. One half to three quarters of predictions made by our methods are in accessible chromatin (Figure 2A). Therefore, a quarter to half of our predictions are inaccessible in the five developmental stages we consider, and even if they are *bona fide* PREs, they would likely go undetected in the experiments that determined the PcG targets that we consider. When comparing *in silico* PRE predictions to experimentally determined PcG targets, we thus focus on PREs in accessible chromatin.

We predict a set of 2908 candidate PREs enriched in biologically relevant signals

To assess the degree to which our predictions recruit PcG proteins and repress or activate chromatin, we acquired genome-wide experimentally determined enrichment signals for three PcG proteins (Pc, Psc and Sfmbt) (13), histone 3 lysine 27 trimethylation (H3K27me3; a mark of Polycomb repressed chromatin) (68), and histone 3 lysine 4 monomethylation (H3K4me1; a mark of Trithorax activated chromatin) (25), from modENCODE (56) (see Materials and Methods).

Of accessible predictions, over half are enriched in H3K27me3 at some point during development, and the majority of these regions are also enriched in at least one PcG protein (Pc, Psc or Sfmbt) (Figure 2A). We extracted the latter subsets for CPREdictor T2017 w. GTGT and SVM-MOCCA (see Materials and Methods), henceforth CPREdictor T2017 w. GTGT HC (1036 high-confidence candidate PREs; Supplementary File 2) and SVM-MOCCA HC (2908 high-confidence candidate PREs; Supplementary Files 4 and 5) respectively. In addition, we extracted predictions enriched in H3K4me1 (1723 candidate TREs for CPREdictor T2017 w. GTGT, 3616 candidate TREs for SVM-MOCCA; Supplementary Files 10 and 11, respectively). The SVM-MOCCA PRE and TRE sets have 2412 candidates in common, supporting the notion of a dual function of PREs as TREs. The four sets constitute collections of candidate PRE/TREs with experimental support in the form of enrichment in biologically relevant signals.

Models of genome-wide PcG target sites increase the agreement between PRE prediction and genome-wide experiments

For independent evaluation of our predictions, we considered two independent published sets of PcG target regions: one determined using ChIP-chip (36) and one using ChIP-seq (35). The Schwartz *et al.* (34) and Kahn *et al.* (36) sets are both based on Sg4 cells and have related sources in terms of authors and institutions. However, whereas the Schwartz *et al.* (34) set is based on peaks of E(z), Psc and Pc, the Kahn *et al.* (36) set is based on peaks of E(z), Trx, Pc and H3K27me3. The Kahn *et al.* (36) set is also larger than the Schwartz *et al.* (34) set (201 versus 170 candidate PREs, respectively, in *Drosophila* genome assembly R6; 165 in the Schwartz *et al.* set when excluding known PREs around the *invected*/*engrailed* and *vestigial* loci). As a result of their relatedness, the Kahn *et al.* (36) and Schwartz *et al.* (34) sets have a high number of overlaps (70.65−83.53% when considering the full sets).

The Enderle *et al.* (35) set is unrelated to the Kahn *et al.* (36) and Schwartz *et al.* (34) sets, determined using a different experimental method (ChIP-seq), cell culture (S2 cells) and factors (Pc, Ph, Psc and Trx-C). The Enderle *et al.* (35) set is an order of magnitude larger than the other sets, at 2274 regions (2265 euchromatic regions). As a result, the Enderle *et al.* (35) set covers most of the Schwartz *et al.* (34) and Kahn *et al.* (36) sets (91.18% and 89.55% of regions, respectively, when considering the full sets). Additionally, we used a set of functionally tested PREs compiled from the literature (69).

Sequence models trained on genome-wide experimentally determined PcG target sites predict a larger fraction of each of the independent experimental sets, compared to the CPREdictor trained with the T2003 set (Figure 2B). SVM-MOCCA predicts the majority of each of these sets (Figure 2B). Out of our predictions in accessible chromatin, over a quarter overlap with regions from the Schwartz, Enderle and Kahn sets (Figure 2C). Of the remainder, the majority are enriched with histone 3 lysine 4 monomethylation, potentially indicative of TREs/PREs in active states (25).

During training, we left out five PREs from the well-studied *vestigial* (*vg*) (28), *invected* (*inv*) (70) and *engrailed* (*en*) (71,72) loci. Of these PREs, CPREdictor trained with the T2003 set predicts only one, whereas CPREdictor trained with the T2017 set predicts three out of five, and SVM-MOCCA predicts all five (Figure 2D). SVM-MOCCA also predicts several other peaks, with no experimental evidence.

We were interested in the degree to which our final predictions conform to the PREs and non-PREs used for training by Ringrose *et al.* (9). We thus acquired genomic coordinates for the T2003 set by BLAST search, and compared overlaps. CPREdictor T2017 w. GTGT and SVM-MOCCA predict 45.45% and 90.91% of the T2003 PREs, respectively, which is a 1.7–3.3-fold increase over CPREdictor T2003, for which this set was used for training. Whereas CPREdictor T2003 predicts none of the T2003 non-PREs, CPREdictor T2017 w. GTGT and SVM-MOCCA predict 18.75% and 56.25%, respectively. Though SVM-MOCCA predicts many of the T2003 non-PREs, SVM-MOCCA HC Core predicts as many T2003 PREs as SVM-MOCCA, but only 18.75% of T2003 non-PREs, the same number as CPREdictor T2017. See Supplementary Figure S10 for an extended evaluation.

Taken together, these results demonstrate that models of genome-wide PcG target sites have larger agreement with independent genome-wide experimental data and functionally verified PREs than models based on the Ringrose *et al.* (9) training set.

We predict a large new set of candidate PcG regulated genes, enriched in transcription factor and signalling functions

Given our much larger set of candidate PRE predictions, it is interesting to identify candidate target genes and their functions and to compare them with previously published sets. Target genes for our predictions were assigned as described in Materials and Methods. Target genes for other publications were extracted or defined also as described in Materials and Methods.

Similar to the prediction of PREs, our methods predict many more target genes than previously published methods (Figure 3A). The majority of predicted PcG target genes has associated PRE predictions either at the promoter or in non-coding sequence, but not both (Figure 3B). Our target gene predictions have higher numbers of overlaps with target genes from genome-wide PcG profiling studies than previously published *in silico* methods (Figure 3C). The sensitivities of our predictions to the Schwartz *et al.* (34) and Enderle *et al.* (35) sets are lower when based on genes (Figure 3C), in comparison to when based on PREs (Figure 2B).

We summarized gene set overlaps with Venn diagrams (Figure 3D). For the Schwartz *et al.* (34), Enderle *et al.* (35) and Kahn *et al.* (36) sets, respectively, 21.82%, 74.63% and 18.18% of each is unique. The majority of the Kahn *et al.* (36) set is in consensus with the other sets, whereas the majority of the Schwartz *et al.* (34) set is in agreement with the Enderle *et al.* (35) set but not the Kahn *et al.* (36) set. The largest target gene agreement is observed between the Enderle *et al.* (35) and Schwartz *et al.* (34) sets, at 319 genes, corresponding to 24.82% of the Enderle *et al.* (35) set and 76.50% of the Schwartz *et al.* (34) set. Accordingly, the sets of experimentally determined PcG target genes that we consider have different sizes and incomplete overlaps. Of published PREdictor gene predictions (9), 43.06% correspond to genes in at least one of the experimentally determined sets. The ratio of SVM-MOCCA predictions that correspond to experimentally determined PcG target genes is smaller, at 17.20%. There are only 12 validated genes that only the PREdictor predicts and SVM-MOCCA does not, and SVM-MOCCA predicts an additional 657 validated PcG target genes that the PREdictor does not. As such, SVM-MOCCA predicts many PcG target genes with experimental support, as well as a large new set of candidate PcG target genes that await experimental verification.

We analyzed PcG target gene predictions for enriched gene ontologies using GOrilla (57). Target genes predicted by SVM-MOCCA are highly enriched in transcription factor functions (Supplementary Figure S11). We compared gene ontology terms enriched in predictions made by SVM-MOCCA with terms enriched in the PREdictor, EpiPredictor (basic) and EpiPredictor (CG) predictions, the Schwartz *et al.* (34) HC Class I and II sets, and the Enderle *et al.* (35) set. The top three terms are enriched in all sets considered and are all related to transcription factor activities. The fourth term, 'Protein binding', is enriched for one of the experimental sets. Six terms are enriched in zero or one other set and comprise functions unrelated to transcription factor activities: 'Calcium ion binding', 'Potassium ion transmembrane transporter activity', 'Cytoskeletal protein binding', 'Actin binding', 'Cell adhesion molecule binding' and 'Protein kinase activity'. The remaining enriched terms correspond to transcription factor and signalling activities (see Supplementary File 9 for complete lists of enriched terms in all sets).

DISCUSSION

Previous approaches to modelling *Drosophila* PREs have used comparatively small sets of functionally characterized PREs and non-PREs for training binary classifiers

(9,31,32). Here, we trained models on published genome-wide sets of PcG-recruiting chromatin regions. Negatives were generated by fourth-order Markov chains trained either on the same set of PcG-recruiting sequences or the entire genome and also taken from coding sequence.

Genome-wide sets of experimentally determined PcG-recruiting regions can be expected to contain false positives, due both to physical chromatin interactions and to experimental conditions. PREs have been observed to make long-range chromatin contacts with promoters, with ChIP signals at both contact points, where then one signal may be only a shadow of the interaction (1,73). A recent Hi-C study by Eagen *et al.* (74) found PRC1 enriched at 26% of chromatin loop anchors, and for loops where not both anchors correspond to PREs, there could thus be additional shadow signals. Furthermore, the majority of PRE ChIP studies rely on cell cultures, and even if assuming optimal experimental conditions and choice of antibodies, cultured cells are not normal cells (75), and genome-wide epigenetic states are likely to differ from those *in vivo*. Furthermore, ChIP only captures protein binding at a certain time in a certain population of cells, and results are thus unlikely to reflect the epigenetic diversity in the entire animal. Additionally, the PcG-recruiting regions we consider are large (3 kb after expansion to account for potential distancing between recruiting sequences and recruited factors). Nonetheless, models trained on PcG-recruiting regions and automatically generated controls generalize well to independent PcG-recruiting regions over cross-validation, with substantially higher PRC AUC than the CPREdictor trained on the set used by Ringrose *et al.* (9) (2.88-fold increase). Thus, our modelling methods are robust against any non-PRE signals that the ChIP-data used for training may contain, and they manage to pick out general features predictive of PcG-recruiting sequences.

Identifying a large, definitive set of genomic non-PREs that is sufficiently PRE-like to use for training sequence models is challenging. We circumvented this problem by automatically generating non-PRE sequences by use of naive PRE models (fourth-order Markov chains), making use of the knowledge that motif pair occurrences are predictive of PREs, while individual motif occurrences are only marginally predictive (9). Thus, the probability of these models generating *bona fide* PREs can be expected to be low, but the sequences they generate have highly similar motif composition to that of PREs. Despite this similarity, our models are able to distinguish them from published PcG target regions, showing that these genome-wide experimentally determined regions are enriched in motif co-occurrence patterns.

We developed a new method for modelling *cis*-regulatory elements, called SVM-MOCCA. SVM-MOCCA distinguishes itself from other PRE-modelling methods by modelling the local motif and dinucleotide occurrence landscape around motif occurrences. Across the board, SVM-MOCCA gave the best generalization to independent PcG-recruiting regions over cross-validation.

The models we trained on genome-wide experimental data and randomly generated controls predict many more PREs genome-wide than previous methods, for the same expected precision of 80%. This is accompanied by our methods predicting a much larger number of experimentally determined PcG target regions than previous methods. We excluded five well-studied PREs at the *vestigial*, *engrailed* and *invected* loci from our training data, both during model testing and for genome-wide prediction, and we predict the majority of these PREs. Our computational approach allowed us to study the importance of the GTGT motif and of other motifs in a genome-wide manner. Adding the GTGT motif results both in increased model generalization and in a higher number of predictions genome-wide, adding to the growing body of evidence that this motif plays an important role in Polycomb recruitment. The inclusion of other published motifs had only little impact on model generalization.

Counterintuitively, models trained using our methods predict more of the PREs used for training by Ringrose *et al.* (9) than does the CPREdictor trained on that very set, for an expected precision of 80% genome-wide (Supplementary Figure S10). A possible explanation for this is that our models have been trained on large sets of non-PRE sequences, and that this makes the models better at distinguishing PcG target sites from genomic background. Models trained with the T2017 set also predict a minimal number of sequences from the non-PRE set used by Ringrose *et al.* (9). SVM-MOCCA predicts over half of the non-PREs used by Ringrose *et al.* (9), but filtering by biological signals and predicting the core predictive regions of the SVM-MOCCA predictions lowers the number of non-PREs predicted to a fifth.

Despite the much larger number of predictions that our models make, and though we predict a large fraction of the PREs in the experimental sets that we consider, none of our sets of predictions completely cover any of the experimentally determined PRE sets. There may be several reasons for this. Our models may lack the sequence features needed in order to accurately model the remaining PREs, such as additional motifs, higher-order motif occurrence combinatorics, strandedness and positioning, or taking local or distal sequence elements into consideration. The experimental sets may also contain regions that are not in fact PREs, but are instead marked by PcG proteins due to physical interactions with PREs, or are enriched due to experimental noise.

As the SVM-MOCCA predictions are 3 kb long, we predicted core PRE fragments. It is interesting to note that the core fragments have fewer overlaps with experimental sets. This means that PcG-enriched regions are close by, and it is possible that experimental signals in some cases have been displaced due to factor mobility. Our observation is also in agreement with the suggestion of Schuettengruber *et al.* (30) that the genome uses 'not only local sequence (high-affinity transcription factor binding sites located at the binding peaks) information to determine PREs, but also integration of regional sequence information [...]' and that the use of such information to predict PREs 'may break the current specificity and sensitivity barriers.' A corollary to this latter notion is the possibility that previous evaluations of PRE prediction have taken regional information (recruitment versus enrichment) into account only insufficiently.

Multiple weaker PREs functioning together has been observed for the *engrailed* gene locus (76). Our core PRE prediction method only finds the sub-region with the strongest

sequence signal enrichment. It may be that some SVM-MOCCA predictions are enriched in multiple weak sequence signals that add up to a significant prediction. If so, ChIP-signals that do not overlap with a predicted core may instead coincide with a separate, weaker PRE sequence signal. It could also be that the position of the final ChIP-peak depends on the structure of the complex of weak PREs and PcG proteins.

We present two high-confidence sets of *D. melanogaster* candidate PRE predictions, based on filtering predictions for enrichment of histone 3 lysine 27 trimethylation and at least one of three PcG proteins (Pc, Psc or Sfmbt). This filtering procedure provides a form of experimental validation of predicted PRE candidates on the basis of previously published ChIP enrichment datasets and is comparable to experimental definitions of PREs from such datasets (34–36). However, our procedure does not define PRE candidates from ChIP enrichment datasets alone, but starts with a set of candidates that were predicted by a well-designed machine-learning model and that share sequence characteristics that have been established to be relevant, both here and in previous work (9,29,30). Furthermore, since with our filtering procedure we treat any type of ChIP enrichment as a necessary but not as a sufficient criterion for PRE-ness, our high-confidence candidates are less prone to potential looping, spreading and displacement artefacts. In fact, one could argue that the presence of a PRE prediction in a region of ChIP enrichment gives credence to that enrichment and indicates the initial Polycomb recruitment site. Even though the high-confidence prediction sets are smaller than the complete prediction sets (1036 versus 3521 predictions for CPREdictor and 2908 versus 6911 for SVM-MOCCA), they have almost as high numbers of overlaps with the experimental sets that we consider (Supplementary Figure S10). As such, we increase precision to the experimentally determined PcG target region sets with low loss of recall. It is worth noting that we used merged ChIP peaks from multiple experiments per factor and that the factors we considered are not only enriched at PREs, making this a modest filtering step. Both high-confidence PRE sets are larger than the Schwartz *et al.* (34) set that the models were trained on, despite the filtering for biologically relevant chromatin signatures. These high-confidence candidate PREs remain to be tested for whether they can maintain target gene transcription states.

Additionally, we predict many PREs outside of the high-confidence sets. A large number of candidate PREs do not overlap with chromatin that is accessible in the developmental stages that we consider. Inaccessible PRE predictions may be functional PREs that recruit PcG/TrxG when chromatin is made accessible. A large number of PRE predictions that do not overlap with experimentally determined PRE sets but are nonetheless in accessible chromatin are enriched for histone 3 lysine 4 monomethylation (H3K4me1). It is possible that these predictions are PRE/TREs in an activated state (25) and that they recruit Polycomb in other contexts. A large proportion (over 82%) of high-confidence PRE candidates are also enriched in H3K4me1, supporting the notion of a dual function of PREs as TREs. Furthermore, the fact that all candidates were predicted by a single machine learning model suggests that PREs and TREs have a common sequence code. The remaining predictions may be false positives, due both to a threshold calibration for an expected precision of 80% (corresponding to an expected 20% of false positives among the positive predictions) and to imperfections in our training sets and models.

An extended overlap analysis (Supplementary Table S2) showed only small differences in high-confidence PRE candidate enrichment between H3K4me1 and H3K4me3, the latter of which has previously been reported to be methylated by TRX but was later shown to be mostly methylated by SET1/COMPASS (reviewed in (77)).

In correspondence with our larger numbers of *D. melanogaster* PRE predictions compared to previously published *in silico* methods, we predict a larger set of candidate PcG/TrxG target genes, with higher numbers of overlaps with published experimentally determined PcG/TrxG target genes. We speculate that, like our predicted PREs themselves, predicted targets that have not previously been identified on the basis of ChIP enrichment, might recruit Polycomb or Trithorax group proteins and associated histone modifications in cell types or in conditions that so far have not been studied with respect to their epigenetic regulatory landscape. Our target gene predictions are highly enriched for transcription factor functions and also for novel potential PcG target gene functions. The sensitivities of predictions to experimentally determined sets are lower when considering PcG target genes than for candidate PREs. This can be attributed to different methods being employed for predicting target genes from regions, as well as different genome annotations used while predicting target genes. Schwartz *et al.* (34) used the Dm2 assembly and Enderle *et al.* (35) used Dm3. Both Schwartz *et al.* (34) and Enderle *et al.* (35) determined PcG target genes based on enrichment of PcG signals proximal to the TSS, rather than based on gene proximity to candidate PREs. Overall, our genome-wide PcG target gene predictions are more sensitive to experimentally determined PcG target genes than are published predictions from previous *in silico* PcG target gene prediction methods.

Although we devoted most of our attention to training with the Schwartz *et al.* (34) candidate PREs, we obtain similar results when training with the Enderle *et al.* (35) and Kahn *et al.* (36) sets (Supplementary File 8), demonstrating that our results are general. Training SVM-MOCCA with the Schwartz *et al.* (34) candidate PREs resulted in 6911 predictions genome-wide, training with the Enderle *et al.* (35) set resulted in 5910 predictions genome-wide, and 5294 of the Schwartz *et al.* (34)–based predictions overlap with Enderle *et al.* (35)–based predictions (CPREdictor results are similar, at lower total numbers of predictions, 3521, 2775 and 2768, respectively). This high overlap indicates the robustness of our approach and might also suggest a potential saturation of PRE prediction.

There are multiple ways in which our work can be expanded upon. The majority of the steps have been written as a computational pipeline, aiding not only the reproducibility of our results, but also the application to other problems. Our methods can be adapted to the modelling of other classes of regulatory sequences and for use in other genomes, given appropriate sets of motifs and genome-wide experimental data. Our high-confidence PRE predictions

are a rich source of candidates for the further study of PRE function, architecture and dynamic behaviour during development.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Takaya Saito and Leonie Ringrose for discussions and for feedback on the article manuscript.

FUNDING

Deutsche Forschungsgemeinschaft, Excellence Initiative, Institutional Strategies [0192854102] (in part). Funding for open access charge: University of Bergen.
Conflict of interest statement. None declared.

REFERENCES

1. Ringrose,L. and Paro,R. (2007) Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development*, **134**, 223–232.
2. Steffen,P. and Ringrose,L. (2014) What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory. *Nat. Rev. Mol. Cell Biol.*, **15**, 340–356.
3. Schuettengruber,B., Bourbon,H., Di Croce,L. and Cavalli,G. (2017) Genome regulation by Polycomb and Trithorax: 70 years and counting. *Cell*, **171**, 34–57.
4. Simon,J., Chiang,A., Bender,W., Shimell,M.J. and O'Connor,M. (1993) Elements of the Drosophila bithorax complex that mediate repression by Polycomb group products. *Dev. Biol.*, **158**, 131–144.
5. Chan,C., Rastelli,L. and Pirrotta,V. (1994) A Polycomb response element in the Ubx gene that determines an epigenetically inherited state of repression. *EMBO J.*, **13**, 2553–2564.
6. Chinwalla,V., Jane,E.P. and Harte,P. (1995) The Drosophila Trithorax protein binds to specific chromosomal sites and is co-localized with Polycomb at many sites. *EMBO J.*, **14**, 2056–2065.
7. Klymenko,T. and Müller,J. (2004) The histone methyltransferases Trithorax and Ash1 prevent transcriptional silencing by Polycomb group proteins. *EMBO Rep.*, **5**, 373–377.
8. Ringrose,L. and Paro,R. (2004) Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.*, **38**, 413–443.
9. Ringrose,L., Rehmsmeier,M., Dura,J.M. and Paro,R. (2003) Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. *Dev. Cell*, **5**, 759–771.
10. Schwartz,Y.B., Kahn,T.G., Nix,D.A., Li,X.Y., Bourgon,R., Biggin,M. and Pirrotta,V. (2006) Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat. Genet.*, **38**, 700–705.
11. Nègre,N., Hennetin,J., Sun,L.V., Lavrov,S., Bellis,M., White,K.P. and Cavalli,G. (2006) Chromosomal distribution of PcG proteins during Drosophila development. *PLoS Biol.*, **4**, e170.
12. Ringrose,L. (2007) Polycomb comes of age: genome-wide profiling of target sites. *Curr. Opin. Cell Biol.*, **19**, 290–297.
13. Schwartz,Y.B. and Pirrotta,V. (2013) A new world of Polycombs: unexpected partnerships and emerging functions. *Nat. Rev. Genet.*, **14**, 853–864.
14. Di Croce,L. and Helin,K. (2013) Transcriptional regulation by Polycomb group proteins. *Nat. Struct. Mol. Biol.*, **20**, 1147–1155.
15. Bauer,M., Trupke,J. and Ringrose,L. (2016) The quest for mammalian Polycomb response elements: are we there yet? *Chromosoma*, **125**, 471–496.
16. Müller,J. and Kassis,J. (2006) Polycomb response elements and targeting of Polycomb group proteins in Drosophila. *Curr. Opin. Genet. Dev.*, **16**, 476–484.
17. Kassis,J. and Brown,J. (2013) Polycomb group response elements in Drosophila and vertebrates. *Adv. Genet.*, **81**, 83–118.
18. Brown,J.L., Mucci,D., Whiteley,M., Dirksen,M.L. and Kassis,J.A. (1998) The Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. *Mol. Cell*, **1**, 1057–1064.
19. Shao,Z., Raible,F., Mollaaghababa,R., Guyon,J.R., Wu,C.t., Bender,W. and Kingston,R.E. (1999) Stabilization of chromatin structure by PRC1, a Polycomb complex. *Cell*, **98**, 37–46.
20. Czermin,B., Melfi,R., McCabe,D., Seitz,V., Imhof,A. and Pirrotta,V. (2002a) Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell*, **111**, 185–196.
21. Müller,J., Hart,C., Fracis,N., Vargas,M., Sengupta,A., Wild,B., Miller,E., O'Connor,M., Kingston,R. and Simon,J. (2002) Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell*, **111**, 197–208.
22. Cao,R., Wang,L., Wang,H., Xia,L., Erdjument-Bromage,H., Tempst,P., Jones,R. and Zhang,Y. (2002) Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science*, **298**, 1039–1043.
23. Kuzmichev,A., Nishioka,K., Erdjument-Bromage,H., Tempst,P. and Reinberg,D. (2002) Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes Dev.*, **16**, 2893–2905.
24. Klymenko,T., Papp,B., Fischle,W., Köcher,T., Schelder,M., Fritsch,C., Wild,B., Wilm,M. and Müller,J. (2006) A Polycomb group protein complex with sequence-specific DNA-binding and selective methyl-lysine-binding activities. *Genes Dev.*, **20**, 1110–1122.
25. Tie,F., Banerjee,R., Saiakhova,A.R., Howard,B., Monteith,K.E., Scacheri,P.C., Cosgrove,M.S. and Harte,P.J. (2014) Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing. *Development*, **141**, 1129–1139.
26. Rickels,R., Hu,D., Collings,C., Woodfin,A., Piunti,A., Mohan,M., Herz,H., Kvon,E. and Shilatifard,A. (2016) An evolutionary conserved epigenetic mark of Polycomb response elements implemented by Trx/MLL/COMPASS. *Mol. Cell*, **63**, 318–328.
27. Kassis,J., Desplan,C., Wright,D. and O'Farrell,P. (1989) Evolutionary conservation of homeodomain-binding sites and other sequences upstream and within the major transcription unit of the Drosophila segmentation gene engrailed. *Mol. Cell Biol.*, **9**, 4304–4311.
28. Okulski,H., Druck,B., Bhalerao,S. and Ringrose,L. (2011) Quantitative analysis of Polycomb response elements (PREs) at identical genomic locations distinguishes contributions of PRE sequence and genomic environment. *Epigenet. Chromatin*, **4**, 4.
29. Ray,P., De,S., Mitra,A., Bezstarosti,K., Demmers,J.A., Pfeifer,K. and Kassis,J.A. (2016) Combgap contributes to recruitment of Polycomb group proteins in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 3826–3831.
30. Schuettengruber,B., Ganapathi,M., Leblanc,B., Portoso,M., Jaschek,R., Tolhuis,B., Lohuizen,M.v., Tanay,A. and Cavalli,G. (2009) Functional anatomy of Polycomb and Trithorax chromatin landscapes in Drosophila embryos. *PLoS Biol.*, **7**, e13.
31. Fiedler,T. and Rehmsmeier,M. (2006) jPREdictor: a versatile tool for the prediction of *cis*-regulatory elements. *Nucleic Acids Res.*, **34**, W546–W550.
32. Zeng,J., Kirk,B.D., Gou,Y., Wang,Q. and Ma,J. (2012) Genome-wide Polycomb target gene prediction in Drosophila melanogaster. *Nucleic Acids Res.*, **40**, 5848–5863.
33. Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.
34. Schwartz,Y.B., Kahn,T.G., Stenberg,P., Ohno,K., Bourgon,R. and Pirrotta,V. (2010) Alternative epigenetic chromatin states of Polycomb target genes. *PLoS Genet.*, **6**, e1000805.
35. Enderle,D., Beisel,C., Stadler,M.B., Gerstund,M., Athri,P. and Paro,R. (2011) Polycomb preferentially targets stalled promoters of coding and noncoding transcripts. *Genome Res.*, **21**, 216–226.
36. Kahn,T.G., Stenberg,P., Pirrotta,V. and Schwartz,Y.B. (2014) Combinatorial interactions are required for the efficient recruitment of pho repressive complex (PhoRC) to Polycomb response elements. *PLoS Genet.*, **10**, e1004495.
37. Tolhuis,B., Muijrers,I., de Wit,E., Teunissen,H., Talhout,W., van Steensel,B. and van Lohuizen,M. (2006) Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. *Nat. Genet.*, **38**, 694–699.

38. Oktaba,K., Guitiérrez,L., Gagneur,J., Girardot,C., Sengupta,A.K., Furlong,E.E. and Jürg,M. (2008) Dynamic regulation by Polycomb group protein complexes controls pattern formation and the cell cycle in Drosophila. *Dev. Cell*, **15**, 877–889.
39. Horak,C.E. and Snyder,M. (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.*, **350**, 469–483.
40. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
41. Cheutin,T. and Cavalli,G. (2014) Polycomb silencing: from linear chromatin domains to 3D chromosome folding. *Curr. Opin. Genet. Dev.*, **25**, 30–37.
42. Xiao,X., Li,Z., Liu,H., Su,J., Want,F., Wu,X., Liu,H., Wu,Q. and Zhang,Y. (2013) Genome-wide identification of Polycomb target genes in human embryonic stem cells. *Gene*, **518**, 425–430.
43. van Heeringen,S.J., Akkers,R.C., van Kruijsbergen,I., Arif,M.A., Hanssen,L.L., Sharifi,N. and Veenstra,G.J.C. (2014) Principles of nucleation of H3K27 methylation during embryonic development. *Genome Res.*, **24**, 401–410.
44. Du,J., Kirk,B., Zeng,J., Ma,J. and Wang,Q. (2018) Three classes of response elements for human PRC2 and MLL1/2-Trithorax complexes. *Nucleic Acids Res.*, **46**, 8848–8864.
45. Chang,Y., King,B., O'Connor,M., Mazo,A. and Huang,D. (1995) Functional reconstruction of trans regulation of the Ultrabithorax promoter by the products of two antagonistic genes, Trithorax and Polycomb. *Mol. Cell Biol.*, **15**, 6601–6612.
46. Tillib,S., Petruk,S., Sedkov,Y., Kuzin,A., Fujioka,M., Goto,T. and Mazo,A. (1999) Trithorax- and Polycomb-group response elements within an Ultrabithorax transcription maintenance unit consist of closely situated but separable sequences. *Mol. Cell Biol.*, **19**, 5189–5202.
47. Brock,H. and van Lohuizen,M. (2001) The Polycomb group–no longer an exclusive club? *Curr. Opin. Genet. Dev.*, **11**, 175–181.
48. Bloyer,S., Cavalli,G., Brock,H. and Dura,J. (2003) Identification and characterization of polyhomeotic PREs and TREs. *Dev. Biol.*, **261**, 426–442.
49. Hoskins,R.A., Carlson,J.W., Wan,K.H., Park,S., Mendez,I., Galle,S.E., Booth,B.W., Pfeiffer,B.D., George,R.A., Svirskas,R. *et al.* (2015) The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res.*, **25**, 445–458.
50. dos Santos,G., Schroeder,A.J., Goodman,J.L., Strelets,V.B., Crosby,M.A., Thurmond,J., Emmert,D.B., Gelbart,W.M. and Consortium,F. (2014) FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.*, **43**, D690–D697.
51. Gramates,L.S., Marygold,S.J., dos Santos,G., Urbano,J.M., Antonazzo,G., Matthews,B.B., Rey,A.J., Tabone,C.J., Crosby,M.A., Emmert,D.B. *et al.* (2017) FlyBase at 25: looking to the future. *Nucleic Acids Res.*, **45**, D663–D671.
52. CBN (1970) IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970. *Biochem. J.*, **120**, 449–454.
53. Brown,J.L. and Kassis,J.A. (2013) Architectural and functional diversity of Polycomb group response elements in Drosophila. *Genetics*, **195**, 407–419.
54. Saito,T. and Rehmsmeier,M. (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, **10**, e0118432.
55. Chang,C.C. and Lin,C.J. (2011) LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27:1–27:27.
56. Celniker,S.E., Dillon,L.A., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
57. Eden,E., Navon,R., Steinfeld,I., Lipson,D. and Yakhini,Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinf.*, **10**, 48.
58. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
59. Saito,T. and Rehmsmeier,M. (2017) Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, **33**, 145–147.
60. Jim,L. (2006) Plotrix: a package in the red light district of R. *R-News*, **6**, 8–12.
61. Chen,H. and Boutros,P.C. (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinf.*, **12**, 35.
62. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.
63. Freese,N.H., Norris,D.C. and Loraine,A.E. (2016) Integrated genome browser: visual analytics platform for genomics. *Bioinformatics*, **32**, 2089–2095.
64. Campbell,G. and Tomlinson,A. (1999) Transducing the Dpp morphogen gradient in the wing of Drosophila: regulation of Dpp targets by brinker. *Cell*, **96**, 553–562.
65. Jaźwińska,A., Kirov,N., Wieschaus,E., Roth,S. and Rushlow,C. (1999) The Drosophila gene brinker reveals a novel mechanism of Dpp target gene regulation. *Cell*, **96**, 563–573.
66. Minami,M., Kinoshita,N., Kamoshida,Y., Tanimoto,H. and Tabata,T. (1999) brinker is a target of Dpp in Drosophila that negatively regulates Dpp-dependent genes. *Nature*, **398**, 242–246.
67. Hauenschild,A., Ringrose,L., Altmutter,C., Paro,R. and Rehmsmeier,M. (2008) Evolutionary plasticity of Polycomb/Trithorax response elements in Drosophila species. *PLoS Biol.*, **6**, e261.
68. Czermin,B., Melfi,R., McCabe,D., Seitz,V., Imhof,A. and Pirrotta,V. (2002b) Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell*, **111**, 185–196.
69. Erceg,J., Pakozdi,T., Marco-Ferreres,R., Ghavi-Helm,Y., Girardot,C., Bracken,A.P. and Furlong,E.E. (2017) Dual functionality of cis-regulatory elements as developmental enhancers and Polycomb response elements. *Genes Dev.*, **31**, 590–602.
70. Cunningham,M.D., Brown,J.L. and Kassis,J.A. (2010) Characterization of the Polycomb group response elements of the Drosophila melanogaster invected locus. *Mol. Cell Biol.*, **30**, 820–828.
71. Americo,J., Whiteley,M., Brown,J.L., Fujioka,M., Jaynes,J.B. and Kassis,J.A. (2002) A complex array of DNA-binding proteins required for pairing-sensitive silencing by a Polycomb group response element from the Drosophila engrailed gene. *Genetics*, **160**, 1561–1571.
72. DeVido,S.K., Kwon,D., Brown,J.L. and Kassis,J.A. (2008) The role of Polycomb-group response elements in regulation of engrailed transcription in Drosophila. *Development*, **135**, 669–676.
73. Bantignies,F. and Cavalli,G. (2011) Polycomb group proteins: repression in 3D. *Trends Genet.*, **27**, 454–464.
74. Eagen,K.P., Aiden,E.L. and Kornberg,R.D. (2017) Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 8764–8769.
75. Cherbas,L. and Gong,L. (2014) Cell lines. *Methods*, **68**, 74–81.
76. De,S., Mitra,A., Cheng,Y., Pfeifer,K. and Kassis,J.A. (2016) Formation of a Polycomb-domain in the absence of strong Polycomb response elements. *PLoS Genet.*, **12**, e1006200.
77. Sneppen,K. and Ringrose,L. (2019) Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent. *Nat. Commun.*, **10**, 2133.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230843680 (print)
9788230853054 (PDF)