Advancing Software Tools for Designing Oligonucleotide FISH Probes: Enabling Visualization of Repetitive DNA in Varied Genomes

Robin Aguilar

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Brian J. Beliveau, Co-Chair
William S. Noble, Co-Chair
David Shechner

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Advancing Software Tools for Designing Oligonucleotide FISH Probes: Enabling Visualization of Repetitive DNA in Varied Genomes

Robin Aguilar

Chairs of the Supervisory Committee:
Brian J. Beliveau
William S. Noble
Department of Genome Sciences

Genomes are organized in a specific and hierarchical manner that influences cell function and fate. The perturbation of genomic stability has been shown to mediate the rise of human diseases. While numerous tools have been developed to better understand the relationship of the non-repetitive genome toward preserving genomic stability, our understanding of the functional consequences of highly repetitive DNA is limited. During my thesis work, I have contributed to the development of software tools that may be used to support fluorescent *in* situ hybridization (FISH) assays to visualize highly repetitive DNA at the scale of diverse and fully assembled genome builds. The work I describe here includes the development of Tigerfish, a software tool to design oligonucleotides that target unique repetitive DNA intervals at the scale of genomes. Additionally, I also curated numerous scientific communications and advocacy resources to facilitate building more inclusive research spaces in genomics within and outside of the University of Washington. Through my work, I also created a curriculum that may be used to teach others in PhD programs about the importance of fostering supportive academic environments for those with diverse lived experiences beyond scientific learning spaces.

Table of Contents

Chapter I: Introduction *7*

1.1 3D genome organization is essential for unraveling complex cellular processes **7**

1.2 Visualizing how genomes are organized using Fluorescence *in situ* Hybridization (FISH) **10**

1.3 Improvements in DNA sequencing provide novel insights on repetitive DNA intervals **18**

1.4 Unraveling Structure-Function Relationships of Repetitive DNA Intervals through FISH **31**

1.5 Genomics research and its implications outside of the lab **33**

Chapter 2: Implementation of computational methods to visualize repetitive DNA intervals in the human genome *34*

2.1 Abstract **35**

2.2 Introduction **35**

2.3 Results **39**

2.4 Discussion **52**

2.5 Methods **53**

2.6 Contributions **60**

Chapter 3: Seeking Justice in Genomics Education and Research *61*

3.1 "Life happens over the course of a PhD." **61**

3.2 Community organizing and a PhD **72**

3.3 DEI initiatives are owed accountability, examination, and re-evaluation. **80**

3.4 Genomics must reflect on what values our teaching spaces practice. **85**

3.5 I created a curriculum that discusses identity and belonging in the biosciences. **87**

Chapter 4: Discussion and Conclusion *89*

4.1 Further refining the Tigerfish workflow to explore diverse repetitive DNA families. **89**

4.2 Creating community resources for open-access repetitive DNA specific oligo probes. **89**

4.3 Interrogating the structure and function relationships of distinct repetitive DNA arrays using interdisciplinary experimental approaches. **90**

4.4 Rethinking what the future of inclusive genomics education could look like in PhD programs. . **90**

Funding *92*

Bibliography *92*

Appendices *120*

Appendix A - Tigerfish Supplementary Data and Software 120

Appendix B - GSAIMS Academic Archive 145

Intro 145
Events 145
Frameworks 162
Monthly Meetings 162
SACNAS/ABRCMS Recruitment 163
GS Recruitment Weekends 163
STEMPals 164
Art Zine 165
GSAIMS Career Symposium 165
Hosting Seminar Speakers 166
Mentorship Program 167
Finances 167
Contact Directory 168
Conclusions 169

Appendix C - Curriculum Outline and Resources 170

About 170
Format of the course 171
Grading 172
Course Guide and Material by Week 172
Classroom Interactions Guidelines and Community Expectations 181
Weekly Discussion Question Framework 182
Final Project Timeline and Submission Guidelines 183
Weekly Reflection Submission Guidelines 184
Weekly Reflection Prompts by Week 185
Supplementary Readings by Week 186

ACKNOWLEDGEMENTS

The completion of my dissertation would not have been possible without the support of many people in my life.

First and foremost, I would like to express my heartfelt gratitude to my thesis committee for dedicating their time and careful expertise to advise me on my scientific pursuits, as well as my personal and career goals. I extend my sincere thanks to my advisors, Brian Beliveau and Bill Noble, for forming a co-mentorship that greatly contributed to my research in their accommodating and welcoming labs. Both of my advisors have exhibited profound creativity, unwavering support, and extensive scientific knowledge. I am genuinely thankful for their mentorship, which has been instrumental in my journey to becoming a better scientist throughout my Ph.D. I deeply appreciate their guidance and support as I navigated the challenges inherent in both graduate school and life beyond it. I am also thankful to Brian for having the courage to welcome me as his first Ph.D. student when we both joined the Department of Genome Sciences in the same year. The transformative changes and sense of community fostered in the Beliveau lab over the past five years have been remarkable, and I cannot express enough gratitude to this lab group for their unwavering support. Above all, I am profoundly grateful for the privilege of calling two labs my home, and I draw inspiration from my lab members who have encouraged me to pursue the science that ignites my passion. I carry this reminder with me wherever my research endeavors may lead.

Throughout my graduate school journey, I have had the privilege of collaborating with many kind and brilliant individuals in Genome Sciences. I cannot thank my lab groups enough, especially Eva Nichols, Conor Camplisson, Yuzhen Liu, Lily Deng, and Elliot Hershberg, who have been exceptional collaborators, mentors, and friends. Together, we worked tirelessly to establish a new lab space in Genome Sciences, and their support has been invaluable.

I want to extend my heartfelt thanks to my dearest friends in Genome Sciences who have celebrated with me through all seasons. Without their unwavering support, this dissertation would not have come to fruition. I am grateful for the communities we have fostered within and beyond Genome Sciences. I also owe a debt of gratitude to the mentors I have encountered across the country and through my advocacy work in science. I would like to acknowledge the significant mentorship I received from Atom Lesiak, Jey McCreight, and Gina Driscoll, whose impact on my life, especially toward the end of my Ph.D., cannot be overstated. I extend this appreciation to my friends and co-workers at Mercury @ Machinewerks, who have been incredibly supportive, humble, and kind as I learned the ropes of bartending during my final year of graduate school. Furthermore, I wish to express my thanks to the ReclaimingSTEM Institute team and the HHMI Gilliam Scholar program for their unwavering support and belief in me as I navigated a career path in science communication and illustration.

I am deeply grateful to my family for standing by my side every step of the way. To my mother and father, I extend my heartfelt thanks for your unconditional love and support throughout my entire life. Your encouragement and support for my dreams have been a driving force, and I will always find inspiration in both of you. I believe that our culture and upbringing have played a pivotal role in shaping the scientist I have become today.

This Ph.D. is also dedicated to those who never had the opportunity to complete their dissertations for various personal reasons, especially those who faced challenges in navigating the academic world. These words were crafted with your stories in mind.

Chapter I: Introduction

1.1 3D genome organization is essential for unraveling complex cellular processes.

1.1.1 Eukaryotic genomes are nonrandomly organized.

Understanding the 3D organization of chromatin within the nucleus is essential for elucidating its function[1]. The linear sequence of the human genome, with approximately 6 billion base pairs spanning roughly 2 meters when fully relaxed, necessitates a non-random and compact organization to fit within the cellular nucleus, which is roughly 5–10 micrometers in diameter in most human cells[2]. Such specific organization is essential in maintaining the physical organization of chromosomes so that proper gene transcription, DNA replication, and DNA repair mechanisms are maintained to preserve genome stability[2]. This specialized chromatin organization plays a pivotal role in mediating and influencing crucial biological processes including gene expression and DNA replication timing across cell types and developmental trajectories [2–5].

Chromatin's 3D architecture involves conserved hierarchical structures[2,3,6]. At the primary level of packaging, DNA wraps around histone proteins to form nucleosomes, marking the first level of condensation in 3D genome organization[7]. Beyond nucleosomes, chromatin undergoes higher-order folding and compaction, leading to the formation of chromatin fibers, loops, and topologically associated domains (TADs)[4,8–10]. These non-random structures are tightly regulated and play vital roles in gene expression, mediating enhancer-promoter interactions, and regulating cell differentiation and development across lineages[3,8]. Additionally, the organization of chromatin is associated with directing DNA replication and repair in human health and disease[3,11–15].

To comprehend the relationship between chromatin's spatial proximity and its states is crucial for understanding how it influences gene expression. Furthermore, it is essential to grasp the

significance of distinct chromatin interactions with nuclear domains and compartments to form a comprehensive view of global genome organization. Metazoan chromatin is broadly classified into euchromatin, comprising gene-rich and transcriptionally active regions, and heterochromatin, which houses transcriptionally silent genes[5]. Heterochromatin was initially identified by its high-intensity DNA staining throughout the cell cycle in iterative fragments[1,16]. Notably, classic studies provided early insight into the repressive role of heterochromatin in gene expression. The effect of heterochromatin on gene silencing was manifested in position effect variegation (PEV) in *Drosophila*, where active genes are transcriptionally silenced upon ectopic placement in juxtaposition to heterochromatin via chromosome rearragements[17]. Another classic example that demonstrates the silencing effect of heterochromatin is the dosage compensation of X-linked genes in mammalian females, which was shown to be the result of a heterochromatinized Barr body[18]. Subsequently, the positional relationship between euchromatin and heterochromatin has demonstrated that regions rich in active genes are primarily organized at the nuclear interior, while regions with inactive genes are positioned towards the nuclear periphery[2,5,19].

1.1.2 Technical developments empowering studies of nuclear architecture.

The past century has witnessed significant advances in sequencing, microscopy, and computational approaches, coinciding with the completion of the first phase of the human genome project and the genome-wide characterization of genetic variations[5,20–22]. These advances have significantly enriched our understanding of the biological properties of nuclear architecture, leading to substantial progress in studying 3D genome organization across diverse cell types and developmental stages[13,23–32].

A groundbreaking milestone in this field was the development of chromosome conformation capture (3C), a nuclear ligation assay used in conjunction with PCR to investigate 3D genome organization. Various 3C derivatives were subsequently introduced with increasing

throughput[33,34]. These techniques have evolved over time to facilitate the enrichment of specific contacts driven by proteins of interest through tools like ChIP-loop[35] and ChIA-PET[36], as well as contacts focused on specific genomic regions (Capture Hi-C)[35]. For instance, 4C[37] was designed to assess "one vs. many" loci contact frequency, while approaches such as Capture-Hi-C[35] and Capture-C[38–42] were employed to resolve "many vs. all" genomic contacts, and Hi-C[43] was utilized to analyze genome-wide contact interactions.

Simultaneously, ligation-independent strategies have been developed, including methods such as Tyramide Signal Amplification (TSA)[44], DNA Adenine Methyltransferase Identification (DAM identification)[45], and Split-Pool Recognition of Interactions by Tag Extension (SPRITE)[46], each contribute uniquely to the understanding of genome structure. TSA enhances the visualization of specific molecules, such as proteins, within the genome, aiding in their localization. DAM identification identifies sites where DNA adenine methylation occurs, shedding light on epigenetic regulation in bacterial genomes. SPRITE, on the other hand, provides comprehensive insight into the 3D spatial interactions of DNA regions, facilitating the mapping of genome architecture and its role in gene regulation.[35,47]. Moreover, advances in super-resolution microscopy and imaging-based assays have enabled the resolution of chromatin conformation in single cells at high resolution and throughput. Notably, live-imaging applications using the CRISPR-Cas9 systems are currently being developed to gain insights into life cycle-dependent chromatin-chromatin contacts[48].

Despite these remarkable strides, the structure-function relationships between euchromatin and heterochromatin in the context of 3D genome organization remain incompletely understood[5,15]. While the role of heterochromatin in repressing repetitive DNA elements and safeguarding genome stability is well established, heterochromatin's contribution to maintaining higher-order folding principles in global genome organization and the roles of distinct repetitive DNA families

remain enigmatic. For instance, open questions remain regarding the mechanisms governing context-specific deposition and maintenance of heterochromatin modifications across species including *Drosophila* and *S.Pombe* and how the existence of evolutionarily conserved H3K9me3 histone modification mediate spatiotemporal regulation of heterochromatin during cell development[5].

Recent accomplishments, such as achieving a gapless human genome assembly and obtaining knowledge of specific sequences, have presented researchers with the means to discern the relationships between active and repressed chromatin hubs responsible for gene expression or repression[5,49,50]. These insights can be obtained by incorporating information from repetitive DNA intervals that were previously omitted from incomplete genome assemblies. Leveraging genomics technologies that integrate microscopy, spatial genomics, and systematic studies across single cells promises to uncover novel insights into how heterochromatin and diverse repetitive DNA families facilitate genome stability through careful regulation of condensation and compaction of genomic regions enriched with heterochromatin throughout development and human disease.

1.2 Visualizing how genomes are organized using fluorescent *in situ* hybridization (FISH)

1.2.1 The ability to visualize cells was foundational to the field of 3D genome organization.

The earliest histochemistry techniques involved using natural and synthetic dyes to identify cellular organelles and components[51]. However, these compounds were non-specific due to their affinities for the various biomolecules present in cells. DNA-specific stains emerged in the 19th century and chromosomes were shown to contain DNA through the specificity of Feulgen stains[52,53]. Through this early work, the field of chromosome biology would become shaped by the development of early visualization techniques. In the 1940s, antibodies were conjugated to fluorochromes without losing their epitope binding specificity[51,54]. In the late 1960s, the earliest *in*

situ hybridizations were performed using probes labeled with radioisotopes[55]. During this time, efforts were also made to improve cell culture and the preparation of metaphase spreads, which posed early challenges for the study of human and mammalian cells[51]. Advancements in cell culture enabled the development of techniques that could selectively visualize portions of chromosomes in fixed preparations using nucleic acids. Ultimately, these pioneering techniques underscored the importance of visual methods in elucidating the role of chromatin in various biological contexts, including the understanding of inheritance, cytogenetics, and nuclear architecture.[51,54–58].

Early ISH methods provided valuable insights into the localization of repetitive elements in genomes[59]. Notably, satellite DNA families served as remarkable sites for FISH probe signals, and the visualization of mouse major satellite sequences underscored the importance of studying chromatin organization variations in genomes[59]. These repetitive sequences proved to be excellent targets in the design of early ISH probes due to the inherent nature of these sequences that allowed enhanced signal amplification due to the consistently repetitive nature of these sequences[59]. Additionally, the exploration of hybridization kinetics empowered the distinction between DNA families, where repetitive DNA was further classified across species and empowered the development of visualization experiments to survey such genomic regions[60,61].

However, these methods faced technical limitations that restricted their application to a wide range of biological targets[55]. The use of ^{3}H as a labeling method for genomic targets of interest posed challenges, such as difficulties in amplifying probe RNA or DNA for non-repetitive targets due to the infrequent radioactive decay of ^{3}H and the generation of probes against non-repetitive sequence targets[59]. Nevertheless, ISH methods saw improvement through the adoption of recombinant DNA techniques, including the incorporation of radiolabeled nucleotides via "nick translation," which enhanced probe production[62].

A key development in the development of these visualization methods was the introduction of “non-isotropic’ labeling in 1977. This method involved the conjugation of a rhodamine label was to a secondary antibody to label polytene chromosomes in *Drosophila melanogaster*[63]. Rhodamine is a fluorescent dye known for its photostable properties and by conjugating it to a secondary antibody, researchers could create a specific, lab stable fluorescent label. This would empower the first report of FISH using a direct label in 1980, where RNA directly labeled on the 3’ end with a fluorophore was used to probe specific DNA sequences[64]. Later, enzymatic incorporation of modified deoxynucleotide bases became widely used in FISH. The use of amino-allyl modified bases, which could be enzymatically incorporated into probe molecules and subsequently conjugated to desired haptens or fluorophores using amine-reactive chemistries, played a pivotal role in the development of *in situ* technologies as it allowed the production of high specific activity probes using simple chemistry[65,66].

In subsequent advances, nick-translated, biotinylated probes and secondary detection using fluorescent streptavidin conjugates were used to detect distinct mRNA and DNA targets[62,67,68]. Early probes produced from clones were large and probes were prepared by growth in a vector or *in vitro* transcription, followed by nick-translation. However, a drawback of these early large probes was that they often contained repeat sequences, making them prone to high background[54].

It's important to acknowledge that while ISH and FISH probes were generated from isolated genomic material, researchers also used RNA and DNA oligos produced by chemical synthesis as probe material[69]. Oligo probes were first used in radioactive ISH using the phosphoramidite method of oligo synthesis described in 1981[70]. Through the use of phosphoramidite as a method of oligo synthesis, the reagents involved in this process were stable under laboratory conditions

and therefore were much more applicable in the synthesis of oligos[71,72]. The improvement in the chemical synthesis of synthetic, single-stranded oligonucleotide (oligo) based probes eventually led to the preparation of hybridization probes carrying enough fluorescent molecules to allow for direct detection as described in experiments throughout the 1980s[70–72].

In addition to these technological advances toward understanding genome organization, it's essential to describe the significance of FISH in a clinical cytogenetic context. An attractive feature of FISH assays is the ability to detect several targets simultaneously in the same sample preparation, thus gathering cytogenetic information on interphase nuclei, making FISH valuable for clinical applications such as karyotyping and detection of chromosomal abnormalities by morphology and genomic copy number evaluations[51,54,58].

The discovery of diverse genomic features such as diverse repetitive DNA families and genomic motifs involved in genome organization, along with the eventual reduction of probe size, advances in microscopy and detector hardware, and the development of mathematical image processing algorithms, paved the way for improvements in microscopy methods to visualize diverse regions of the genome at varying scales[73–78]. Significant advances included the detection of chromosome targets to distinguish between all human chromosomes using computed interpretation of a 5-color scheme and, later, the multiplexed visualization of mRNA[79,80].

Until the development of 3C techniques and related high throughput derivatives, FISH served as the primary method for determining nuclear organization and chromatin conformation[35,54]. The sensitivity and resolution of the assay remained a limiting factor until further technological improvements were made with innovations in probe design through the ability to multiplex experiments to target many sites in parallel readout through iterative secondary hybridization[81]. Therefore, the tradeoff of sensitivity and resolution was a limiting factor that influenced the extent

to which early studies on genome organization could be made. Paired in tandem with 3C techniques, FISH offered spatial information and resolution that was complementary to the quantitative information provided by 3C methods, leading ultimately toward a more complete picture of genome organization in diverse cell types[2].

1.2.2 The introduction of synthetic DNA oligonucleotides

Currently, a diverse array of FISH assays can be applied to chromosomes at various scales. For instance, assays can be applied to megabase, submegabase, and even smaller resolutions of ≥1 kb for DNA FISH and individual transcripts for RNA FISH transcripts, respectively[54]. Developments in two-dimensional FISH (2D FISH), 3D-FISH, cryo-FISH, and their derivatives have made it possible to visualize DNA segments or whole chromosomes in relation to other foci and chromosomal components, demonstrating features like the chromatin loops and higher order chromatin territories[82–88].

One important technical development has been the introduction of synthetic DNA oligos as probe materials[89]. The advent of low-cost, high-complexity oligonucleotide synthesis, enabled by the proliferation of massively parallel DNA synthesis, has allowed synthetic, rationally and computationally designed oligonucleotides to overcome many limitations of cloning-dependent approaches[5,90]. Oligo probes offer advantages such as specific thermodynamic properties and the ability to contain exogenous sequences serving as secondary labels and readouts to complement the target oligo[56,91,92]. These advantages have led to the introduction of growing sets of ‘spatial genomics’ and ‘spatial transcriptomics’ methods. These methods use complex “probe sets” comprising many distinct oligo species, combined with iterative rounds of secondary hybridization, to visualize numerous genomic regions and RNA species within the same cell or tissue sample[93–98]. Technical innovations in parallel synthesis have enabled microscopy

experiments that can multiplex over a wide range of samples while maintaining target specificity[29,78,80,96,99].

Further developments in super-resolution microscopy approaches provide unprecedented views of chromatin structure across various scales, from whole chromosomes to the observation of cis-regulatory elements[31,100,101]. Multiplexed methods allow for the capture of thousands of potential interactions simultaneously, enabling unbiased image-based discovery over large populations of cells and providing robust statistical power for quantitative validations. These advances offer insights into various relationships between chromatin structure, epigenetics, and the nature of topologically associated domains (TADs) and chromatin subcompartments[9,102–104]. The implementation of tools in this field will provide additional spatial information on the localization of proteins and genomic motifs essential to cell function and nuclear architecture.

1.2.3 Computational tools facilitating the design of oligo probes.

Oligo-based FISH probes have become widely used in the study of animals, plants, and bacteria due to their significant advantage in exploring biological questions. This demand led to the development of computational resources for designing oligo probes at the genomic scale[90]. These resources are designed to identify the best probe candidates for targeting genomic arrays of interest, maximizing probe-to-target number, while ensuring uniformity of hybridization patterns and avoiding off-target binding. These computational tools also offer uniformity in design, enabling high-throughput and multitarget applications. Critical aspects of probe design, such as probe length, melting temperature, and the screening for the formation of secondary structures and dimers, are considered to prevent issues like hairpin loops, which can cause probes to fail to bind. Computational tools also aid in designing probes that can be applied to resolve karyotypic

differences between samples, facilitate super-resolution studies, and validate accuracy of reference genomes[104–106].

Many probe design tools share common themes in their frameworks. Candidate probes are first identified and then screened for specificity to predict any potential off target binding. Specificity screening typically involves using alignment programs like BLAST[107] or Bowtie2[108,109] to search for regions with high sequence similarity to the candidate probes. Alternatively, *k*-mer counting programs like Jellyfish[110] are used to assess whether the candidate probes contain substrings (*k*-mers) with high abundance in the genome of interest. After this specificity screening, candidate probes with predicted off-target binding are filtered out, and a final set of target-specific oligo probes is obtained.

Among the various tools available, PROBER[111], Chorus[112], mathFISH[113], OligoMiner[114], iFISH[115], ProbeDealer[116], and PaintSHOP[117] have provided robust frameworks for diverse FISH applications. PROBER utilizes *k*-mer matching to eliminate probes that overlap with repetitive sequences and addresses the formation of dimers in probe design. It focuses on 100-2000 bp oligo targets in specific chromosome regions and has been used for tiling FISH probe design in humans[111,118]. The mathFISH program specializes in evaluating the predesign of oligo probes using thermodynamic mathematical models to assess the *in silico* performance of a probe sequence along its target sequence. This tool has been used for FISH studies of interphase nuclei during class switch recombination in humans[113]. Chorus and Chorus2 are pipelines that integrate RepeatMasker[119] to deal with repetitive sequences. They use BLAT[120] or BWA[121] to identify single-copy oligos, avoiding repetitive sequences. Single-copy probes are further filtered using Primer3[122] based on melting temperature (T_m) to ensure high target specificity. Chorus2, in its updated version, incorporates *k*-mer based methods to increase probe specificity. Applications of the Chorus2 framework have been used in plant cytogenetic studies and chromosome evolution

research[123–125]. OligoMiner streamlines the process of integrating essential variables in probe design, including probe length and thermodynamic analyses, for diverse FISH experiments that require unique oligo probe parameter profiles. PaintSHOP uses the updated backend OligoMiner scripts and provides a platform for designing genome-scale oligonucleotides and planning multiplexed experiments[114,117]. iFISH serves as a platform for designing FISH probes using a pre-designed genome-wide set of oligo probes from the human genome as a database. It selects a handful of optimal probes over target genomic intervals, resulting in the visualization of chromosome territories and the quantification of chromosome localization in human cells[115]. ProbeDealer designs oligos, while considering thermodynamic data like GC content and secondary structure and performs specificity checks using BLAST. This tool has a GUI and has been utilized for chromatin tracing and RNA FISH experiments[116]. A common theme in these frameworks is the focus on single-copy oligo probes for experiments such as chromosome-specific painting. To design oligo probes fulfilling the requirements of single-copy oligo-FISH, it is crucial to exclude repetitive sequences present in the target genome[90].

Repetitive sequences can cause unwanted background in *in situ* hybridization experiments due to their high copy number[90,98]. Therefore, computational oligo probe design methods typically try to avoid selecting candidate probes in repetitive sequences or filter candidate probes that align many times to the genome or contain highly abundant *k*-mers. In this way, computational approaches can produce repositories of tens of millions of oligo probes for large and complex genomes. However, a substantial fraction of each large genome typically remains uncovered by probes due to the presence of repetitive sequences.

Repetitive intervals can serve as highly robust and effective FISH targets, providing large, bright signals at low cost. Due to the inherent nature of these repetitive targets, oligo probes that target such repetitive sequences can take advantage of these tandem arrays to amplify strong oligo

signal over megabases of similar sequence using just one or a few probe species. While computational tools exist to identify tandem repeat regions and select candidate chromosome-specific imaging oligo probes for experimental validation[126–128], none of these approaches provide a scalable way to assess the predicted *in situ* behavior of oligo probes targeting repetitive DNA in the background of large and complex genomes. Additionally, designing specific probes against individual repeat intervals, such as alpha satellite repetitive intervals with a high degree of similarity between chromosomes, presents a unique challenge. Distinguishing between scaffolds for repetitive DNA studies would be particularly informative for understanding how individual repetitive intervals are organized within the nucleus and may influence genome stability. Computational methods have not yet incorporated frameworks to ensure that selected probe sequences will not lead to off-target binding on repetitive intervals between scaffolds with homologous sequence structure. The development of such resources, which is the focus of Chapter 2 of this thesis, will allow for novel insights into the numerous biological roles that repetitive sequences play in genome organization and stability.

1.3 Improvements in DNA sequencing provide novel insights on repetitive DNA intervals.

1.3.1 The discovery of repetitive DNA elements.

The discovery of mobile transposable elements (TEs) in maize by Barbara McClintock was one of the seminal studies that demonstrated that TEs play a role in the evolution of eukaryotic gene regulation[129]. Her work demonstrated that TEs were responsible for differentially controlling the time and type of activity of individual genes. Such foundational work was also further developed by other researchers in the field, including Britten and Kohne in the late 1960s, who contributed to insights on the complex repetitive nature of metazoan and plant genomes[60]. Distinct classes of repetitive DNA were identified using analyses of reannealing kinetics roughly two decades before the development of DNA sequencing[130]. When DNA is denatured and given the opportunity to

reanneal, single copy DNA anneals more slowly than repetitive DNA because the DNA strand complementary to any given strand is found at low concentrations compared to sequences that are present multiple times[131]. In C_0t analysis, DNA from multicellular eukaryotes anneals in distinct components, where slow annealing single copy DNA is roughly half of the genome and the remaining fractions include middle repetitive dispersed elements and highly repetitive tandem repeats[130,132]. These highly repetitive fractions were also detectable by buoyant density gradient centrifugation of isolated mouse genomic DNA. Using a cesium chloride (CsCl) gradient, DNA bands were identified by base composition. The term “satellite DNA” describes DNA with a base composition that forms a satellite band above (AT-rich DNA) or below (GC-rich DNA) the primary band that contains the bulk of the genomic sample[130].

Both C_0t analysis and CsCl gradient studies would empower repetitive DNA studies in the 1960s and 1970s, because these techniques presented researchers with the ability to selectively fractionate and characterize the complexity and abundance of repetitive DNA sequences. Additionally, further identification of repetitive DNA families’ properties was made possible through dot-blot and Southern blot techniques[133–136]. Using these two methods, detection of shared satellite DNA families between species to establish phylogenetic relationships and patterns of sequence variation within and between species were possible, respectively. *In situ* hybridization techniques also allowed for visual determination of highly repetitive satellite DNA sequences[59]. Including the visualization of mouse major satellite repeats and would empower future studies of repetitive sequences to acquire spatial information about these repeat families[55,59]. The development of polymerase chain reaction (PCR) enabled the isolation of repeats of a satellite DNA family from a species or from related species using primers to uncover variation in sequences between individual species[137]. In concert, the innovation of methods would facilitate the discovery of the locations of most highly repetitive DNA arrays, the lengths of such sequences, and the functional roles of satellite DNAs in some evolutionary contexts[131]. Several other types of

elements were identified, including short and long interspersed elements (SINEs and LINEs)[130,138], pericentromeric and centromeric satellite repeats[139], telomeric repeats[140], retroviral sequences[141] and short tandem repeats (STRs)[142–145]. With the advancement of genomics by next-generation sequencing technologies and bioinformatics tools for repeat analysis (e.g., Tandem Repeat Finder[146], RepeatExplorer[147], TAREAN[148] etc.), the cataloging of the repetitive regions gained momentum, which together with genetic and cell biological approaches has revealed that many repetitive DNA families are involved in distinct cellular functions including mediating gene expression and being involved in dynamic cellular evolutionary processes and speciation[5,149,150].

1.3.2 Distinctions in repetitive DNA families in eukaryotic genomes.

Overall, repeat classes in eukaryotic genomes fall into two categories: 1) interspersed repeat families, including SINE, LINE, and ALU repeats, which occur over short, interspersed intervals within larger blocks of non-repetitive DNA sequence, and (2) long tandem repeats including alpha satellite and human satellite repeat families.

1.3.2.1 Interspersed Repeats

TEs are a prevalent feature in eukaryotic genomes, characterized by their ability to independently replicate and integrate into the host genome[151]. Their influence on chromosome structure and gene content has been well documented, yet the magnitude of their impact remains a subject of ongoing investigation. The presence of a significant proportion of TEs participating in various molecular interactions and regulatory activities within eukaryotic genomes suggests that transposition may have profound implications for shaping evolutionary networks. TEs, which function as genomic parasites, rely on the host cells' machinery to express their genes effectively. To achieve this, TEs have evolved cis-regulatory sequences that closely resemble host promoters. For instance, endogenous retroviruses (ERVs) harbor cis-regulatory promoters and

RNA Polymerase II (Pol II) promoters within their long terminal repeats (LTRs)[152]. In contrast, LINEs, a distinct class of retrotransposons, feature an internal Pol II located in the 5' untranslated region (UTR) along with an antisense promoter[153]. Additionally, short interspersed nuclear elements (SINEs) originate from cellular genes transcribed by Pol III, thus allowing full-length SINEs to efficiently recruit Pol III[153]. These diverse TE families employ various mechanisms, such as target-primed reverse transcription and the utilization of LINE replication machinery, to facilitate their propagation and mobility within the genome.

1.3.2.2 Long Tandem Repeats

Centromeres are crucial sites on chromosomes where spindles attach during cell division. They play a role in separating replicated chromosomes during mitosis and meiosis. The process is conserved across diverse species, yet centromeric satellites have been shown to evolve rapidly and contain vast variations even among closely related species[154]. In primates, centromeres are made up of alpha satellites, which are abundant repeats in the human genome that consist of a 171-bp monomer subunit and can span megabases[155]. There are twenty suprachromosomal families of alpha satellites, where each is organized based on their sequence similarity[156,157]. The nucleotide sequence identities of alpha satellites can vary between 70% and 90% when comparing monomeric repeats of the same unit[156]. The similarity is higher than 95% when a monomeric unit is compared to a counterpart monomer (found at the same position) in another multiple unit[156]. Different chromosome-specific alpha satellite subfamilies have been observed in humans, and they are characterized by the higher order repeat (HOR) structure. This monomeric subunit has been shown to have experienced a number of sequence and structural variations throughout human evolution, as it has been shown that old and new world monkeys and gibbons lack chromosome-specific subfamilies[154,157]. Human centromeres are mainly composed of six classes of repetitive DNA: α-satellite, β-satellite, G-satellite, and three shorter motifs termed HSATI, HSATII, and HSATIII[156].

In mice, chromosomes have centromeric satellite arrays which are located near the proximal chromosomal ends. In *Mus musculus*, centromeres are relatively homogenous within the species and consist of 120-bp minor satellite (MiSat) arrays[158]. Centromeric satellites MS3 and MS4 also colocalize with MiSats[159]. *M.spretus*, which diverged from *M.musculus* around 1-2 million years ago, have retained some MiSats distributed throughout the centromeric and pericentromeric domains[160,161]. These differences in satellite repeats suggest variations in centromere evolution between different mouse species. Similarly, satellite repeat variation has been observed in *Drosophila* as a key marker of speciation in early evolutionary diversion events[150]. Plant centromeric regions are often occupied by retroelements belonging to the Ty3/gypsy endogenous retrovirus family[162,163]. These retroelements contain integrases with chromodomains which allow them to integrate properly into centromeric regions. This has been observed in maize with the CentC satellite repeat unique and the pAL1 repeat unit in *Arabidopsis thaliana*, respectively[164–166].

Pericentromeric regions border centromeres and are shown to be crucial for preventing the premature separation of sister chromatids. They are abundantly made up of pericentromeric satellites, and in interphase nuclei these satellites are found organized as chromocenters, which are observed broadly in eukaryotes, including mammals, flowering plants, and *Drosophila*[59,167–169]. Chromocenters are highly coiled genomic regions within the cell nucleus and are distinctive structures involved in the clustering of heterochromatin[170]. Proper chromocenter formation is essential for the maintenance of DNA damage repair, and gene regulation in *Drosophila*[169].

Telomeric repeats are mostly microsatellites, while subtelomeric repeats are classified as satellites. DNA polymerases replicate DNA in a 5-to-3' direction and cannot fully replicate the ends of chromosomes, causing telomeres to shorten with each replication. Telomerase, found in most eukaryotes, remedies this shortening by adding new telomeric repeats[171–174]. Loss of

telomere maintenance can lead to human diseases like bone marrow failure, cancer, and premature aging[175]. Telomerase is active in both germline and stem cells in humans[171].

1.3.3 The telomere-to-telomere era of genome sequencing.

Despite our advancing understanding of the role of linear sequence variation in many repetitive DNA families, satellite DNAs have remained significantly understudied across taxa. This is primarily due to the limitations of widely used sequencing technologies and software tools in exploring variation[5]. For instance, while Illumina sequencing provides data with low error rates at a low cost, the short-read library preparation techniques involve a PCR step that becomes impractical for large sequences with extreme GC content[176]. Additionally, when short read sequencing is used on highly repetitive portions of genomes, there is significant ambiguity of mapping short reads to nearly identical genomic reads that can occur at multiple locations across chromosomes. Consequently, this hinders the inclusion of large repetitive regions in genome assemblies due to read length and biased underrepresentation of satellite DNA. Additionally, many computational tools used to study variation in both repetitive and non-repetitive DNA rely on reference assemblies to understand differences in sequence structure. Therefore, while computational tools such as Tandem Repeat Finder[146] and RepeatMasker[119] offer opportunities to examine motifs present in highly repetitive DNA sequences, this information solely isn't enough to understand satellite DNA variation without a comprehensive genome assembly paired with additional genomic validation strategies.

The development of long-read sequencing technologies[177] in tandem with repeat assembly methods has enabled complete generation and assemblies of diverse repetitive DNAs, including human centromeric HOR arrays that were previously included in incomplete forms in genome builds. This progress is marked by the availability of long reads with high consensus base quality, known as high-fidelity (HiFi) sequencing data from Pacific Biosciences[178], and ultralong data

reads from nanopore sequencing by Oxford Nanopore Technologies[179]. Simultaneously, the improvement of automated satellite DNA assembly and validation techniques, coupled with molecular biology innovations in pulsed-field gel electrophoresis and Southern blotting, has led to the curation of telomere-to-telomere (T2T) genome assemblies[50,180]. Notably, the current T2T reference genome was derived from an effectively haploid cell line originating from a complete hydatidiform mole (CHM13hTERT or CHM13)[154].

Approximately 54% of the T2T-CHM13v2 genome comprises diverse repetitive DNA elements. Within this genome sequence, the largest class of repetitive elements is composed of interspersed repeats which are hypothesized to be present at diverse frequencies due to distinct genomic integration mechanisms[5,181]. Non-mobile DNA sequences, including pericentromeric and centromeric satellite repeats, STRs, and VNTRs, occur at high frequency in the human genome[182]. As our knowledge of the complexity and diversity of repetitive DNA elements increases among individuals, it is likely to provide valuable insights into the functions and roles of the repetitive genome in phenotypic variation in the coming years. The recent releases of the first complete human genome assembly offer rich opportunities to evaluate complete alpha satellite assembly variation and confirm the expectations of original models of centromeric organization and structural variation[50]. Current findings, derived from comparing variants directly adjacent to human centromere arrays or within some centromere-spanning haplotypes have been observed from several clinical studies. Further efforts to expand variant maps in centromeric regions are challenging, even with emerging releases of high-quality reference maps and will require new methods to identify and describe candidate disease causal variants that are predicted to be associated with satellite DNAs[154,182–184]. Current evidence has also suggested that alpha satellite arrays evolve through layered expansions, with the inner kinetochore protein CENP-A tending to associate with the most recently evolved sequences, thus shifting the kinetochore to new loci, where old loci shift and decay[154].

However, because centromeric satellite repeat copy number and sequence variants are expected to vary considerably, the findings described within CHM13 do not offer a comprehensive view of the extent of sequence and epigenetic variation in the population[154,156]. Current evidence points toward signatures of positive selection on select chromosomes[185]. Human diversity cohorts, such as the 1000 Genomes Collection[186], have previously demonstrated substantial variation in alpha satellite array lengths within the population, differing by a factor of 5–10, and even among homologous chromosomes within the same individual[156]. Furthermore, cytogenetic studies also suggest that such variation may contribute to predispositions to cancer, infertility, and chromosomal aneuploidies[49,154]. Therefore, an extensive analysis of alpha satellite arrays may provide additional insights regarding how these regions evolve over time and how their variation shapes patterns of inner kinetochore proteins and elements involved in centromere stability.

1.3.4 Repetitive DNAs are central to a diverse set of cellular functions.

1.3.4.1 Repetitive DNA elements in human disease

TEs have been well documented to cause disease through two core mechanisms: insertional mutagenesis and chromosomal rearrangements[187]. Specifically*, de novo* germline TE insertions that disrupt normal gene function have been implicated in over one hundred inherited diseases[151]. The transposition and TE mediated chromosomal rearrangements have also been observed in several forms of cancers[188]. Similarly, satellite DNA elements offer fragile sites that are particularly sensitive to variation. Karyotypic abnormalities are often a hallmark of cancer cells, which includes increased rates of aneuploidy, genomic rearrangements, deletions, fragmentations, and duplications in the genome[189]. Chromosome instability is often correlated with the tumor severity of cancer prognosis. Human centromeres often serve as sites where aberrant genomic rearrangements are observed. Despite the fact that distinct rearrangements in cancer have been

observed, the mechanisms of this centromeric variation and malignancy are not well understood[189–191].

Karyotypic abnormalities are a hallmark feature of many cancer cells. These abnormalities include increased rates of aneuploidy, rearrangements, deletions, fragmentations, and duplications[192]. It has been shown that the severity of such chromosome instability is correlated with a poor prognosis[193]. Human centromeres have been demonstrated to be sites of aberrant rearrangements in many tumors, and this chromosome instability has also been attributed to centromeric fission[190,191]. Furthermore, studies have shown an increased rate of centromeric recombination and other abnormalities compared to healthy cells. Although the frequency of centromeric rearrangements in cancer is appreciated, the mechanism of malignancy is poorly understood[189,194]. It is hypothesized that perturbation of the epigenetic landscape at select genomic sites may interfere with transcriptional dysregulation[195]. Rearrangements within centromeric DNA may disrupt local heterochromatin and contribute to long-range changes in general gene expression[189]. This further suggests that changes in the chromatin state of the region also impact aberrant rates of transcription and malignant transformation. Another facet of this dysregulation includes the observed overexpression of CENP-A across tumors. The increase in CENP-A often correlates with cancer severity, and several studies have observed that CENP-A overexpression can lead to its incorporation within ectopic interstitial loci outside of the alpha satellites arrays[196–199].

As high quality reference maps continue to be produced as HPRC[200] and T2T efforts allow for the completion of additional sequenced genomes, new methods must be developed to identify and describe causal variants that are associated with satellite DNA and diverse TEs across human disease. Such efforts will help improve our understanding of disease-associated variants and their relationship with genome instability.

1.3.4.2 Repetitive DNA elements in evolution

Repetitive DNA sequences play a crucial role in generating novel genic and regulatory elements, and they are believed to be involved in evolutionary processes such as meiotic drive and speciation. However, the extent of their involvement in these processes is still being widely explored. For example, transposons in the human genome have been inactive for the last 500 million years, but some retrotransposable elements, including LINE and SINE repeats, remain active in the pericentromere of most human chromosomes[201,202]. Additionally, centromeric elements might have origins in retrotransposable elements, like the CENP-B box and CENP-B protein[202]. These elements have been identified in the centromeric region and even within core HOR sequences and among epialleles.

Satellite DNA arrays have been found to undergo high evolutionary turnover due to stochastic events and selective pressures[155,165,189]. The rapid evolution of centromeric satellites has led to the emergence of species-specific satellite sequences, which have been useful in identifying species phylogeny[157,203]. Comparative studies between human and nonhuman genomes reveal that centromeric satellites evolve at an accelerated pace due to various mutational processes, including concerted gene evolution, saltatory amplification, unequal crossover, and non-allelic homologous recombination[131,155,182,187,204,205]. The dynamic variation in both sequence and copy number among species, even among close relatives, is striking, highlighting the rapid evolution of these genomic repeats in the population[150,206]. However, testing these models on a genome-wide scale has been challenging due to technical and genome assembly limitations in assessing repetitive regions properly.

The library model for satellite DNA evolution suggests that closely related species should share conserved satellite DNA families inherited from a common ancestor[207,208]. However, each species

amplifies these satellite DNA families differently. This conservation of sequence has been observed in several species over long evolutionary periods, indicating that low-copy number satellites are dispersed among large arrays of major satellites throughout the heterochromatic block[207]. The library model has been confirmed across various organisms, including plants, mammals, nematodes, and insects[204,209–211]. Satellite DNA repeats are thought to be generated and amplified through mechanisms involving activity from transposable elements and replication of extrachromosomal tandem repeat circles via rolling-circle replication and reinsertion into the genome through unequal crossing over[131,212].

Another proposed model involves the concerted evolution of satellite DNA. The content of satellite DNA can vary significantly between species, even those closely related. This divergence has been observed in speciation events in *Drosophila*, where some satellite families are shared by several species in a genus[213]. It has been found that rapidly evolving gene pairs can contribute to hybrid failure and speciation, known as Bateson–Dobzhansky–Muller incompatibilities[214]. In concerted evolution, monomers of a satellite DNA family can become homogenized within a species, resulting in identical tandem repeats, as seen in mouse centromeres, or higher-order repeats[215].

Other notable models propose mechanisms for the evolution of satellite DNAs. For instance, Crow and Kimura hypothesized that the distribution of repeat numbers per genome should reach an equilibrium point, stabilized via unequal crossing over and stabilizing selection for optimal array size[216]. Smith's computational modeling approaches suggest that DNA sequences not constrained by selection tend to become repetitive over time through unequal crossover processes[217]. On the other hand, simulations of repeated out-of-register crossover between sister chromatids suggest that non-repetitive sequences will eventually exhibit satellite periodicity[217].

While molecular drive may contribute to satellite DNA evolution under relaxed selection, centromeres remain vulnerable to intense selection due to their critical role in centromere integrity and proper cell function. The centromere drive model was proposed to explain the rapid evolution of centromeric DNA sequences and the essentiality of centromere proteins[218,219]. In female mouse meiosis, only one of the four gametic products is included in the egg, while the other three are lost to the polar body[219,220]. This results in the stronger centromere variant preferentially segregating to the egg pole, and centromere proteins evolve to suppress centromere drive.

As additional genome assemblies are completed, researchers will have a better opportunity to understand the dynamic process of satellite DNA evolution, and population-wide validation of proposed evolutionary models may be examined.

1.3.4.3 Repetitive DNA elements in 3D genome organization

Crosstalk between repetitive DNA and heterochromatin is also emerging in the 3D genome organizational context. While centromeric chromatin is relatively accessible to facilitate kinetochore assembly, satellite repeats found within pericentromeres are constitutively heterochromatinized by histone H3K9me2/3, Hp1, and CpG DNA methylation[221]. Additionally, while pericentromeric heterochromatin is targeted for constitutive repression, it has been shown that non-coding RNAs mediate deposition of H3K9me3 at these pericentromeric repeats[222]. These findings have led to several working models showing how the function of heterochromatin at pericentromeres influences the maintenance of global genome stability.

Although such repetitive element associations have been well established in linear chromatin fiber, new insights are being established with respect to heterochromatin and its involvement in higher-order folding patterns of chromatin. Studies exploring the activation of DNA repair pathways within clustered pericentromeric repeats in mice have suggested that the spatial

positioning of genomic loci with double strand breaks outside of chromocenters promotes repair of pericentromeric repeats, thereby highlighting the interplay of higher-order folding of heterochromatin with the repetitive genome[223]. Repeat expansion disorders exhibit unstable STR expansions, but the specific properties of the repeat tracts vary across diseases, including the repeat unit sequence and length[224]. In fragile X syndrome, STR expansion disrupts TAD boundaries, where STRs colocalize at TAD boundaries[142,143]. These data suggest a link between STR instability events and molecular features at some TAD boundaries, but it remains uncertain whether instability drives genome misfolding or vice versa.

Another notable example of the role of repetitive DNAs in directly mediating valuable cellular structures and functions is the discovery of nucleolar organizer regions (NORs). The nucleolus was first identified in the 19th century, and later studies in the 20th century uncovered how this cellular feature relates chromosome organization[225]. *In situ* hybridization demonstrated that NORs are the sites of ribosomal genes on acrocentric chromosomes[226]. These NORs contain tandem rDNA arrays that are responsible for composing the nucleolus, which serves as a site for the synthetic processing of ribosomal proteins and rRNA[227]. Variations in the localization and transcriptional activity of NORs have been well documented in many species[226,228,229]. NOR chromosomes are also commonly involved in Robertsonian translocations, and such breakages may serve as a hallmark of acrocentric associations[228,230].

With respect to interspersed repeats, more than 95% of mammalian retroelement loop anchor CTCF sites originate from SINEs, LINEs, and LTRs[231]. CTCF is recognized as a core architectural protein and is well recognized as a transcriptional factor[4]. Additionally, in mice, SINEs are enriched with CTCF motifs. Current models propose that SINEs and LINEs may influence the folding of the human genome by redistributing CTCF binding sites, which play a crucial role in maintaining the higher-order folding of the genome[5,231]. These models suggests that the

mobilization of SINE sequences during evolution may have significant effects on genome folding patterns in a CTCF-dependent manner.

1.4 Unraveling Structure-Function Relationships of Repetitive DNA Intervals through FISH

1.4.1 Heterochromatin preserves genome stability by repressing repetitive sequences.

Advances in our understanding of repetitive DNA and heterochromatin have shaped our understanding of the role of repetitive DNA families in 3D genome organization. Since the initial discovery of heterochromatin, key principles of heterochromatin function and maintenance have been associated with diverse repetitive DNA families[16]. Heterochromatin has been shown to play a critical role in preserving genome stability through the repression of repetitive DNA elements and ensuring that proper chromosome condensation prior to mitosis occurs[5,43,232]. This form of targeted repression is crucial to countering the propensity for instability events that repetitive DNA elements are prone to, including stepwise expansions, duplications, inversions, and recombination[5]. Likewise, distinct heterochromatin signatures demonstrate context-specific constitutive and developmentally regulated patterns of maintenance across species, time, and space. Many crucial questions remain regarding which mechanisms govern the previously described deposition and maintenance of H3K9me3 heterochromatin modifications and how such evolutionarily conserved features of heterochromatin play distinct roles in spatiotemporal regulation of heterochromatin in mammalian development[9].

Given the non-random distribution of repetitive DNA elements along the genome, the structural relationship of repetitive DNA intervals must be elucidated to discern key players and genome organizers involved in a diverse array of biological phenomena, where the organization of these genomic regions contains constitutive heterochromatin. Such studies have the potential to

elucidate mechanisms between types of repetitive DNA interval arrays, including distinct satellite DNA intervals that occupy specific genome loci that mediate crucial genetic functions for centromere maintenance and stability. Likewise, cross-functional studies involving other technologies used to validate chromatin contacts or identify relevant protein-protein interactions at the scale of genomes may be used in parallel to provide a more holistic understanding of the spatial and biological factors associated with such crucial cellular processes associated with the preservation of the 3D genome's organization.

1.4.2 Introducing tools to unravel genome organizing mechanisms of repetitive DNAs.

In Chapter 2, I describe the development a computational tool that enables the design of oligo probes to target abundantly repetitive DNA intervals. I present Tigerfish[233], a computational tool for designing oligos with high specificity for target repeat intervals, the ability to target evolutionarily similar repeat intervals of interest, and the ability to provide a robust predictive framework for quantitative *in silico* binding predictions that evaluates genome oligo probe binding behavior. I use Tigerfish to design a genome-wide panel of oligo probes which are imaged targeting all human chromosomes in metaphase. My supplementary datasets contain oligos unique to repetitive DNA intervals throughout the human genome. Tigerfish may also be applied broadly to diverse genome builds and across species. As genomic assemblies improve in quality and precision through efforts including the T2T and HPRC, the accessibility of oligo datasets unique to individuals and species will be essential toward future studies seeking to better understand the spatial relationships of heterochromatin-rich repetitive DNA intervals. In Chapter 4, I introduce FISHTank, a proposed framework for a database containing repetitive DNA-specific oligo sequences across individuals and model organism genomes. FISHTank, once implemented, will empower diverse studies that explore biological questions associated with the roles of target repetitive DNA intervals.

1.5 Genomics research and its implications outside of the lab

1.5.1 A call for equitable science communication resources that discuss genomics and society.

The significance of precise genome assemblies and their implications in understanding repetitive DNA arrays have been the focus of my research for the past five years as a biologist specializing in this field and as a technology developer. Through the advent of 'telomere-to-telomere' genome assemblies, a new frontier in genomics research has emerged. Access to more centromere assemblies presents opportunities to explore genetic rearrangements of centromeric and heterochromatic satellite DNAs in aging and cancers, shedding light on the epigenetic and transcriptional landscape of human centromeric regions. However, understanding the diverse biological phenomena observed in centromeres at the population level requires datasets beyond the single CHM13 reference genome. Multi-modal sequencing across diverse populations and generations can provide invaluable insights into the variation of human satellite DNA and its impact on disease and genome stability.

As we pursue diverse genomic datasets, ethical considerations become paramount. The commodification of data and the lack of proper consent protocols can perpetuate historical injustices and marginalize underrepresented populations. Efforts to diversify genomic databases must empower communities to govern their data, ensuring transparency and fair usage. Initiatives like the "All of Us"[234] research program aim to foster diversity and equity in advancing precision medicine, but potential exploitation of data from underrepresented populations raises concerns about equitable benefits and protections.

Additionally, geneticists must address the misuse of scientific research by far-right groups to propagate racist ideologies[235]. By unequivocally rejecting the concept of biological distinctions

among "races" and ensuring comprehensive data collection that includes underrepresented populations, researchers can combat the resurgence of scientific racism[236]. The representation of marginalized scholars is vital, but it comes with a cost, and academia must foster accessible, equitable, and supportive environments to empower such individuals as influential leaders.

1.5.2 Introducing curricular and community resources toward inclusive academic climates.

In the context of my academic journey, I have sought to create inclusive and supportive community spaces within genomics research. Through my research and advocacy, I aim to shed light on the challenges faced by scholars with marginalized identities. The lack of formal training in addressing race, gender, sexuality, and other intersecting identities within genomics necessitates efforts to empower researchers and diversify genomics databases.

In Chapter 4, I provide a comprehensive account of my work in bioscience advocacy, including the historical archive of the Genome Sciences Association for the Inclusion of Marginalized Students (GSAIMS) and a curriculum focused on how identity and belonging shape outcomes in bioscience training programs. By sharing my experiences and knowledge, I aim to contribute to more equitable practices in academia, empowering researchers to combat scientific racism and promote fair and just applications of genomics research in society.

Chapter 2: Implementation of computational methods to visualize repetitive DNA intervals in the human genome.

This thesis chapter describes work submitted to bioRxiv describing the Tigerfish[233] software.

Aguilar, R., Camplisson, C.K., Lin, Q., Miga, K.H., Noble, W.S. and Beliveau, B.J., 2023. Tigerfish designs oligonucleotide-based in situ hybridization probes targeting intervals of highly repetitive DNA at the scale of genomes. bioRxiv, pp.2023-03.

My role in this work was developing the software concept and implementation alongside BJB and WSN. In addition to developing and testing the pipeline, I was responsible for performing experimental and computational validation of oligo probes developed by Tigerfish. I also performed the image analysis and paper construction for this project. CKC assisted with experimental PaintSHOP validation. KHM and QL provided valuable directions on software feedback.

2.1 Abstract

Fluorescent *in situ* hybridization (FISH) is a powerful method for the targeted visualization of nucleic acids in their native contexts. Recent technological advances have leveraged computationally designed oligonucleotide (oligo) probes to interrogate >100 distinct targets in the same sample, pushing the boundaries of FISH-based assays. However, even in the most highly multiplexed experiments, repetitive DNA regions are typically not included as targets, as the computational design of specific probes against such regions presents significant technical challenges. Consequently, many open questions remain about the organization and function of highly repetitive sequences. Here, we introduce Tigerfish, a software tool for the genome-scale design of oligo probes against repetitive DNA intervals. We showcase Tigerfish by designing a panel of 24 interval-specific repeat probes specific to each of the 24 human chromosomes and imaging this panel on metaphase spreads and in interphase nuclei. Tigerfish extends the powerful toolkit of oligo-based FISH to highly repetitive DNA.

2.2 Introduction

Fluorescent *in situ* hybridization (FISH) is a powerful technique that can reveal the spatial positioning and abundance of DNA and RNA molecules in fixed samples with subcellular resolution. Since their introduction in 1969[55], ISH and later FISH[63,64,237] methods have been

refined to improve their detection efficiency and sensitivity[238]. One important technical development has been the introduction of synthetic DNA oligonucleotides (oligos) as a source of probe material[239]. Oligo-based probes offer important advantages over more traditional probes deriving from isolated genomic material, as oligo probes can be designed to have specific thermodynamic properties and programmed to contain stretches of exogenous sequences that can serve as 'readout' domains via the 'secondary' hybridization of a labeled, complementary oligo. These advantages have led to the introduction of a growing set of 'spatial genomics' and 'spatial transcriptomics' methods that use complex 'probe sets' of many distinct oligo species[95,97,98,240] in combination with iterative rounds of secondary hybridization to visualize dozens or more genomic regions[241–244] and thousands or more RNA species[245–247], respectively, in the same cell or tissue sample.

The rapid adoption of oligo probes as a source of FISH probe material has also catalyzed the parallel development of computational tools for oligo probe design. These tools—which include OligoArray[248], PROBER[111], Chorus[249], mathFISH[113], OligoMiner[114], iFISH[115], ProbeDealer[116], Chorus2[112], and PaintSHOP[250]—aim to identify short windows of genomic sequence that have suitable thermodynamic and sequence properties to serve as FISH probes. Once identified, 'candidate' probes are next screened for specificity to predict whether they will have off-target sites in addition to their intended target. This specificity screening typically relies on using alignment programs such as BLAST[107] or Bowtie2[109] to search for regions with high sequence similarity to the candidate probes, the use of *k*-mer counting programs such as Jellyfish[110] to assess whether the candidate probes contain *k*-mers (*i.e.*, substrings) with high abundance in the genome of interest, or a combination of both approaches. After this specificity screening, candidate probes with predicted off-target binding are filtered and a final set of target-specific oligo probes is returned.

A key advantage of oligo probes is that they can be designed specifically to avoid targeting repetitive sequences. Repetitive sequences are frequent sources of unwanted background when performing *in situ* hybridization experiments due to their high copy number, and a set of "suppressive hybridization" methods using unlabeled repetitive DNA from the *C_0t-1* fraction[61] as a blocking agent have been introduced to abrogate this background when using probes derived directly from genomic material[57,251,252]. Such blocking agents are generally not needed when using oligo probes, however, as computational oligo probe design methods either avoid discovering candidate probes in sequence annotated as being repetitive by tools like RepeatMasker[111,114,248–250,253] or purposefully filter candidate probes that align many times to the genome[112–116,248–250] or contain highly abundant *k*-mers[112,114,115,250]. As a result, while computational oligo probe design tools are able to operate at the scale of whole plant and mammalian genomes to produce repositories of tens of millions of oligo probes[115,250], a substantial fraction of large and complex genomes remains intentionally uncovered due to the presence of repetitive sequences.

Repetitive DNA accounts for ~50% of the human and mouse genomes and often even higher percentages in the genomes of plants[50,61,254]. Broadly, repetitive DNA falls into two categories: 1) Interspersed repeats such as SINE, LINE, and ALU elements that often occur as short, spatially isolated intervals within larger blocks of non-repetitive sequence[254]; 2) long tandem repeat arrays such as alpha satellite, human satellites 1–3, and the 45S ribosomal DNA at which a single monomer is repeated many times to form multi-megabase intervals of repetitive sequence that are frequently located in pericentromeric regions and on the short arms of acrocentric chromosomes[50,182]. Collectively, repetitive DNA sequences are central to a set of diverse and essential cellular and organismal functions, including the recruitment of the chromosome segregation machinery during mitosis, the encoding of essential information such as the 47S rRNA[255] and the replication-dependent histone genes[256], and the protection of chromosome ends[56]. Moreover, repetitive sequences are an important source of novel genic and regulatory

sequences[257] and are hypothesized to be actively involved in potent evolutionary processes such as meiotic drive and speciation[258]. Thus, more detailed studies of highly repetitive DNA regions and their transcription products through targeted assays such as FISH may help uncover the mechanisms by which these mysterious regions exert their influence on important biological processes.

When desired, repetitive intervals make highly robust and effective FISH targets, as one or a few probe species can bind many times and thus produce a very large, bright signal at low cost. Indeed, all of the initial ISH targets were repetitive[55,167], and repetitive targets continue to be used routinely for diagnostic assays such as aneuploidy detection via interphase chromosome enumeration[259]. However, the deployment of probes against repetitive targets either requires the isolation and experimental validation of cloned genomic material or *a priori* knowledge of experimentally validated oligo sequences. Computational approaches have been introduced to identify tandem repeat regions in worm[126] and plant systems[127,128] to select candidate chromosome-specific imaging oligo probes for experimental validation. However, neither these approaches nor computational tools designed to target non-repetitive regions provide a computationally scalable way to assess the predicted *in situ* behavior of oligo probes targeting repetitive DNA in the background of large and complex genomes.

Here, we introduce Tigerfish, a computational ecosystem tailored for the design and characterization of oligo probes targeting intervals of repetitive DNA at the genome scale. Tigerfish provides all functionality needed for discovering repetitive regions *de novo*, designing candidate probes, and performing deep *in silico* profiling of predicted binding activity. Tigerfish is open source, freely available, supported by extensive documentation and tutorials, and ships with a dedicated set of utilities to make it easier for users to visualize the predicted experimental outcomes of their designs. We showcase the utility of Tigerfish by designing and experimentally

validating at least one interval-specific repeat probe for all 24 human chromosomes on metaphase spreads and augment these data by performing interphase enumeration of chromosomal copy number in human primary lymphocytes for all 24 human chromosomes. Finally, we provide a comprehensive catalog of probes and their predicted associated binding specificities that have been discovered by Tigerfish in the fully assembled human CHM13 genome released by the Telomere-to-Telomere Consortium[50]. As our knowledge of the complete sequence of highly repetitive regions and how these regions vary amongst individuals and populations continues to increase from efforts such as the Human Pangenome Project[260] and Vertebrate Genomes Project[261], we anticipate that Tigerfish will play a key role in a number of applications including genome assembly variation, *in situ* karyotyping, and biological discovery.

2.3 Results

2.3.1 Oligo probe design with Tigerfish

Tigerfish is a computational pipeline composed of a collection of Python scripts embedded in an automated Snakemake workflow[262] and is designed to be executed in a command line environment. No direct knowledge of programming is required to run Tigerfish, and this bioinformatic workflow can be deployed on any modern Windows, Macintosh, or Linux system. Tigerfish is open-source, freely available via GitHub (https://github.com/beliveau-lab/TigerFISH), and depends on Bowtie2[109], NUPACK[263], Jellyfish,[110] SamTools[264], Biopython[265], Scikit-learn[266], and chromoMap[267]. Tigerfish is also supported by extensive documentation (https://beliveau-lab-tigerfish.readthedocs-hosted.com). In order to run Tigerfish, users must include the full sequence of the genome assembly in which probe design is to be performed in FASTA format[268] and also provide an accompanying “chrom.sizes” file that details the scaffolds present in the assembly and their lengths in base pairs. Users must also edit a small configuration file in which the locations of

relevant files and scripts can be specified and parameter choices for the probe discovery can be specified.

Tigerfish can be run in one of three execution modes (**Fig. 1**). In the first, termed “Repeat Discovery Mode”, users list genomic scaffolds where *de novo* repeat discovery and probe design is to be performed in the configuration file. Repeat Discovery Mode uses a *k*-mer counting strategy to identify repetitive DNA regions *de novo* by identifying intervals that contain *k*-mers with high abundance in the genome (**Methods**). Users can tune the size of the search window and the magnitude of the *k*-mer count values needed for an interval to be flagged as repetitive, thereby controlling the nature of the repeat regions identified. Tigerfish may also be run in “Probe Design Mode” in instances where the genomic interval(s) a user wants to target for probe design are already known. In this case, the user must provide an additional BED-formatted file[269] that specifies the genomic coordinates for interval(s) to perform probe design against. Lastly, “Probe Analysis Mode” generates a new set of *in silico* binding predictions for probes contained in an existing Tigerfish output file. Tutorials providing a comprehensive walkthrough of these three modes, along with an example of implementing Tigerfish in the human CHM13 genome on a satellite repeat, can be found at https://beliveau-lab-tigerfish.readthedocs-hosted.com.

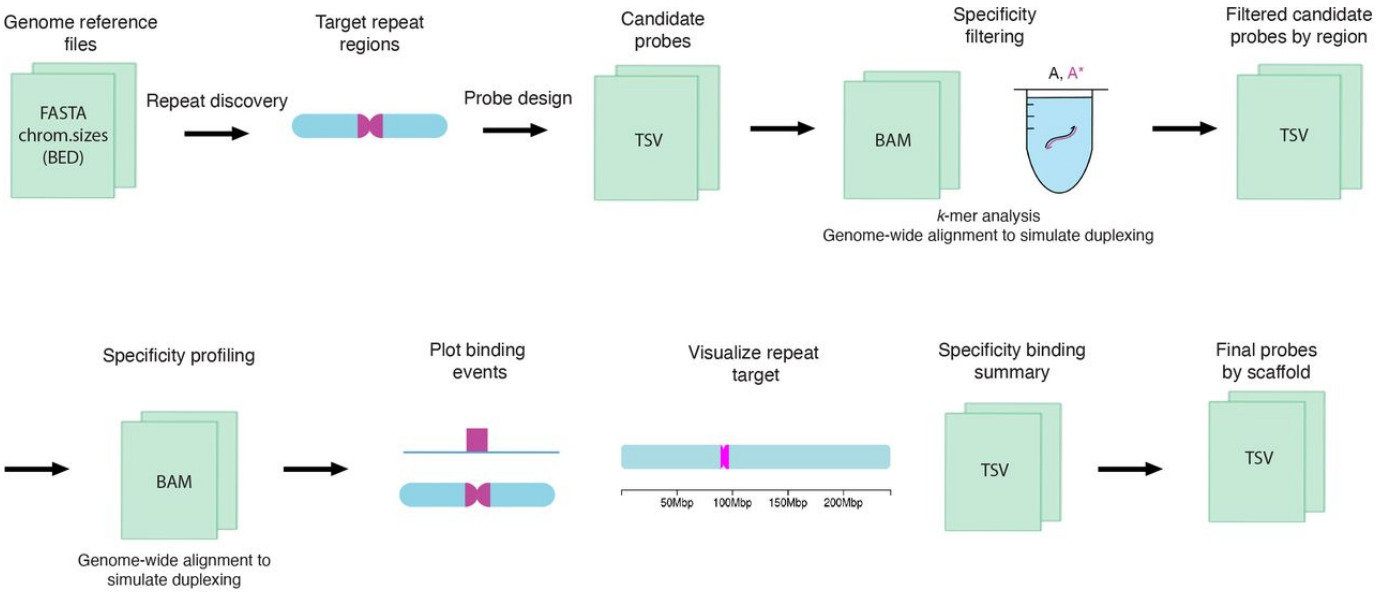


Fig. 1. The Tigerfish workflow. Schematic overview of the inputs, major processing steps, and outputs of the Tigerfish probe design pipeline.

When using Repeat Discovery Mode or Probe Design Mode, Tigerfish designs candidate oligo probes for each genomic interval passed forward (Repeat Discovery Mode) or specified in the user-provided BED (Probe Design Mode). Candidate probe discovery is performed using a modified version of the 'blockParse.py' script from OligoMiner[114] that screens the provided sequences for windows with desirable sequence and thermodynamic properties (**Methods**). To maximize the chance that the optimal probe or set of probes will be identified, Tigerfish mines the entire repeat region for candidate probes, which can result in redundant and even duplicate candidate probe sequences being returned. To minimize the amount of downstream computation needed, duplicates are removed and the candidate probes for each region are then rank-ordered to prioritize candidates that contain *k*-mers with elevated abundance specifically in the target interval from which they were designed (**Methods**), as such candidates are more likely to have many on-target binding sites while having minimal binding elsewhere in the genome.

To return a final probe set, Tigerfish begins with the top-ranked candidate probe for each target interval and performs deep *in silico* specificity profiling. The selected candidate probe is aligned to the genome with very sensitive settings (**Methods**) and up to 500,000 alignments are returned. The genomic sequence of each alignment site is then extracted and put into a virtual test tube to simulate how likely binding would be with the input candidate probe in FISH conditions using NUPACK[263]. Finally, Tigerfish processes the result of these simulations and calculates the number of predicted on- and off-target binding sites for each candidate probe (**Methods**). Users can specify several parameters to tune performance at this step, including the maximum number of allowed off-target binding sites per probe, the minimum number of required on-target binding sites per probe, and the maximum number of probes in the final set (**Methods**, **Supplementary Note 1**). If needed, Tigerfish will continue analyzing the predicted binding specificities of candidates from the rank-ordered list until either the user-supplied criteria are met, or all possible candidate probes are considered. The final output of Tigerfish includes a text file containing all final probes and their aggregate on- and off-target binding predictions, a summary table that lists all target intervals for which probes were designed and their aggregate on- and off-target binding predictions for the probes that map to each interval, and a set of auxiliary files that provide more detailed information about the predicted binding profiles of the probes. Users can also optionally populate chromoMap ideograms that depict the chromosomal locations of probe binding for the probe or set of probes designed against each target interval (**Fig. 1**). Example input and output files for full test runs of Tigerfish in Repeat Discovery Mode, Probe Design Mode, and Probe Analysis Mode can be found within **Supplementary Software**.

2.3.2 Probe discovery at the scale of human genomes

In order to demonstrate the scalability of Tigerfish, we set out to perform genome-wide *de novo* repeat interval identification and probe design for all 24 chromosomes in the human telomere-to-telomere CHM13v2 + HG002 chrY assembly[50] using Repeat Discovery Mode. In order to

showcase how users can tune parameters to optimize their design for different types of repeat regions, we performed our genome-scale runs with two sets of parameter groupings: 1) a 'conservative' set that prioritizes identifying large intervals of highly repetitive sequence such as those found at pericentromeres; 2) a 'permissive' set that aims to discover smaller, interspersed intervals of repetitive DNA (**Supplementary Data 1**) in addition to larger intervals found in the 'conservative' set (**Supplementary Data 2**). The genome-wide probe design runs designed probe sets for 263 intervals, of which 235 intervals were only identified with the 'permissive' parameter settings and 28 intervals were identified by both parameter settings. We found that Tigerfish was able to generate at least one interval-specific probe or probe set for all 24 chromosomes, prominently covering the pericentromeric and subtelomeric regions of most chromosomes. The Tigerfish probes mostly fell into regions not already covered by existing PaintSHOP probes[250] designed with non-repetitive intervals in mind (**Fig. 2a**) and predominantly mapped to annotated satellite and simple repeat regions (**Fig. 2b**). We found that the repeat intervals identified spanned a broad range of sizes ranging from 411 bp – 34.3 Mb (median: 3.6 kb) for the group identified using the 'permissive' settings and from 37.6 kb – 34.2 Mb (median: 2.7 Mb) for the group identified using the 'conservative' settings (**Fig. 2c**). Collectively, these probes and probe sets cover 164.5 Mb of the human T2T CHM13v2 HG002 chrY assembly after accounting for any differences between the size of the interval inputted for design and the effective size of the interval covered by the output probes (**Supplementary Fig. 1**). Our *in silico* specificity profiling also revealed a broad distribution of predicted binding activities for the probes or probe sets covering the 263 intervals, ranging from 25–30,972 target sites in the 'permissive' group (median: 236.9 target sites) and 500–30,972 targets sites in the 'conservative group (median: 20,165.2 target sites) (**Fig. 2d**). When factoring in the size of the target intervals, we observed target site densities of 0.017–798.6 target sites per kb (median: 47.9 target sites per kb) for the 'permissive' group and 0.64–475.9 target sites per kb (median: 6.4 target sites per kb) for the 'conservative' group (**Fig. 2e**).

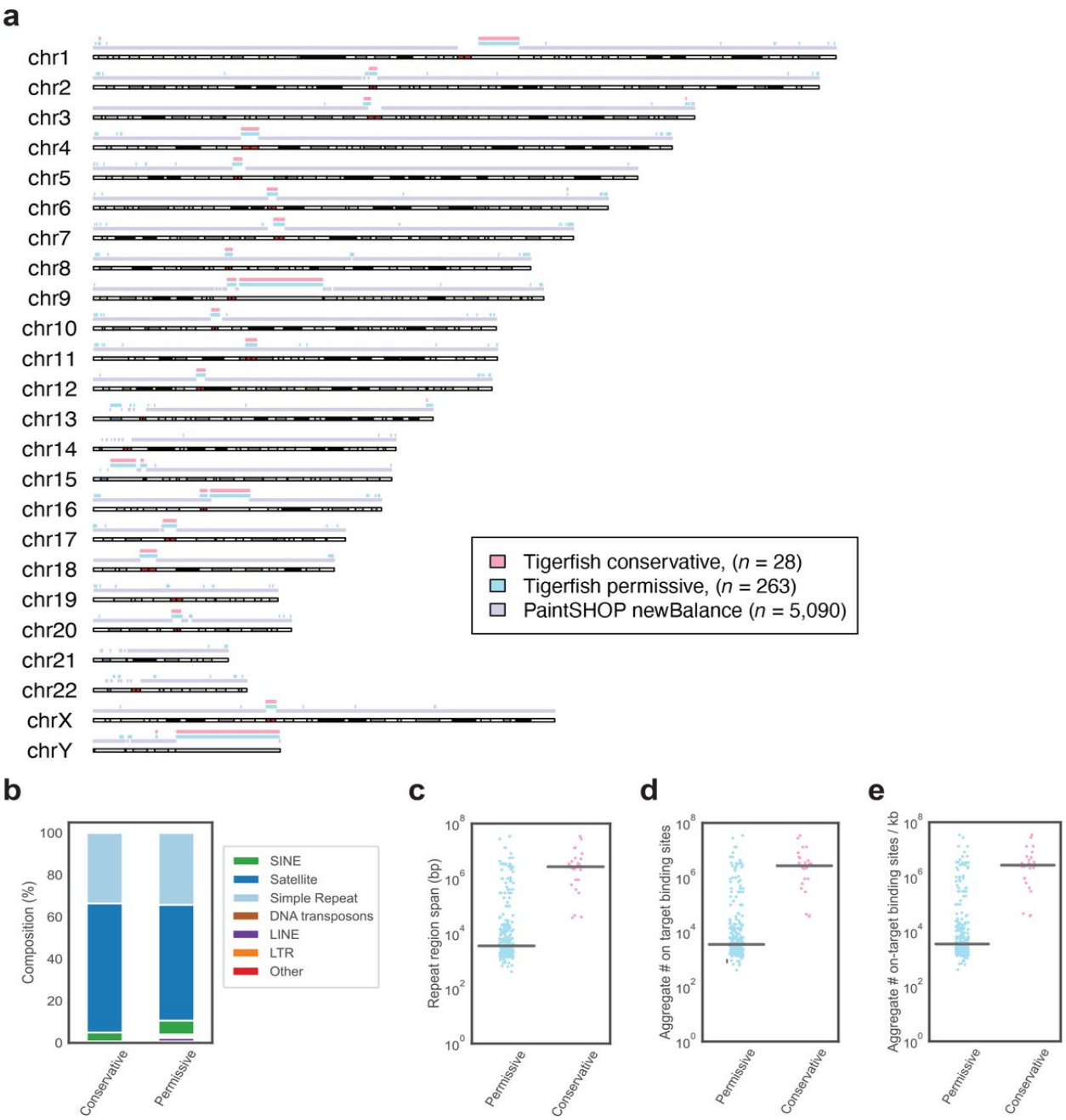


Fig. 2. Genome-scale probe design with Tigerfish. **a**, Schematic visualization of intervals for which Tigerfish probe sets were identified using conservative (pink) or permissive (teal) parameters and intervals covered by existing PaintSHOP probes designed using parameters suitable for non-repetitive targets (lilac). **b**, The distribution of RepeatMasker annotations for intervals identified and processed by Tigerfish using conservative and permissive settings. **c**, Length distributions of

the regions identified and targeted by Tigerfish using conservative and permissive parameters. **d**, The aggregate number on-target binding predictions for probe sets designed by Tigerfish using conservative and permissive parameters. **e**, The aggregate number on-target binding predictions per kilobase for probe sets designed by Tigerfish using conservative and permissive parameters.

2.3.3 Validating Tigerfish probes *in situ*

To evaluate how effectively the *in silico* design approach of Tigerfish translates to performance *in situ*, we designed and conducted a series of FISH experiments. Specifically, we set out to investigate whether Tigerfish was able to generate a panel of FISH probes targeting repetitive DNA intervals specific to each of the 24 human chromosomes, as such a panel would have utility in diagnostic and chromosomal enumeration assays. In order to showcase the versatility of the different Tigerfish run modes, our panel consisted of a mix of probes designed against regions identified using "Repeat Discovery Mode" and regions selected manually based on their RepeatMasker[253] annotations using "Probe Design Mode" (**Supplementary Data 3 and Supplementary Data 4**). The panel spanned a range of target sizes (10 kb – 4.5 Mb, mean = 1.3 Mb) and predicted on-target binding activities (477.5 – 7,228, mean = 2,418.1; **Table 1**).

Imaging Coordinates	On-target	Off-target	Imaging Repeat Length (Mb)
chr1:134680000-134800000	7228.4	191.2	0.12
chr2:92330000-94670000	3174.8	243.0	2.34
chr3:91730000-92590000	505.4	161.4	0.86
chr4:52140000-53070000	1074.6	51.0	0.93
chr5:47650000-48150000	1675.0	352.1	0.5
chr6:58540000-61060000	2356.1	73.6	2.52
chr7:60410000-63720000	4977.8	251.6	3.31
chr8:44250000-46320000	1964.4	933.4	2.07
chr9:44960000-47230000	2049.0	261.9	2.27
chr10:39640000-40710000	1637.4	25.3	1.07
chr11:51040000-54420000	3908.9	110.8	3.38
chr12:34780000-37060000	2022.9	108.8	2.28
chr13:111520000-111570000	927.9	644.1	0.05
chr14:99470000-99490000	477.5	1188.7	0.02
chr15:8550000-8680000	4162.8	1608.9	0.13
chr16:48950000-48980000	3812.7	807.1	0.03
chr17:23890000-27420000	3912.9	512.0	3.53
chr18:15970000-20430000	4576.7	3319.3	4.46
chr19:21000000-21060000	1408.6	261.4	0.06
chr20:27580000-27630000	950.1	174.9	0.05
chr21:44760000-44780000	761.1	440.1	0.02
chr22:18540000-18550000	1347.7	295.7	0.01
chrX:58910000-59080000	1518.8	146.5	0.17
chrY:20960000-21230000	1603.9	175.5	0.27

Table 1. Description of the 24-target Tigerfish probe set panel.

In order to verify that our Tigerfish probes were binding to their intended genomic targets, we implemented an experimental scheme in which the Tigerfish probe set targeting a given interval was co-hybridized with a set of 1,000 probes designed by PaintSHOP[250] that targeted a 200 kb non-repetitive interval on the target chromosome, with the Tigerfish and PaintSHOP probe sets being labeled with spectrally distinct fluorophores (**Fig. 3a**, **Supplementary Data 4**). We used this experimental design to perform a series of 24 two-color FISH experiments on 46, XY human primary metaphase chromosome spreads (**Fig. 3b**). Using this approach, we confirmed that our metaphase FISH produced the predicted staining patterns for all 24 combinations of Tigerfish and PaintSHOP probe sets (**Fig. 3c**, **Supplementary Fig. S2**–**S5**).

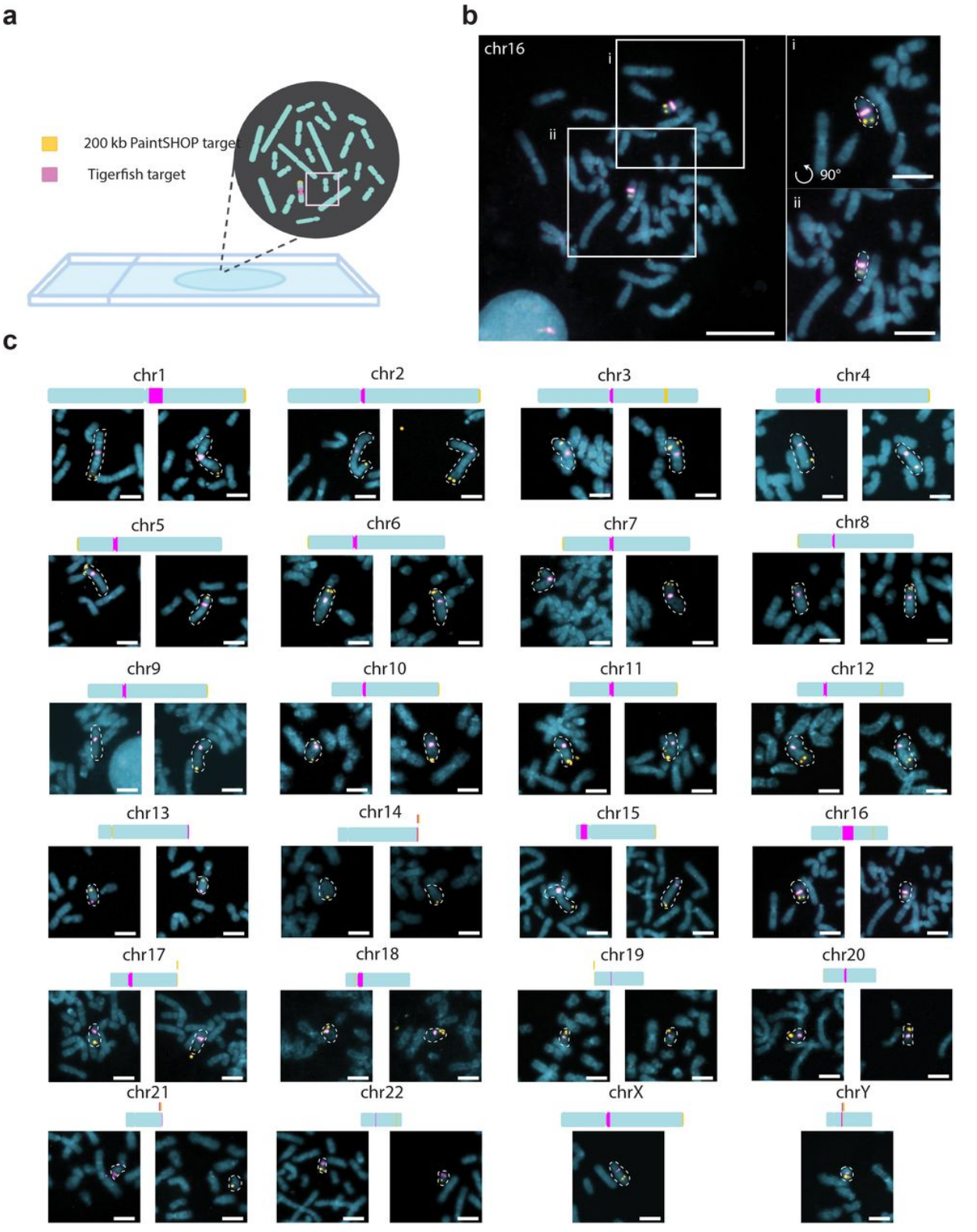


Fig. 3. *In situ* validation of Tigerfish probes. **a**, Schematic overview of the experimental design used to validate Tigerfish probe sets on metaphase chromosome spreads also labeled with probe

sets targeting non-repetitive DNA designed by PaintSHOP. **b**, Representative full field of view (left) and zoomed insets (right) showing Tigerfish (magenta) and PaintSHOP (yellow) probe sets targeting chr16. **c**, Zoomed crops depicting Tigerfish (magenta) and PaintSHOP probes targeting the indicated chromosomes. For the autosomes, each image pair was obtained from the same metaphase spread. The X and Y chromosome images were obtained from separate 46, XY spreads and thus only have one chromosome each. Please see Supplementary Figs. S8–S11 for the full spread images. Images are maximum intensity projections in Z. Scale bars, 5 μm (zoomed crops) or 20 μm (fields of view).

To augment our metaphase data, we also performed a series of 24 interphase FISH experiments on 46, XY primary human lymphoblasts using the same Tigerfish and PaintSHOP probe set combinations to visually enumerate chromosomal copy number (**Fig. 4a**). Specifically, we imaged >40 cells for each experiment and quantified the number of observed Tigerfish and PaintSHOP foci in the 3D volume of the nucleus (**Fig. 4b**, **Supplementary Fig. S6**–**S9**). Our analysis of the resulting data revealed a strong agreement between the two types of probe set (78.4% concordance, n = 1,061), with both approaches predominantly displaying 2 foci per nucleus (PaintSHOP: 781/1061, 73.6%; Tigerfish: 922/1061, 86.9%) and identifying a range of foci (1–4) per nucleus consistent with our previous studies using oligo-based probes for enumeration[100,114,240] (**Fig. 4c**). Taken together, our metaphase and interphase FISH experiments demonstrate the specificity and utility of Tigerfish for visualizing the positioning and abundance of highly repetitive DNA intervals *in situ*.

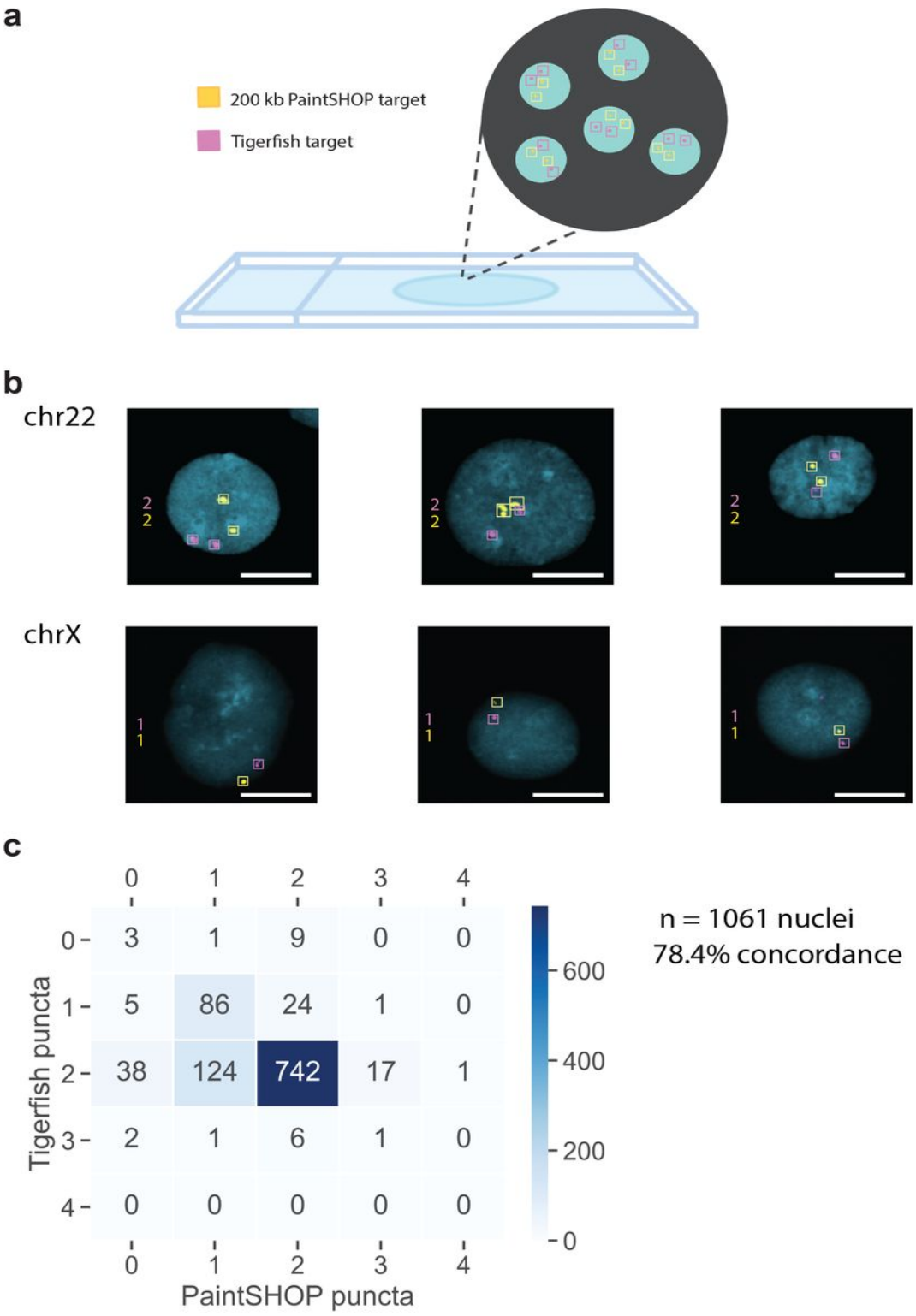


Fig. 4. Chromosome enumeration in interphase nuclei. **a**, Schematic overview of the experimental design used to perform chromosome enumeration using Tigerfish probe sets in 46, XY interphase nuclei also labeled with probe sets targeting non-repetitive DNA designed by PaintSHOP. **b**,

Representative images of nuclei labeled with Tigerfish probe sets (magenta) and PaintSHOP probe sets (yellow) targeting intervals on chr22 (top row) or chrX (bottom row). **c**, Heatmap displaying the observed distribution of Tigerfish and PaintSHOP puncta per nucleus. Images are maximum intensity projections in Z. Scale bars, 10 μm.

2.3.4 Computational requirements to run Tigerfish.

To evaluate the computational resources required to run Tigerfish at the scale of mammalian genomes, we collected a series of benchmarking data during our probe design runs on the full human CHM13v2 + HG002 chrY assembly using the 'permissive' and 'conservative' parameter settings. Our analyses focused on four key usage metrics: 1) he "wall clock" run time, which reflects the overall duration of the run from start to finish; 2) the amount of active CPU processing time needed to complete the run; 3) the maximum amount of virtual memory used, which represents the sum total of physical (RAM) and swap (hard disk) memory allocations; 4) the maximum amount of physical memory used, which reflects the RAM component of the virtual memory pool. As Tigerfish uses Snakemake[262] for parallelization, we were able to record data about these four metrics on a per-interval basis for all 263 intervals identified collectively by the 'permissive' and 'conservative' parameter settings. In line with the broad range of observed target interval sizes and target site numbers of the 263 intervals (**Fig. 2f–h**), we also found a wide distribution of resource usage values. Our analyses revealed that probe design against most target intervals finished quickly, with a median run time of 6.9 hours (range: 1.8 min – 50.2 hr) and a median CPU time of 5.1 hours (range: 0.6 min – 4.9 hr) (**Fig. 5a, b**). Moreover, Tigerfish generally required only modest amounts of memory for software designed to be run on a computing cluster, with a median max virtual memory allocation of 24.8 GB (range 12.3–44.8 GB) and a median max physical memory allocation of 20.3 GB (range 8.2–40.8 GB) (**Fig. 5c, d**).

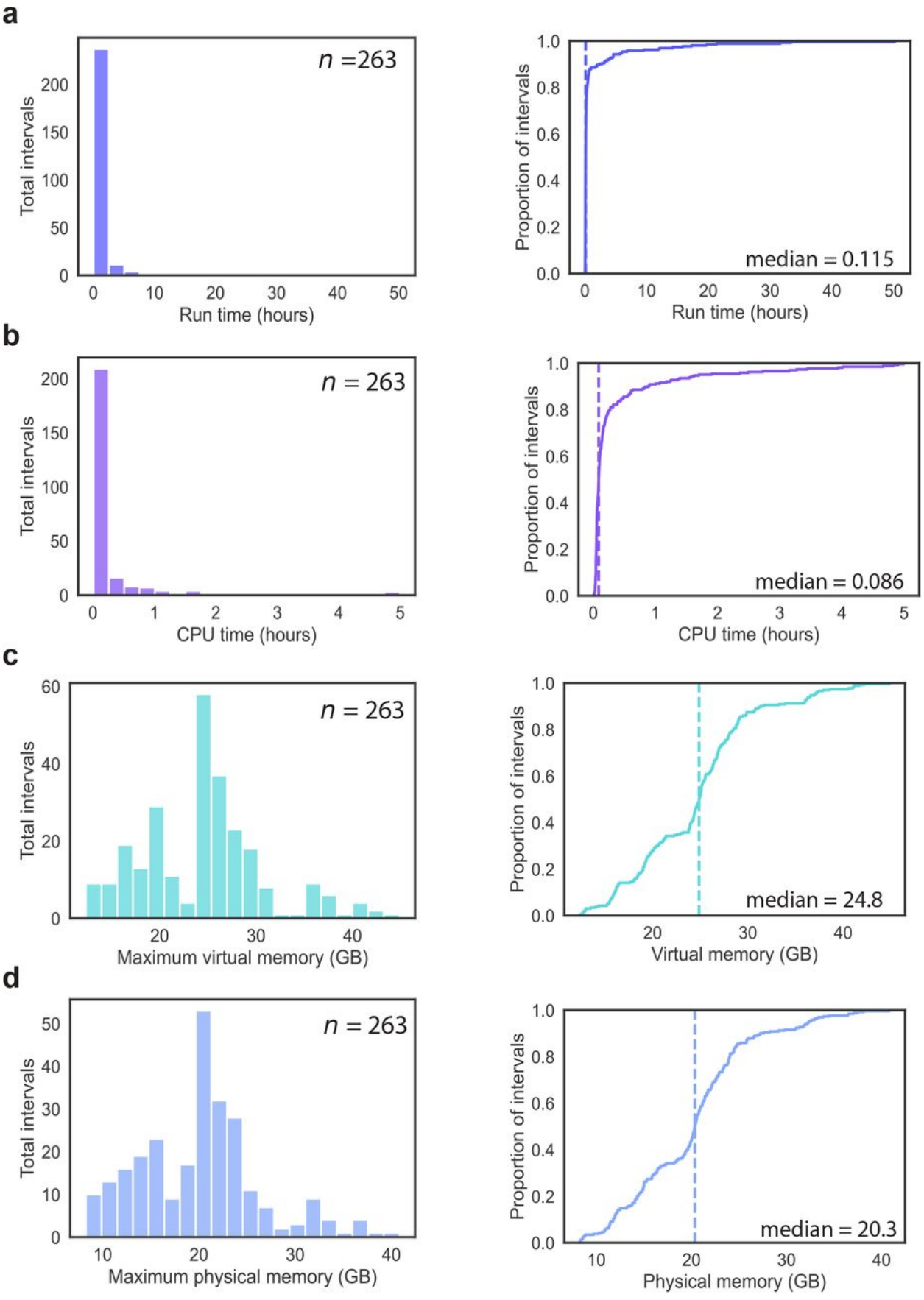

a
n =263
Total intervals
Run time (hours)
Proportion of intervals
median = 0.115
Run time (hours)
b
n = 263
Total intervals
CPU time (hours)
Proportion of intervals
median = 0.086
CPU time (hours)
c
n = 263
Total intervals
Maximum virtual memory (GB)
Proportion of intervals
median = 24.8
Virtual memory (GB)
d
n = 263
Total intervals
Maximum physical memory (GB)
Proportion of intervals
median = 20.3
Physical memory (GB)

Fig. 5. Resource requirements for genome-scale Tigerfish probe design. **a**, Distribution (left) and empirical cumulative distribution (right) of the wall-clock runtime recorded for running the 263 conservative and permissive intervals. **b**, Distribution (left) and empirical cumulative distribution (right) of the CPU runtime recorded for running the 263 conservative and permissive intervals. **c**, Distribution (left) and empirical cumulative distribution (right) of the maximum recorded virtual memory allocation for running the 263 conservative and permissive intervals. **d**, Distribution (left) and empirical cumulative distribution (right) of the maximum recorded physical memory allocation for running the 263 conservative and permissive intervals. Vertical dashed lines in the cumulative distribution plots correspond to the median values.

Given the observed spread in the resource usage values, we hypothesized that the resource requirements might vary as a function of the size of the target interval. Indeed, stratifying the benchmarking data into three groups based on span of the target interval revealed that the group of intervals less than 100 kb in span had a median run time of 5.8 minutes (range: 1.8 min – 43.8 min, n = 211) and the group of intervals between 100 kb and 1 Mb had run a median run time of 21.6 minutes (range: 6 min – 59.4 min, n = 20), with the group of intervals >1 Mb in span having a considerably longer median run time of 4.6 hours (range: 16.2 min – 50.1 hr, n = 32) (**Supplementary Fig. 10**). We did not observe a similar trend with virtual memory or physical memory usage, as all three length groups had nearly identical memory requirements (**Supplementary Fig. 10**). Taken together, our benchmarking results indicate that Tigerfish can readily be deployed on computing clusters or powerful individual computers to identify repetitive intervals and design probes specific to these intervals at the scale of genomes.

2.4 Discussion

Tigerfish is a freely available computational platform that facilitates the design of oligo-based FISH probes against intervals of repetitive DNA at the scale of genomes. The Tigerfish pipeline

establishes a paradigm for the deep specificity analysis of probes targeting repetitive sequences, which in turn enables users to establish criteria by which to select and empirically evaluate the effectiveness of oligos targeting such regions. Once designed, Tigerfish probes can readily be augmented with any of the powerful toolkits available for oligo-based FISH, including signal amplification approaches such as SABER[270], HCR[271], and RCA[272] and multiplexing approaches such as DNA MERFISH[29] and DNA seqFISH[244]. Moreover, Tigerfish offers users a great number of tunable parameters, providing flexibility to tailor the probe design process for different types of repetitive intervals and different genome compositions and complexities. We have demonstrated the efficacy of Tigerfish by performing genome-scale probe discovery in a fully assembled human genome and provided extensive experimental validation on both spread metaphase chromosomes and in interphase nuclei for the specificity of Tigerfish probes. Tigerfish is supported by extensive documentation and tutorials and can perform complex probe discovery tasks against the most challenging intervals of genomic DNA using only modest computational resources. We anticipate Tigerfish will play a key role in the experimental validation and biological investigation of repetitive DNA intervals as more fully assembled human, vertebrate, plant, and other model organism genomes continue to be introduced.

2.5 Methods

2.5.1 Genome sequences used for probe set design.

The CHM13 genome assembly versions 1.0, 1.1, and 2.0 were downloaded without repeat masking from the T2T consortium at https://github.com/marbl/CHM13.

2.5.2 Pipeline construction and implementation

Tigerfish is written in Python 3.7.8 with dependencies that include Biopython 1.77[265], Bowtie 2.3.5.1[109], NUPACK 4.0[263], BEDtools 2.29.2[273], Numpy 1.18.5[274], Pandas 1.0.5[275], pip 20.1.1,

pybedtools 0.8.1[273,276], sam2pairwise 1.0.0[277], samtools 1.9[264], scikit-learn 0.23.1[266], scipy 1.5.0[278], zip 3.0, matplotlib 3.3.4[279], seaborn 0.11.1[280], pytest 6.2[281], and Jellyfish 2.2.10[110]. All Tigerfish probe collections were generated using a pipeline implemented with Snakemake 7.19[262]. Dependencies that implement Python libraries can be found via the tigerfish.yml, snakemake_env.yml, and chromomap_env.yml files that are used to execute Tigerfish as a Snakemake[262] pipeline. These scripts and their dependencies are documented on Tigerfish's GitHub repository. These environments are also available in the **Supplementary Software**. Scripts were executed locally in an OS X Anaconda Python 3.7[282] environment or in a CentOS Linux environment on the Department of Genome Science 'Grid' Cluster at the University of Washington.

2.5.3 Whole genome probe discovery

Genome assemblies in FASTA format without repeat masking were used when building Jellyfish[110] files and Bowtie2[109] indices, and were used as input files for probe discovery. Jellyfish hash size was set to approximate the size of the genome assembly so that files were generated using the command, "jellyfish count -s 3300M -m 18".

2.5.4 Identification of *k*-mer enriched sequences

Tigerfish identifies repeat regions in Repeat Identification mode by using a sliding window of a specified size (*window*, W) flagging all counts exceeding a user-specified value (*threshold, T)*. The sum of the counts within the sliding window are divided by the length of the window so that if the user-specified composition score (*composition, C)* is exceeded, Tigerfish will identify windows of the genome where k-mer counts which map to abundantly repetitive sequences. Here, users may also specify at what base position they wish to start searching for repeats, which is described as a *file_start* parameter. Alternatively, if the user provides coordinates of target regions (i.e.,

defined_coords=True and repeat_discovery=False), then the user must also provide the name of the scaffold. In this case, Tigerfish skips the 'repeat_ID.py' script entirely to proceed with oligo probe design. For whole genome mining in CHM13, the sliding window was implemented with parameters described in **Supplementary Data 5**.

2.5.5 Designing oligo probes.

Tigerfish implements logic as described in the OligoMiner[114] framework for probe design using the bed file generated during Repeat Identification mode or from a user-provided BED file. Here, a FASTA file containing all regions of interest is used to design valid probe sequences using parameters values for probe length, percent G+C content (GC%) and adjusted melting temperature T_m calculated using nearest neighbor thermodynamics[114]. The modified blockParse script described in OligoMiner was used to mine probe candidates ranging in length from (*min_length, max_length*) 25-50 nt and T_m (*min_temp, max_temp*) between 42–52ºC.

2.5.6 Predicting probe specificity.

The *k*-mer binding proportion (*enrich_score*, K_b) was determined by obtaining the proportion of two computed values, *copy_num* and *total_genome_binding*. The aggregate count of all *k*-mers for any given probe sequence within its respective repeat target is described as *copy_num*, or R_m . The aggregate count of all *k*-mers for any given probe within the entire queried genome is described as *total_genome_binding*, or (H_m). Thus, the *k*-mer binding proportion was computed as R_m/H_m. Probes with shared *k*-mer composition similarity above the *mer_cutoff* proportion are omitted from downstream filtering. Probes are ranked in descending order within each repeat region by Normalized Rank (Nr) = $(R_m/(\max(R_m)*c1)) + (K_b/(\max(K_b)*c2))$, where c1 (*c1_val*) and c2 (*c2_val*) are user-specified constants. The *mer_cutoff* proportion is determined by storing *k*-mers of ranked probes and profiling all consecutive candidate probes to see if the proportion of their *k*-mer composition exceeds that of the *mer_cutoff*. Users may modify *enrich_score*,

copy_num, *c1_val*, *c2_val*, and *mer_cutoff* within the config.yml file. The parameters chosen for the conservative and permissive datasets are reported in **Supplementary Data 5**.

2.5.7 Computing *in silico* binding predictions

Bowtie2 was run on each probe sequence against the human genome using the following parameters (--local -N 1 -R 3 -D 20 -i C,4 --score-min G,1,4, -L 15, k 500000). The parameters -L (*seed_length*) and -k (*bt2_alignment*) may be modified by users within the Tigerfish config.yml. Probe alignments are returned as a BAM file for each probe sequence, which is then processed from the resulting SAM file using SAMtools[264]. Using this SAM file, sam2pairwise[277] is used to return derived alignment sequences. With these provided pairs of probe sequence and derived alignment sequence, NUPACK 4.0[263] computes the predicted thermodynamic likelihood that each alignment pair will form duplexes under FISH conditions[114]. The NUPACK model summarizing these conditions is described as (material='dna', celsius=69.5, sodium=0.39, magnesium=0.0, ensemble='stacking'). Candidate probes are only added to the final probe set if they do not share predicted probe binding greater than the value *max_pdups_binding*.

The on-target alignment score (On_T) is determined by taking the sum of all predicted duplexing scores for derived alignments that are found within the repeat target. Off-target alignment scores are computed by taking the aggregate sum of all predicted duplexing scores from derived alignments that are found outside the repeat target (Off_T). The predicted *in silico* on-target binding proportion (*binding_prop*) for each oligo is then computed as $On_T/(On_T+Off_T)$. Genome bins (genome_windows) are generated using BEDtools makewindows, and BEDtools intersect is applied to all reported sam2pairwise genome alignments to identify potential off-target binding signals. All predicted duplexing scores are aggregated over windows, which are binarized to map binding signals to the repeat target and all other genomic regions where binding events are predicted. Probes with an aggregate Off_T over any given non-target genome bin that exceeds the

parameter *off_bin_thresh* are culled from the candidate probe set. Users may modify the parameters seed_length, bt2_alignment, genome_windows, binding_prop, and off_bin_thresh. There are additional parameters that may be used to control permissiveness of filtering in the alignment_filter.py script. Users may control the desired aggregate on-target sum for any set of probes designed against a repeat region (*target_sum*), the minimum on-target value for any desired candidate probe (*min_on_target*), and maximum desired candidate probes to be returned in any target repeat region (*max_probe_return*). Parameters chosen for conservative and permissive datasets may be viewed in **Supplementary Data 5**.

2.5.8 Visualizing candidate probe *in silico* binding

Bowtie2 alignments are derived for individual or pools of probes against a repeat region where predicted thermodynamic binding is computed over a given size of genomic bins generated by BEDtools (*thresh_windows*). These predicted thermodynamic binding events are summarized by scaffolds and are used to determine the size of the imaging target window for bins containing binding events that are greater than the parameter within the repeat region target (align_thresh). The sum of predicted duplexing values are aggregated over computed genomic bins and normalized using the MinMaxScalar function of scikit-learn[283], where the range of values is mapped from 0 to 255 to summarize predicted binding over genomic bins. chromoMap[284] in R is used to generate summary ideograms of probe target signals as an optional step in Probe Analysis Mode.

2.5.9 Read the Docs

A Read the Docs web page (https://beliveau-lab-tigerfish.readthedocs-hosted.com) was created to provide detailed documentation of our tool. The intention of hosting our work on Read the Docs was to provide sufficient background and resources for individuals from all computational backgrounds to be able to leverage Tigerfish for their own work. Here we provide installation

information, simple tutorials for testing the Tigerfish install, a glossary of all parameters that may be modified by users, summaries of our default parameters, and frequently asked questions.

2.5.10 Computational benchmarking

Speed calculations were computed using the Snakemake benchmark feature. Each scaffold in the CHM13v2 + HG002 chrY assembly was run as its own individual cluster job in parallel for the repeat discovery steps, and the resulting intervals identified for probe design were also processed in parallel. Benchmarking was performed on a Dell PowerEdge R840 server node equipped with 4 Intel Xeon Gold 6252 2.1 GHz 24-core CPUs (192 total job threads) and 1.5 TB of DDR4 PC4-23400 2933 Mhz ECC RAM running CentOS 7.9 Linux.

2.5.11 PER concatemerization

Here, 100 µl reactions were prepared for both Tigerfish probes and PaintSHOP bridge sequences with a final concentration of 1x PBS, 10 mM $MgSO_4$, 400–1,000 U ml^{-1} Bst DNA Polymerase (large fragment), 120,000 units/ml (NEB M0275M), 100 nM of Clean G hairpin, 50 nM - 1 µM of hairpin and water to 90 µL. After incubation for 15 min at 37ºC, 10 µM oligo probe(s) were added and the reaction was incubated for another 2 hours with another 20 min at 80ºC to heat-inactivate the polymerase. PER extension solutions were directly diluted into FISH solutions. Lengths of the concatemers were evaluated by diluting 6.7 µL of the in vitro reaction with 3.3 µL 6X TriTrack. Samples were then run on a 10% TBE-Urea denaturation gel (ThermoFisher EC68755BOX) for

10 min alongside 1 kb Plus DNA Ladder to estimate length and imaged with SYBR Gold channel and then imaged after a 15 min incubation.

2.5.12 DNA-SABER-FISH on spread metaphase chromosomes

PaintSHOP bridge oligos and Tigerfish primary probes were extended using the PER as previously described[270]. Dry microscope slides containing human 46, XY metaphase spreads (from AppliedGenetics Laboratories) were immersed in 2× SSCT + 70% (vol/vol) formamide at 70°C and incubated for 90 s in Coplin jars. Slides were then transferred and incubated in ice-cold 70% (vol/vol) ethanol, ice-cold 90% (vol/vol) ethanol, and ice-cold 100% (vol/vol) for 5 minutes each. Slides were then air dried after incubation in 100% ethanol. A hybridization solution consisting of 2X SSCT, 50% formamide, 10% (wt/vol) dextran sulfate, 40 ng/µL RNase A (EN0531; Thermo Fisher), and resuspended PER-extended PaintSHOP bridge oligos (20 pmol total), amplified ssDNA primary probes (25 pmol total), and PaintSHOP bridge library (60 pmol total) which were dried at 60ºC for 30 minutes using a SpeedVac concentrator. The solution was sealed using a 22 x 22-mm #1.5 coverslip using rubber cement. Samples hybridized overnight at 45ºC in a humidified chamber. Samples were then washed for 15 min in 2X SSCT at 60ºC and then twice for 5 mins with room temperature 2X SSCT. Samples were then incubated in a secondary hybridization containing 5X PBST, 10% dextran sulfate, 10 µM fluorescent oligos for 1 hour at 37ºC. Slides were then washed three times with 1X PBST at 37ºC. After air drying slides, samples were mounted with SlowFade Gold + DAPI and sealed beneath a 22 x 30 - mm #1.5 coverslip using nail polish.

2.5.13 Microscopy

Microscopy was performed using a Yokogawa CSU-W1 SoRa spinning disc confocal unit attached to a Nikon Eclipse Ti-2 chassis. Excitation light was emitted at 30% of maximal intensity

from 405 nm, 488 nm, 561 nm, or 640 nm lasers housed inside of a commercial Nikon LUNF 405/488/561/640NM launch. Laser excitation was delivered via a single-mode optical fiber into the CSU-W1 SoRa unit. Excitation light was then directed through a microlens array disc and a 'SoRa' disc containing 50 μm pinholes and directed to the rear aperture of a 100x N.A. 1.49 Apo TIRF oil immersion objective lens by a prism in the base of the Ti2. Emission light was collected by the same objective and passed via a prism in the base of the Ti2 back into the SoRa unit, where it was relayed by a 1x lens through the pinhole disc and directed into the emission path by a quad-band dichroic mirror (Semrock Di01-T405/488/568/647-13x15x0.5). Emission light was then spectrally filtered by one of four single bandpass filters (DAPI: Chroma ET455/50M; ATTO 488: Chroma ET525/36M; ATTO 565: 27 Chroma ET605/50M; Alexa Fluor 647: Chroma ET705/72M) and focused by a 1x relay lens onto an Andor Sona 4.2B-11 camera with a physical pixel size of 11 μm, resulting in an effective pixel size of 110 nm. The Sona was operated in 30 16-bit mode with rolling shutter readout and exposure times of 300 ms. Images were processed in ImageJ and Fiji[285,286] and Adobe Photoshop.

2.6 Contributions

R.A., W.S.N., and B.J.B. conceived the study. R.A., C.K.C., and Q.L. wrote and optimized software code. R.A. and C.K.C. performed experiments. R.A., W.S.N., and B.J.B. wrote the manuscript. K.H.M., W.S.N., and B.J.B. supervised the work.

Chapter 3: Seeking Justice in Genomics Education and Research

3.1 “Life happens over the course of a PhD.”

3.1.1 Finding questions in what my research meant beyond lab.

As a biologist specializing in repetitive DNA and a technology developer, I have dedicated the past five years to pondering the significance of precise genome assemblies in designing unique oligo probes specifically tailored to repetitive DNA arrays. During my PhD, I was fortunate enough to embark on an incredible journey. In my very first year of graduate school, I had the opportunity to delve into the rich history of the human genome project[201]. Simultaneously, I was among the pioneering group of researchers who had the privilege of utilizing the fully assembled CHM13 human genome for our research in technology development[287]. This experience left an indelible mark on me, allowing me to collaborate with scientists who shared a common goal: to understand the repetitive DNA megabase gaps in genomic sequences that had been omitted from genomic assays due to limitations in sequencing and mapping techniques.

I vividly recall engaging in discussions with Dr. Mark Diekhans and Dr. Karen Miga, who are both researchers and collaborators at UCSC and I were able to have early access genome browser tracks for CHM13, even as the assembly of some acrocentric satellite arrays was still ongoing. Witnessing Tigerfish effectively identify and map probes to these satellite DNA arrays, I couldn't help but feel immensely fortunate to be a graduate student at UW, involved in the realm of tech development, and contributing to a profoundly timely and impactful thesis project in the field of repetitive DNA biology.

The advent of “telomere-to-telomere” genome assemblies brings forth a new frontier in genomics research, accompanied by a fresh set of challenges. As more centromere assemblies become

accessible, a multitude of biological questions surrounding the genetic rearrangement of centromeric and heterochromatic satellite DNAs in aging and cancers can finally be addressed[189,288–290]. Obtaining sequence-specific information about these satellite DNA repeats will enable high-resolution investigations into the epigenetic and transcriptional landscape of human centromeric regions across individuals. Although it is widely recognized that these satellite sequences exhibit considerable structural variability within the population, our understanding of the extent of this variation and its impact on centromere identity remains limited. Unraveling this enigmatic 'centromere paradox' at the population level necessitates future studies that extend beyond the single CHM13 benchmark reference genome[49,291]. Multi-modal sequencing datasets encompassing individuals from diverse populations and spanning multiple generations will likely yield profound insights into the variation of human satellite DNA and its implications for epigenetic dysregulation, aberrant transcription, and centromere protein associations in disease.

Undoubtedly, creating high-quality, base-level maps of human centromeres from individuals in the population will require collaborative efforts as the field moves away from relying solely on a 'gold standard' model of a single genome reference. However, one of the greatest challenges in centromere biology and genome assembly is rooted in ethics and equity. As I worked on generating metaphase spreads from CHM13, I found myself pondering the origin of the cell line and wanting to understand the process of collecting and selecting samples for genomics studies. CHM13 cells were initially derived from a hydatidiform mole at Magee Women's Hospital as part of a research study[180]. Due to the unique characteristics of this cell line, CHM13 proved to be an ideal candidate for "telomere-to-telomere" genome assembly, as it represents an effectively haploid genome sequence. However, for future studies involving larger populations, researchers will need to devise innovative methods to unravel the distinct differences and contributions among haplotypes[200]. Most importantly, researchers must carefully consider the implications and impacts of how these samples are collected from the population. More specifically, active efforts must be

made that addresses and endows leadership from populations that have been historically excluded from broader medical and genomics studies that ensures mechanisms of data privacy and autonomy are respected.

3.1.2 Placing value to human DNA.

The global significance of human genome sequences as a valuable resource that has been capitalized by pharmaceutical and genomic testing companies. Notably, direct-to-consumer genetic testing companies like 23andMe have benefitted from this trend, selling access to their database containing digital sequence information from approximately 5 million individuals to GlaxoSmithKline for $300 million[234]. It comes as no surprise that 23andMe has also forged partnerships with pharmaceutical companies, enabling them to license drugs for development using the wealth of information derived from their vast user platform. However, it is crucial to acknowledge that 88% of people included in large-scale studies of human genetic variation are individuals of European ancestry[292]. As pharmaceutical corporations strive to enhance their genomic databases to deepen their understanding of genotype-phenotype associations, innovative approaches are needed to improve the data quality of existing genomic repositories. One promising avenue involves expanding existing biobanks by incorporating genetic information from populations that have not yet undergone sequencing or have historically been excluded from such population studies. By including genetic data from underrepresented populations, valuable insights can be gained into locally prevalent, population-specific variants, which could prove invaluable in identifying novel drug targets. However, best practices are being developed iteratively over time and tools are being created to support such research efforts often with increasing expertise and leadership from Indigenous scholars, representatives of descendent or interested communities, and other key stakeholders and organizations such as ASHG[293]. Despite this, even as new best practices emerge, foundational principles are often points of discussion throughout the genomics community and organizations such as ASHG who are offering platforms

and opportunities to discuss how genomics can empower individuals from populations that have been underrepresented in genomics efforts[293–295]. Such principles and efforts highlight demonstrating respect for individuals and communities, respecting tribal sovereignty, and engaging in ethical communication and consultation practices in genomics research[293].

Programs like the "All of Us" research program are dedicated to ensuring that at least 50% of its participants come from underrepresented minority populations[296]. The aim of this decision is to foster diversity and equity in the pursuit of advancing precision medicine. However, there are valid concerns that these efforts could inadvertently disempower communities that have historically been excluded from large-scale studies of human genetic variation[234,297]. Past initiatives, such as the Human Genome Diversity Project[298] and the 1000 Genomes Project[186], both government-funded sequencing endeavors, serve as examples of open-source data initiatives that have been exploited for profit, generating nearly a billion dollars in revenue for pharmaceutical and ancestry-testing companies[186,298,299]. If the "All of Us" program follows the same unrestricted data-access and sharing protocols, there will be no safeguards in place to prevent the commodification of DNA from underrepresented minority populations, particularly those of African and Indigenous ancestry. While participants in these studies may benefit from the development of new pharmaceutical treatments, it remains unclear whether any of these drugs or therapies, derived from insights gained through these populations, will directly benefit members of those communities through subsidized medications, royalties, or intellectual property rights.

3.1.3 The legacy of Henrietta Lacks on bioethical implications in genomics research

The commodification of data and the existence of policies allowing open-source access to this information exacerbate the long-standing history of marginalization and disempowerment faced by underrepresented groups in the field of medicine. The story of Henrietta Lacks serves as a poignant example of the racial inequities deeply ingrained within the US research and healthcare

system[300]. Lacks, an African American woman, passed away in 1951 from an aggressive form of cervical cancer. During her treatment at Johns Hopkins Hospital, doctors took samples of her cancerous cell line without her knowledge or consent, and these samples were subsequently shared with researchers. These cells, now known as HeLa cells, displayed an extraordinary ability to survive and essentially became an immortal cell line[301]. They have since played a pivotal role in countless breakthroughs in biological research and have even been instrumental in the development of COVID-19 vaccines[302].

Despite the immense profits reaped by biotechnology companies utilizing her cells, Thermo Fisher Scientific has been the first company to come to a settlement with the Lacks family[303]. Shockingly, her genome and medical records have even been published online without proper consent. The Lacks family is actively collaborating with scientists to establish more stringent regulations governing the use of human specimens. However, there is an enormous amount of work to be done in conjunction with public health agencies, including increased scrutiny by institutional review boards (IRBs), to dismantle the disparities that persist in basic research and academic spaces[300]. The goal is to ensure that the rights and dignity of individuals and communities are upheld, rectifying the injustices of the past and shaping a more equitable future in the realm of medical research.

3.1.4 Misappropriation of genomics research by white supremacists

Beyond medicine, it's imperative that geneticists reassess how research is conducted and communicated. In May 2022, 10 Black people were murdered by a white supremacist shooter at Tops Grocery Store in Buffalo, NY, who referenced a figure from a population genomics study in his 180-page diatribe to justify his abhorrent actions[235,304]. However, even prior to this tragic event, researchers had already observed patterns of online discussions related to population genomics research among proponents of white nationalist ideologies[235]. The misappropriation of scientific

research has played a central role in fueling racism and inciting violent acts by far-right groups, tracing back to the aftermath of the Second World War[305,306].

While geneticists are becoming increasingly aware, particularly on social media platforms like Twitter and Facebook, that scientific findings have been co-opted by far-right groups in the resurgence of white supremacy, concrete actions to address these issues remain elusive. To further complicate matters, scholars with far-right ideologies are leveraging the principles of open science to promote scientific racism, disseminating their distorted narratives on preprint servers and even within peer-reviewed journals. Notably, researchers such as Carlson *et al.* have documented a troubling surge in the misrepresentation and dissemination of scientific data on social media platforms and forums like 4chan, exploited by far-right individuals to propagate racist ideologies[235]. This underscores the urgent need for geneticists to actively confront these challenges. Efforts must be made to combat the misinterpretation and misuse of scientific research, while also ensuring that scholarly outputs are appropriately scrutinized and regulated.

To enhance scientific efforts in countering the harmful appropriation of scientific research, several examples of improvement measures include:

1. Developing improved standards for data visualization that convey information effectively and are less prone to misinterpretation.

2. Ensuring that scientific teams are led by individuals from groups historically harmed by weaponized science.

3. Providing opportunities for geneticists to participate in workshops and courses that explore ethics and best practices, with the aim of creating graduate training programs and studies that promote equity for individuals with historically excluded identities in biosciences.

These steps will be crucial in addressing and mitigating the negative impacts of scientific research appropriation. By proactively addressing these issues, geneticists can work towards dismantling the ties between scientific knowledge and the propagation of racial hatred, fostering a more inclusive and responsible scientific community.

Human geneticists must recognize and actively combat the influence of white supremacy. It is crucial for scientists, scientific societies, and journals to unequivocally reject the notion that human "races" possess biological distinctions, emphasizing instead that race is a socially constructed concept with a historical and political context[307,308]. Most often, researchers in population genomics describe that human genetic variation follows a continuous gradient, and that the current patterns of genome diversity can be explained by the migration and intermixing of populations throughout human history. However, it is essential to acknowledge that certain findings in human and population genetics have inadvertently provided fodder to support the notion of biological distinctions among races[309,310].

In this context, it is important to revisit the 1000 Genomes Project and the Human Genome Diversity Project. These initiatives, when considered, reveal a significant disparity in the representation of data from populations outside of Africa and those of non-Indigenous ancestry[234]. This raises the need for introspection and critical examination of the data collection practices employed, as they have inadvertently contributed to an imbalanced representation of human genetic diversity. Geneticists must actively work to rectify this imbalance by broadening the scope of data collection efforts to be more inclusive and representative. By doing so, they can ensure that future research accurately reflects the rich diversity of human populations, dismantling any false biological claims associated with race and contributing to a more just and equitable scientific landscape. By conscientiously considering the data that geneticists choose for analysis and implementing more comprehensive data visualization frameworks, we can take initial steps to

prevent the misrepresentation of scientific research. Moreover, in the same way that scientists are expected to articulate the positive societal impacts of their work, it is equally vital to confront the potential harmful consequences and consider how research findings can be misused to propagate scientific racism. Ultimately, claims of genetic superiority among certain individuals lack any scientific foundation. As scientists, it is our imperative to firmly oppose the resurgence of scientific racism and take a clear stance against it.

3.1.5 What can be done to make genomics researchers more accountable for inequities in science?

In addition to exploring the direct impact of genomics research on society and medicine, I have delved into the repercussions of biases on researchers with underrepresented identities and diverse life experiences. The fact that underrepresented populations have been largely overlooked in genomics research in health and medicine raises important questions about the disparities in training experiences of individuals with underrepresented identities in genomics and bioscience PhD programs and the potential impact on their training outcomes. A practical and effective solution to diversify genomics databases involves empowering marginalized communities to actively participate in governing data, biological samples, and the use of digital tools that determine sample usage rights. This approach ensures transparency and integrity in handling digital sequence information. The concept of Indigenous data sovereignty captures this practice, wherein Indigenous people have control over data from their populations[234,311]. Through community trusts and partnerships with biomedical research institutions, subsidized access to successful drug development approaches can be provided, ultimately reinvesting in the participating communities. Extending these parallels to our own academic programs and research departments raises a crucial question: How can we foster accessible, equitable, and supportive academic environments that enable scholars with marginalized identities to contribute as influential leaders in initiatives that empower communities through their scientific expertise? To

address this question, I draw upon the lessons I have learned from my personal journey as a PhD student.

Before I joined Genome Sciences, one of my undergraduate advisors casually mentioned something to me that has stuck with me ever since. We were having coffee together just a few days before my graduation when she shared with me that, “Life happens over the course of a PhD." But what resonated even more was her next statement, "Everyone's journey is unique, and it's up to you to discover the questions that truly fulfill you. Yes, there will be challenging days, but those are the moments when you must pause and introspect. Reflect on how you arrived at this point." What amazed me about her mentorship was her ability to empathize through storytelling. She bravely shared her experiences of feeling isolated as a woman of color in the scientific field, and she entrusted me with her words because we shared those visible identities at the time.

As I reflect upon my journey as a fifth-year grad student, I find myself pondering the moments that brought me to this point and the profound changes that have unfolded, not only in my academic endeavors but also in my personal life. Amidst the inevitable ups and downs, I am immensely grateful for the abundance of kindness I have encountered and the valuable relationships I have fostered since starting graduate school. It is worth noting that along the way, I had the opportunity to embrace a new name and begin hormonal transition as I navigated my gender identity. Simultaneously, I have passionately pursued a career in science communication and illustration, immersing myself in a realm where art seamlessly intertwines with communicating research stories.

For a period, I found myself contemplating the best approach to introduce this chapter, questioning whether it should even be included in my thesis. However, through extensive self-reflection on my experiences in mentorship and advocacy, I realized the importance of articulating

the process behind curating these valuable resources during my tenure as a graduate student in Genome Sciences. Personally, I believe that my work in scientific research and my lived experiences are inherently intertwined, primarily due to the visibility of my identities within the field of Genome Sciences and the broader domain of genomics.

While the representation of marginalized scholars holds immense significance, it is crucial to acknowledge the costly burden that accompanies visibility[312]. To be frank, my journey within Genome Sciences has encompassed both positive and challenging aspects, prompting me to write this chapter as a means to convey that my experiences in genomics research are not isolated incidents for individuals with visible minoritized identities[313]. Moreover, I believe that sharing this story can help others comprehend that my graduate experience transcended the realm of scientific research alone. The curation of affirming resources and the advocacy for my own needs as a trainee within Genome Sciences became indispensable for my survival in academia[314–316]. Without the support of the communities that embraced me and those to which I contributed; this chapter would not have come to fruition.

My intention was to transform these insights into tangible coursework, recognizing a deficiency during my time in Genome Sciences—namely, the absence of spaces or courses that addressed the intersections of race, gender, sexuality, disability, and immigration within the realm of genomics research. To me, this signified a gap in the formal training provided by the PhD program. Ultimately, we shouldn't disregard the profound impact of visible and invisible identities on the experiences and outcomes of marginalized scholars within academic contexts.

By neglecting to recognize insights from individuals with diverse lived experiences as scientists, we severely bias the outcomes of our research, restricting its scope. As mentioned earlier, the societal implications of genomics research hold tremendous significance for individuals from

underrepresented backgrounds. It is imperative that we actively support and elevate the voices of those with diverse lived experiences, enabling them to assume leadership roles in fostering inclusive research environments. This includes the establishment of frameworks for data security, science communication, and equitable learning spaces. Only through these concerted efforts can we strive towards a society that benefits from the advances in genomics research in a fair and equitable manner.

In this section, I aim to share my research, perspectives, and personal experiences that have shaped various aspects of my academic career. I do so with the intention of shedding light on the challenges that my unique story has encountered within an R1 academic training environment. By sharing these experiences, my goal is to provide contextual understanding and highlight the connections between biases and experiences within academic spaces and the societal impacts of genomics research that I previously described. I firmly believe that academic learning environments play a pivotal role in fostering more equitable practices and facilitating collaborations with government agencies and institutions that strive for equitable applications of bioscience research.

Creating more equitable training environments represents a powerful step toward empowering researchers to combat scientific racism. By equipping them with the necessary background knowledge and resources, they can confidently communicate their findings to non-scientific audiences and uplift underrepresented groups as active collaborators and leaders in this era of data sovereignty and open-source genomic data. In my view, educating others about the historical inequities within academic bioscience training programs and institutions is an essential part of this solution.

Therefore, the outcomes I aim to achieve through this chapter are threefold:

1. Provide a comprehensive account of my work in the bioscience advocacy space, particularly focusing on individuals with marginalized backgrounds.
2. Present a historical archive document of the Genome Sciences Association for the Inclusion of Marginalized Students (GSAIMS), outlining the frameworks for executing trainee-led community events.
3. Produce a complete 10-week curriculum that can be incorporated into graduate training programs in genomics, offering insights into how identity and a sense of belonging can shape outcomes in PhD bioscience training programs. This curriculum is, to my knowledge, the first of its kind to be specifically designed for graduate students that describes the context of academic environments and social movements as it relates to experiences in the biosciences.

3.2 Community organizing and a PhD.

3.2.1 Seeking role models and spaces of belonging.

During my first year of graduate school, I struggled with finding a sense of belonging and community. In addition to having to re-explain my pronouns and gender to faculty and trainees in Genome Sciences often, I found myself tokenized for my lived experiences. Because I am also Latinx, the intersectionality of my experiences further amplified the feelings of isolation that I experienced. Some experiences that I encountered included being outed by my deadname in a grant writing course, being asked intrusive questions about my gender and transition during departmental happy hours, being harassed by faculty members upon speaking up about my own challenging experiences and finding myself being excluded from study sessions among some of my own cohort mates. During my first year of graduate school, I also had to work as an online ESL teacher to support myself because of the financial expense of relocating to Seattle. This

experience really demonstrated how financial privilege can play out in academic settings because most of my cohort was able to rely on their family's financial support for such a move.

Furthermore, after raising feedback about my own personal experiences in the program, it became clear that Genome Sciences lacked a community space that acknowledged the challenges that marginalized scholars face in graduate programs. At the time, I explored Women in Genome Sciences (WiGS) and Genomics Salon as possible community groups but noted clear shortcomings in the spaces that they offered. WiGS' leadership was led predominantly by white cis-women and the name alone didn't feel welcoming to someone like me who is trans and genderfluid. Despite their work in expanding their space to more non-white trainees and researchers, in 2018 WiGS' definitions of feminism didn't feel expansive or reflective of my own background. Likewise, Genomics Salon discussed topics related to ethics in genomics, but often failed to host speakers and events that discussed intersectionality and race in an appropriate context that marginalized scholars with diverse identities face.

The Genome Sciences Association for the Inclusion of Marginalized Students (GSAIMS) was created out of necessity in January 2019. I co-founded the organization with another graduate student in my cohort who also observed that these community resources were necessary for current and future trainees. We created an organizational structure that would allow others to be involved in collaborating on making events possible. While we encountered pushback from some trainees about the need for such an organization, we were overwhelmingly supported by many trainees and faculty in Genome Sciences. In fact, many of our initial events had a supportive turnout which sparked conversations about the importance of acknowledging diversity and inclusion within our program.

To date, GSAIMS has organized 35+ events in Genome Sciences, and has provided 250+ hours of community centered programming over the course of four active years. We were awarded a Seed Grant from UW in 2020 and are currently supported by the Genome Sciences Data Science Training Grant for our department symposium organizing. Furthermore, I produced a timeline of programming that GSAIMS leadership team members were involved in during our active dates as an organization [attached as an appendix]. GSAIMS became a valuable resource and community hub for marginalized trainees in Genome Sciences who sought the support and mentorship of other trainees.

However, it's important to share that while GSAIMS has experienced support, we've also had our lows as an organization. Effectively, many of our trainees in leadership positions encountered academic burnout from having to navigate the pandemic and other external stressors due to challenging academic environments. Because of this, GSAIMS and many trainee-led organizations will likely remain as frameworks that future trainees may implement. However, this underscores the importance of graduate departments needing to create environments that offset the burden of advocacy that marginalized scholars face.

Ultimately, trainee led DEI efforts cannot be successful without frameworks for continuity and widespread support from the community who benefits from these resources[317,318]. Because trainees within Genome Sciences often remain in the program for a six year period with many required obligations during this period, it's clear that marginalized trainees will face an unfair amount of pressure to sustain the organizations that support their well being[319]. I will discuss further alternatives that may be implemented to further support trainees and offset this burden (see 'DEI initiatives are owed accountability, examination, and re-evaluation').

3.2.2 Navigating healthcare, the pandemic, and my intro to science communication

In November 2019, I began a two-year long medical odyssey that greatly shaped my grad school experience. Beginning in November, I caught what seemed to be a cold that I struggled to recover from. By February 2020, my condition had significantly worsened with severe breathing issues and joint pain. Because of the onset of the pandemic, I went home in March 2020, and it was nearly impossible to schedule an appointment with any pulmonary specialists until July 2020 following my general exam. Up until this time, I was being treated with antibiotics and corticosteroids which minimally improved my condition. After years of working with pulmonary specialists and rheumatologists, I found myself bouncing between an autoimmune diagnosis and a 'long COVID' recovery regimen.

I'm sharing this story because during my illness, I was also attempting to continue my own thesis research, complete my requirements for my general exam, and lead GSAIMS and DEI initiatives within Genome Sciences. Over Zoom, it was impossible to really tell that I had been navigating an invisible disability and barrier that kept me from engaging in my work as fully as what I could have for several years because I had to navigate doctors' appointments among other obligations.

This experience only broadened my perspectives toward becoming more informed on DEI approaches that accounted for greater accessibility for researchers with disabilities[320]. This experience with getting sick during my graduate program over the course of a pandemic made me reflect deeply on what I wanted to commit my energy and graduate training toward. Ultimately, my bandwidth was stretched to my limits, and I had to reckon with focusing on areas of my training that would energize me if I wanted to stay in graduate school. During this time, I had many affirming conversations with mentors who helped me discuss belonging and identity in graduate school and their stories empowered me to share my own experiences.

Additionally, I found myself in a unique position whereas a scientist, my own loved ones who primarily spoke Spanish sought expertise about COVID-19 vaccines. I recalled reading resources that my family members in LA county had shared with me about vaccines and found myself disappointed with the poor translations and graphics used to explain how these vaccines worked. By engaging with my family members and translating this information into more approachable language, I found that I was able to engage with them about my own work in science. Not only was this experience affirming, but it reminded me how sorely needed it is to have scientists engage with the public in accessible and meaningful ways. Furthermore, it empowered my own community members to feel heard and listened to by a scientist who looked like them and shared a common background.

This experience became foundational to my own beginnings as a science communicator with a public facing platform. I was encouraged to join science Twitter at the beginning of my PhD to make community and to build connections with other scientists. However, it was only until after July 2020 and following the publication of my *Nature Careers* piece[316] when I began to share my experiences in academia that I garnered a larger following of scientists, non-scientists, technical illustrators, and others from a variety of science media platforms. While this platform offered the benefits of some shared community spaces, it also brought on challenges. Some of these challenges included harassment from scientists and strangers in my direct messages. Eventually, I had to take some time away from social media and science communication in this public format to prioritize and care for my own well-being. Despite this, I was routinely invited to speak with a number of podcasts, media outlets, deliver workshops through organizations such as the ReclaimingSTEM Institute[321], and serve as a guest speaker or panelist at a number of institutions. One of the most striking interactions I had, however, were from the eight aspiring scientists and students who independently reached out to me seeking a mentor who was also queer and trans.

From this experience, I began to see how critical it was for researchers to use their platforms to advocate for those with marginalized identities and to build resilient communities within STEM and beyond the ivory tower. I found myself becoming the mentor I wish I had access to when I was considering a career as a scientist. This ultimately also showed me the importance of what representation can mean to people who are looking for mentors who can share their experiences in a career that they are passionate about. This led me to ultimately continue my own work in the STEM advocacy and science communication space, where I aimed to focus on storytelling and unpacking academic norms that historically excluded those from diverse lived experiences. Broadly, I hope to continue my work in science communication because I think that science is never the result of a sole genius, but rather the story of collaborative teams, ideas, and a lot of time is spent navigating failure. I want to share these breakthroughs with both experts and non-science audiences to inspire others on how incredible and engaging bioscience research can be.

3.2.3 The summer of 2020 introduced many researchers to Diversity, Equity, and Inclusion

Witnessing the BLM uprisings in the summer of 2020 following the murder of George Floyd proved to be an emotionally challenging time in Genome Sciences. Three days after protests began in Minneapolis, GSAIMS organized a vigil via zoom which was facilitated by Dr. Atom Lesiak in response to the police violence that was triggering for many of our trainees. An overwhelming number of people showed up on this zoom call and many department members found themselves entering this space and noticing that this was the first time that they had been made aware of anti-Black violence in Seattle and across the country.

This summer and the months following these events proved to be a period of acknowledgement and learning for many people within Genome Sciences. In June 2020, GSAIMS produced a public letter condemning anti-Black police brutality and called for greater action from the wider Genome

Sciences community to support Black, Indigenous, and scientists of color who disproportionately experience discrimination in the field[322,323]. Following the call to action, GSAIMS meetings had received an overwhelming number of new attendees. We observed going from the usual five team members on a Zoom call to sometimes over fifteen attendees who wanted to get a better understanding of what events we were planning for the department.

This hypervisibility in our work unfortunately led to further tokenization of our experiences and labor. GSAIMS was cited on multiple active departmental grants that described the department as being a diverse and inclusive learning space for trainees despite this being far different from the realities that trainees had experienced in this program[316,324,325]. Furthermore, there was little engagement and participation in GSAIMS programming from the greater faculty community which was also reflected in the actions of research scientists and trainees in the program. In the long term, departmental DEI book clubs and initiatives were founded with the intention of learning about biases in academia, but these spaces ultimately don't address these challenges at an institutional level. At some point, the learning must stop and action must begin with the spaces that we participate in[324].

The experiences GSAIMS encountered following 2020 were not unique, and similar occurrences could be seen by organizers of Black in Neuro[326]. Additionally, in my own personal academic journey following the summer of 2020, my expertise in scientific research became synonymous with navigating higher education with marginalized identities. My science Twitter account erupted with followers after the summer of 2020 because I came forward about some of my own experiences with challenges in higher education. While finding communities of scholars internationally who shared such similar stories felt validating, I knew that ultimately these challenges that we faced were deeply systemic and ultimately rooted in the history of white supremacy that has been present in the ivory tower. My own experiences with navigating DEI

only motivated me to research the history of genomics as it intersected with race and education in the ivory tower.

3.2.4 Navigating how my research expertise was perceived.

Throughout my PhD, it was much more common for me to be invited to give talks and seminars about my own work in community building and DEI than that of my own scientific expertise. Below, I've listed all talks that I've been invited to that were related to DEI work compared to those that were solely about my research.

Scientific Invited Talks

- T2T / HPRC Towards a Complete Reference of Human Genome Diversity, Remote
- SACNAS National Conference, Honolulu, HI
- Pairing Meeting, Cleveland Clinic, Cleveland, OH

DEI Invited Talks

- Pride in Research Event, MIT Whitehead Institute, Remote
- Nanostring Queer Researchers Panel, Seattle, WA
- Keynote speaker, SUPERB Scholarship program, Counce, TN
- Gender diversity in STEM Panelist - Australian Conference of Undergraduate Research, Remote
- Reclaiming STEM Workshop Host, Remote
- Coming Together Weekend, DePauw University, Greencastle, IN

While these talks were incredible opportunities for me to engage with the community at different institutions, this trend is common among graduate students who are involved in DEI work who also have public facing platforms. Additionally, it is more common for non-white trainees who are

involved in DEI work to be perceived and seen as experts in this domain over their own scientific research. According to research demonstrated in the text *Black, Brown, Bruised*, it is shown that faculty members of color are more likely to be required to deliver additional service toward mentorship to support marginalized scholars[324]. This additional service labor is also present for undergraduates, graduate students, and postdoctoral researchers with intersecting marginalized identities and can significantly impact retention. Ultimately, in higher education, we need to recognize that DEI related labor is overlooked, undervalued, and at times not recognized as of scientific importance. From my own experiences and those of others that are cited, DEI efforts benefit departments through grants to support scholars and to offer tangible community spaces and mentoring opportunities.

3.3 DEI initiatives are owed accountability, examination, and re-evaluation.

3.3.1 My involvement in broad departmental DEI efforts

Before June 2020, the term 'DEI' was commonly used in the humanities and social sciences to describe approaches for delivering educational materials. Since then, institutions have frequently adopted 'DEI' to establish committees aimed at creating more equitable environments within graduate training programs[319,327]. However, many existing models of “DEI” programs inadvertently perpetuate the very issues they seek to address. In this chapter, I aim to address these concerns and observations, drawing from my experiences in Genome Sciences and broader national trends. But first, let's unpack the definitions of 'DEI'.

'DEI' stands for "diversity, equity, and inclusion." In this context, “diversity” refers to the range of identities present within an organization or institution, “equity” entails fair treatment and equal opportunities for all members to pursue their goals, and “inclusion” involves fostering a welcoming culture where everyone feels respected and valued[324]. In academia and STEM programs, DEI

initiatives aim to foster a greater sense of belonging, community, and support for all members of the department. In Genome Sciences, I became involved in the formation of the first DEI committee shortly after June 2020, where I contributed to crafting the department's online statement.

> *"The Department of Genome Sciences is dedicated to creating a welcoming and inclusive environment. Our objective is to cultivate a space for learning and collaboration where diversity is acknowledged and celebrated. At Genome Sciences, we strive to establish supportive environments for marginalized individuals in higher education and society at large. To achieve this, we acknowledge our ongoing responsibility to comprehend, confront, and challenge systems of privilege and disadvantage in higher education, such as those based on race, color, creed, caste, religion, national origin, citizenship, sex, age, marital status, sexual orientation, gender identity or expression, disability, veteran status, or socioeconomic status."*

During this period, I also collaborated closely with others to develop a DEI activity for the virtual department retreat in Fall 2020 and contributed to the creation of more inclusive guidelines for Research Report feedback. However, being one of the few marginalized individuals on the DEI committee proved to be exhausting as many other members were in the process of learning and unlearning their own privileges derived from participation in academic norms[236,317,319,324]. While the committee was instrumental in initiating discussions within the department to improve recruitment and retention of trainees, institutional-level DEI committees have been shown to perpetuate systemic issues[319,324,328,329].

One common critique of DEI committees is their tendency to center white emotions and feelings in anti-racism efforts. In academia, it is often junior faculty members who serve on DEI

committees, despite already being overworked and unable to allocate sufficient energy toward addressing ongoing issues of racism and discrimination within their departments[313,324]. Additionally, it is crucial to recognize how current systems often fail marginalized groups through cumbersome Title IX processes and reporting tools that prioritize protecting perpetrators of abuse rather than providing support and reconciliation for those affected by discrimination[329,330]. Hence, it is critical for institutions and academic departments to hire educators who possess expertise and a long-term vision for maximizing support for all members of the department.

3.3.2 Let's unpack 'The Minority Tax'.

The concept of the Minority Tax refers to the additional burden of time and resources placed on minority individuals to advocate for and represent their communities[331]. To ensure the effectiveness of DEI efforts, it is crucial to establish norms that recognize advocacy work as valuable for both community and scientific progress[324]. Within Genome Sciences, most researchers engaged in community advocacy and outreach projects are women and gender minorities with intersecting marginalized identities. This trend is also evident across higher education, as demonstrated in Peña, 2022[331].

Academia has traditionally held a limited perspective on what defines "success." Historical norms in the ivory tower have perpetuated the idea of a solitary researcher or investigator solving complex problems through innovative methods and techniques, serving as a measure of success[325]. However, this problem-solving and learning style fails to consider the needs of diverse scholars who thrive in multifaceted learning environments and benefit from mentorship[332,333]. Moreover, this model no longer accommodates scholars with interdisciplinary interests extending beyond academic contexts[322].

The labor involved in advocacy work within academia is often inadequately compensated, both in terms of career advancement and financial remuneration. Marginalized researchers frequently find themselves having to create and support organizations that uplift their communities, all while leading research projects and collaborations to advance their scientific careers[326]. This disparity is not experienced by those with privilege. In academia, white, cisgender, heterosexual, and able-bodied men are more likely to advance their scientific careers because they are not burdened by such marginalization and are not tasked with the additional emotional and physical labor required to navigate academia and address environmental challenges[312,320].

Consequently, current DEI initiatives and programs often face limited participation from trainees and faculty due to the penalty associated with time spent away from scientific research[312]. Additionally, trainees often have limited time within academic programs and are unable to effect lasting changes in academic environments unless a significant amount of time and energy is allocated, often at the expense of their scientific research. As a result, these trainees, despite their breadth of experience navigating higher education, may ultimately be pushed out of academia because they do not meet the publication requirements for prestigious postdoctoral fellowships or investigator positions.

Until all members of the community find ways to engage in actions that prioritize accessibility and equity for marginalized scholars, inequities that predominantly affect those engaged in this type of community work will persist. I firmly believe that establishing norms that emphasize the importance of scientific contributions within the field of genomics research and society at large is essential. In the curriculum I have developed, I deliberately selected articles and literature that highlight the diverse lived experiences of scholars in academia and provide a framework for contextualizing our own work in genomics within broader society.

3.3.3 The art installation

The installation of the most recent woodwork art in Genome Sciences, which took place in 2022, incurred a cost of $29,982. I find it important to mention this price because, amidst the pandemic, numerous trainees in Genome Sciences faced various challenges that could have been alleviated with internal financial support if it had been available. With that in mind, I would like to present a list of ways in which Genome Sciences could have utilized this installation fee to contribute to 'DEI' initiatives in the program:

1. Subsidized internet payments for trainees who needed to work remotely.
2. Establishment of a work-from-home support fund that could cover expenses such as standing desks and technology to facilitate virtual learning.
3. Subsidized mental health payments to support trainees seeking therapy and wellness resources.
4. Offering relocation fellowships for those who had to relocate during the pandemic and were required to live with or support loved ones.
5. Subsidized support for trainees in need of emergency medical assistance.

By reallocating resources toward these initiatives, Genome Sciences could have addressed pressing concerns faced by trainees and provided necessary support during challenging times.

3.3.4 Participating in ‘DEI’ conversations is the bare minimum.

All trainee-led organizations and departmental initiatives require constant critical evaluation and feedback to address the most relevant trends needed by a community. If only a select few consistently participate in these conversations, meaningful initiatives cannot be developed that align with the trajectory or long-term goals of a community. Similar to Dr. Beronda Montgomery’s teaching philosophy, academic communities need frequent check-ins to ensure their healthy and

effective implementation[330]. Faculty must model this behavior for trainees, demonstrating that racism and all forms of discrimination will not be tolerated in the program and that the departmental climate is crucial for improving the progress and outcomes of scientific research. Even in smaller environments, lab spaces can be made inclusive through leadership styles that are motivated to solve conflicts effectively and empathetically.

I believe that the development of empathy and community-building skills can begin at the graduate level, particularly for those who are motivated to become principal investigators at any institution. Throughout the graduate and postdoctoral training process, there are very few resources and workshops available to assist researchers in transitioning into the role of a group leader. Consequently, many faculty members must learn leadership skills on the fly, drawing from their past mentorship experiences. This can lead to challenges in conflict resolution, mentorship, and active listening skills down the line. It is important for faculty members to have ongoing opportunities for training to support DEI initiatives that align with other aspects of their careers. When lab spaces reflect values of accessibility and equity, and trainees feel comfortable discussing challenges they may face and having their concerns respected, it fosters a healthy and accountable academic environment that affirms everyone's backgrounds.

3.4 Genomics must reflect on what values our teaching spaces practice.

3.4.1 The first year of graduate school is critical in training program outcomes.

During my first year of graduate school, I had anticipated encountering a greater number of research articles that acknowledge the complex and challenging history of genomics within the context of eugenics and white supremacy. While a few articles did recognize that scientific research in the past has both benefited from and exploited marginalized communities in health and medicine, I had expected a more extensive discussion on the current impacts of research

and the ongoing realities faced by Black, Indigenous, and people of color in both science and society. Given that our program is an R1 training program, I believe it is of utmost importance for scholars to be aware of the disparities present in higher education, health, and STEM fields, enabling them to effectively communicate their research in culturally competent and impactful ways. In this section, I reflect on my personal experiences in Genome Sciences and emphasize the significance of a curriculum that acknowledges the importance of identity and a sense of belonging in graduate school, as it plays a critical role in shaping the future of scholars in genomics research.

Based on my previous personal experiences, the academic and cohort environment plays a critical role in determining the level of support graduate students receive during their education. The presence of a sense of belonging and a strong community is crucial for both personal and professional growth. It is of utmost importance that trainees can explore their scientific interests while receiving support from their peers and the overall environment. During the initial stages of their training, first-year trainees are particularly sensitive to the prevailing norms, which are shaped by the behavior of senior graduate students and faculty members. As a result, it is essential to establish and prioritize values related to DEI as well as advocacy right from the beginning of trainees' careers. This ensures that they are well-informed about the available community resources.

3.4.2 Challenging courses shouldn’t impact the well-being of trainees.

While the current coursework is designed to be challenging and thought-provoking, it should not be so overwhelming that it drives trainees to the point of wanting to withdraw from graduate programs. I personally encountered this issue in multiple courses while also managing my rotation projects and dealing with environmental stressors, which nearly caused me to drop out of the program several times. It is crucial for faculty members to reflect on why they push their trainees

in such manners. If they have experienced mistreatment in their own past academic settings, it is important that they break the cycle and refrain from perpetuating this behavior onto future students.

3.4.3 Trainees need better mentorship in obtaining alt-academic careers.

Graduate students and postdoctoral researchers should be provided with networking opportunities and workshops that address the wide range of career options available to them beyond academia. To achieve this, it is necessary to implement department-wide initiatives that invite guest speakers to conduct workshops, lead journal clubs, and deliver seminars, offering guidance to graduates regarding the various workplace cultures they may encounter in different settings. These initiatives can foster a greater sense of belonging and understanding among trainees as they establish networks with researchers outside of academia[333]. Furthermore, it is important for trainees to understand how their work can have an impact beyond the academic sphere. Outreach initiatives can serve as a bridge between research, coursework, and real-world significance. In addition to fostering a sense of community and leadership by engaging with the public, these initiatives can present unique opportunities to pilot projects such as yEvo[334], inspiring local students to explore research from scientists.

3.5 I created a curriculum that discusses identity and belonging in the biosciences.

3.5.1 How long did this take?

Based on my personal experiences, I have reflected upon and identified the resources that have helped me gain a better understanding of the shortcomings in current academic learning environments. As a result, I have developed a curriculum that delves into the topic of white supremacy in scientific research, explores the impact of social movements during the summer of 2020 on bioscience graduate programs regarding belonging and identity (titled "Interrogating Belonging and Identity in Bioscience Graduate Programs"), and proposes ways in which graduate

programs can foster more inclusive and equitable learning environments for individuals from diverse backgrounds and varied life experiences. The culminating project of this curriculum involves an opportunity to design a pilot outreach initiative or create a community resource that can be presented or published to benefit other bioscience trainees. This process of curriculum development took four years and involved many conversations with various individuals, including Dr. Atom Lesiak, Dr. Hannah Jordt, colleagues at SACNAS conferences, and colleagues at HHMI through the Gilliam Fellowship program.

3.5.2 Who is this for?

This curriculum is designed for individuals who are pursuing graduate or postdoctoral training or hold a faculty position in genomics. The resources included in this curriculum are publicly accessible, and the coursework guides are structured around a 10-week quarter system. Classes are scheduled to meet three times a week, making it convenient for teaching purposes. A significant portion of the curriculum involves discussion and research-based prompts and exercises, which facilitate collaboration among students and encourage stimulating discussions throughout the course.

3.5.3 Where can this work be used?

This work has the potential to be extended into a semester-long course and can be further refined to be piloted at a conference. The materials for this work are publicly accessible and can be widely distributed to those who are interested in learning about this form of advocacy and teaching. The curriculum itself is included as an appendix to this thesis, which will be added accordingly.

Chapter 4: Discussion and Conclusion

4.1 Further refining the Tigerfish workflow to explore diverse repetitive DNA families.

In Chapter 2, the Tigerfish workflow is explained. It is a versatile platform with extensive user documentation, designed to let users of all coding abilities interact with the tool. This interaction enables the creation of repetitive DNA sequences (oligos) for various genomics experiments on different genome versions. While Tigerfish can already analyze diverse repetitive DNA segments linked to specific DNA families, future software developments could enhance its capacity to design probes for smaller repetitive DNA sections. These refined probes could be used alongside microscopy methods to amplify signals on a smaller scale of DNA. These combined methods might reveal LINEs, SINEs, and other repeat families scattered across different species' genomes at varying frequencies. These advancements would yield a more comprehensive understanding of how distinct repeat families contribute to maintaining overall genomic stability and influencing evolution in diverse species.

4.2 Creating community resources for open-access repetitive DNA specific oligo probes.

Certainly, as more individuals' and species' genomes become fully assembled ('T2T'), new opportunities will arise for studies that delve deeper into the roles of satellite DNA repeat tracts within human and model organism genome variations. Tigerfish can play a pivotal role in enabling such studies by providing unique probes designed for specific repetitive DNA sequences. These probes can serve as valuable markers, allowing researchers to differentiate between genomic samples from various individuals within a population. This distinction may lead to a better understanding of the significance of centromeric variation. To facilitate this, I have developed preliminary designs for a database intended to be a collaborative resource for researchers interested in using FISH methods to explore distinctive repetitive DNA sequences among

individuals and model organisms. This database prototype, known as “FISHtank”, offers a user-friendly preview of features that could empower researchers to devise experiments and gain insights into potential repetitive DNA sequences of interest. The wireframe for this database was created using Figma and is also included as supplementary figure 11 for future reference and implementation within the Beliveau lab.

4.3 Interrogating the structure and function relationships of distinct repetitive DNA arrays using interdisciplinary experimental approaches.

To further contextualize the roles of specific repetitive DNA families and their relationships to genome organization, it will be important to gather data from multimodal approaches that can further interrogate the relationships between genomic position and function. One such approach that can offer such distinct context would be the use of Tigerfish in parallel with O-MAP[335] to better understand which proteins proximal to RNA molecules near pericentromeric DNA arrays are colocalized across cell states and trajectories. Due to the versatility of Tigerfish, several technology development applications may be utilized to understand the relationships of specific biological motifs across an array of phenomena. For instance, Tigerfish may be used to uncover the role of specific acrocentric chromosome repeat arrays in the maintenance of the NOR under diverse healthy and disease states such as cancer cells. Investigating such perturbations further may reveal the organizational importance of such repetitive DNA families in the maintenance of genomic stability. Through ongoing work in the Beliveau and Noble labs, Tigerfish may serve as a valuable framework that can provide further clarity toward emerging studies in repetitive DNA that impact how genomic structure is preserved.

4.4 Rethinking what the future of inclusive genomics education could look like in PhD programs.

In August 2023, I engaged in a conversation with a mentor that prompted me to deeply consider the significance and necessity of Diversity, Equity, and Inclusion (DEI) efforts within academic contexts. The mentor posed a thought-provoking question: "How do you imagine the future of 'DEI' within an academic setting?" This question resonates with me because it encourages contemplation on which members of the community are included or left out of DEI discussions and activities. Moreover, the question highlights who is predominantly responsible for the labor involved in DEI initiatives and the underlying intentions of how this work can be integrated within academic departments and beyond.

In my perspective, I foresee a future where the creation of positive, welcoming, and supportive academic environments won't require the label of 'DEI' work. My aspiration is that decisions regarding inclusivity and accessibility will naturally form integral parts of the culture within any learning environment. Thus, I envision a future where such efforts are deemed significant and are seamlessly integrated into academic research practices, without necessitating a specific label. These supportive actions should ideally become intrinsic to the fabric of academic research culture, evident in all aspects of academic leadership and community involvement.

Through the frameworks I've presented and the curriculum I've developed, I anticipate that these resources will offer a glimpse into diverse methods by which communities can embrace more inclusive practices. Ideally, these initiatives will extend the impact of our scientific research beyond the laboratory bench. Ultimately, my hope is that this endeavor will inspire others to pause, look ahead, and envision a more sustainable future that fosters the well-being and success of everyone who embarks on the journey of genomics research.

Funding

This work was supported by the National Institutes of Health (under grants 1R35GM137916 to B.J.B., UM1HG011531 to W.S.N., and 1R01HG011274 to K.H.M) and the Brotman Baty Institute for Precision Medicine (under a Catalytic Collaboration award to B.J.B.). R.A. was supported by a National Science Foundation Graduate Research Fellowship Program Award and a Howard Hughes Medical Institute Gilliam Fellowship for Advanced Study. C.K.C. was supported by NIH training grant 5T32HG000035.

Bibliography

1. Heitz, E. *Das heterochromatin der moose*. (Bornträger, 1928).
2. Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiol. Mol. Biol. Rev.* **79**, 347–372 (2015).
3. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).
4. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
5. Haws, S. A., Simandi, Z., Barnett, R. J. & Phillips-Cremins, J. E. 3D genome, on repeat: Higher-order folding principles of the heterochromatinized repetitive genome. *Cell* **185**, 2690–2707 (2022).
6. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17**, 487–500 (2016).
7. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251–260 (1997).
8. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).

9. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
10. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
11. Bickmore, W. A. & van Steensel, B. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–1284 (2013).
12. Sexton, T. & Cavalli, G. The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049–1059 (2015).
13. Therizols, P. *et al.* Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science* **346**, 1238–1242 (2014).
14. Gonzalez-Sandoval, A. *et al.* Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. *Cell* **163**, 1333–1347 (2015).
15. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).
16. Becker, J. S. *et al.* Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. *Mol. Cell* **68**, 1023-1037.e15 (2017).
17. Muller, H. J. & Altenburg, E. The Frequency of Translocations Produced by X-Rays in Drosophila. *Genetics* **15**, 283–311 (1930).
18. Barr, M. L. & Bertram, E. G. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature* **163**, 676 (1949).
19. Fortuny, A. & Polo, S. E. The response to DNA damage in heterochromatin domains. *Chromosoma* **127**, 291–300 (2018).
20. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
21. Boettiger, A. & Murphy, S. Advances in Chromatin Imaging at Kilobase-Scale Resolution.

Trends Genet. **36**, 273–287 (2020).

22. Nurk, S. *et al.* The complete sequence of a human genome. *bioRxiv* 2021.05.26.445798 (2021) doi:10.1101/2021.05.26.445798.
23. Cremer, T. *et al.* Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. *Hum. Genet.* **62**, 201–209 (1982).
24. Manuelidis, L. Individual interphase chromosome domains revealed by in situ hybridization. *Hum. Genet.* **71**, 288–293 (1985).
25. Schardin, M., Cremer, T., Hager, H. D. & Lang, M. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Hum. Genet.* **71**, 281–287 (1985).
26. Cremer, M. *et al.* Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. *Methods Mol. Biol.* **463**, 205–239 (2008).
27. Branco, M. R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* **4**, e138 (2006).
28. Rosin, L. F., Nguyen, S. C. & Joyce, E. F. Condensin II drives large-scale folding and spatial partitioning of interphase chromosomes in Drosophila nuclei. *PLoS Genet.* **14**, e1007393 (2018).
29. Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659.e26 (2020).
30. Fritz, A. J., Sehgal, N., Pliss, A., Xu, J. & Berezney, R. Chromosome territories and the global regulation of the genome. *Genes Chromosomes Cancer* **58**, 407–426 (2019).
31. Nguyen, H. Q. *et al.* 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nat. Methods* **17**, 822–832 (2020).
32. Payne, A. C. *et al.* In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371**, (2021).
33. Cullen, K. E., Kladde, M. P. & Seyfred, M. A. Interaction between transcription regulatory

regions of prolactin chromatin. *Science* **261**, 203–206 (1993).

34. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
35. Jerkovic, I. & Cavalli, G. Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* **22**, 511–528 (2021).
36. Fullwood, M. J., Wei, C.-L., Liu, E. T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* **19**, 521–532 (2009).
37. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
38. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
39. Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat. Genet.* **47**, 1179–1186 (2015).
40. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
41. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
42. Davies, J. O. J. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).
43. Sati, S. & Cavalli, G. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma* **126**, 33–44 (2017).
44. Stack, E. C., Wang, C., Roman, K. A. & Hoyt, C. C. Multiplexed immunohistochemistry, imaging, and quantitation: a review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods* **70**, 46–58 (2014).
45. Greil, F., Moorman, C. & van Steensel, B. [16] DamID: Mapping of In Vivo Protein–Genome Interactions Using Tethered DNA Adenine Methyltransferase. in *Methods in Enzymology* vol.

410 342–359 (Academic Press, 2006).

46. Quinodoz, S. A. *et al.* SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. *Nat. Protoc.* **17**, 36–75 (2022).
47. Sexton, T. *Spatial Genome Organization: Methods and Protocols*. (Springer Nature, 2022).
48. Shaban, H. A. & Seeber, A. Monitoring the spatio-temporal organization and dynamics of the genome. *Nucleic Acids Res.* **48**, 3423–3434 (2020).
49. Miga, K. H. Centromere studies in the era of “telomere-to-telomere” genomics. *Exp. Cell Res.* **394**, 112127 (2020).
50. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
51. van der Ploeg, M. Cytochemical nucleic acid research during the twentieth century. *Eur. J. Histochem.* **44**, 7–42 (2000).
52. Flemming, W. Beitrage zur Kenntniss der Zelle und ihrer Lebenserscheinungen. *Arch. Mikrosk. Anat. (1865)* **16**, 302–436 (1879).
53. Chieco, P. & Derenzini, M. The Feulgen reaction 75 years on. *Histochem. Cell Biol.* **111**, 345–358 (1999).
54. Levsky, J. M. & Singer, R. H. Fluorescence in situ hybridization: past, present and future. *J. Cell Sci.* **116**, 2833–2838 (2003).
55. Pardue, M. L. & Gall, J. G. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proc. Natl. Acad. Sci. U. S. A.* **64**, 600–604 (1969).
56. Moyzis, R. K. *et al.* A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 6622–6626 (1988).
57. Lichter, P., Cremer, T., Borden, J., Manuelidis, L. & Ward, D. C. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. *Hum. Genet.* **80**, 224–234 (1988).
58. Trask, B. J. Human cytogenetics: 46 chromosomes, 46 years and counting. *Nat. Rev. Genet.*

3, 769–778 (2002).

59. Pardue, M. L. & Gall, J. G. Chromosomal Localization of Mouse Satellite DNA. *Science* vol. 168 1356–1358 Preprint at https://doi.org/10.1126/science.168.3937.1356 (1970).
60. Britten, R. J. & Kohne, D. E. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* **161**, 529–540 (1968).
61. Britten RJ and Kohne DE. Repeated Sequences in DNA. *Science* **161**, 529–540 (1968).
62. Rigby, P. W., Dieckmann, M., Rhodes, C. & Berg, P. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *J. Mol. Biol.* **113**, 237–251 (1977).
63. Rudkin, G. T. & Stollar, B. D. High resolution detection of DNA–RNA hybrids in situ by indirect immunofluorescence. *Nature* **265**, 472–473 (1977).
64. Bauman, J. G. J., Wiegant, J., Borst, P. & van Duijn, P. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Exp. Cell Res.* **128**, 485–490 (1980).
65. Wiegant, J. *et al.* In situ hybridisation with fluoresceinated DNA. *Nucleic Acids Res.* **19**, 3237–3241 (1991).
66. Langer, P. R., Waldrop, A. A. & Ward, D. C. Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 6633–6637 (1981).
67. Manuelidis, L., Langer-Safer, P. R. & Ward, D. C. High-resolution mapping of satellite DNA using biotin-labeled DNA probes. *J. Cell Biol.* **95**, 619–625 (1982).
68. Singer, R. H. & Ward, D. C. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 7331–7335 (1982).
69. Xxxii, P. Synthesis of a dithymidine dinucleotide containing a 3’-5’internucleotide linkage.

Michelson AM, Todd AR. *J. Chem. Soc.* (1955).

70. Matteucci, M. D. & Caruthers, M. H. Synthesis of deoxyoligonucleotides on a polymer support. *J. Am. Chem. Soc.* **103**, 3185–3191 (1981).
71. Beaucage, S. L. & Caruthers, M. H. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* **22**, 1859–1862 (1981).
72. Caruthers, M. H. A brief review of DNA and RNA chemical synthesis. *Biochem. Soc. Trans.* **39**, 575–580 (2011).
73. Willard, H. F. & Waye, J. S. Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**, 192–198 (1987).
74. Lawrence, J. B., Singer, R. H., Villnave, C. A., Stein, J. L. & Stein, G. S. Intracellular distribution of histone mRNAs in human fibroblasts studied by in situ hybridization. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 463–467 (1988).
75. Arndt-Jovin, D. J., Robert-Nicoud, M., Kaufman, S. J. & Jovin, T. M. Fluorescence digital imaging microscopy in cell biology. *Science* **230**, 247–256 (1985).
76. Palotie, A., Heiskanen, M., Laan, M. & Horelli-Kuitunen, N. High-resolution fluorescence in situ hybridization: a new approach in genome mapping. *Ann. Med.* **28**, 101–106 (1996).
77. Carrington, W. A. *et al.* Superresolution three-dimensional images of fluorescence in cells with minimal light exposure. *Science* **268**, 1483–1487 (1995).
78. Brown, J. *et al.* Identification of a subtle t(16;19)(p13.3;p13.3) in an infant with multiple congenital abnormalities using a 12-colour multiplex FISH telomere assay, M-TEL. *Eur. J. Hum. Genet.* **8**, 903–910 (2000).
79. Schröck, E. *et al.* Multicolor spectral karyotyping of human chromosomes. *Science* **273**, 494–497 (1996).
80. Speicher, M. R., Gwyn Ballard, S. & Ward, D. C. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat. Genet.* **12**, 368–375 (1996).
81. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. High density synthetic

oligonucleotide arrays. *Nat. Genet.* **21**, 20–24 (1999).

82. Speicher, M. R. & Carter, N. P. The new cytogenetics: blurring the boundaries with molecular biology. *Nat. Rev. Genet.* **6**, 782–792 (2005).
83. Volpi, E. V. & Bridger, J. M. FISH glossary: an overview of the fluorescence in situ hybridization technique. *Biotechniques* **45**, 385–6, 388, 390 passim (2008).
84. Chambeyron, S. & Bickmore, W. A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev.* **18**, 1119–1130 (2004).
85. Eskeland, R. *et al.* Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Mol. Cell* **38**, 452–464 (2010).
86. Mahy, N. L., Perry, P. E., Gilchrist, S., Baldock, R. A. & Bickmore, W. A. Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. *J. Cell Biol.* **157**, 579–589 (2002).
87. Volpi, E. V. *et al.* Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *J. Cell Sci.* **113 (Pt 9)**, 1565–1576 (2000).
88. Williamson, I. *et al.* Anterior-posterior differences in HoxD chromatin topology in limb development. *Development* **139**, 3157–3167 (2012).
89. Fodor, S. P. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991).
90. Liu, G. & Zhang, T. Single Copy Oligonucleotide Fluorescence In Situ Hybridization Probe Design Platforms: Development, Application and Evaluation. *Int. J. Mol. Sci.* **22**, (2021).
91. Matera, A. G. & Ward, D. C. Oligonucleotide probes for the analysis of specific repetitive DNA sequences by fluorescence in situ hybridization. *Hum. Mol. Genet.* **1**, 535–539 (1992).
92. Dernburg, A. F. *et al.* Perturbation of nuclear architecture by long-distance chromosome interactions. *Cell* **85**, 745–759 (1996).
93. Dirks, R. W. *et al.* Simultaneous detection of different mRNA sequences coding for

neuropeptide hormones by double in situ hybridization using FITC- and biotin-labeled oligonucleotides. *Journal of Histochemistry & Cytochemistry* vol. 38 467–473 Preprint at https://doi.org/10.1177/38.4.2108203 (1990).

94. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts in situ. *Science* **280**, 585–590 (1998).
95. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
96. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
97. Yamada, N. A. *et al.* Visualization of fine-scale genomic structure by oligonucleotide-based high-resolution FISH. *Cytogenet. Genome Res.* **132**, 248–254 (2011).
98. Boyle, S., Rodesch, M. J., Halvensleben, H. A., Jeddeloh, J. A. & Bickmore, W. A. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. *Chromosome Res.* **19**, 901–909 (2011).
99. Tian, T., Wang, Y., Wang, H., Zhu, Z. & Xiao, Z. Visualizing of the cellular uptake and intracellular trafficking of exosomes by live-cell microscopy. *J. Cell. Biochem.* **111**, 488–496 (2010).
100. Beliveau, B. J. *et al.* Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nat. Commun.* **6**, 7147 (2015).
101. Ni, Y. *et al.* Super-resolution imaging of a 2.5 kb non-repetitive DNA in situ in the nuclear genome using molecular beacon probes. *Elife* **6**, (2017).
102. Sigal, Y. M., Zhou, R. & Zhuang, X. Visualizing and discovering cellular structures with super-resolution microscopy. *Science* **361**, 880–887 (2018).
103. Huang, B., Wang, W., Bates, M. & Zhuang, X. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science* **319**, 810–813 (2008).

104. Nir, G. *et al.* Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet.* **14**, e1007872 (2018).

105. Braz, G. T. *et al.* Comparative Oligo-FISH Mapping: An Efficient and Powerful Methodology To Reveal Karyotypic and Chromosomal Evolution. *Genetics* **208**, 513–523 (2018).

106. Šimoníková, D. *et al.* Chromosome Painting Facilitates Anchoring Reference Genome Sequence to Chromosomes In Situ and Integrated Karyotyping in Banana (Musa Spp.). *Front. Plant Sci.* **10**, 1503 (2019).

107. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

108. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

109. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

110. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

111. Navin, N. *et al.* PROBER: oligonucleotide FISH probe design software. *Bioinformatics* **22**, 2437–2438 (2006).

112. Zhang, T., Liu, G., Zhao, H. & Braz, G. T. Chorus2: design of genome-scale oligonucleotide-based probes for fluorescence in situ hybridization. *Plant Biotechnol.* (2021).

113. Yilmaz, L. S., Parnerkar, S. & Noguera, D. R. mathFISH, a web tool that uses thermodynamics-based mathematical models for in silico evaluation of oligonucleotide probes for fluorescence in situ hybridization. *Appl. Environ. Microbiol.* **77**, 1118–1122 (2011).

114. Beliveau, B. J. *et al.* OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2183–E2192 (2018).

115. Gelali, E. *et al.* iFISH is a publically available resource enabling versatile DNA FISH to study

genome architecture. *Nature Communications* vol. 10 Preprint at https://doi.org/10.1038/s41467-019-09616-w (2019).

116. Hu, M. *et al.* ProbeDealer is a convenient tool for designing probes for highly multiplexed fluorescence in situ hybridization. *Sci. Rep.* **10**, 22031 (2020).
117. Hershberg, E. A. *et al.* PaintSHOP enables the interactive design of transcriptome- and genome-scale oligonucleotide FISH experiments. Preprint at https://doi.org/10.1101/2020.07.05.188797.
118. Dorman, S. N., Shirley, B. C., Knoll, J. H. M. & Rogan, P. K. Expanding probe repertoire and improving reproducibility in human genomic hybridization. *Nucleic Acids Res.* **41**, e81 (2013).
119. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **25**, 4–10 (2009).
120. James Kent, W. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
121. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
122. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115–e115 (2012).
123. Xin, H. *et al.* Chromosome painting and comparative physical mapping of the sex chromosomes in Populus tomentosa and Populus deltoides. *Chromosoma* **127**, 313–321 (2018).
124. Albert, P. S. *et al.* Whole-chromosome paints in maize reveal rearrangements, nuclear domains, and chromosomal relationships. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 1679–1685 (2019).
125. Liu, X. *et al.* Dual-color oligo-FISH can reveal chromosomal variations and evolution in Oryza species. *Plant J.* **101**, 112–121 (2020).
126. Adilardi, R. S. & Dernburg, A. F. Robust, versatile DNA FISH probes for chromosome-specific repeats in Caenorhabditis elegans and Pristionchus pacificus. *G3* **12**, (2022).

127. Tang, S. *et al.* Developing New Oligo Probes to Distinguish Specific Chromosomal Segments and the A, B, D Genomes of Wheat (Triticum aestivum L.) Using ND-FISH. *Front. Plant Sci.* **9**, 1104 (2018).
128. Lei, J. *et al.* Development of oligonucleotide probes for FISH karyotyping in Haynaldia villosa, a wild relative of common wheat. *The Crop Journal* **8**, 676–681 (2020).
129. Mcclintock, B. Intranuclear systems controlling gene action and mutation. *Brookhaven Symp. Biol.* 58–74 (1956).
130. Kit, S. Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *J. Mol. Biol.* **3**, 711–716 (1961).
131. Thakur, J., Packiaraj, J. & Henikoff, S. Sequence, Chromatin and Evolution of Satellite DNA. *Int. J. Mol. Sci.* **22**, (2021).
132. Peterson, D. G. *et al.* Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* **12**, 795–807 (2002).
133. Willard, H. F. Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.* **37**, 524–532 (1985).
134. Rudd, M. K. & Willard, H. F. Analysis of the centromeric regions of the human genome assembly. *Trends Genet.* **20**, 529–533 (2004).
135. Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. Genomic and genetic definition of a functional human centromere. *Science* **294**, 109–115 (2001).
136. Rudd, M. K., Schueler, M. G. & Willard, H. F. Sequence organization and functional annotation of human centromeres. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 141–149 (2003).
137. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. 1986. *Biotechnology* **24**, 17–27 (1992).
138. Schmid, C. W. & Deininger, P. L. Sequence organization of the human genome. *Cell* **6**, 345–358 (1975).

139. Tyler-Smith, C. & Brown, W. R. Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.* **195**, 457–470 (1987).

140. Blackburn, E. H. & Gall, J. G. A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *J. Mol. Biol.* **120**, 33–53 (1978).

141. Martin, M. A., Bryan, T., Rasheed, S. & Khan, A. S. Identification and cloning of endogenous retroviral sequences present in human DNA. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 4892–4896 (1981).

142. La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).

143. Oberlé, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252**, 1097–1102 (1991).

144. Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).

145. Yu, S. *et al.* Fragile X genotype characterized by an unstable region of DNA. *Science* **252**, 1179–1181 (1991).

146. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

147. Novák, P., Neumann, P., Pech, J., Steinhaisl, J. & Macas, J. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* **29**, 792–793 (2013).

148. Novák, P. *et al.* TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **45**, e111 (2017).

149. George, C. M. & Alani, E. Multiple cellular mechanisms prevent chromosomal rearrangements involving repetitive DNA. *Crit. Rev. Biochem. Mol. Biol.* **47**, 297–313 (2012).

150. Chang, C.-H. *et al.* Islands of retroelements are major components of Drosophila centromeres. *PLoS Biol.* **17**, e3000241 (2019).
151. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
152. Mager, D. L. & Stoye, J. P. Mammalian Endogenous Retroviruses. *Microbiol Spectr* **3**, MDNA3-0009–2014 (2015).
153. Richardson, S. R. *et al.* The influence of LINE-1 and SINE retrotransposons on mammalian genomes. in *Mobile DNA III* 1165–1208 (ASM Press, 2015).
154. Miga, K. H. & Alexandrov, I. A. Variation and Evolution of Human Centromeres: A Field Guide and Perspective. *Annu. Rev. Genet.* **55**, 583–602 (2021).
155. Garrido-Ramos, M. A. Satellite DNA: An Evolving Topic. *Genes* **8**, (2017).
156. Logsdon, G. A. & Eichler, E. E. The Dynamic Structure and Rapid Evolution of Human Centromeric Satellite DNA. *Genes* **14**, (2022).
157. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. Alpha-satellite DNA of primates: old and new families. *Chromosoma* **110**, 253–266 (2001).
158. Kipling, D., Ackford, H. E., Taylor, B. A. & Cooke, H. J. Mouse minor satellite DNA genetically maps to the centromere and is physically linked to the proximal telomere. *Genomics* **11**, 235–241 (1991).
159. Kuznetsova, I. S., Prusov, A. N., Enukashvily, N. I. & Podgornaya, O. I. New types of mouse centromeric satellite DNAs. *Chromosome Res.* **13**, 9–25 (2005).
160. Garagna, S. *et al.* Genome distribution, chromosomal allocation, and organization of the major and minor satellite DNAs in 11 species and subspecies of the genus Mus. *Cytogenet. Cell Genet.* **64**, 247–255 (1993).
161. Boursot, P., Auffray, J.-C., Britton-Davidian, J. & Bonhomme, F. The evolution of house mice. *Annu. Rev. Ecol. Syst.* **24**, 119–152 (1993).
162. Wang, Y. *et al.* Euchromatin and pericentromeric heterochromatin: comparative composition

in the tomato genome. *Genetics* **172**, 2529–2540 (2006).

163. Novák, P. *et al.* Genome-wide analysis of repeat diversity across the family Musaceae. *PLoS One* **9**, e98918 (2014).

164. Ananiev, E. V., Phillips, R. L. & Rines, H. W. Chromosome-specific molecular organization of maize (Zea mays L.) centromeric regions. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 13073–13078 (1998).

165. Melters, D. P. *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).

166. Lermontova, I., Sandmann, M. & Demidov, D. Centromeres and kinetochores of Brassicaceae. *Chromosome Res.* **22**, 135–152 (2014).

167. Jones, K. W. Chromosomal and nuclear location of mouse satellite DNA in individual cells. *Nature* **225**, 912–915 (1970).

168. Fransz, P., De Jong, J. H., Lysak, M., Castiglione, M. R. & Schubert, I. Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14584–14589 (2002).

169. Jagannathan, M., Cummings, R. & Yamashita, Y. M. A conserved function for pericentromeric satellite DNA. *Elife* **7**, (2018).

170. Brändle, F., Frühbauer, B. & Jagannathan, M. Principles and functions of pericentromeric satellite DNA clustering into chromocenters. *Semin. Cell Dev. Biol.* **128**, 26–39 (2022).

171. Blackburn, E. H. Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. *FEBS Lett.* **579**, 859–862 (2005).

172. Bebikhov, D. V. Repeating sequences, organizing the telomeric region of chromosomes from the eukaryotic genome. *Genetika* **29**, 373–387 (1993).

173. Vítková, M., Král, J., Traut, W., Zrzavý, J. & Marec, F. The evolutionary origin of insect telomeric repeats, (TTAGG) N. *Chromosome Res.* **13**, 145–156 (2005).

174. Riha, K. & Shippen, D. E. Telomere structure, function and maintenance in Arabidopsis.

Chromosome Res. **11**, 263–275 (2003).

175. Makarov, V. L., Hirose, Y. & Langmore, J. P. Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. *Cell* **88**, 657–666 (1997).

176. Lower, S. S., McGurk, M. P., Clark, A. G. & Barbash, D. A. Satellite DNA evolution: old ideas, new approaches. *Curr. Opin. Genet. Dev.* **49**, 70–78 (2018).

177. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* vol. 37 1155–1162 Preprint at https://doi.org/10.1038/s41587-019-0217-9 (2019).

178. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).

179. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

180. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).

181. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).

182. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).

183. Miga, K. H. Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes* **10**, (2019).

184. Cechova, M. & Miga, K. H. Comprehensive variant discovery in the era of complete human reference genomes. *Nat. Methods* **20**, 17–19 (2023).

185. Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3**, e90 (2007).

186. Siva, N. 1000 Genomes project. *Nat. Biotechnol.* **26**, 256 (2008).

187. Lupski, J. R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* **1**, e49 (2005).

188. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. L. All y'all need to know 'bout retroelements in cancer. *Semin. Cancer Biol.* **20**, 200–210 (2010).

189. Black, E. M. & Giunta, S. Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes* **9**, (2018).

190. Mitelman, F., Mertens, F. & Johansson, B. A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nat. Genet.* **15**, 417–474 (1997).

191. Padilla-Nash, H. M. *et al.* Jumping translocations are common in solid tumor cell lines and result in recurrent fusions of whole chromosome arms. *Genes Chromosomes Cancer* **30**, 349–363 (2001).

192. Kim, T.-M. *et al.* Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res.* **23**, 217–227 (2013).

193. Ricke, R. M. & van Deursen, J. M. Aneuploidy in health, disease, and aging. *J. Cell Biol.* **201**, 11–21 (2013).

194. Giunta, S. & Funabiki, H. Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1928–1933 (2017).

195. Fournier, A. *et al.* 1q12 chromosome translocations form aberrant heterochromatic foci associated with changes in nuclear architecture and gene expression in B cell lymphoma. *EMBO Mol. Med.* **2**, 159–171 (2010).

196. Tomonaga, T. *et al.* Overexpression and mistargeting of centromere protein-A in human primary colorectal cancer. *Cancer Res.* **63**, 3511–3516 (2003).

197. Amato, A., Schillaci, T., Lentini, L. & Di Leonardo, A. CENPA overexpression promotes genome instability in pRb-depleted human cells. *Mol. Cancer* **8**, 119 (2009).

198. Li, Y. *et al.* ShRNA-targeted centromere protein A inhibits hepatocellular carcinoma growth. *PLoS One* **6**, e17794 (2011).

199. Zhang, W. *et al.* Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nat. Commun.* **7**, 12619 (2016).
200. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
201. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
202. Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? *Trends Genet.* **23**, 183–191 (2007).
203. Csink, A. K. & Henikoff, S. Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.* **14**, 200–204 (1998).
204. Meštrović, N. *et al.* Structural and functional liaisons between transposable elements and satellite DNAs. *Chromosome Res.* **23**, 583–596 (2015).
205. Parks, M. M., Lawrence, C. E. & Raphael, B. J. Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biol.* **16**, 72 (2015).
206. Lee, Y. C. G. *et al.* Pericentromeric heterochromatin is hierarchically organized and spatially contacts H3K9me2 islands in euchromatin. *PLoS Genet.* **16**, e1008673 (2020).
207. Mestrović, N., Plohl, M., Mravinac, B. & Ugarković, D. Evolution of satellite DNAs from the genus Palorus--experimental evidence for the “library” hypothesis. *Mol. Biol. Evol.* **15**, 1062–1068 (1998).
208. Fry, K. & Salser, W. Nucleotide sequences of HS-alpha satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. *Cell* **12**, 1069–1084 (1977).
209. del Bosque, M. E. Q., Navajas-Pérez, R., Panero, J. L., Fernández-González, A. & Garrido-Ramos, M. A. A satellite DNA evolutionary analysis in the North American endemic dioecious plant Rumex hastatulus (Polygonaceae). *Genome* **54**, 253–260 (2011).
210. del Bosque, M. E. Q., López-Flores, I., Suárez-Santiago, V. N. & Garrido-Ramos, M. A. Satellite-DNA diversification and the evolution of major lineages in Cardueae (Carduoideae Asteraceae). *J. Plant Res.* **127**, 575–583 (2014).

211. Feliciello, I., Picariello, O. & Chinali, G. The first characterisation of the overall variability of repetitive units in a species reveals unexpected features of satellite DNA. *Gene* **349**, 153–164 (2005).

212. Quesada del Bosque, M. E., López-Flores, I., Suárez-Santiago, V. N. & Garrido-Ramos, M. A. Differential spreading of HinfI satellite DNA variants during radiation in Centaureinae. *Ann. Bot.* **112**, 1793–1802 (2013).

213. Wei, K. H.-C., Grenier, J. K., Barbash, D. A. & Clark, A. G. Correlated variation and population differentiation in satellite DNA abundance among lines of Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 18793–18798 (2014).

214. Cutter, A. D. The polymorphic prelude to Bateson–Dobzhansky–Muller incompatibilities. *Trends Ecol. Evol.* **27**, 209–218 (2012).

215. Packiaraj, J. & Thakur, J. DNA satellite and chromatin organization at house mouse centromeres and pericentromeres. *bioRxiv* (2023) doi:10.1101/2023.07.18.549612.

216. Kimura, M. & Crow, J. F. THE NUMBER OF ALLELES THAT CAN BE MAINTAINED IN A FINITE POPULATION. *Genetics* **49**, 725–738 (1964).

217. Smith, G. P. Evolution of Repeated DNA Sequences by Unequal Crossover. *Science* **191**, 528–535 (1976).

218. Iwata-Otsubo, A. *et al.* Expanded Satellite Repeats Amplify a Discrete CENP-A Nucleosome Assembly Site on Chromosomes that Drive in Female Meiosis. *Curr. Biol.* **27**, 2365-2373.e8 (2017).

219. Chmátal, L. *et al.* Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Curr. Biol.* **24**, 2295–2300 (2014).

220. Fishman, L. & Saunders, A. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science* **322**, 1559–1562 (2008).

221. McKinley, K. L. & Cheeseman, I. M. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* **17**, 16–29 (2016).

222. Velazquez Camacho, O. *et al.* Major satellite repeat RNA stabilize heterochromatin retention of Suv39h enzymes by RNA-nucleosome association and RNA:DNA hybrid formation. *Elife* **6**, (2017).
223. Lehnertz, B. *et al.* Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr. Biol.* **13**, 1192–1200 (2003).
224. Orr, H. T. & Zoghbi, H. Y. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* **30**, 575–621 (2007).
225. Gall, J. G. The human nucleolus organizer regions. *Genes Dev.* **33**, 1617–1618 (2019).
226. Hirai, H. Chromosome Dynamics Regulating Genomic Dispersion and Alteration of Nucleolus Organizer Regions (NORs). *Cells* **9**, (2020).
227. Stimpson, K. M., Sullivan, L. L., Kuo, M. E. & Sullivan, B. A. Nucleolar organization, ribosomal DNA array stability, and acrocentric chromosome integrity are linked to telomere function. *PLoS One* **9**, e92432 (2014).
228. Bury, L. *et al.* Alpha-satellite RNA transcripts are repressed by centromere–nucleolus associations. *Elife* **9**, e59770 (2020).
229. Kato, A., Lamb, J. C. & Birchler, J. A. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13554–13559 (2004).
230. Picart-Picolo, A., Picault, N. & Pontvianne, F. Ribosomal RNA genes shape chromatin domains associating with the nucleolus. *Nucleus* **10**, 67–72 (2019).
231. Choudhary, M. N. K. *et al.* Publisher Correction: Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol.* **21**, 28 (2020).
232. Allshire, R. C. & Ekwall, K. Epigenetic Regulation of Chromatin States in Schizosaccharomyces pombe. *Cold Spring Harb. Perspect. Biol.* **7**, a018770 (2015).
233. Aguilar, R. *et al.* Tigerfish designs oligonucleotide-based in situ hybridization probes targeting intervals of highly repetitive DNA at the scale of genomes. *bioRxiv* (2023)

doi:10.1101/2023.03.06.530899.

234. Fox, K. The Illusion of Inclusion - The “All of Us” Research Program and Indigenous Peoples’ DNA. *N. Engl. J. Med.* **383**, 411–413 (2020).

235. Carlson, J., Henn, B. M., Al-Hindi, D. R. & Ramachandran, S. Counter the weaponization of genetics research by extremists. *Nature* **610**, 444–447 (2022).

236. Gouvea, J. S. Addressing Racism in Human Genetics and Genomics Education. *CBE Life Sci. Educ.* **21**, fe5 (2022).

237. Langer-Safer, P. R., Levine, M. & Ward, D. C. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 4381–4385 (1982).

238. Lawrence, J. B. & Singer, R. H. Quantitative analysis of in situ hybridization methods for the detection of actin gene expression. *Nucleic Acids Res.* **13**, 1777–1799 (1985).

239. Lewis, M. E., Sherman, T. G. & Watson, S. J. In situ hybridization histochemistry with synthetic oligonucleotides: strategies and methods. *Peptides* **6 Suppl 2**, 75–87 (1985).

240. Beliveau, B. J. *et al.* Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proceedings of the National Academy of Sciences* **109**, 21301–21306 (2012).

241. Wang, S. *et al.* Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598–602 (2016).

242. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, (2018).

243. Mateo, L. J. *et al.* Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* vol. 568 49–54 Preprint at https://doi.org/10.1038/s41586-019-1035-4 (2019).

244. Takei, Y. *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021).

245. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature methods* vol. 11 360–361 (2014).

246. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

247. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92**, 342–357 (2016).

248. Rouillard, J. M., Zuker, M. & Gulari, E. OligoArray 2.0: Design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.* **31**, 3057–3062 (2003).

249. Han, Y., Zhang, T., Thammapichai, P., Weng, Y. & Jiang, J. Chromosome-Specific Painting in Cucumis Species Using Bulked Oligonucleotides. *Genetics* **200**, 771–779 (2015).

250. Hershberg, E. A. *et al.* PaintSHOP enables the interactive design of transcriptome- and genome-scale oligonucleotide FISH experiments. *Nat. Methods* **18**, 937–944 (2021).

251. Landegent, J. E., Jansen in de Wal, N., Dirks, R. W., Baao, F. & van der Ploeg, M. Use of whole cosmid cloned genomic sequences for chromosomal localization by non-radioactive in situ hybridization. *Hum. Genet.* **77**, 366–370 (1987).

252. Pinkel, D. *et al.* Fluorescence in situ hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 9138–9142 (1988).

253. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015. http://www.repeatmasker.org.

254. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).

255. Gonzalez, I. L. & Sylvester, J. E. Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics* **27**, 320–328 (1995).

256. Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J. & Maltais, L. J. The human and mouse replication-dependent histone genes. *Genomics* **80**, 487–498 (2002).

257. Franke, V. *et al.* Long terminal repeats power evolution of genes and gene expression

programs in mammalian oocytes and zygotes. *Genome Res.* **27**, 1384–1394 (2017).

258. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).

259. Riegel, M. Human molecular cytogenetics: From cells to nucleotides. *Genetics and Molecular Biology* vol. 37 194–209 Preprint at https://doi.org/10.1590/S1415-47572014000200006 (2014).

260. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).

261. Paez, S. *et al.* Reference genomes for conservation. *Science* **377**, 364–366 (2022).

262. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

263. Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173 (2011).

264. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

265. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* vol. 25 1422–1423 Preprint at https://doi.org/10.1093/bioinformatics/btp163 (2009).

266. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

267. Anand, L. & Rodriguez Lopez, C. M. ChromoMap: an R package for interactive visualization of multi-omics data and annotation of chromosomes. *BMC Bioinformatics* **23**, 33 (2022).

268. Lipman, D. J. & Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441 (1985).

269. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

270. Kishi, J. Y. *et al.* SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nat. Methods* **16**, 533–544 (2019).
271. Choi, H. M. T., Beck, V. A. & Pierce, N. A. Next-generation in situ hybridization chain reaction: Higher gain, lower cost, greater durability. *ACS Nano* **8**, 4284–4294 (2014).
272. Banér, J., Nilsson, M., Mendel-Hartvig, M. & Landegren, U. Signal amplification of padlock probes by rolling circle replication. *Nucleic Acids Res.* **26**, 5073–5078 (1998).
273. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* vol. 26 841–842 Preprint at https://doi.org/10.1093/bioinformatics/btq033 (2010).
274. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
275. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010). doi:10.25080/majora-92bf1922-00a.
276. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
277. LaFave, M. C. & Burgess, S. M. sam2pairwise version 1.0. 0. 2014.
278. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
279. Hunter. Matplotlib: A 2D Graphics Environment. **9**, 90–95 (2007).
280. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
281. Krekel, H. *et al.* pytest 5.3. 2. Preprint at (2004).
282. Anaconda, I. Anaconda Software Distribution. *Computer software* (2014).
283. Garreta, R. & Moncecchi, G. *Learning scikit-learn: Machine Learning in Python*. (Packt Publishing Ltd, 2013).
284. Anand, L. & Rodriguez Lopez, C. M. chromoMap: An R package for Interactive Visualization and Annotation of Chromosomes. Preprint at https://doi.org/10.1101/605600.
285. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image

analysis. *Nat. Methods* **9**, 671–675 (2012).

286. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

287. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

288. Tyler-Smith, C. *et al.* Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. *Nat. Genet.* **5**, 368–375 (1993).

289. Qu, G.-Z., Dubeau, L., Narayan, A., Yu, M. C. & Ehrlich, M. Satellite DNA hypomethylation vs. overall genomic hypomethylation in ovarian epithelial tumors of different malignant potential. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* vol. 423 91–101 Preprint at https://doi.org/10.1016/s0027-5107(98)00229-2 (1999).

290. Peng, J. C. & Karpen, G. H. Epigenetic regulation of heterochromatic DNA stability. *Current Opinion in Genetics & Development* vol. 18 204–211 Preprint at https://doi.org/10.1016/j.gde.2008.01.021 (2008).

291. Maloney, K. A. *et al.* Functional epialleles at an endogenous human centromere. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 13704–13709 (2012).

292. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun Biol* **2**, 9 (2019).

293. Wagner, J. K. *et al.* Fostering Responsible Research on Ancient DNA. *Am. J. Hum. Genet.* **107**, 183–195 (2020).

294. Bardill, J. *et al.* Advancing the ethics of paleogenomics. *Science* **360**, 384–385 (2018).

295. Prendergast, M. E. & Sawchuk, E. Boots on the ground in Africa's ancient DNA 'revolution': archaeological perspectives on ethics and best practices. *Antiquity* **92**, 803–815 (2018).

296. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

297. Jackson, L., Kuhlman, C., Jackson, F. & Fox, P. K. Including Vulnerable Populations in the Assessment of Data From Vulnerable Populations. *Front Big Data* **2**, 19 (2019).

298. Cavalli-Sforza, L. L. The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* **6**, 333–340 (2005).

299. Garrison, N. A. *et al.* Genomic Research Through an Indigenous Lens: Understanding the Expectations. *Annu. Rev. Genomics Hum. Genet.* **20**, 495–517 (2019).

300. Henrietta Lacks: science must right a historical wrong. *Nature Publishing Group UK* http://dx.doi.org/10.1038/d41586-020-02494-z (2020) doi:10.1038/d41586-020-02494-z.

301. Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).

302. Wolinetz, C. D. & Collins, F. S. Recognition of Research Participants' Need for Autonomy: Remembering the Legacy of Henrietta Lacks. *JAMA* **324**, 1027–1028 (2020).

303. Rendleman, D. & Roberts, C. Memorandum of Amici Curiae Doug Rendleman & Caprice Roberts in Support of Plaintiff: Estate of Henrietta Lacks v. Thermo Fisher Scientific. (2022).

304. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).

305. Rose, S. Racism refuted. *Nature Publishing Group UK* http://dx.doi.org/10.1038/274738a0 (1978) doi:10.1038/274738a0.

306. Dawkins, R. Selfish genes in race or politics. *Nature Publishing Group UK* http://dx.doi.org/10.1038/289528a0 (1981) doi:10.1038/289528a0.

307. Jorde, L. B. & Wooding, S. P. Genetic variation, classification and "race." *Nat. Genet.* **36**, S28–S33 (2004).

308. Rotimi, C. N. Are medical and nonmedical uses of large-scale genomic markers conflating genetics and "race"? *Nat. Genet.* **36**, S43–S47 (2004).

309. Yudell, M., Roberts, D., DeSalle, R. & Tishkoff, S. Taking race out of human genetics. *Science* **351**, 564–565 (2016).

310. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).

311. Kukutai, T. & Taylor, J. *Indigenous data sovereignty: Toward an agenda*. (ANU press, 2016).
312. Cech, E. A. The intersectional privilege of white able-bodied heterosexual men in STEM. *Sci Adv* **8**, eabo1558 (2022).
313. Jimenez, M. F. *et al.* Underrepresented faculty play a disproportionate role in advancing diversity and inclusion. *Nat Ecol Evol* **3**, 1030–1033 (2019).
314. Stanton, J. D. *et al.* Drawing on Internal Strengths and Creating Spaces for Growth: How Black Science Majors Navigate the Racial Climate at a Predominantly White Institution to Succeed. *CBE Life Sci. Educ.* **21**, ar3 (2022).
315. Ostrove, J. M. & Long, S. M. Social Class and Belonging: Implications for College Adjustment. *rhe* **30**, 363–389 (2007).
316. Aguilar, R. Breaking the binary by coming out as a trans scientist. *Nature* **591**, 334+ (2021).
317. Davis, L. K. Human genetics needs an antiracism plan. *Sci. Am.*
318. Dukes, A. How to better support Black trainees in the biomedical sciences. *Nat. Med.* **26**, 1674 (2020).
319. Singleton, K. S., Murray, D.-S. R. K., Dukes, A. J. & Richardson, L. N. S. A year in review: Are diversity, equity, and inclusion initiatives fixing systemic barriers? *Neuron* **109**, 3365–3367 (2021).
320. Reinholz, D. L. & Ridgway, S. W. Access Needs: Centering Students and Disrupting Ableist Norms in STEM. *CBE Life Sci. Educ.* **20**, es8 (2021).
321. Valdez-Ward, E., Ulrich, R. N., Marcus, T. S. & Cat, L. A. Reclaiming STEM: Bringing your Identity and Culture to STEM. in vol. 2019 U33B-05 (ui.adsabs.harvard.edu, 2019).
322. Bernard, D. L., Lige, Q. M., Willis, H. A., Sosoo, E. E. & Neblett, E. W. Impostor phenomenon and mental health: The influence of racial discrimination and gender. *J. Couns. Psychol.* **64**, 155–166 (2017).
323. Correction: Language Matters: Considering Microaggressions in Science. *CBE Life Sci. Educ.* **19**, co2 (2020).

324. McGee, E. O. *Black, Brown, Bruised: How Racialized STEM Education Stifles Innovation*. (Harvard Education Press, 2021).

325. Lesiak, A. J. I'm a trans scientist - here's my advice for navigating academia. *Nature* (2023) doi:10.1038/d41586-023-00923-3.

326. Murray, D.-S. *et al.* Black In Neuro, Beyond One Week. *J. Neurosci.* **41**, 2314–2317 (2021).

327. Taffe, M. A. & Gilpin, N. W. Equity, diversity and inclusion: Racial inequity in grant funding from the US National Institutes of Health. eLife, 10, Article e65697. Preprint at (2021).

328. Davis, F. M. DEI conversations: more than a box-ticking exercise. *Nat. Rev. Mol. Cell Biol.* **24**, 238 (2023).

329. Prescod-Weinstein, C. *The disordered cosmos: A journey into dark matter, spacetime, and dreams deferred*. (Hachette UK, 2021).

330. Montgomery, B. L. *Lessons from Plants*. (Harvard University Press, 2021).

331. Peña, L. G. *Community as Rebellion: A Syllabus for Surviving Academia as a Woman of Color*. (Haymarket Books, 2022).

332. Liu, S.-N. C., Brown, S. E. V. & Sabat, I. E. Patching the "leaky pipeline": Interventions for women of color faculty in STEM academia. *Arch. Sci. Psychol.* **7**, 32–39 (2019).

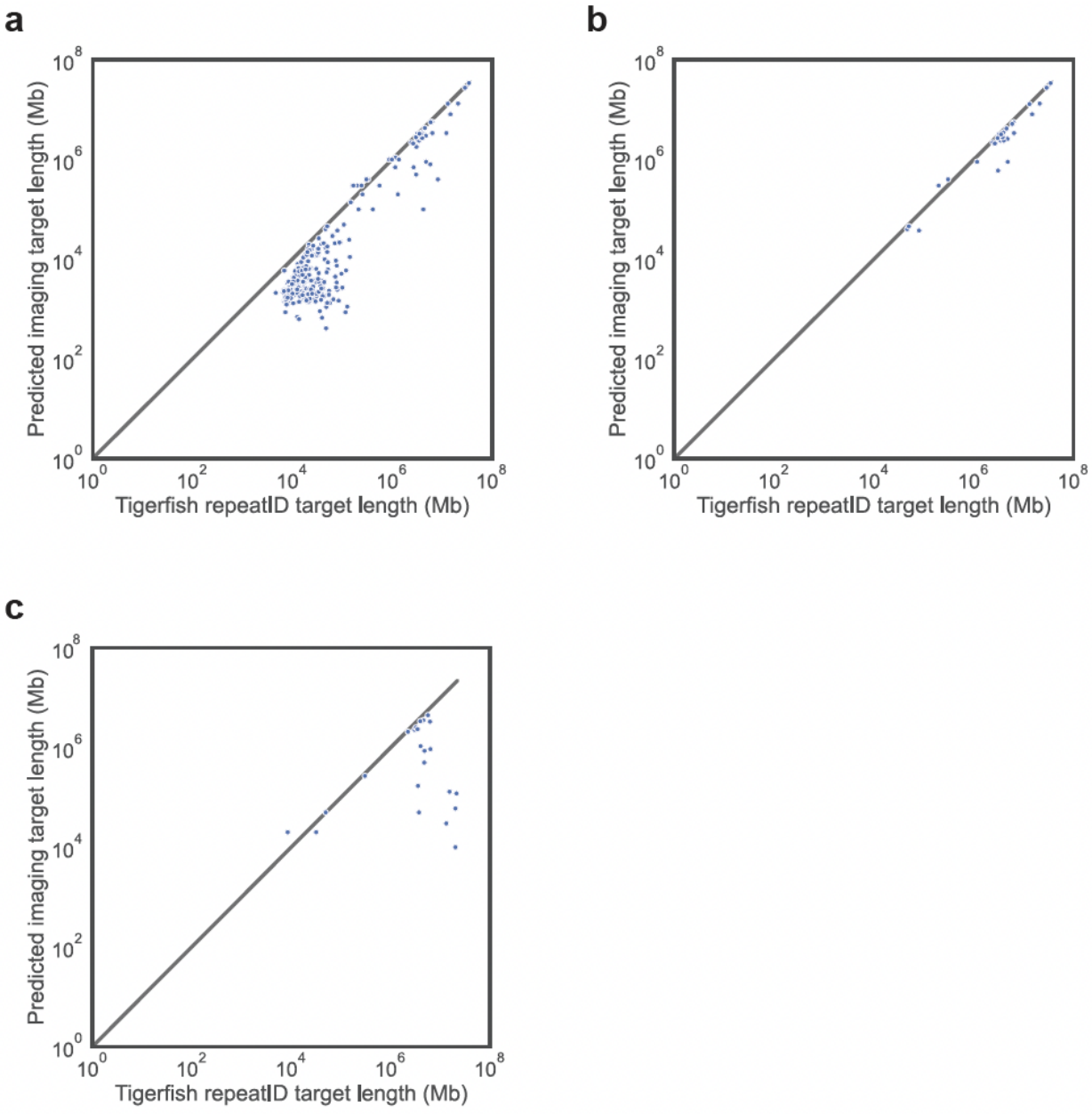
333. Chatterjee, D. *et al.* Career self-efficacy disparities in underrepresented biomedical scientist trainees. *PLoS One* **18**, e0280608 (2023).

334. Taylor, M. B. *et al.* yEvo: experimental evolution in high school classrooms selects for novel mutations that impact clotrimazole resistance in Saccharomyces cerevisiae. *G3* **12**, jkac246 (2022).

335. Tsue, A. F. *et al.* Oligonucleotide-directed proximity-interactome mapping (O-MAP): A unified method for discovering RNA-interacting proteins, transcripts and genomic loci in situ. *bioRxiv* (2023) doi:10.1101/2023.01.19.524825.

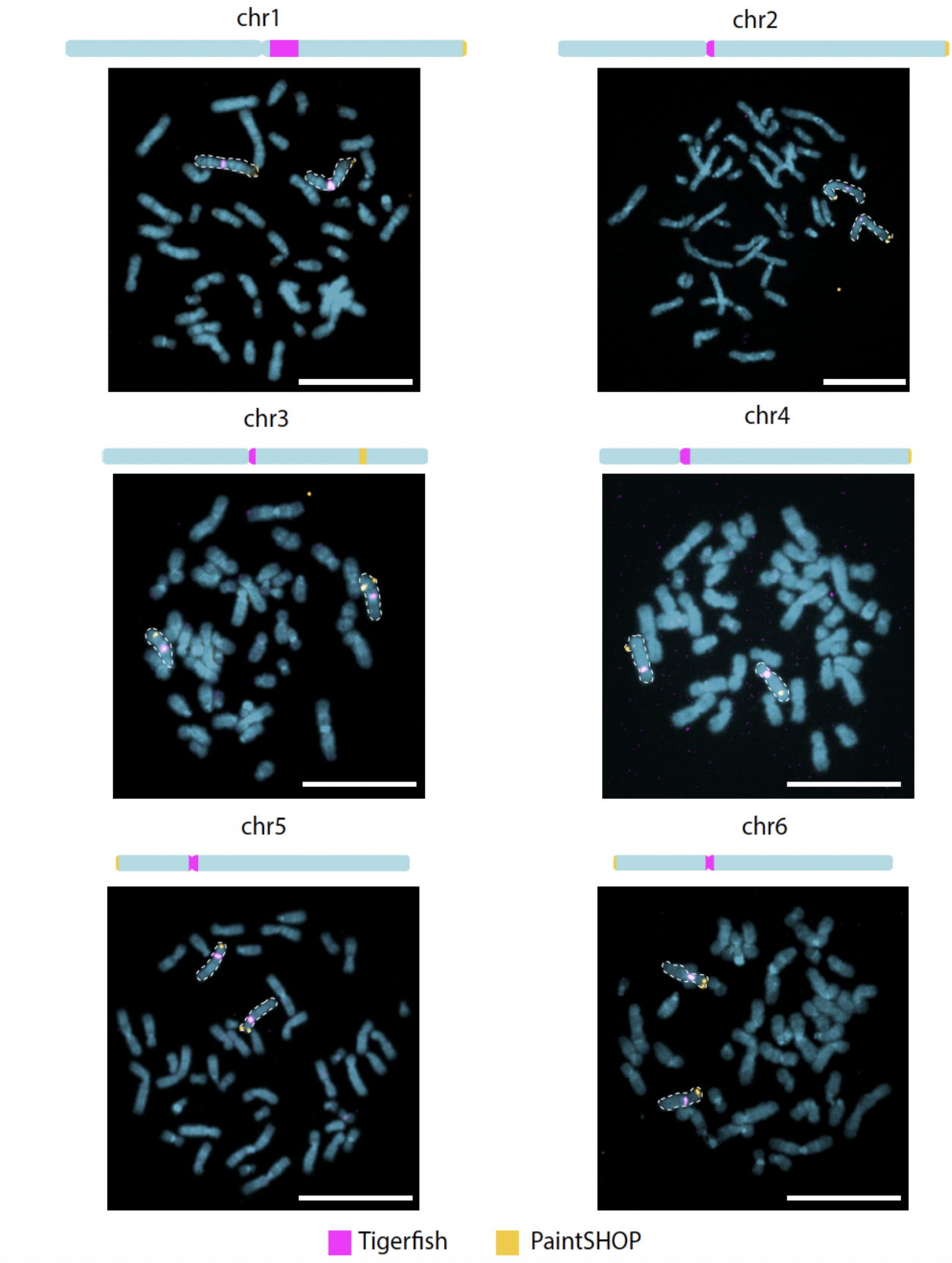
Appendices

Appendix A - Tigerfish Supplementary Data and Software

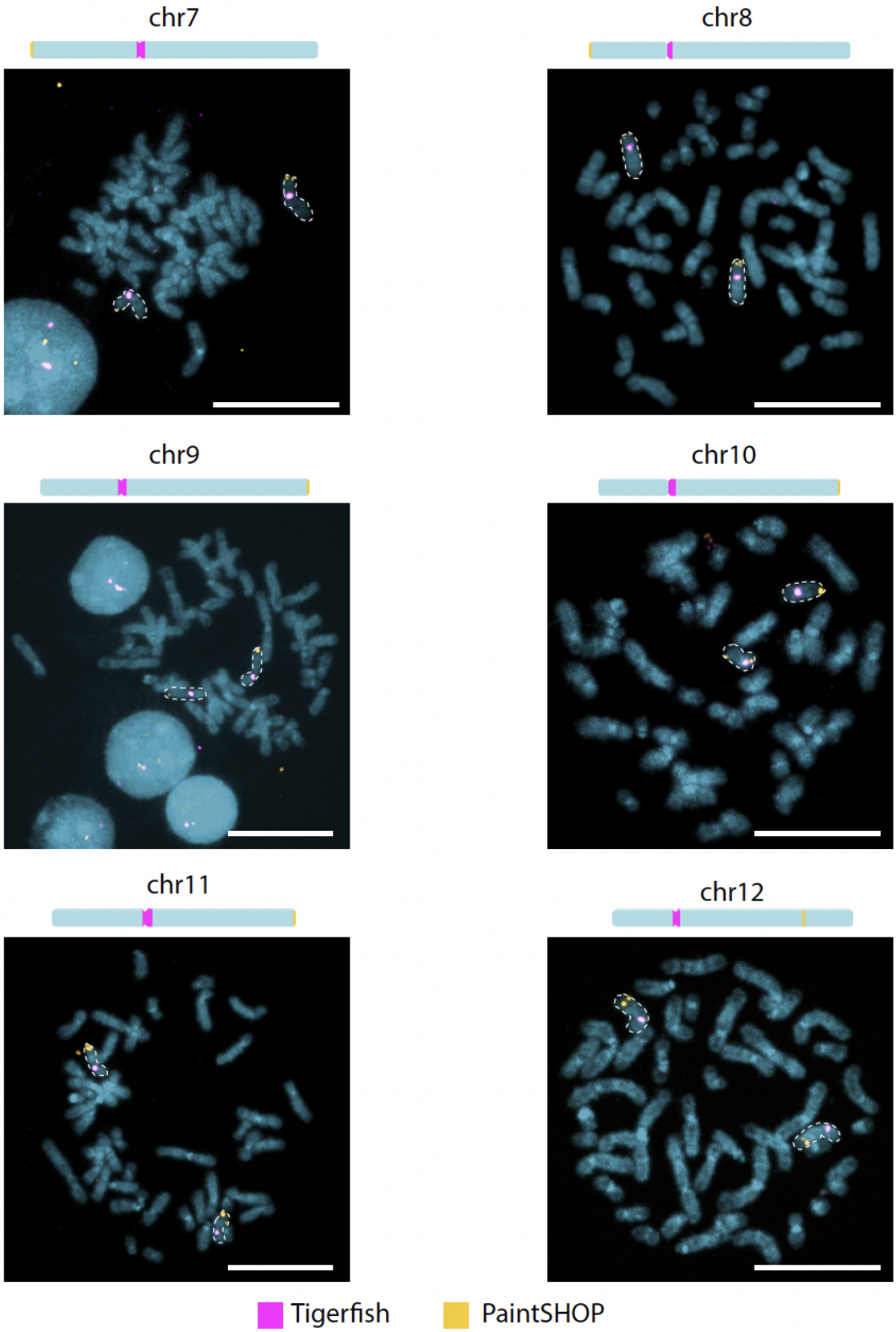


Supplementary Fig. 1 | Effective lengths of Tigerfish target intervals. a, Scatter plot depicting the effective lengths of intervals identified and processed successfully for probe design using permissive parameters (Y-axis) and the length of these intervals when first identified at the

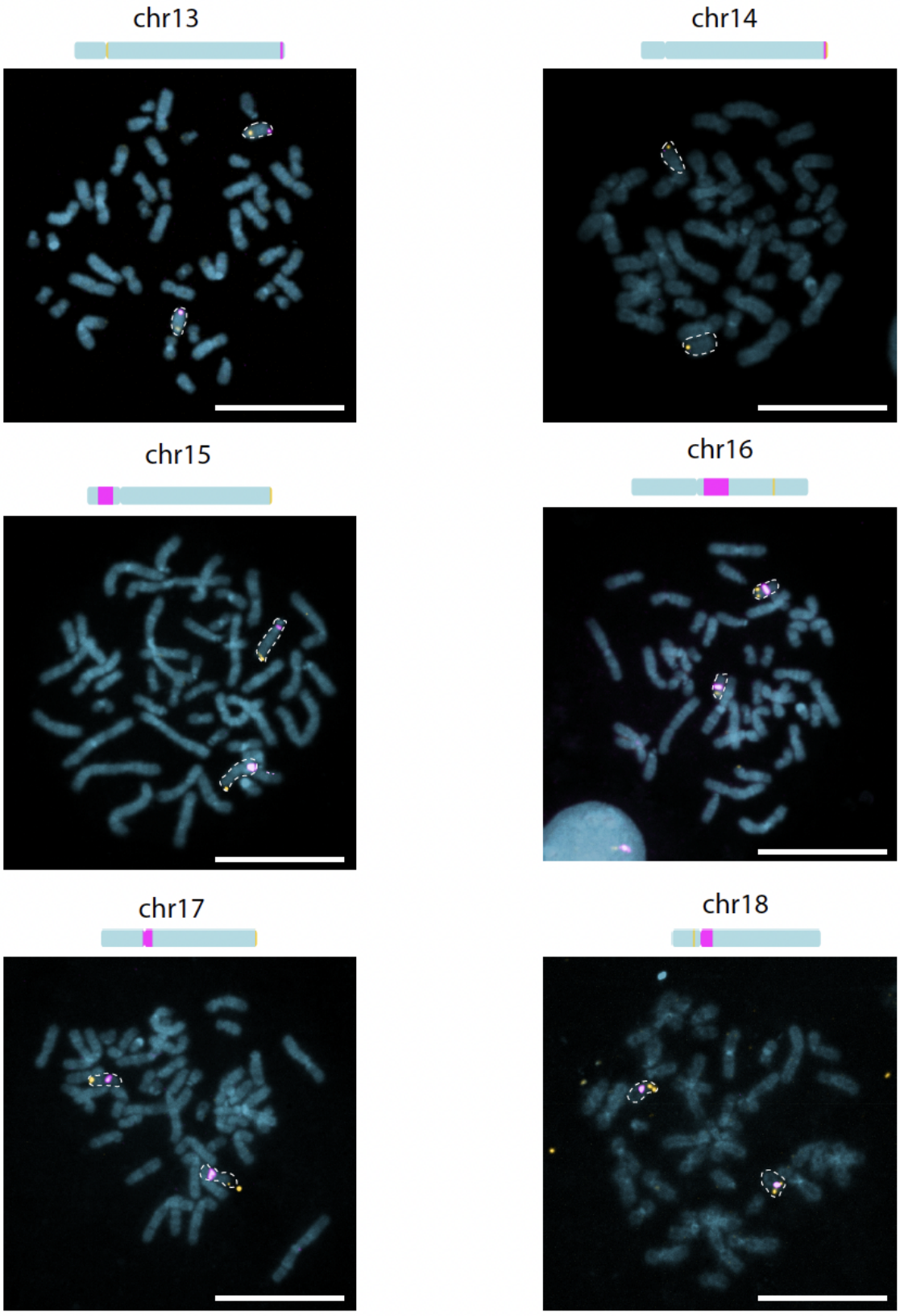
“Repeat discovery” step. b, Scatter plot depicting the effective lengths of intervals identified and processed successfully for probe design using conservative parameters (Y-axis) and the length of these intervals when first identified at the “Repeat discovery” step. c, Scatter plot depicting the effective lengths of intervals inputted for probe design (Y-axis) and the length of these intervals when first identified at the “Repeat discovery” step for the 24-target panel used for in situ validation experiments.


chr1
chr2
chr3
chr4
chr5
chr6
Tigerfish
PaintSHOP

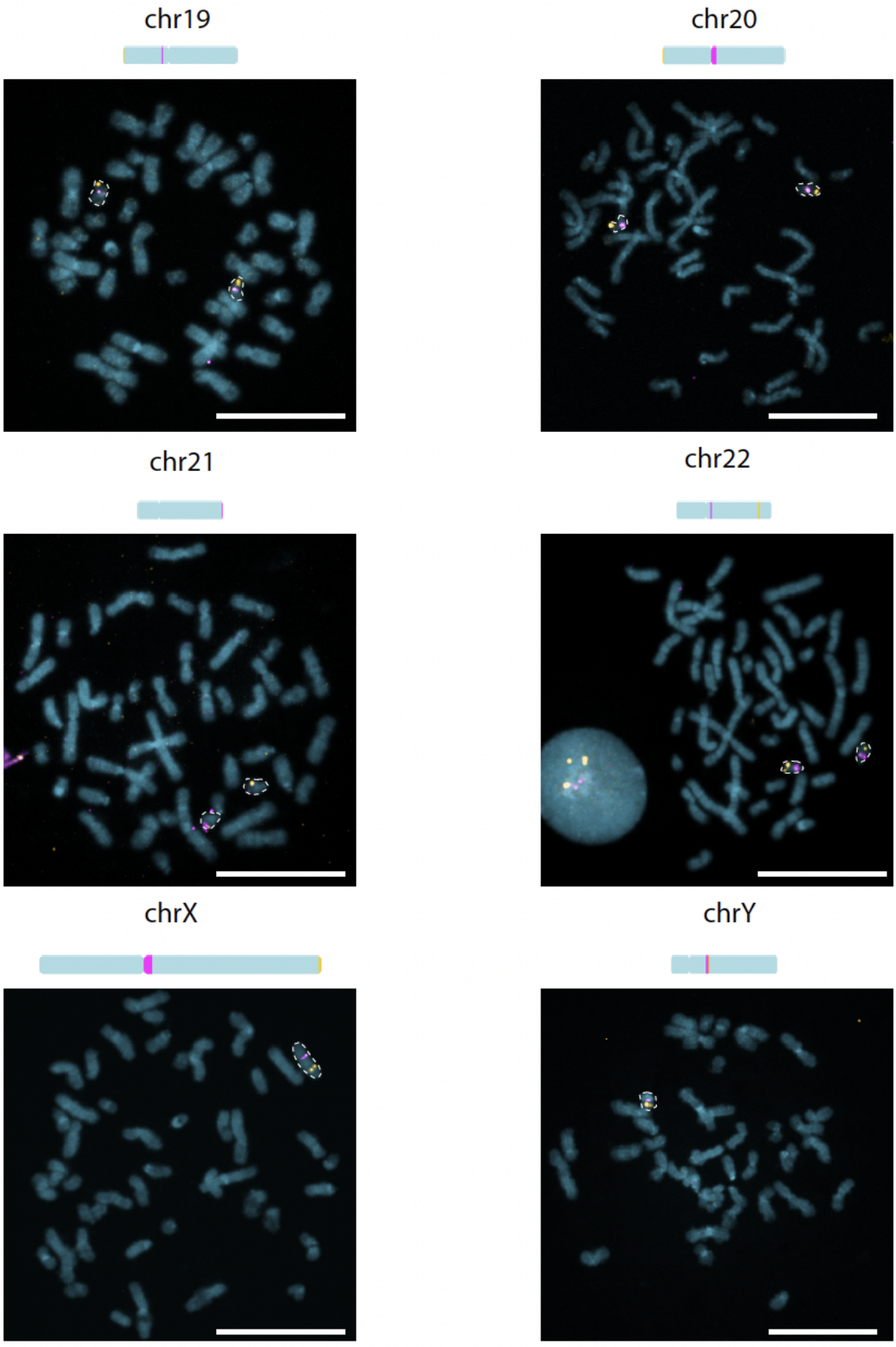
Supplementary Fig. 2 | Full-field metaphase spreads for chromosomes 1–6. Full-field images of the metaphase spreads from which the crops depicted in Figure 4b originated showing the staining pattern of the indicated Tigerfish (magenta) and PaintSHOP (yellow) probe sets. Images are maximum intensity projections in Z. Scale bars, 20 µm.


chr7
chr8
chr9
chr10
chr11
chr12
Tigerfish
PaintSHOP

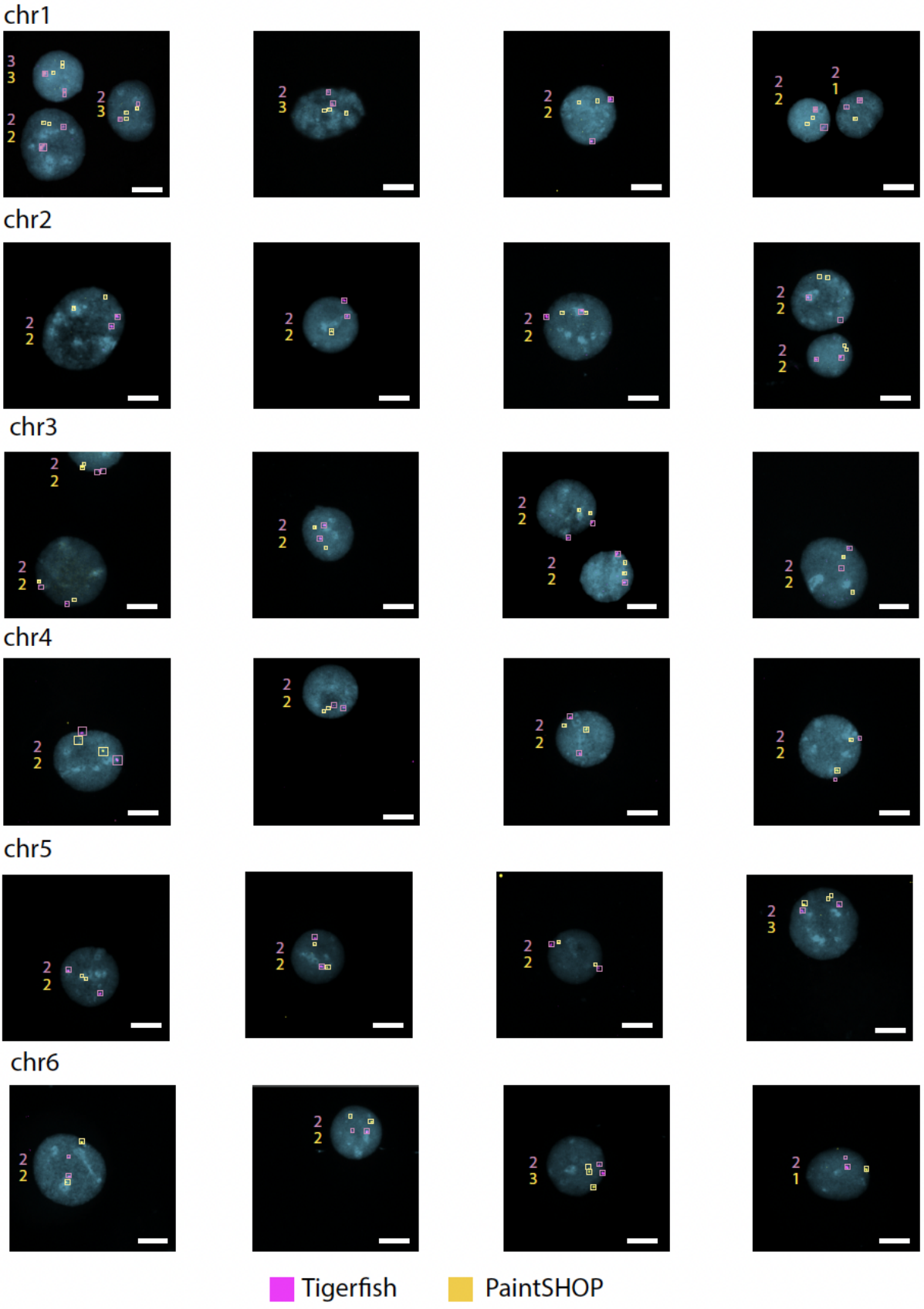
Supplementary Fig. 3 | Full-field metaphase spreads for chromosomes 7–12. Full-field images of the metaphase spreads from which the crops depicted in Figure 4b originated showing the staining pattern of the indicated Tigerfish (magenta) and PaintSHOP (yellow) probe sets. Images are maximum intensity projections in Z. Scale bars, 20 µm.


chr13
chr14
chr15
chr16
chr17
chr18
Tigerfish
PaintSHOP

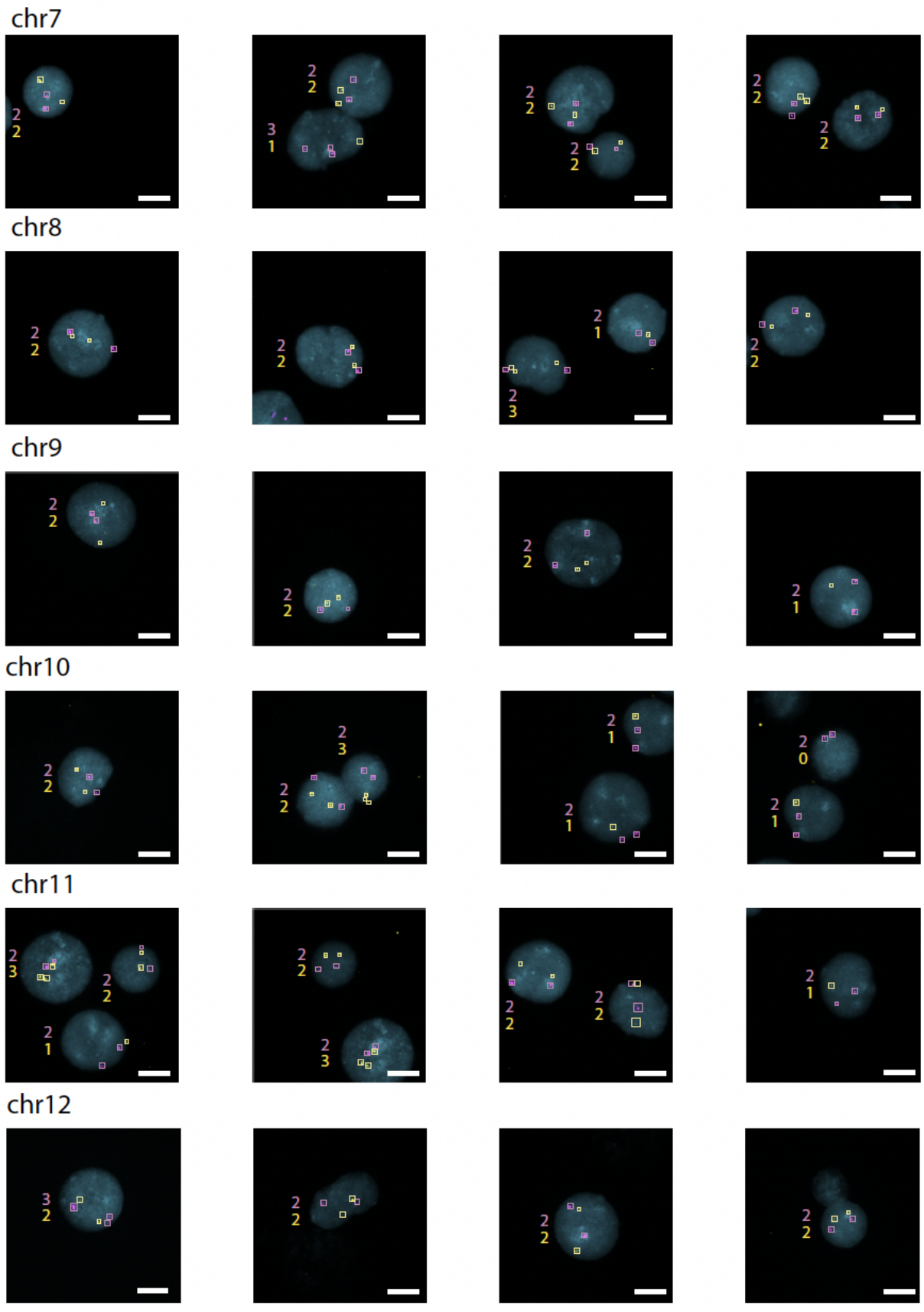
Supplementary Fig. 4 | Full-field metaphase spreads for chromosomes 13–18. Full-field images of the metaphase spreads from which the crops depicted in Figure 4b originated showing the staining pattern of the indicated Tigerfish (magenta) and PaintSHOP (yellow) probe sets. Images are maximum intensity projections in Z. Scale bars, 20 µm.


chr19
chr20
chr21
chr22
chrX
chrY
Tigerfish
PaintSHOP

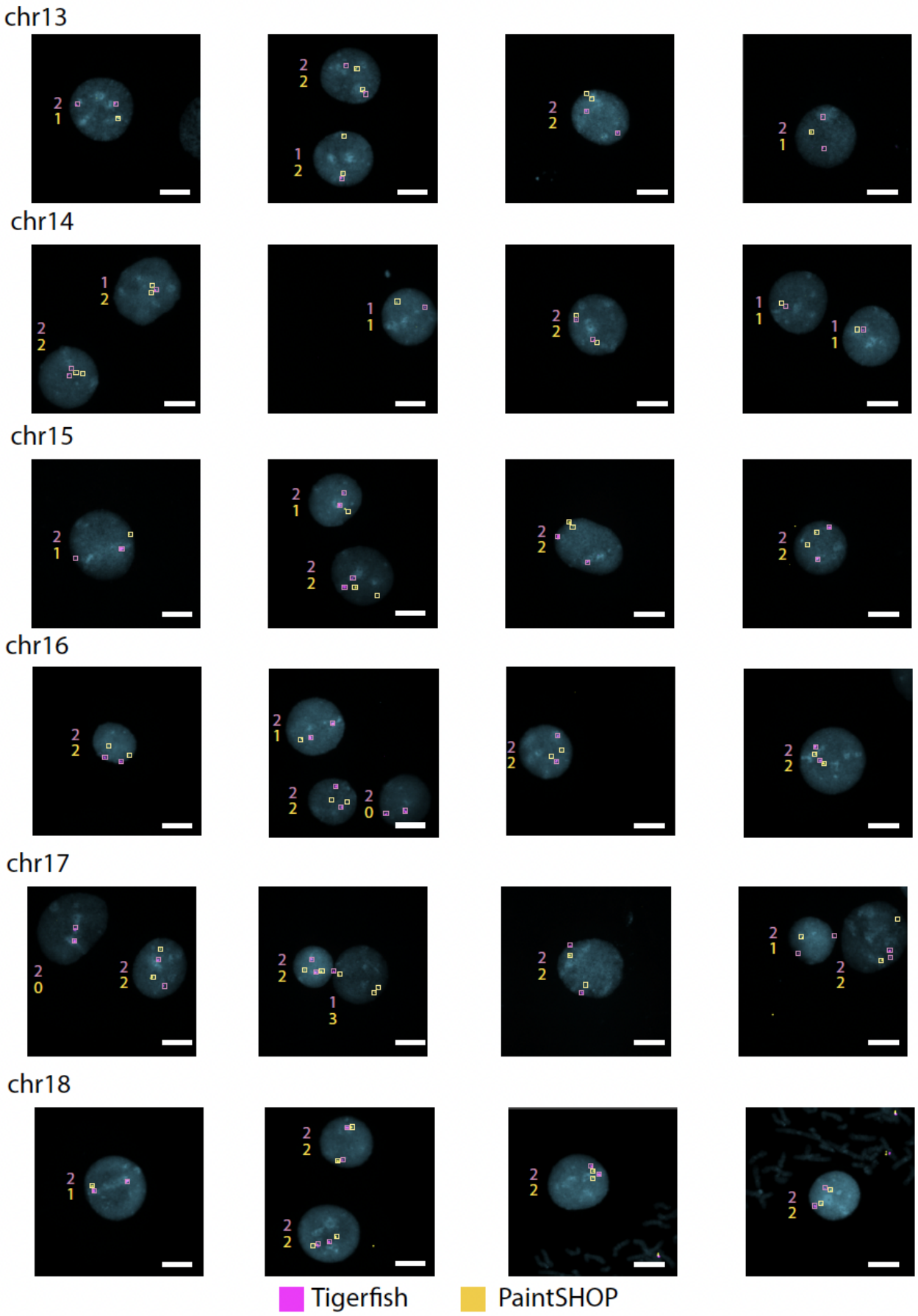
Supplementary Fig. 5 | Full-field metaphase spreads for chromosomes 19–Y. Full-field images of the metaphase spreads from which the crops depicted in Figure 4b originated showing the staining pattern of the indicated Tigerfish (magenta) and PaintSHOP (yellow) probe sets. Images are maximum intensity projections in Z. Scale bars, 20 µm.


chr1
chr2
chr3
chr4
chr5
chr6
Tigerfish
PaintSHOP

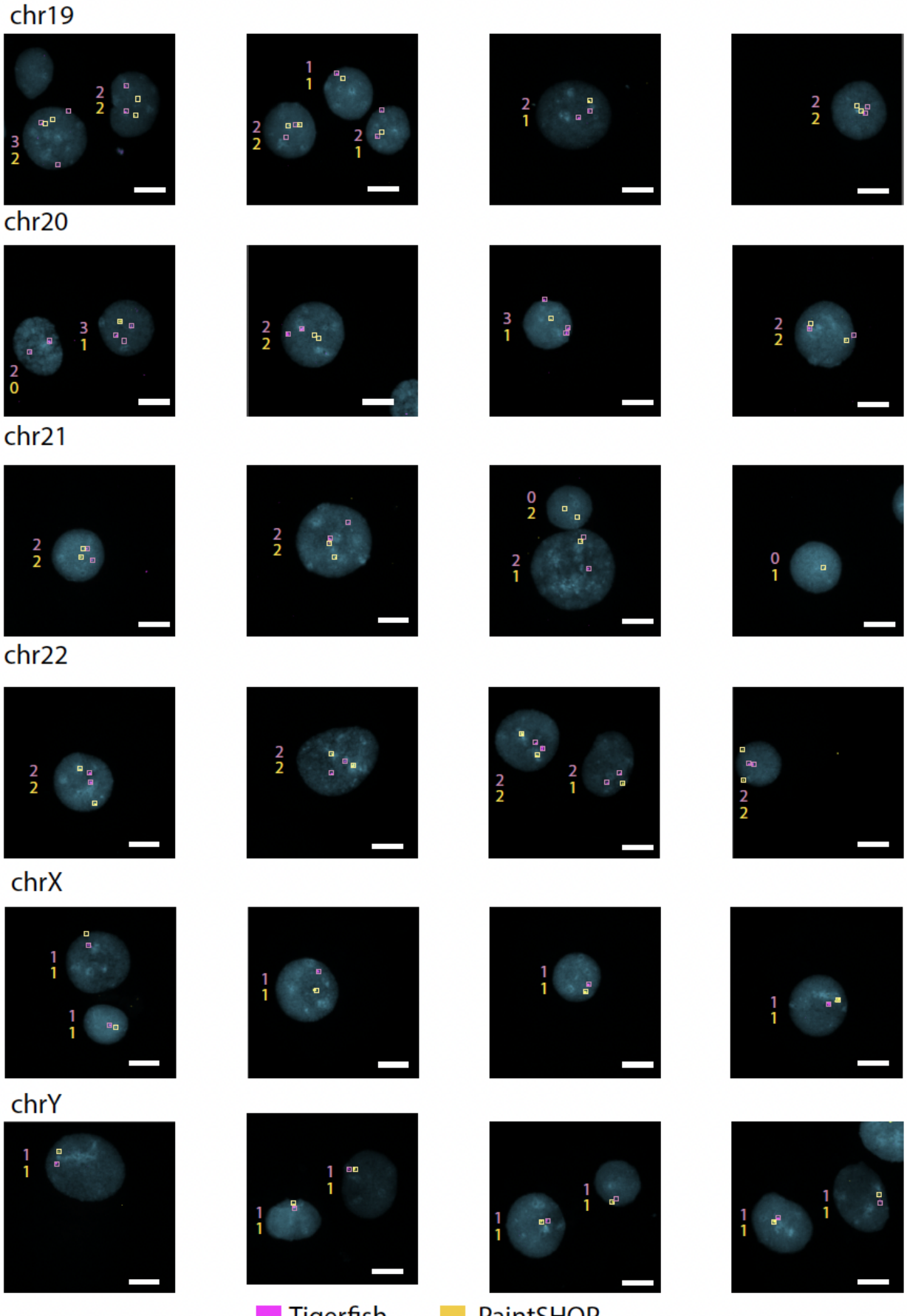
Supplementary Fig. 6 | Representative enumeration images for chr1–6. Four representative images of interphase nuclei and corresponding puncta counts for the specified Tigerfish (magenta) and PaintSHOP (yellow) probe sets. Images are maximum intensity projections in Z. Scale bars, 10 μm.


chr7
chr8
chr9
chr10
chr11
chr12
Tigerfish
PaintSHOP

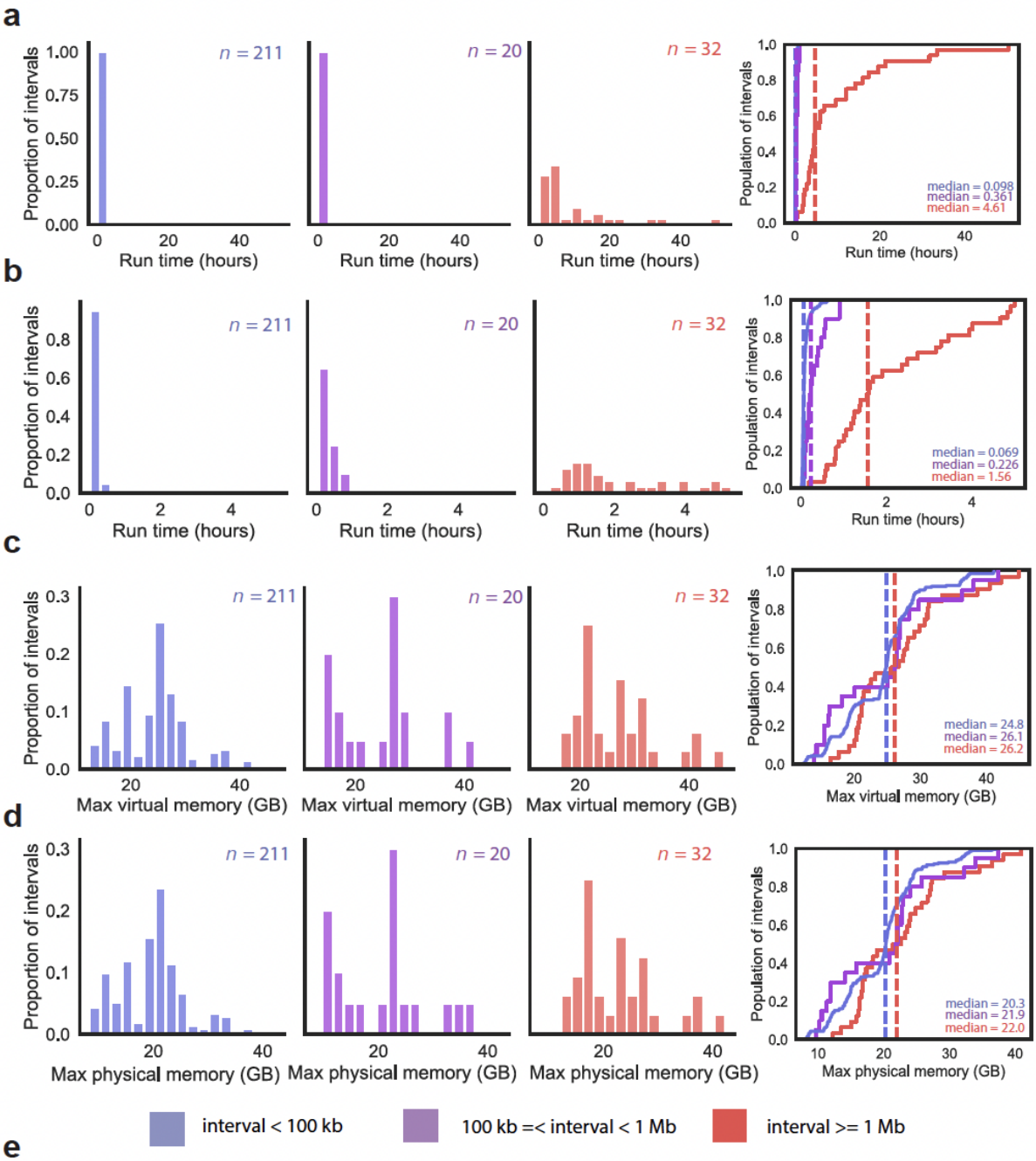
Supplementary Fig. 8 | Representative enumeration images for chr7–12. Four representative images of interphase nuclei and corresponding puncta counts for the specified Tigerfish (magenta) and PaintSHOP (yellow) probe sets. Images are maximum intensity projections in Z. Scale bars, 10 μm.


chr13
chr14
chr15
chr16
chr17
chr18
Tigerfish
PaintSHOP

Supplementary Fig. 8 | Representative enumeration images for chr13–18. Four representative images of interphase nuclei and corresponding puncta counts for the specified Tigerfish (magenta) and PaintSHOP (yellow) probe sets. Images are maximum intensity projections in Z. Scale bars, 10 μm.


chr19
chr20
chr21
chr22
chrX
chrY
Tigerfish
PaintSHOP

Supplementary Fig. 9 | Representative enumeration images for chr19–Y. Four representative images of interphase nuclei and corresponding puncta counts for the specified Tigerfish (magenta) and PaintSHOP (yellow) probe sets. Images are maximum intensity projections in Z. Scale bars, 10 μm.



Interval length	Min runtime	Max runtime	Min CPU time	Max CPU time	Min virtual memory	Max virtual memory	Min physical memory	Max physical memory
I < 100 kb	0.03	0.73	0.01	0.64	12.35	41.14	8.25	37.06
1 Mb > I >= 100 kb	0.10	0.99	0.06	0.91	14.00	41.71	9.70	37.47
I >= 1 Mb	0.27	50.18	0.19	4.99	16.39	44.87	12.24	40.84

Supplementary Fig. 10 | Tigerfish computational requirements by repeat region length. a, Histograms (left) and empirical cumulative distributions (right) of the wall-clock runtime recorded for running the 263 conservative and permissive intervals stratified by the target interval length.

b, Histograms (left) and empirical cumulative distributions (right) of the CPU runtime recorded for running the 263 conservative and permissive intervals stratified by the target interval length. c, Histograms (left) and empirical cumulative distributions (right) of the maximum recorded virtual memory allocation for running the 263 conservative and permissive intervals. d, Histograms (left) and empirical cumulative distributions (right) of the maximum recorded physical memory allocation for running the 263 conservative and permissive intervals. Vertical dashed lines in the cumulative distribution plots correspond to the median values. E, Summary statistics for the values presented in panels a–d.

Input(s)	Snakemake step	Output(s)	Option	Usage
FASTA	**generate_jf_count*** Generates a Jellyfish index file to count *k*-mers genome wide.	counts.jf	mer_val fasta_file	*Required*. *k*-mer size *Required*. Genomic reference FASTA
FASTA	**generate_bt2_indices*** Generates genome wide Bowtie2 (Bt2) index to align probes.	Bt2 Indices	fasta_file	*Required*. Genomic reference FASTA
counts.jf FASTA	**generate_jf_idx*** Generates *k*-mer count index files for each scaffold.	counts.txt index.txt scaffold FASTA	fasta_file mer_val sample	*Required*. Genomic reference FASTA *Required*. *k*-mer size *Required*. Scaffold name(s)
BED	**split_bed**** Takes BED file and splits coordinates into distinct files.	BED BED ...	bed_file sample	*Optional*. BED file with repeat target coordinates. *Required*. Scaffold name(s)
counts.txt index.txt	**repeat_ID**** Identifies repeat regions over queried chromosomes.	BED	sample file start window threshold composition mer_val	*Required*. Scaffold name(s) *Required*. Base position where repeat search begins. *Required*. Size of *k*-mer search window (W). *Required*. Min *k*-mer count value (T). *Required*. Proportion of elevated *k*-mers within W (C). *Required*. *k*-mer size
BED (repeat_ID) or BED (split_bed)	**design_probes**** Designs probes from identified repeat coords or user provided BED file.	probes.tsv region FASTA	fasta_file sample min_len max_len min_temp max_temp	*Required*. Genomic reference FASTA *Required*. Scaffold name(s) *Required*. Min size of candidate probe (bp) *Required*. Max size of candidate probe (bp) *Required*. Min melting temp of probe (C) *Required*. Max melting temp of probe (C)
probes.tsv counts.txt region FASTA	**kmer_filter** Computes each probe's on-target and off-target *k*-mer counts, then sorts probes by on-target count.	k_mer_sort.tsv	c1_val c2_val mer_val	*Required*. Constant to rank probes by copy_num. *Required*. Constant to rank probes by enrich_score *Required*. *k*-mer size
k_mer_sort.tsv	**probe_mer_filter** Filters probes based on probe *k*-mer similarity and target binding.	k_mer_filter.tsv	enrich_score copy_num mer_cutoff mer_val	*Required*. Min proportion a probe's *k*-mers must bind within a repeat region target. *Required*. Total sum of any probe's *k*-mers within a target repeat region. *Required*. Any probes within a target repeat exceeding this proportion of shared *k*-mers will be filtered *Required*. *k*-mer size
chrom.sizes	**generate_genome_bins:** Takes reference genome and creates bins.	alignment_bin.BED threshold_bin.BED	genome_windows chrom_sizes_file thresh_window	*Required*. Size of genome bins for alignment. *Required*. chrom.sizes file of queried genome. *Required*. Size of threshold bins to flag to determine imaging target coordinates.

** Denotes optional step if proper file paths/options are specified in config.yml*
*** Denotes step specific to either probe_design or repeat_ID run modes*

Input(s)	Snakemake step	Output(s)	Option	Usage
k_mer_filter.tsv	**make_chrom_dir** Seperates probe files by repeat region.	repeat.txt		
repeat.txt Bt2 indices alignment_bin.BED	**alignment_filter** Filters candidate probes using Bt2 and NUPACK to predict in silico probe binding.	filtered_probes.tsv	target_sum bt2_alignments seed_length model_temp max_pdups min_on_target max_probe_return off_bin_thresh align_thresh ref_flag	*Required*. Total on-target sum of all probes desired. *Required*. Bt2will return alignments up to this val. *Required*. Controls Bt2 seed length. *Required*. Temperature of NUPACK predict model. *Required*. Probes in final dataset with predicted binding greater than this value will be filtered. *Required*. Min on target score for any given probe. *Required*. Max probes to be returned/repeat. *Required*. Off-target threshold over any non-target genomic bin. Required. Min binding sites required to flag a binned region as significant toward probe binding. Required. Provides intermediate output files.
filtered_probes.tsv	**merge_alignment_filter** Aggregates all candidate probes by scaffold.	chrom_probes.tsv		
chrom_probes.tsv	**split_rm_alignments** Creates a new directory containing all candidate probe files by repeat.	scaffold/region.tsv		
scaffold/region.tsv	**align_probes** Takes candidate probes corresponding to compute target specificity. SAM derived sequences are used to compute *in silico* binding.	region_align.txt	bt2_alignments seed_length model_temp mer_val	*Required*. Bt2 will return alignments up to this val. *Required*. Controls Bt2 seed length. *Required*. Temperature of NUPACK predict model. binding greater than this value will be filtered. *Required*. Any probes within a target repeat exceeding this proportion of shared *k*-mers will be filtered
region_align.txt	**derived_beds** Creates BED file from SAM derived sequence alignment for each candidate probe.	derived_align.BED		
region_align.txt	**get_region_bed** Creates BED file for the target repeat region.	repeat.BED		
derived_align.BED repeat.BED	**bedtools_intersect** Performs two bedtools intersects: 1. Derived alignments against threshold genome bins. 2. Repeat region against threshold genome bins.	derived_BEDtools.txt repeat_BEDtools.txt		

** Denotes optional step if proper file paths/options are specified in config.yml*
*** Denotes step specific to either probe_design or repeat_ID run modes*

Input	Snakemake step	Output	Option	Usage
derived_BEDtools.txt repeat_BEDtools.txt region_align.txt chrom.sizes	**get_alignments** Computes genome wide binding summaries for all probes within a target repeat region.	binding_map.png thresh_summary.txt binding_quant.txt	align_thresh	*Required*. Binding sites required to flag a bin as significant toward probe signal.
filtered_probes.tsv binding_map.png thresh_summary.txt binding_quant.txt	**map_region_coords** Adds imaging target coordinates to the candidate probes file. Creates probe binding summary.	final_probes.tsv		
final_probes.tsv probe_summary.txt	**merge_mapping** Aggregates all repeat regions into a single file by scaffold.	probes_merged.tsv		
probes_merged.tsv	**summary** Summarizes probe binding and count by repeat region target.	probe_summary.txt		

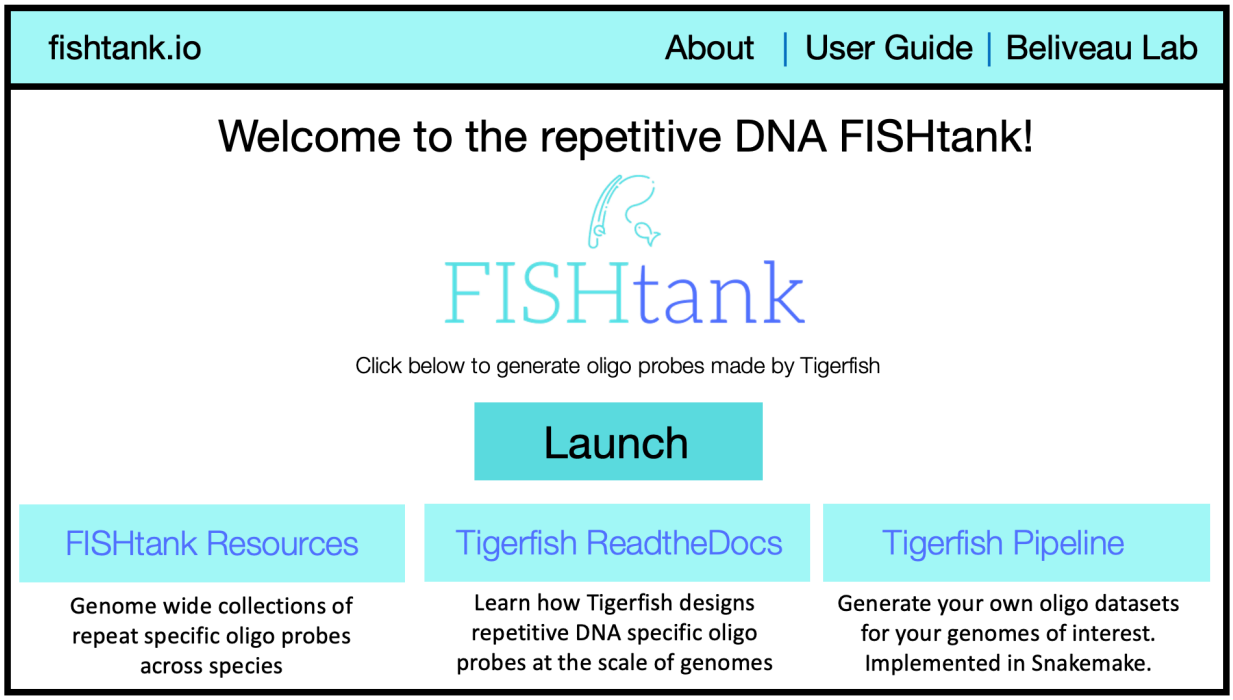
** Denotes optional step if proper file paths/options are specified in config.yml*
*** Denotes step specific to either probe_design or repeat_ID run modes*

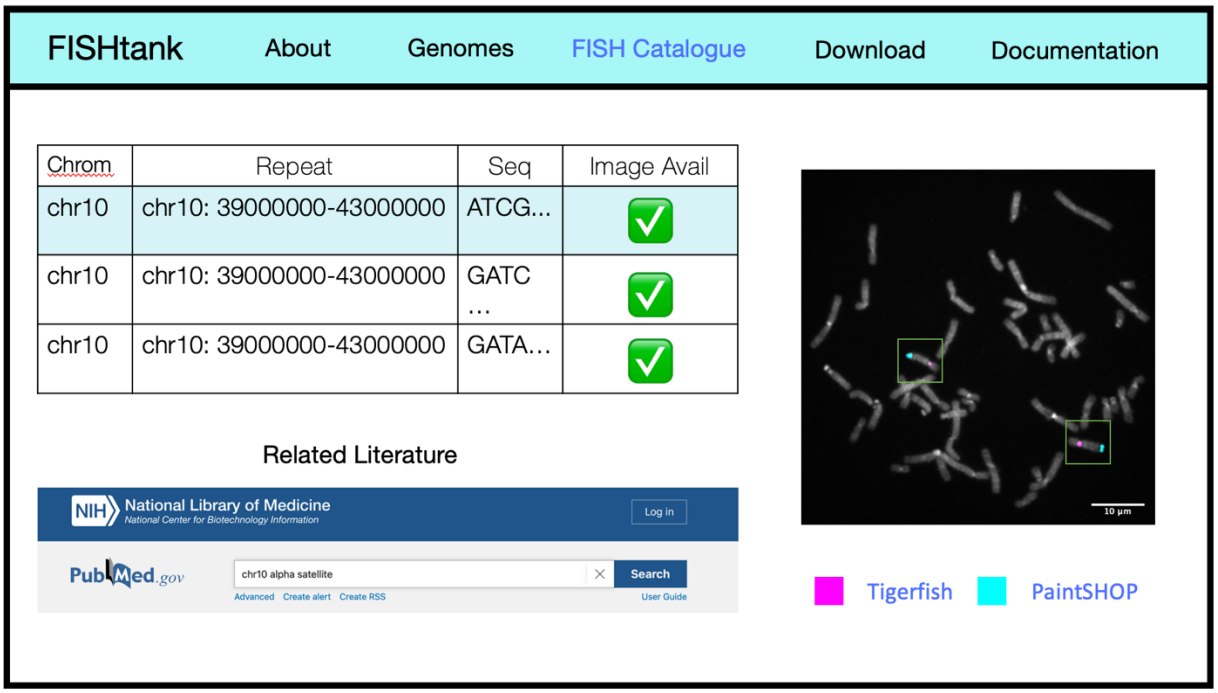
Input	Snakemake step	Output	Option	Usage
filtered_probes.tsv	**gather_repeat_regions** Takes filtered candidate probes and splits them by scaffold. Input file should have one repeat per scaffold.	split_probes.txt	sample	*Required*. Scaffold name(s)
split_probes.txt	**align_cand_probes** Takes candidate probes corresponding to compute target specificity. SAM derived sequences are used to compute *in silico* binding.	region_align.txt	bt2_alignments seed_length model_temp mer_val	*Required*. Bt2 will return alignments up to this val. *Required*. Controls Bt2 seed length. *Required*. Temperature of NUPACK predict model. binding greater than this value will be filtered. *Required*. Any probes within a target repeat exceeding this proportion of shared *k*-mers will be filtered
region_align.txt	**derived_cand_beds** Creates BED file from SAM derived sequence alignment for each candidate probe.	derived_align.BED		
region_align.txt	**get_cand_region_bed** Creates BED file for the target repeat region.	repeat.BED		
derived_align.BED repeat.BED	**bedtools_cand_intersect** Performs two bedtools intersects: 1. Derived alignments against threshold genome bins. 2. Repeat region against threshold genome bins.	derived_BEDtools.txt repeat_BEDtools.txt		
derived_BEDtools.txt repeat_BEDtools.txt region_align.txt chrom.sizes	**get_cand_alignments** Computes genome wide binding summaries for all probes within a target repeat region.	binding_map.png thresh_summary.txt binding_quant.txt	align_thresh	*Required*. Binding sites required to flag a bin as significant toward probe signal.
chrom.sizes repeat.BED	**generate_cand_chromomap*** Creates a karyoplot using chromoMap of the repeat target.	chromomap.HTML		
filtered_probes.tsv binding_map.png thresh_summary.txt binding_quant.txt	**map_cand_region_coords** Adds imaging target coordinates to the candidate probes file. Creates probe binding summary.	final_probes.tsv		
final_probes.tsv probe_summary.txt	**merge_cand_mapping** Aggregates all repeat regions into a single file by scaffold.	probes_merged.tsv		
probes_merged.tsv	**summary** Summarizes probe binding and count by repeat region target.	probe_summary.txt		

** Denotes optional step if proper file paths/options are specified in config.yml*
*** Denotes step specific to either probe_design or repeat_ID run modes*

Supplementary Note 1. Tigerfish inputs and outputs.

Supplementary Fig 11. FISHTank Figma Mockups





Appendix A Note: Supplementary Data Items 1- 5 and Supplementary Software Items are appended as complementary files.

Appendix B - GSAIMS Academic Archive

Intro

GSAIMS (founded in 2019) is a trainee-led and organized student organization that supports, uplifts, and mentors trainees in Genome Sciences and specifically those who have backgrounds and identities that are underrepresented in STEM at the University of Washington. This group fundamentally created space and resources specifically for trainees with marginalized identities in STEM and offered much needed community for trainees, faculty, and any participants interested in general programming and events.

Monthly meetings and events are open to all department members unless otherwise described (ex. Trainees only, etc.). More information on our constitution, leadership positions, and archive of events can be found on our Dropbox.

This document is intended to describe all the events that GSAIMS organized while active in the Department of Genome Sciences. Additionally, this document offers frameworks for executing all outreach activities that GSAIMS created, accounts of important historical events within the department during this time frame, and outlines hours spent, attendee participation in events, important individuals involved in GSAIMS, and general conclusions on this type of advocacy space in a graduate academic program.

Events

Below summarizes events between the active years of GSAIMS.

Fall 2019

GSAIMS First Year Welcome Orientation and Bake Off

Date
09.17.2019

Purpose
To welcome the first-year grad students through an orientation to introduce them to guidelines, campus resources, and more. There was a social hour where the first years were able to meet their assigned student mentors and faculty mentors. Following this event there was a department wide bakeoff competition with desserts.

Outcomes
Overall, this event was successful from an implementation standpoint. A lot of people showed up for the mentorship socials and the bakeoff. The bake off probably had the greatest turnout of any event (in part because this was also right before the COVID-19 pandemic). It felt like there were at least 50+ people there. Total hours – 8 hours.

First year slice of life Q&A

Date
10.02.2019

Purpose
To provide first year students with a space where they can ask grad students who have been in the program any questions about their first year. The first-year mentors were also invited to chat with their students about their rotations and their experiences at the GS retreat that year.

Outcomes
This event was attended by nearly all first-year students and their mentors. Overall conversation was good, but I think one area of feedback that we noticed is that it would have been helpful to have more ice breakers. It also was a bit difficult to help some of the first years talk to older grads because both groups were kind of already in their own social spaces. Total hours – 2 hours.

Hoppy Hour

Date
11.07.2019

Purpose
To host an informal happy hour hangout at Burke Gilman Brewing.

Outcomes
Event was highly successful, we had 18+ attendees with a wide range of grad students and faculty. Total hours – 3 hours.

GSAIMS goes to SACNAS

Date
10.28.2019 - 11.01.2019

Purpose

To share information about GS at the SACNAS conference which took place in Honolulu.

Outcomes
There were a lot of good conversations with recruits that year and there was a lot of good energy at the conference! Total hours tabling – 12 hours.

GSAIMS goes to ABRCMS

Date
11.13.2019 - 11.16.2019

Purpose
To share information about GS at the ABRCMS conference which took place in Anaheim, CA.

Outcomes
We only sent one representative to this conference this year, but it seemed like from his perspective there were also a lot of good conversations here. Total hours tabling – 12 hours.

GSAIMS is invited to participate in selecting seminar speakers

Date
11.19.2019

Purpose
We were reached out to by faculty to see if we would be interested in inviting seminar speakers and helping host their visit to the program. This was a good opportunity because we were able to invite several national speakers to give a seminar.

Outcomes
This would lead to several invited speakers coming through the department which is documented further in this archive. Total hours - 1 hour.

Festive Feedback

Date
12.19.2019

Purpose
To give first year students a way to have some snacks, drinks, and a space to work on their rotation talks with low stakes feedback and encouragement.

Outcomes
It seemed like there was a good spread of first years who were able to bounce ideas from other grad students about their presentation style so overall this event was a success. Total hours – 2 hours.

Winter 2020

Chat with Lisa Peterson

Date
01.10.2020

Purpose
To chat with Lisa Peterson (she/her) about a program that she is running (as of 2020) that was funded by NHGRI for underrepresented students in the sciences. This program was mainly for undergraduate students, and she wanted to talk to Robin about running GSAIMS and what they noticed with GSAIMS and mentorship program outcomes.

Outcomes
This meeting was productive because we were able to make connections with another organization that was also learning the ropes with navigating institutional policies. Nothing much came of this meeting in following years, but it was a good resource and connection here at UW. Please see contact directory for all individuals GSAIMS maintained contact with over this active period. Total hours – 1 hour.

GSAIMS applies to the SEED Grant

Date
11.2019 - 02.2020

Purpose
To gain funding from UW to apply funding to get workshops for external speakers.

Outcomes
This grant was funded in spring 2020 and the funds were used to support our events for the 2020 - 2021 year. Funds acquired – $5k. Total hours - 8 hours.

Imposter Syndrome and Mindfulness Event

Date
02.05.2020

Purpose
This workshop was led by Fievel Jack Finley (he/him) who was working as an LMT at UW, and he was willing to lead a workshop on navigating imposter syndrome as well as discussions on how it impacts graduate students. This workshop was intended to be supportive for students who were having challenges with navigating the long dark that is Seattle in the winter.

Outcomes
Overall, there weren't many attendees for this workshop (approximately 6 attendees). However, from the feedback that was gathered, it would have been helpful to have made the space more accommodating for discussion because this talk was led as more of a presentation. Total hours – 2 hours.

GSAIMS participates in grad student recruitment

Date
Last week of February, Second week of March 2020

Purpose
To speak with incoming graduate students about GSAIMS and tell them what our past events have been like.

Outcomes
GSAIMS did get several questions this year and many students said they enjoyed the content of the presentation. Total hours – 2 hours.

March 20, 2020

Campus began moving to complete remote learning due to the COVID-19 pandemic. Events were pretty much paused until later in the spring quarter.

Spring 2020

Coffee Chat with Dr. Tim Thornton (remote)

Date
05.20.2020

Purpose
To have a career coffee chat with Dr. Thornton (he/him) about how he landed his job as a professor in UW Biostat and what were the highs and lows of his current work.

Outcomes
This was a well-attended zoom event and was the first of many remote events. Attendance was roughly 8 people. Total hours – 2 hours.

Pride Happy Hour (remote)

Date
06/10/2020

Purpose
To connect with GS community members to celebrate pride remotely. During this event we had a variety of breakout rooms for different activities: Watching a show, crafting, chatting, etc.

Outcomes
This was when folks were early into learning Zoom, so it was a bit of a struggle figuring out the breakout rooms and testing things like Netflix party. Otherwise, there was a great turnout of +15 people. Total hours – 2 hours.

Ghibli movie night (remote)

Date
06/26/2020

Purpose

To watch a movie remotely together on zoom. We aired My Neighbor Totoro!

Outcomes
It helps when you can watch a higher quality movie together. I think we ended up using a 4k version of the video that someone had downloaded and then proceeded to screen share. Total hours – 2 hours.

Summer 2020

Real Life Summer Book Club (remote)

Date
07.08.2020 - 08.31.2020

Purpose
To read the Book “Real Life” which is Brandon Taylor’s debut novel where he writes about a Black and queer character who navigates the challenges of systemic racism in a biology graduate program.

Outcomes
This book club was well received with 20 participants. Over the course of the summer, this number did drop off to about 10 because I think a lot of people's bandwidth was taken up. Please see in the framework section more info on book club organizing led by GSAIMS. Total hours – 10 hours.

Fall 2020

Seed Grant Series: Dr. LaShawnDa Pittman (remote)

Date
10.12.2020

Purpose
This event started off the Seed grant series that we were awarded funding for. Here, we featured the research of Dr. Pittman (she/her) who's an assistant professor at UW in American Ethnic Studies. Her research is specifically on grandparent caregiving in the African American community and in Black women's reproductive health and justice.

Outcomes
This was a well-attended seminar with roughly 12 attendees. Folks asked questions and were excited to learn more about the series. Total hours – 2 hours.

Spirited Away Movie Night (remote)

Date
10.23.2020

Purpose
To host a remote Halloween social by watching Spirited Away.

Outcomes
This was also well attended, and it was helpful to have the 4k movie download. At this point we had a bit more practice with zoom so overall this was a good event. Total hours – 2 hours.

Seed Grant Series: Dr. Kate Mittendorf (remote)

Date
11.02.2020

Purpose

Our second seed grant series seminar featuring Dr. Mittendorf (they/them). They are a research scientist at Kaiser in Vancouver, WA as well as a scientific illustrator and advocate for folks with disabilities.

Outcomes
This was a well-attended event with roughly 10 attendees. People enjoyed hearing about their research in cancer genomics. Total hours – 2 hours.

WiGS + GSAIMS Nautilus Biotech Workshop (remote)

Date
11.18.2020

Purpose
A workshop led by folks at Nautilus biotech to learn more about biotech careers, interview and CV tips, etc. This was co-hosted by WiGS.

Outcomes
This was very well attended, +25 attendees. People enjoyed the interactivity, ability to ask questions, and learn more about the application timeline process. Total hours – 2 hours.

Wellness Workshop with Dr. Andrea Salazar-Nuñez (remote)

Date
12.07.2020

Purpose
To give trainees the ability to talk about mindfulness and grad student mental health with a trained LMT who works primarily with students with marginalized identities.

Outcomes
This was a smaller event attended by 5 people. Attendees liked how aware the speaker was about microaggressions and she offered cultural responses toward finding healing. Total hours – 2 hours.

Winter 2021

Seed Grant Series: Dr. Yi (Jenny) Xiao (remote)

Date
01.25.2021

Purpose
To speak with Dr. Xiao (she/her) about her research in social identities and how this impacts perception as it relates to race and lived experience. She is an assistant professor at UW Tacoma.

Outcomes
This talk had 15+ attendees and people were engaged with asking her questions, so it seemed like a good success! Total hours - 2 hours.

Seed Grant Series: Dr. La Tasha Levy (remote)

Date
02.11.2021

Purpose
Dr. La Tasha Levy (she/her) is an Assistant Professor in the Department of American Ethnic Studies. She discussed her work and research on the historical roots of #BlackLivesMatter in Media and Popular Culture.

Outcomes
This talk was attended by 8 attendees and had good discussion. Something that resonated with a lot of attendees was how she discussed how challenging it was to write her thesis well and woes related to publishing. She was a great person to connect with! Total hours - 2 hours.

Spring 2021

WiGS + GSAIMS Happy Hour (remote)

Date
03.05.2021

Purpose
To host a remote social happy hour with WiGS. We used gathertown to play games and chat.

Outcomes
We had 15+ attendees. Gathertown is an interesting platform for users, and I think it went through a lot of bug patches over the pandemic. As of 2021 it seemed like a pretty good and fun option for remote socials. Total hours - 3 hours.

Seminar with Dr. Tobias Mann (remote)

Date
04.12.2021

Purpose
To speak with Dr. Mann (he/him) about his experiences working with Adaptive and advice on working in biotech.

Outcomes
This was a well-attended remote seminar with 15+ attendees. A lot of folks had questions on his experiences in Genome Sciences and transferable skills applicable to industry. Total hours – 2 hours.

E-portfolio and storytelling workshop (remote)

Date
05.11.2021

Purpose
This was hosted by Robin Aguilar (they/them) and Atom Lesiak (they/xe) to share tips and pointers for designing personal websites that are useful for hosting portfolios, etc.

Outcomes
This was attended by 6 people and people really liked the intent of this workshop. Participants asked a lot of questions, design pointers, and some people were following along with getting started on making their websites during the workshop. Total hours - 3 hours.

Cops off Campus Coalition (in-person)

Date
05.23.2021

Purpose
To support Decriminalize UW at the George Washington Statue for coffee and art.

Outcomes
This has 3 attendees from GS, but the on-campus contingent had a lot of attendees. Total hours – 3 hours.

Socially Distant Social (in-person)

Date
05.28.2021

Purpose
This was to have an in-person outdoor social with snacks at Gasworks.

Outcomes
There were roughly 8 people in attendance! It was one of the first socials we had since the start of the pandemic, so it felt good seeing people in person again. Total hours – 2 hours.

Seed Grant Seminar: Dr. Kaela Singleton (remote)

Date
06.09.2021

Purpose
This was to have a remote seminar with Dr. Singleton (she/her) who studies neuroscience and is an assistant professor at Agnes Scott.

Outcomes
Her talk was attended by 6 attendees and people were able to ask her a lot of questions about her work. She was also great at talking about her research and people seemed engaged. Total hours – 1 hour.

Summer 2021

Summer book club: The Disordered Cosmos (remote)

Date
07/14/2021 - 09/01/2021

Purpose
To read The Disordered Cosmos by Dr. Chanda Prescod-Weinstein (she/they) which discusses her research in physics as well as her perspectives and research on Black feminism in science.

Outcomes
This book club didn't have nearly the amount of attendance as what we received in the summer of 2020. However, there were 10 people who registered with books, and we regularly received 5-6 people to discuss book chapters each week. This book club was also remote. Total hours – 10 hours.

Fall 2021

First and Second Year Mentorship Mixer co-hosted by the Grad Council (in-person)

Date
09.30.2021

Purpose
To welcome first years into the department.

Outcomes
This was well attended by nearly all the first years and this event was also in person. We used Vista Cafe for this space. Total hours – 2 hours.

Halloween Hangout with WiGS (in-person)

Date
10.27.2021

Purpose
To have an informal Halloween social gathering co-hosted by WiGS.

Outcomes
There was a decent turnout and I think at this point many student orgs were noticing general burnout caused by stressors pandemic related. Regardless, pumpkin painting, snacks, and costume party was a success, there were 20+ attendees. Total hours – 2 hours.

Grad School Panel co-hosted by WiGS (remote)

Date
11.02.2021

Purpose
To help undergrads learn more about graduate school programs and the application process.

Outcomes
There were very few attendees who registered for the panel (1-2 students). This is a challenging area to navigate when creating events like this for undergrads - it seems like GS has a hard reach for these types of events. It could have also been since pandemic stressors have remained high, and a lot of people experienced very real zoom fatigue during these months. Total hours – 1 hour.

Applying to Grad School Workshops co-hosted with WiGS (remote)

Date
11.16.2021

Purpose
To help undergraduate students work on graduate school applications remotely.

Outcomes
We also only had 1-3 students come with essays that they wanted feedback on. See outcomes of events above as something to keep in mind when organizing events like this with undergraduates. Total hours – 2 hours.

GS Seminar: Dr. Robert Fernandez (in-person)

Date
12.01.2021

Purpose
To learn from Dr. Fernandez (he/him) and his research in C. elegans, neuroscience, and his mentoring through Cientifico Latino.

Outcomes
This visit was well planned and there were 3-4 attendees at his meeting with students and a pretty good department turnout. Total hours – 4 hours.

Winter 2022

First and Second Year Mentorship Mixer (in-person)

Date
02.23.2022

Purpose
To hold space for first- and second-year students to talk about adjusting to grad school.

Outcomes
There was good turnout for this event with 12+ attendees. We used the Vista patio as the venue for this event. Total hours – 2 hours.

Spring 2022

AAAS Mass Media Fellowship workshop with Dr. Evelyn Valdez-Ward (remote)

Date
04.08.2022

Purpose
To learn from a science communicator Dr. Valdez-Ward (she/her) about what her experiences were with applying and participating in the AAAS fellowship program.

Outcomes
This was one of our least attended sessions ever. This session was recorded and shared to our group drive though. Total hours – 2 hours.

HHMI Invited Speaker with Dr. Irene Y. Chen (in-person)

Date
05.05.2022

Purpose
This was one of the first speaker sessions invited by the HHMI Gilliam Fellowship where Dr. Chen (she/her) was able to talk about her research as a postdoc.

Outcomes
This event was reportedly well attended and supported by the department. Few exec members from GSAIMS were able to attend. Total hours – 4 hours.

GS Seminar with Dr. Andrea Gomez (in-person)

Date
05.25.2022

Purpose
To learn from assistant professor Dr. Gomez (she/her) about what her research and new lab is up to at UC Berkeley.

Outcomes
Similarly, her talk was well attended, and we also had a pretty good discussion at dinner with roughly 8 attendees. Total hours – 3 hours.

Summer 2022

HHMI Seminar with Dr. Nick Altemose (in-person)

Date
06.23.2022

Purpose
To hear about Dr. Altemose's (he/him) research and his perspectives on being a new assistant professor at Stanford.

Outcomes
His lunch talk and seminar talk were both very well attended! Total hours – 2 hours.

Pride Social (in-person)

Date
06.28.2022

Purpose
To celebrate Pride with food and company.

Outcomes
This was a highly successful event with 30+ people attending and having conversations in Vista Cafe. The thing with a lot of these events to keep in mind though is that a lot of people are likely to come if there's food and snacks but not a lot of people are willing to help organize these types of events. Total hours – 2 hours.

Summer book club - Lessons from Plants (in-person)

Date
07.13.2022 - 08.31.2022

Purpose
To read the book Lessons from Plants by Dr. Beronda Montgomery (she/her) to learn more about her research in plant genomics as well as her insights that she's gathered from plants towards being a better mentor.

Outcomes
This book club was attended by a total of 4 people despite having 8 people register. We did in-person meetings for this book club, had weekly meetings, and tried to meet outside to help with social distancing. One thing to think about is to make sure to get events for book clubs on the calendar early. I think this is the first year where we had conflicts with another student group that was also leading a summer book club which may have caused some people to feel overwhelmed with summer events. Total hours – 10 hours.

Fall 2022

First and Second Year Mentorship Mixer (in-person)

Date
09.30.2022

Purpose
To have a space where the first- and second-year grad students can talk about their experiences and offer guidance.

Outcomes
We ordered pizza, snacks, and drinks for this event and it seemed like there was good attendance in person. We didn't have every first year or second year attendance but there was a decent turnout. Total hours – 2 hours.

Grad student town hall #1 (in-person)

Date
09.28.2022

Purpose
To discuss important grad student concerns that we wanted to address in a town hall meeting.

Outcomes
This event was facilitated by the grad student reps for this academic year who are Syd Sattler (she/her), David Lee (he/him), and Atom Lesiak (they/them). We helped order pizza and participated in the event to discuss pressing topics that grad students were interested in improving in the department. Intergenerational flow of knowledge was a big topic of importance here. Total hours – 2 hours.

SACNAS recruiting (in-person)

Date
10.26.2022-10.30.2022

Purpose
To recruit and talk to potential students interested in UW Genome Sciences to learn how to apply to grad school.

Outcomes
We received 20+ trainees who were interested in learning more about the program. Total hours – 10 hours.

Halloween (in-person)

Date
11.01.2022

Purpose
To celebrate Halloween and Day of the Dead with lots of snacks and company.

Outcomes
This seemed to be well attended by 12+ people. Total hours – 2 hours.

Winter 2023

Grad student town hall #2 (in-person)

Date
01.11.2023

Purpose
To discuss grad student concerns and info for getting involved in the graduate student union.

Outcomes
This event was facilitated by Luana Paleologu (she/her) and Syd Sattler (she/her) to discuss successes of the student union, how to get involved, as well as discuss other graduate student concerns. Total hours – 2 hours.

STEMPals (in-person)

Date
09.28.2022 - current

Purpose
To connect high school students with GS grads to talk about their experiences as researchers.

Outcomes
Please read the frameworks section.

Art Zine (in-person)

Date
01.18.2022 - current

Purpose

To build a creative community by making a zine that features the art and creative works of anyone who wants to get involved in Foege.

Outcomes
Please read the frameworks section.

Frameworks

This section describes initiatives and common events that we hosted and how we navigated logistics. This can be important for re-creating these events and navigating them in the future.

Monthly Meetings

GSAIMS runs monthly meetings that are open to anyone in the department. Primarily exec members join in, but we have hosted guests and trainees who also want to learn how to get more involved.

This is the general structure of our meetings:

- Round table check-ins
- Upcoming events
- Brainstorming for future events in the quarter(s)

Typically, meetings last 30 mins - 1 hour. Recorded meeting minutes were posted to our slack channel after each meeting for folks who couldn't make it to monthly meetings.
Over the course of the pandemic, we kept GSAIMS meetings to zoom since they were more accessible for folks whose labs were also at Fred Hutch. Zoom links were sent to Brian Giebel at the start of the academic year so they could exist on the Genome Sciences calendar as a recurring event.

SACNAS/ABRCMS Recruitment

Beginning in 2019, GS began sending reps from GSAIMS to SACNAS and ABRCMS conferences to help with recruitment. Conferences we attended are highlighted above in events but are summarized below:

- SACNAS 2019
- ABRCMS 2019
- SACNAS 2020
- SACNAS 2022
- ABRCMS 2022 (attended by Brian Giebel)

If future trainees from GSAIMS wish to be involved in attending these conferences for recruitment these are some notes that were gathered about improving the experience.

- Instead of a large standing banner which is provided by the department admin, it would be helpful to have a tablecloth that can be printed with the GS logo on it. It makes it much nicer for traveling.
- Swag. Like pens, stickers, candy, mints, notepads, sticky notes, etc. would help bring more students to the table.
- Bring hand sanitizer to the table!
- Reach out to students directly at their posters.

GS Recruitment Weekends

Recruitment weekends happen twice in the winter quarter. Typically, GSAIMS participates by giving a 5 minute presentation to recruits. This presentation can be found in the GSAIMS Dropbox and should be updated annually with current and upcoming events.

STEMPals

This was an outreach project intended to pair GS trainees with high school classrooms to send pen pal letters.

The way this was organized was to create a google form (template hosted on our Dropbox) that gathered emails from trainees in GS interested. Dr. Atom Lesiak (the GS outreach director) was able to connect with local Seattle schools who were interested in this program and was able to gather 7 classrooms who were interested in this program.

Once participants' info was gathered (there was a 1-month deadline to fill out the form), those individuals were contacted to get a better sense of their preferences and to share the time commitment involved. They filled out a final form to provide their preferences which then allowed Atom to contact classrooms with student pairs. When we did this there were 7 classrooms and 12 GS participants.
Letters will be sent once a quarter and there will be a small thank you celebration social where we write final letters and enjoy small plates.

Overall, because of the pandemic and SPS teachers also went on strike earlier in the year, this program would likely do better with a bit more planning in advance with contact to each of the teachers of interest. While some participants did get responses from their classrooms, others didn’t receive any contact from their classrooms. To work on this in the future, I think teachers need to be checked on a few months in advance and a few weeks before the program launches to make sure that they are still on board. This turned out to be a great concept outreach project, but it would be helpful to have more admin support to handle school contact logistics.

Art Zine

This was a community building project intended to create a printed GS zine to celebrate and share other creative pursuits that folks in GS were up to. It involved organizing a few social meetups over the winter quarter that this was implemented, setting up a date for editors to put the zine together, and hosting a social where printed zines were available.

To get this started a google form was set up to gather the info of people who were interested. This poll closed after one month. Then those folks were contacted with more information on how they could be involved with submitting their works and what dates the socials were happening.
A department wide post was shared with the social dates anyway as well as a zine submission form.
At each social there were snacks in a reserved conference room in Foege for folks to just hangout and work on stuff.

Once all the works were submitted after the deadline, the editing team put the zine together on Canva and submitted it to a local printing group. It was helpful beforehand to send out an invitation to the department wide social to get RSVPs and a rough headcount of how many people wanted a printed zine.
The zine turned out to be a great success and a lot of people wanted a PDF copy to read which was nice. It was also cheap to print ~30 good quality copies from Mixam ($100). I would highly recommend this event again as it was successful being able to bring people together. In terms of putting the zine together, Canva was used and just two editors were needed to put it all together over a couple days.

GSAIMS Career Symposium

This was an event that invited five speakers to talk about their careers outside of academia. There was mixed attendance with a total of about 30-40 attendees and 40+ attendees at the dinner reception. The logistics for organizing this event involved:

- Contacting each of the speakers and securing their honorarium, talk info, and their travel logistics were taken care of which was done via google form.
- Contacting catering for pastries and coffee the morning of the talks (Brian Giebel).
- Contacting the Indian restaurant for a catering order (Brian Giebel).
- Creating posters and flyers for the event.
- Advertising said posters and flyers (Brian Giebel).
- Contact GSIT to record the talk and work with A/V setup.
- Making sure that Vista cafe was reserved to serve food and alcohol.
- Getting an alcohol permit.

A benefit to hosting this event on a Friday was that this was coupled with the department social hour to have it be a happy hour also.

Hosting Seminar Speakers

To invite a seminar speaker, everyone was contacted roughly 2-3 months before the event date. Each speaker was also told upfront that they would be given an honorarium. Our honorariums ranged from $250-500 for their time. Posters were also distributed on slack 1-2 weeks in advance and information was given to Brian Giebel to coordinate zoom links or in-person details. If this was a department seminar, Brian Giebel would create the schedule and we would take the speaker out to dinner at the end of their seminar.

Zoom

- We'd check to make sure captions were an option and chat was enabled.
- We would send links out to speakers and Brian Giebel well in advance of the event.
- We would introduce the speaker and save 5-8 minutes for questions.

In-person

- Typically, there were student led meetings with seminar speakers in person and we would also intro and MC the seminar speaker.
- The department budget for seminar speaker dinners was generous and we could generally let 4-6 people attend. Please check in with Brian Giebel on what these numbers are for each year.

Mentorship Program

Please see our Dropbox which summarizes our mentorship program. Here, faculty and mentee pairs were made as well as trainee and mentee pairs. Quarterly socials brought grad students and first year students together through check-in activities to help folks identify their thesis labs. Access to this Dropbox folder may be accessed through admin in the Genome Sciences Department and Chair of the Genome Sciences DEI committee.

Finances

GSAIMS manages an annual budget. Our template for budgets can be found on our dropbox. Below includes descriptions of our funding sources and their quantities:

- 2019-2020: Genome Sciences
    - $2500
- 2020 - 2021: Seed grant
    - $5000

- 2021 - 2022: Genome Sciences
    - $2500
- 2022 - 2023: Genome Sciences and Data Science Grant
    - $7000

For all budget related expenses, Maureen Larsen, Brian Giebel, and Bill Noble (for DEI Committee related items as of 2023) are the go-to contacts.

Contact Directory

Individuals involved with GSAIMS leadership and organizing

Robin Aguilar
Phoebe Parrish
Zorian Thornton
Gesine Cauer
Leah Anderson
Dr. Ken Jean-Baptiste
Dr. Claudia Espinoza
Dr. Bianca Ruiz
Dr. Eva Nichols
Dr. Alberto Rivera
Dr. Atom Lesiak
Brian Giebel
Maureen Larsen

Individuals who participated in GSAIMS events

Lisa Peterson
Fievel Jack Finley
Dr. Tim Thornton
Dr. LaShawnDa Pittman
Dr. Kate Mittendorf
Andrea Salazar-Nuñez
Dr. Yi (Jenny) Xiao
Dr. La Tasha Levy
Dr. Tobias Mann
Dr. Kaela Singleton

Dr. Robert Fernandez
Dr. Evelyn Valdez-Ward
Dr. Andrea Gomez
Dr. Irene Y. Chen
Dr. Nick Altemose
Dr. Jey McCreight
Dr. Leonora Martínez Núñez
Dr. Chazeman Jackson
Dr. Karen Peterson
Jenny Montooth, M.A

Conclusions

GSAIMS programming accounted for over 250 hours of organizing which does not include event set-up, purchasing snacks, items, and supplies from local stores, and more.

Running and leading student-led organizations takes a substantial amount of effort, time, and commitment and there are several lessons that involved team members have learned over the years with managing GSAIMS.

For trainee led initiatives to be successful, they need buy-in, support, and active engagement with department members. This means that beyond involved trainees, faculty members and staff as community members who benefit from these groups through their grants and recruitment efforts **must** be actively involved. Over the years and throughout the 2020 BLM uprisings GSAIMS saw waves of interest come and go as 'DEI' efforts were phased in and out of practice. Even more so, GSAIMS work was cited by department members for funding with little to no support through attendance of events.

While trainee led initiatives and programs are incredibly valuable toward improving student life through mentorships, much needed space, and for building resources, they cannot be successful unless there is continuous and ongoing support, thorough leadership, and accountability from

community members who benefit from these spaces. Even more so, these spaces cannot solely fix systemic and institutional racism, misogyny, ableism, and many forms of discrimination that exist in academic spaces. It's crucially important to address that these inequities in STEM need to be challenged, addressed, and acknowledged.

Even more so, this type of organizing labor absolutely merits to be described as deeply valuable scientific contributions. Without these types of student groups, many people would struggle to do well in their labs and research spaces. At the time that GSAIMS was created, trainees did not receive any additional compensation for this labor. This is something that faculty and staff should absolutely note and should advocate for experts to facilitate that these student groups are supported and that marginalized students are not burdened with taking the initiative of creating spaces where they are already burdened. I hope that this type of labor is fairly compensated in the future.

This leads to the importance of continuity and implementation. DEI efforts in academic spaces would be much more effective if there were trained staff and faculty members who were thinking about the lasting vision of these efforts full-time. Committees that challenge students and tax faculty who are already overworked cannot lead to practical and tangible solutions that will benefit trainees who are marginalized.

Appendix C - Curriculum Outline and Resources

Course Title: Interrogating Belonging and Identity in Bioscience Graduate Programs

About

As scientists, we must acknowledge that our research findings can have real-world applications and consequences. Genomics research has had a significant impact on discussions surrounding racial and gender discrimination. Unfortunately, poorly communicated scientific findings have contributed to the continuation of eugenics in medicine and present-day racism in academic research. Additionally, bioscience research spaces can be challenging and isolating for researchers with marginalized identities due to various biases present in academic institutions. This course aims to explore the intersections of identity and belonging, highlighting their crucial roles in the success and outcomes of pursuing graduate studies in the biosciences. Given the broad nature of topics like identity, community building, equity, and justice in higher education, this course serves as an introductory exploration of how these factors intersect with race, gender, sexual orientation, socioeconomic status, ethnicity, nationality, and disability in academic spaces.

The goals of this course are twofold. Firstly, it aims to provide participants with a platform to discuss the impacts and implications of marginalization in academic spaces. Secondly, it seeks to empower trainees to integrate principles of equity and justice into their research labs and academic programs as active community members.

Furthermore, trainees will have the opportunity to collaborate on community-centered projects that uplift and support individuals with shared lived experiences and learning goals. These projects will serve as bridges, connecting aspiring scientists and non-scientist community members who are interested in accurate and informative scientific research outcomes.

Format of the course

This course primarily focuses on reading and discussion, supplemented with in-class materials such as discussions from assigned readings and recorded talks from academic researchers. At the beginning of the course, participants will be randomly assigned to reading and project groups. Each group will be responsible for leading weekly discussions on assigned paper topics. The

course is designed to span 10 weeks for a course that meets three times a week for a size of 12-20 participants. However, the course can be adapted and expanded to serve as a semester-long course, thanks to the availability of extensive supplementary resources.

Grading

All course participants start the course with an A and grades are based on class participation, weekly online discussion posts from readings, and a final group course project and featured presentation.

Course Guide and Material by Week

Week 1 - Introduction, Course Guidelines, Discussion, DEI work in academic spaces

Day 1 - Monday

Assigned readings:

- Murray, De-Shaine, et al. "Black in neuro, beyond one week." *Journal of Neuroscience* 41.11 (2021): 2314-2317.
- Singleton, Kaela S., et al. "A year in review: Are diversity, equity, and inclusion initiatives fixing systemic barriers?." *Neuron* 109.21 (2021): 3365-3367.
- Harrison, Colin, and Kimberly D. Tanner. "Language matters: Considering microaggressions in science." *CBE—Life Sciences Education* 17.1 (2018): fe4.

In class handouts:

- Syllabus
- Community Guidelines

Day 2 - Wednesday

Assigned readings:

- Cech, Erin A. "The intersectional privilege of white able-bodied heterosexual men in STEM." *Science Advances* 8.24 (2022): eabo1558.
- Carlson, Jedidiah, et al. "Counter the weaponization of genetics research by extremists." *Nature* 610.7932 (2022): 444-447.

In class handouts:

- Discussion Set 1 (Instructor)

Day 3 - Friday

Assigned readings:

- Jimenez, Miguel F., et al. "Underrepresented faculty play a disproportionate role in advancing diversity and inclusion." *Nature ecology & evolution* 3.7 (2019): 1030-1033.
- Taffe, M. A., and N. W. Gilpin. "Equity, diversity and inclusion: Racial inequity in grant funding from the US National Institutes of Health. eLife, 10, Article e65697." (2021).
- Billmyre, Katherine K., et al. "Meiosis in Quarantine discussions lead to an action plan to increase diversity and inclusion within the genetics community." *PLoS genetics* 17.7 (2021): e1009648.

In class:

- Discussion Set 2
- Introduction to final projects
- Week 1 reflections due

Week 2 - Medical Racism and Population Genomics, Science communication

Day 1

Assigned readings:

- Gouvea, Julia Svoboda. "Addressing Racism in Human Genetics and Genomics Education." *CBE—Life Sciences Education* 21.4 (2022): fe5.
- Wedow, Robbee, Daphne O. Martschenko, and S. Trejo. "Scientists must consider the risk of racist misappropriation of research." *Scientific American* (2022).
- Cerdeña, Jessica P., Vanessa Grubbs, and Amy L. Non. "Genomic supremacy: the harm of conflating genetic ancestry and race." *Human Genomics* 16.1 (2022): 1-5.

In class:

- Discussion set 3 - Group 1

Day 2

Assigned readings:

- Davis, L. K. "Human genetics needs an antiracism plan." *Scientific American* 17 (2021): 2021.
- Byeon, Yen Ji Julia, et al. "Evolving use of ancestry, ethnicity, and race in genetics research—A survey spanning seven decades." *The American Journal of Human Genetics* 108.12 (2021): 2215-2223.

In class:

- Discussion set 4 - Group 2

Day 3

Assigned readings:

- Stanton, Julie Dangremond, et al. "Drawing on internal strengths and creating spaces for growth: How Black science majors navigate the racial climate at a predominantly white institution to succeed." *CBE—Life Sciences Education* 21.1 (2022): ar3.
- Skloot, Rebecca. *The immortal life of Henrietta Lacks*. Broadway Paperbacks, 2017. (Chapters 23 - 26)

In class:

- Discussion set 5 - Group 3
- Week 2 reflections due

Week 3 - Mentorship outcomes in STEM higher education

Day 1

Assigned readings:

- Dukes, Angeline. "How to better support black trainees in the biomedical sciences." *Nature Medicine* 26.11 (2020): 1674-1674.
- Harp, Djana, et al. "Race and gender inequalities in medicine and biomedical research." *Critical research on sexism and racism in STEM fields*. IGI Global, 2016. 115-134.
- Nguyen, Mytien, et al. "Variation in research experiences and publications during medical school by sex and race and ethnicity." *JAMA network open* 5.10 (2022): e2238520-e2238520.

In class:

- Discussion set 6 - Group 4

Day 2

Assigned readings:

- Ostrove, Joan M., and Susan M. Long. "Social class and belonging: Implications for college adjustment." *The Review of Higher Education* 30.4 (2007): 363-389.
- Sensoy, Özlem, and Robin DiAngelo. "“We are all for diversity, but...”: How faculty hiring committees reproduce whiteness and practical suggestions for how they can change." *Harvard Educational Review* 87.4 (2017): 557-580.

In class:

- Discussion set 7 - Group 1

Day 3

Assigned readings:

- Montgomery, Beronda L. *Lessons from plants*. Harvard University Press, 2021. - Final chapter

In class:

- Discussion set 8 - Group 2
- Week 3 reflections due

Week 4 - Indigenous Data Sovereignty and Race in Genomics, Community Advocacy

Day 1

Assigned readings:

- Claw, Katrina G., et al. "A framework for enhancing ethical genomic research with Indigenous communities." *Nature communications* 9.1 (2018): 2957.
- Fox, Keolu. "The illusion of inclusion—The “All of Us” research program and indigenous peoples’ DNA." *New England Journal of Medicine* 383.5 (2020): 411-413.
- Tsosie, Krystal S., Keolu Fox, and Joseph M. Yracheta. "Genomics data: the broken promise is to Indigenous people." *Nature* 591.7851 (2021): 529-530.

In class:

- Discussion set 9 - Group 3
- Watch a talk by Dr. Keolu Fox

Day 2

Assigned readings:

- Fox, Keolu, Kartik Lakshmi Rallapalli, and Alexis C. Komor. "Rewriting human history and empowering indigenous communities with genome editing tools." *Genes* 11.1 (2020): 88.
- Mackey, Tim K., et al. "Establishing a blockchain-enabled Indigenous data sovereignty framework for genomic data." *Cell* 185.15 (2022): 2626-2631.

In class:

- Discussion set 10 - Group 4

Day 3

Assigned readings:

- TallBear, Kim. *Native American DNA: Tribal belonging and the false promise of genetic science*. U of Minnesota Press, 2013. - Chapters 1 and 2

In class:

- Discussion set 11 - Group 1
- Week 4 reflection due

Week 5 - Latinx in genomics, Careers beyond academia

Day 1

Assigned readings:

- Peña, Lorgia García. *Community as rebellion: A syllabus for surviving academia as a woman of color*. Haymarket Books, 2022. - Chapters 1 and 3
- De Ver Dye, Timothy, et al. "Participation in genetic research among Latinx populations by Latin America birth-residency concordance: a global study." *Journal of Community Genetics* 12.4 (2021): 603-615.

In class:

- Discussion set 12 - Group 2

Day 2

Assigned readings:

- McGee, Ebony Omotola. *Black, brown, bruised: How racialized STEM education stifles innovation*. Harvard Education Press, 2021. - Chapters 3 and 5
- Liu, S.N.C., Brown S.E.V., & Sabat, I. E. (2019). Patching the “Leaky Pipeline”: Interventions for Women of Color Faculty in STEM Academia. *American Psychological Association, 7*, 32-39. http://dx.doi.org/10.1037/arc0000062

In class:

- Discussion set 13 - Group 3

Day 3

Assigned readings:

- Davies, W.S., *et al.* (2021). Promoting inclusive metrics of success and impact to dismantle a discriminatory reward system in science. *PLoS Biology*, *19(6),* e3001282. https://doi.org/10.1371/journal.pbio.3001282
- Deepshikha Chatterjee, Gabrielle A. Jacob, Susi Sturzenegger Varvayanis, Inge Wefes, Roger Chalkley, Ana T. Nogueira, Cynthia N. Fuhrmann, Janani Varadarajan, Nisan M. Hubbard, Christiann H. Gaines, Rebekah L. Layton, Sunita Chaudhary. Career Self-Efficacy Disparities in Underrepresented Biomedical Scientist Trainees. *bioRxiv 2022.10.21.512368*; doi: https://doi.org/10.1101/2022.10.21.512368 (Preprint)

In class:

- Discussion set 14 - Group 4
- Week 5 reflection due

Week 6 - Gender and Sexuality in Genomics and STEM

Day 1

Assigned readings:

- Bryce E. Hughes. (2018) Coming out in STEM: Factors affecting retention of sexual minority STEM students. Science Advances, 4:3, eaao6373, DOI: 10.1126/sciadv.aao6373
- Aguilar, Robin. "Breaking the binary by coming out as a trans scientist." *Nature* 591.7849 (2021): 334-336.
- Freeman, Marc. "Ben Barres: neuroscience pioneer, gender champion." *Nature* 562.7727 (2018): 492-493.

In class:

- Discussion set 15 - Group 1

Day 2

Assigned readings:

- Maloy J, Kwapisz MB, and Hughes BE. (2022). Factors influencing retention of transgender and gender nonconforming students in undergraduate STEM majors. *CBE – Life Sciences Education. 21*(1). https://doi.org/10.1187/cbe.21-05-0136
- Lesiak, Atom J. "I'm a trans scientist-here's my advice for navigating academia." *Nature*.

In class:

- Discussion set 16 - Group 2
- Watch a talk with Dr. Ben Barres

Day 3

Assigned readings:

- Jeremy B. Yoder & Allison Mattheis. (2016) Queer in STEM: Workplace Experiences Reported in a National Survey of LGBTQA Individuals in Science, Technology, Engineering, and Mathematics Careers. Journal of Homosexuality, 63:1, 1–27, DOI: 10.1080/00918369.2015.1078632
- Eric V. Patridge, Ramon S. Barthelemy, Susan R. Rankin. (2014) Factors Impacting the Academic Climate for LGBQ STEM Faculty. Journal of Women and Minorities in Science and Engineering, 20:1, 75–98, DOI:

In class:

- Discussion set 17 - Group 3
- Week 6 reflections due
- Final project proposals due

Week 7 - Black Feminism and Mentorship, Black in the Ivory

Day 1

Assigned readings:

- Alexander, Q. R., & Hermann, M. A. (2016). African-American women's experiences in graduate science, technology, engineering, and mathematics education at a predominantly white university: A qualitative investigation. *Journal of Diversity in Higher Education, 9*(4), 307-322. http://dx.doi.org/10.1037/a0039705
- Bernard, D. L., Lige, Q. M., Willis, H. A., Sosoo, E. E., & Neblett, E. W. (2017). Impostor phenomenon and mental health: The influence of racial discrimination and gender. *Journal of counseling psychology*, *64*(2), 155–166. https://doi.org/10.1037/cou0000197

In class:

- Discussion set 18 - Group 4

Day 2

Assigned readings:

- Hoppe, T. A., Litovitz, A., Willis, K. A., Meseroll, R. A., Perkins, M. J., Hutchins, B. I., Davis, A. F., Lauer, M. S., Valantine, H. A., Anderson, J. M., & Santangelo, G. M. (2019). Topic choice contributes to the lower rate of NIH awards to African-American/black scientists. *Science advances*, *5*(10), eaaw7238.
- The Disordered Cosmos by Dr. Chanda Prescod-Weinstean, last two chapters.

In class:

- Watch a talk by Dr. Chanda Prescod-Weinstein
- Discussion set 19 - Group 1

Day 3

Assigned readings:

- Cox K.L., Elliott K. R., Harris T.M. (2021). Creating supportive environments in academia for Black scientists to thrive. *The Plant Cell 33*(7)*: 2112-2115.* https://doi.org/10.1093/plcell/koab125 (Letter to the Editor)

- Bonham V.L., Green E.D., and Perez-Stable E.J.(2018). Examining How Race, Ethnicity, and Ancestry Data Are Used in Biomedical Research. *JAMA. 320*(15):*1533-1534.* https://doi.org/10.1001/jama.2018.13609 (Viewpoint)

In class:

- Discussion set 20 - Group 2
- Week 7 reflections due

Week 8 - Disability

Day 1

Assigned readings:

- Reinholz, Daniel L., and Samantha W. Ridgway. "Access needs: Centering students and disrupting ableist norms in STEM." *CBE—Life Sciences Education* 20.3 (2021): es8.
- Schneiderwind, Joseph, and Janelle M. Johnson. "Broadening the Equity Lens for STEM Teacher Education: The Invisibility of Disability." *AAAS bulletin* (2021).

In class:

- Discussion set 21 - Group 3

Day 2

Assigned readings:

- My leave of absence from grad school changed my perspective on taking a break. Science (2022). Jacqueline Forson.
- As a Ph.D. student with an expensive chronic disease, low stipends make academia untenable. Science (2022). Ahmed Elbassiouny.
- Wong, Alice, ed. *Disability visibility: First-person stories from the twenty-first century*. Vintage, 2020. Pages 39-53

In class:

- Discussion set 22 - Group 4

Day 3

Assigned readings:

- Pfeifer M.A., Reitar E.M., Cordero J.J., and Stanton J.D. (2021). Inside and Out: Factors That Support and Hinder the Self-Advocacy of Undergraduates with ADHD and/or Specific Learning Disabilities in STEM. *CBE -- Life Sciences Education. 20*(2). https://doi.org/10.1187/cbe.20-06-0107
- Wong, Alice, ed. *Disability visibility: First-person stories from the twenty-first century*. Vintage, 2020. Pages 232 - 242

In class:

- Discussion set 23 - Group 1
- Group proposal feedback due
- Week 8 reflections due

Week 9 - Justice and trainee led organizing in higher education

Day 1

Assigned readings:

- How faculty hiring sustains inequity in academia. Science Careers (2022). Viviane Callier.
- Academic mentors wield great power. We need to feel safe talking about abuses. Science Careers (2022). Adaira Landry.

In class

- Discussion set 24 - Group 2

Day 2

Assigned readings:

- Academic bullying is too often ignored. Here are some targets' stories. Science Careers (2021). Katie Langin.
- Hu, J. "NSF graduate fellowships disproportionately go to students at a few top schools." *Science* (2019).

In class

- Discussion set 25 - Group 3

Day 3

Assigned readings

- Kimmerer, Robin. *Braiding sweetgrass: Indigenous wisdom, scientific knowledge and the teachings of plants*. Milkweed editions, 2013. Picking Sweetgrass section.

In class

- Discussion set 26 - Group 4
- Week 9 reflection due

Week 10 - Final project presentations

Day 1

In class
- Group 1 presentation
- Group 2 presentation

Day 2

In class
- Group 3 presentation
- Group 4 presentation

Day 3

In class
- Course reviews and evaluations
- Week 10 reflection due

Classroom Interactions Guidelines and Community Expectations

These guidelines were adapted from the University of Michigan's 'Guidelines for Classroom Interaction (from CRLT) which references:

Brookfield, S. D., & Preskill, S. (2012). *Discussion as a Way of Teaching: Tools and Techniques for Democratic Classrooms.* Wiley.

Sensoy, Ö., & DiAngelo, A. (2014). Respect Differences? Challenging the Common Guidelines in Social Justice Education. *Democracy and Education, 22*(2), Article 1. Available at: https://democracyeducationjournal.org/home/vol22/iss2/1

Classroom Interactions

- **Share responsibility for including all voices in the conversation.** If you tend to have a lot to say, make sure you leave sufficient space to hear from others. If you tend to stay quiet in group discussions, challenge yourself to contribute so others can learn from you.
- **Listen respectfully.** Don't interrupt, turn to technology, or engage in private conversations while others are speaking. Use attentive, courteous body language. Comments that you make (whether asking for clarification, sharing critiques, or expanding on a point) should reflect that you have paid attention to the previous speakers' comments.

- **Be open to changing your perspectives based on what you learn from others.** Try to explore new ideas and possibilities. Think critically about the factors that have shaped your perspectives. Seriously consider points-of-view that differ from your current thinking.
- **Understand that we are bound to make mistakes in this space**, as anyone does when approaching complex tasks or learning new skills. Strive to see your mistakes and others' as valuable elements of the learning process.
- **Understand that your words have effects on others.** Speak with care. If you learn that something you've said was experienced as disrespectful or marginalizing, listen carefully and try to understand that perspective. Learn how you can do better in the future.
- **Understand that others will come to these discussions with different experiences from yours**. Be careful about assumptions and generalizations you make based only on your own experience. Be open to hearing and learning from other perspectives.
- **Understand that there are different approaches to solving problems.** If you are uncertain about someone else's approach, ask a question to explore areas of uncertainty. Listen respectfully to how and why the approach could work.

Community Expectations

1. Confidentiality. We want to create an atmosphere for open, honest exchange.
2. We will not demean, devalue, or "put down" people for their experiences, lack of experiences, or difference in interpretation of those experiences.
3. We will trust that people are doing the best they can.
4. Challenge the idea and not the person. If we wish to challenge something that has been said, we will challenge the idea or the practice referred to, not the individual sharing this idea or practice.
5. Step Up, Step Back. Be mindful of taking up much more space than others. On the same note, empower yourself to speak up when others are dominating the conversation.

Weekly Discussion Question Framework

Following each course, there will be readings assigned related to each weekly reading topic. Some of these readings will draw from past weekly readings and themes and discussion of these intersections and themes are highly encouraged!

Responsibilities for each group-led discussion:

Each group member must contribute at least <u>one</u> insight, question, or theme that they found of importance from the assigned readings. This can be related to an article or feature that resonated with them, questions they still have about the reading, applications to the present program, observations from lived experiences to what's described in text, etc. It's possible that not every group member's question will be discussed during the in-class discussion session and each group's question set will be submitted towards individual course participation credit and should be submitted before the class discussion on Canvas.

Each group is responsible for introducing the course discussion for the day with their own independent questions that were taken from the previous assigned readings. These questions are intended to summarize the reading and offer perspectives and insights that were taken away during the core readings to facilitate discussion among the rest of the course participants. Others from different groups will share their own insights, takeaways, and observations that they also found were important to them for that reading.

Group participants don't need to worry about moderating discussion, as this is the responsibility of the instructor who will also participate with insights and clarification on topics during discussion.

Example questions from Week 1, Day 1 Assigned Readings:

To introduce the format and flow of general in-class discussions, the instructor will give an example during the first class week of how to consider preparing discussion questions.

Week 1

Day 1

- When did you first learn what a science based DEI committee was?
- What are factors that can make community centered spaces sustainable in biosciences?

Day 2

- What makes intersectionality an important component of DEI work and initiatives?
- Can you think of mediums and ways in which researchers can take an active stance to counter misinformation from their findings?

Final Project Timeline and Submission Guidelines

The final course project will be an opportunity for groups to collectively work on a topic that can have community impact and or publications that can be of value to other researchers. Here, groups will come together to write a pitch of a project of interest that they have in mind to execute:

Examples

1. **How-to topics**. How to use X tool to make scientific illustrations that are accessible.
2. **Tutorials**. Step-by-step guide to launch a personal portfolio website using GitHub with features useful for highlighting X skills.
3. **Medium articles**. "A collaborative approach to studying for qualifying exams."
4. **Perspective pieces to larger research journals**. "How to make a community space for marginalized scientists without burdening trainees?"

5. **Outreach and/or community projects**. Guidelines and form to organize a community science art zine.

Project proposals will include an abstract about pitch, the estimated number of hours to complete the project, anticipated delivery for launch time (website launch, hypothetically when an outreach project would be launched, etc.), and description of responsibilities of group members. If a group is pitching to a research journal and needs help coming up with emails and journal contacts, the instructor can help offer introductions to editors. Likewise, if groups wish to execute their outreach projects beyond the scope of the course, the instructor will facilitate in making connections to mentor students in this area.

The project proposals are due: On the last day of the 6th week of the course.

Other groups will offer anonymous feedback to one another regarding each project and questions that they have about the project and its applications. This is to give each group feedback on their project delivery and feasibility for the final presentation.

This project feedback is due: On the last day of the 8th week of the course

Final projects are introduced to the course as 30 min group presentations. These talks can include interactive components to showcase the articles, article pitches, tutorials, plans for proposed outreach projects, etc. Projects will be graded for completion and group participation. Feedback for each project will be given that addresses how well each group assessed the benefits, costs, and limitations of their projects and their implications and impact for other community members when they create and implement their works.

The projects themselves are due: On the last day of the 9th week of the course. Final group presentations will happen the last week of the course when course feedback is due.

Weekly Reflection Submission Guidelines

At the beginning of each week, a reflection question will be posted on Canvas that will be due on the last day of the class week.

Each of these questions are graded by the instructor and will be graded based on participation and completeness. Full credit will be given to responses that draw from previous readings and in-class discussions. The expected length of a response is ½ - one page in length but reflections will be graded for quality of the response. Reflections can also include lessons learned from the week, and/or valuable insights from other mediums outside of the course that would like to be shared that are related to past or future course topics.

Weekly Reflection Prompts by Week

Week 1

What are the costs and benefits of academic researchers leading DEI initiatives? What solutions do you think are possible in creating more accessible and equitable academic environments?

Week 2

What are ways in which researchers can combat misappropriation of research and in public facing platforms? What aspects of science communication do you believe are crucial toward building trust with communities?

Week 3

What are mentorship attributes that you value? Can you identify how your needs align with your goals in graduate school based on some of the frameworks described from readings this week?

Week 4

Describe something that you learned about Indigenous data sovereignty and data privacy. What are challenges you currently see within academic programs that cause barriers toward partnership with Indigenous communities?

Week 5

What are ways in which a science community can be accountable for 'the minority tax'? Do you think there are costs and benefits to URM REU (underrepresented minority research experiences for undergraduates) programs with recruitment of students with marginalized identity.

Week 6

How does intersectionality play a role in DEI discussions related to sexual orientation and gender in biosciences?

Week 7

What takeaways did you find from 'The Disordered Cosmos' with respect to Dr. Prescod-Weinstein's experiences as a Black agender woman in physics compared to the stories that you've read from researchers in the biosciences?

Week 8

What are ways in which intersectionality and disability are overlooked in DEI work?

Week 9

Reflect on the reading you learned the most from this quarter.

Supplementary Readings by Week

Week 1

- Harper, Jordan, and Adrianna Kezar. "Leadership Development for Racially Minoritized Students: An Expansion of the Social Change Model of Leadership." *Journal of Leadership Education* 20.3 (2021).

Week 2

- Strassle, Paula D., et al. "COVID-19–related discrimination among racial/ethnic minorities and other marginalized communities in the United States." *American Journal of Public Health* 112.3 (2022): 453-466.

Week 3

- Tong, Yunhe. "As a nonnative speaker, I struggled to write scientific papers in English: Here's how I learned." (2022).
- Teves, Sheila. "How I learned to embrace my identity as an academic with immigrant working-class roots." Science Careers. Working Life (2021).
- Ginther, Donna K., et al. "Race, ethnicity, and NIH research awards." *Science* 333.6045 (2011): 1015-1019.
- Byars-Winston, Angela, et al. "Race and ethnicity in biology research mentoring relationships." *Journal of Diversity in Higher Education* 13.3 (2020): 240.

Week 4

- Devasmita Chakraverty (2022). A Cultural Impostor? Native American Experiences of Impostor Phenomenon in STEM. *CBE—Life Sciences Education. 21*(1). https://doi.org/10.1187/cbe.21-08-0204

Week 5

- Hablemos Genomics: Engaging Latinos in the Future of Genomic Science

- Morgan, A., Clauset, A., Larremore, D., LaBerge, N., & Galesic, M. (2021). Preprint. Socioeconomic Roots of Academic Faculty. *SocArXiv.* https://doi.org/10.31235/osf.io/6wjxc

- Christine Pfund, Fátima Sancheznieto, Angela Byars-Winston, Sonia Zárate, Sherilynn Black, Bruce Birren, Jenna Rogers, and David J. Asai. (2022). Evaluation of a Culturally Responsive Mentorship Education Program for the Advisers of Howard Hughes Medical Institute Gilliam Program Graduate Students. *CBE—Life Sciences Education 2022 21:3. https://doi.org/10.1187/cbe.21-11-0321*

Week 6

- Else, Holly. "Largest-ever survey exposes career obstacles for LGBTQ scientists." *Nature* (2021).

Week 7

- One mentor isn't enough. Here's how I built a network of mentors
- Lerma, V., Hamilton, L.T., Nielsen, K. (2020) Racialized Equity Labor, University Appropriation and Student Resistance. *Social Problems*, *67*(2): 286–303, https://doi.org/10.1093/socpro/spz011 (Full manuscript)

- Daniels H.A., Grineski S.E., Collins T.W, and Frederick A.H. (2019). Navigating Social Relationships with Mentors and Peers: Comfort and Belonging among Men and Women in STEM Summer Research Programs. *CBE Life Sciences Education 18*:2. https://doi.org/10.1187/cbe.18-08-0150
- Lloreda CL. (2022). Racial and gender disparities in publishing start early for doctors and scientists. *Science.* https://doi,org/10.1126/science.caredit.adf5132 (Careers Editorial)
- I'm a Black scientist, tired of facing racism and exclusion from academia. Science Careers. Keisha Hardeman (2023).
-

Week 8

- Wong, Alice, ed. *Disability visibility: First-person stories from the twenty-first century*. Vintage, 2020. Pages 271 - end.

Week 9

- Kenneth D. Gibbs, and Kimberly A. Griffin. (2017) What Do I Want to Be with My PhD? The Roles of Personal Values and Structural Dynamics in Shaping the Career Interests of Recent Biomedical Science PhD Graduates *CBE – Life Sciences Education* 12(4). https://doi.org/10.1187/cbe.13-02-0021
- Hinton, A. O., Jr, Termini, C. M., Spencer, E. C., Rutaganira, F., Chery, D., Roby, R., Vue, Z., Pack, A. D., Brady, L. J., Garza-Lopez, E., Marshall, A. G., Lewis, S. C., Shuler, H. D., Taylor, B. L., McReynolds, M. R., & Palavicino-Maggio, C. B. (2020). Patching the Leaks: Revitalizing and Reimagining the STEM Pipeline. *Cell*, *183*(3), 568–575. https://doi.org/10.1016/j.cell.2020.09.029

Vita

Robin Aguilar was born in Downey, California, on June 19, 1996. After growing up in East Los Angeles, CA, they graduated from Sultana High School in the High Desert of Southern California in the spring of 2014 and commenced their undergraduate studies at DePauw University later that year. During this period, they initiated their academic research career in the biochemistry lab of Dr. Daniel Gurnon, collaborating with the Rare Genomics Institute. Throughout their undergraduate journey, they pursued internships at Stanford University, the University of Geneva, and the Department of Genome Sciences in the labs of Dr. Michele Calos, Dr. Marie Cohen, and Dr. Jay Shendure, respectively. In the spring of 2018, Robin graduated from DePauw with a B.A. in Biochemistry and dual minors in Computer Science and Spanish. Later that same year, they embarked on their graduate education in the Department of Genome Sciences at the University of Washington, initially as an NSF Fellow and later as a Gilliam Fellow. During this time, they conducted research in the laboratories of Dr. Brian Beliveau and Dr. William Noble and defended their thesis on October 31, 2023.