1-1-2013

Measuring Transcription Directly From Our Chromosomes

Marshall Levesque
University of Pennsylvania, marshall.levesque@gmail.com

Measuring Transcription Directly From Our Chromosomes

Abstract
Our genome is organized into DNA segments called chromosomes. Alterations to the typically invariant number and composition of chromosomes are hallmarks of serious disease like cancer. Understanding how rearranging chromosomes affects chromosomal behavior and ultimately leads to disease requires chromosome-specific gene expression measurements, but current tools are insufficient. This thesis describes tools for measuring transcription while discriminating which copy of a gene the RNA comes from. The ability to take these measurements in single cells enabled us to measure changes in transcription on translocated chromosomes or from the maternal vs. paternal chromosomes.

Firstly, we introduce intron chromosomal expression FISH (iceFISH), a multiplex imaging method for measuring transcription and chromosome structure simultaneously on single chromosomes. We find substantial differences in transcriptional frequency between genes on a translocated chromosome and the same genes in their normal chromosomal context in the same cell. Correlations between genes on a single chromosome pointed toward a *cis* chromosome-level transcriptional interaction spanning 14.3 megabases.

Chromosomes also come in nearly identical pairs and gene expression is a mixture of RNA transcribed from the maternal or paternal copies. The infrequent sequence differences between parental copies can have serious implications for the viability of cell or organism but detecting single nucleotide differences is difficult, making these behaviors nearly impossible to study in detail. We present a high efficiency fluorescence *in situ* hybridization method for detecting single nucleotide variants (SNVs) on individual RNA transcripts, both exonic and intronic. We used this method to quantify allelic expression at the population and single cell level, and also to distinguish maternal from paternal chromosomes in single cells.

The findings we present in this thesis have far-reaching implications for understanding the transcriptional effects of translocations, and the tools described in this thesis are widely applicable to studying gene regulation and developing in vitro diagnostics.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Bioengineering

First Advisor
Arjun Raj

Second Advisor
Andrew Tsourkas

Keywords
Chromosomes, fluorescence microscopy, RNA FISH, Single nucleotide variants, Translocations

Subject Categories
Biomedical | Cell Biology

MEASURING TRANSCRIPTION DIRECTLY FROM OUR CHROMOSOMES

Marshall James Levesque

A DISSERTATION

in

Bioengineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

Dr. Arjun Raj, Assistant Professor of Bioengineering

Graduate Group Chairperson

Dr. Daniel A. Hammer, Bioengineering Graduate Group Chair

Dissertation Committee

Dr. Gerd Blobel, Professor, Pediatrics

Dr. Mark Goulian, Professor, Biology & Physics

Dr. Ravi Radhakrishnan, Associate Professor, BE & CBE

Dr. Andrew Tsourkas, Associate Professor, BE

ACKNOWLEDGEMENT

I would like to thank my advisor Dr. Arjun Raj for his unwavering attention and support during my time working in his lab. His guidance and enthusiasm made everyday exciting in anticipation of seeing new details of our world for the first time.

The members of the Raj lab are an incredible bunch that always speak their mind and contribute to moving the lab and the science forward.

Biosearch Technologies provided the Raj lab with a lot of reagents during my time in the lab, thus allowing me to explore ideas that may have been out of reach without their support.

Many other people at Penn made my work and time enjoyable. Students, technicians, administrators, facilities... Smiling faces in the workplace were always appreciated.

My buddy Brendan Purcell was my brother that guided me through adjusting to life on the East Coast and learning to love Philly. Thanks B.

Lastly I would like to thank my beautiful wife Reiko for always supporting me and my occasional request to stop into lab on a weekend. And putting up with me staying up late to work on code or me being spaced out at home thinking about the day's experimental results.

I'll always look back at my time during grad school remembering the adventures in and outside of lab.

ABSTRACT

MEASURING TRANSCRIPTION DIRECTLY FROM OUR CHROMOSOMES

Marshall James Levesque

Dr. Arjun Raj

Our genome is organized into DNA segments called chromosomes. Alterations to the typically invariant number and composition of chromosomes are hallmarks of serious disease like cancer. Understanding how rearranging chromosomes affects chromosomal behavior and ultimately leads to disease requires chromosome-specific gene expression measurements, but current tools are insufficient. This thesis describes tools for measuring transcription while discriminating which copy of a gene the RNA comes from. The ability to take these measurements in single cells enabled us to measure changes in transcription on translocated chromosomes or from the maternal vs. paternal chromosomes.

Firstly, we introduce intron chromosomal expression FISH (iceFISH), a multiplex imaging method for measuring transcription and chromosome structure simultaneously on single chromosomes. We find substantial differences in transcriptional frequency between genes on a translocated chromosome and the same genes in their normal chromosomal context in the same cell. Correlations between genes on a single chromosome pointed toward a *cis* chromosome-level transcriptional interaction spanning 14.3 megabases.

Chromosomes also come in nearly identical pairs and gene expression is a mixture of RNA transcribed from the maternal or paternal copies. The infrequent sequence differences between parental copies can have serious implications for the viability of cell or organism but detecting single nucleotide differences is difficult, making these behaviors nearly impossible to study in detail. We present a high efficiency fluorescence *in situ* hybridization method for detecting single nucleotide variants (SNVs) on individual RNA transcripts, both exonic

and intronic. We used this method to quantify allelic expression at the population and single cell level, and also to distinguish maternal from paternal chromosomes in single cells.

The findings we present in this thesis have far-reaching implications for understanding the transcriptional effects of translocations, and the tools described in this thesis are widely applicable to studying gene regulation and developing in vitro diagnostics.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF PATENTS

Raj A, **Levesque MJ** "A method for detecting chromosome structure and gene expression simultaneously in single cells" WIPO Patent Application WO/2012/106711

Levesque MJ, Raj A, "A method for detecting single nucleotide variants in single cells or single molecules" submitted March 2012

CHAPTER 1 : Introduction

RNA is the messaging molecule transcribed from regions of our DNA called genes. After leaving a cells nucleus, RNA is ultimately translated into proteins that perform the functions of the cell. This whole process is called gene expression. The cell regulates levels of RNA as it needs more or less of each gene product. Scientists often study gene expression by growing millions of cells in a dish, breaking them up into a slurry, extracting out the molecules of interest, and end up measuring numbers like average RNA per cell. Measuring population-averaged RNA levels is sufficient to answer some questions about how a cell operates. However, each cell in the genetically identical cell population can behave differently due to variability in gene expression[63]. Population-based measurements mask this diversity of behaviors of individual cells.

With advances in technology, it is possible to characterize the previously 'hidden' differences between individual cells. Quantitatively characterizing this heterogeneity revealed new details in behaviors like intracellular signaling [21], noise in gene expression[16], and the dynamics of RNA transcription[60]. The work described in this thesis involved building and applying new tools to measure single cell transcriptional behaviors. We extend quantitative, single cell transcriptional measurements down to the single chromosome source of RNA.

Transcription is partially regulated by the linear arrangement of genes along chromosomes as well as the spatial positioning of chromosomes in the three-dimensional cell nucleus[53]. To investigate these ideas, we developed an assay called iceFISH capable of measuring chromosome structure and active transcription simultaneously in single cells. Using iceFISH, we uncovered chromosome-specific differences in transcriptional regulation between normal vs. rearranged chromosome copies. Contrary to conclusions in other studies[52, 33], spatial positioning of chromosomes and their genes did not correlate with transcriptional activity. We also showed that genes on the same chromosome copy display interactions despite being separated by significant genomic distance.

Our findings of gene copies behaving differently within the same cell led us to pursue a tool capable of measuring the allelic origin of individual RNA molecules. Alleles are non-identical copies of a gene or genetic locus. However, alleles can differ by only a single nucleotide, making it difficult to discriminate these minor differences in a quantitative manner for measuring RNA levels. After developing an assay capable of distinguishing single nucleotide differences, we characterized allelic imbalance for gene copies at the cell population and single cell level. We demonstrated how a cell population can be balanced in expressing both alleles of a gene, but on the single cell level show a bias towards one allele or the other. As an extension of previously mentioned work in detecting chromosome structure, we labeled numerous parental alleles simultaneously and identified parental origin of chromosomes along with their structure in the nucleus.

The following text provides an overview of literature covering the role of chromosome structure and its role in regulating transcription. We then summarize ideas and evidence of how the two copies of each gene in a cell may or may not produce equal amounts of RNA.

1.1. Genomic structure and its role in regulating gene expression

DNA is a linear molecule and as a genome it occupies the three-dimensional space of a cell's nucleus. In humans, the 6 billion nucleotide pairs (3 billion multiplied by 2 to account for both parents) that make up the genome would reach around 2 meters when stretched out straight. All this DNA is packaged into a nucleus often smaller than 10μm in diameter. The spatial arrangement of the DNA in the nucleus, across a range of length scales and complexity, is thought to regulate transcription, and thus cellular function, in many ways. This regulation affects the ability for cells to differentiate into different cell types[3, 25], to remember cellular state over time and cell division[26, 56], to co-regulate the expression of genes that depend on each other[69, 20], and designate one parental copy as the main source of expression[48]. This structure-function relationship suggests that disruption of the normal, healthy genome structure can lead to disease. The following is a survey of the literature on genomic organization and its role in regulating gene expression.

1.1.1. Chromosome copy number, dosage, aneuploidy

Human genomic DNA is organized into 23[1] segments called chromosomes. Each chromosome also comes in two copies, one from each parent. For a cell to divide, chromosomes are replicated, condense into tube-like structures, line up at the center of the dividing cell, and then evenly distribute into the two new daughter cells. After division, chromosomes decondense and fill up a newly formed nucleus. The number of chromosomes and the linear arrangement of genes that make up each chromosome are nearly flawlessly maintained over the countless cell divisions over one's life and the production of gametes to pass DNA to progeny. Disruptions in chromosome copy number or stitching together parts of different chromosomes are hallmarks of serious disease like cancer and developmental disorders. This strongly suggests that the structural composition of our genome is imperative for proper function.

Variation in chromosome copy number is a gross modification of genomic structure called aneuploidy and has serious implications for cellular function. The gain or loss of a chromosome copy correlates with an increase or decrease in expression of the genes on the chromosome, respectively . Expression changes roughly track copy number 1:1 so therefore having 3 copies of a chromosome, a trisomy, leads to 1.5X expression of the chromosome's genes that are normally 2 copies in a diploid cell[4]. This phenomenon is referred to as gene dosage. Fitness of the cell or organism is most always negatively affected by aneuploidy[70]. Serious developmental diseases like down-syndrome are caused by trisomies[64] and systematically induced trisomies show transcriptome-wide gene expression changes[57]. Detrimental effects of unbalanced chromosome copy number are thought to result from imbalanced stoichiometry of gene products. This argument hinges on the fact that some heteromeric complexes are composed of proteins encoded on different chromosomes and a difference in chromosome number changes the amount of one component, thus throwing off the balance in building the heteromeric protein structures[70].

[1] 22 plus the X or Y sex chromosome

1.1.2. Chromosome and gene positioning

Over more than a decade, views have evolved over what influences chromosome positioning in the nucleus. Most of what we know about chromosome organization in the nucleus came through the application of fluorescently labeled nucleic acid probes to fixed tissues, targeting genomic loci or whole chromosomes, and imaged using microscopy. These techniques are generally referred to as DNA FISH (fluorescence *in situ* hybridization).

Using whole chromosome probes[9], studies showed that the chromosomes occupy relatively distinct territories in the interphase nucleus, with smaller chromosomes toward the nuclear periphery[13]. This idea was explored in more detail by considering the composition of chromosomes, and radial positioning in the nucleus was attributed to gene density and replication timing within a chromosome[37]. Despite being mostly contained to their own territories, all 46 chromosomes packed into the nucleus do intermingle[80]. One is then presented with the questions of whether chromosomes come into contact with others more frequently than random. In studies that sought to better understand the causes of common cancers, researchers showed that the frequency of inter-chromosomal contacts within the nucleus is correlated with frequency of translocation events[7, 74, 89]. Overall, these studies established the existence of chromosome territories in the interphase nucleus, but left open how individual genes are arranged in and around those territories.

DNA FISH methods can be applied with genomic resolution down to detecting locations of individual genes or loci. Studies showed how the spatial positioning of genes within a chromosome territory can correlate with and potentially regulate transcription[55]. In general, the nuclear periphery is occupied by less active genes while active genes prefer the center of the nucleus[22, 28]. Focused regions of the nucleus with hotspots of transcriptional activity are referred to as transcription factories and house collections of active RNA polymerase[58, 54]. These factories can also contain shared transcription factors that bring together co-regulated genes from different chromosomes[69]. Besides the chromosome territories themselves, nuclear structure itself can play a regulatory role as shown in the example

of inactive portions of chromosomes rapidly activating transcription after dissociating with the nuclear lamina[59].

Studying the organization of DNA in the nucleus using microscopy has single cell resolution but suffers from both low-throughput and poor genomic resolution. Biochemical methods based on a technique called chromosome conformation capture (3C) measure DNA contact probability in a cell population with detailed genomic resolution [75, 14] and provide the opportunity to survey genome-wide using high-throughput DNA sequencing[47, 35, 83]. The problem with 3C-based tools is that they only produce gene-pair contact probabilities from the population, thus averaging all interactions. If a chromosome exists primarily in two distinct spatial conformations, the contacts measured from the two conformations would blend together into one set of signals. Another issue with using gene-pair interactions from the population is the example of studying three genes, A, B, and C. All gene-pair interactions might be detected, but one could never confidently claim that all three are in contact with each other at one time in single cells or if they are mutually exclusive. Most problematically, it is difficult to associate observed interactions with gene activity and impossible measure relative spatial position in the nucleus.

The lack of transcriptional activity measurements is a major downside to both DNA FISH and 3C-based methods. In order to characterize the influence of spatial arrangement of chromosomes on transcription, ideally one would observe both simultaneously. Additionally, to study the regulatory effects of altered genomic structure, such as translocations, we need to be able to assign transcriptional activity to either the normal or edited chromosome copies. This is our motivation in developing the iceFISH assay detailed in Single chromosome transcriptional profiling using iceFISH.

1.2. Transcription from different copies of a gene in single cells

A normal human cell contains two copies of its chromosomal DNA, one from each parent. Between the two copies, infrequent nucleotide differences can be sufficient to alter the levels

of transcribed RNA or the functionality of the translated protein product. Most approaches to measuring RNA levels, however, cannot distinguish the mixture of RNA molecules produced from both copies. Lack of interest is not why scientists overlook which allele an RNA comes from. The structural and thermodynamic differences of single nucleotide changes are too subtle for most measurement techniques to detect. If tools were sensitive enough to quantitatively measure RNA levels for each allele, especially at the level of a single cell, many unaddressed topics in transcriptional regulation become reasonable questions to explore. We highlight a number of these allele-specific gene expression topics in this section.

1.2.1. Variability in imprinted gene expression

Through a process of chemical modifications to the DNA during gametogeneis, some genomic regions are marked for parental-specific transcription[2]. This process is called genomic imprinting and is essential to fetal development, placental physiology, and driving the development of certain cell types[19]. Imprinted gene expression is also often seen in a tissue-specific manner[81], mainly in the placenta, testes, and neural cell lineages. To fully characterize this tissue-specific expression of imprinted genes, we must assign transcription to individual cells, thus making *in situ* techniques an attractive strategy.

From a clinical perspective, loss of imprinting due to genetic mutations is the cause of a number of serious diseases. One example is BeckwithWiedemann Syndrome (BWS)[40] where transcription is seen from both parental copies from a normally imprinted locus. Patients with BWS display overgrowth or asymmetry in their physical features and have a significantly higher risk of developing tumors[15]. The affected genes express balanced, biallelic RNA levels according to measurements from a cell population. These population measurements make it impossible to distinguish whether these loss of imprinting mutations lead to a 50/50 split in the population of cells with exclusive expression of maternal or paternal RNA, or each individual cell splitting its transcriptional output 50/50 between paternal and maternal copies. This is an incredibly well-suited opportunity for an allele-specific, single cell RNA measurement tool to provide the details needed to understand

mechanisms of a disease.

1.2.2. Allelic imbalance in non-imprinted genes

Beyond the all-or-none allelic expression of imprinting, which includes only a few hundred genes in humans[36], single cells or the cell population may transcribe RNA for a gene with a bias for the maternal or paternal copy. If one copy is defective in some way, an allelic imbalance as subtle as 60/40 favoring the defective copy could have detrimental physiological consequences. One simple cause for this type of expression pattern is having a single nucleotide polymorphism (SNP) in a transcription factor recognition site on one allele, thus altering transcriptional rate from that allele[65]. It is also plausible that SNPs within the gene do not alter transcription, but complex, inherited genetic differences scattered throughout the genome lead to the imbalance[84]. At the level of a single cell, an allelic imbalance could also be a transient manifestation through the stochastic nature of transcription, where genes are transcribed infrequently and in "bursts"[60]. If an RNA degrades fast enough while being produced in bursts of transcription from the same allele, a cells RNA would then have a strong bias for one version of a gene. This single cell imbalance is imperceptible using RNA from the cell population. Therefore, studying this behavior necessitates a quantitative, single-cell method of measuring allele-specific transcription.

This section outlined motivations for measuring allele-specific RNA levels on the single cell level. We succeeded in developing a technique to perform these types of studies *in situ* and describe the method and applications in Single cell allele-specific expression using SNP FISH.

CHAPTER 2 : Single chromosome transcriptional profiling reveals chromosome-level regulation of gene expression

2.1. Background

Researchers generally believe that the transcription of a gene's DNA into RNA is controlled by the interaction of regulatory proteins with DNA sequences proximal to the gene itself. At the same time, genes are organized by the thousands into chromosomes, raising the possibility that the structure or organization of chromosomes themselves may influence transcription. Indeed, there are several examples of complex regulatory interactions within clusters of genes[41, 68] and between segments of DNA separated by lengths up to a couple megabases[8]; however, little is known about how organization at the chromosome length scale affects gene expression. There are several hints that such chromosome-specific regulation may exist, such as the distinct banding patterns characteristic of particular chromosomes[42] and evidence for large-scale chromosomal rearrangements leading to disease[78], but the lack of tools to measure transcription along individual chromosomes has hampered the ability to directly test such hypotheses.

Here, we describe a microscopy-based method called iceFISH that enabled us to generate per-chromosome transcriptional profiles of 20 genes simultaneously along individual copies of human chromosome 19 in single cells. Using this tool, we first examined how chromosomal translocations altered transcription, finding substantial differences in transcriptional frequency between the majority of genes located on a translocated piece of chromosome 19 and the two normal copies of chromosome 19 present in HeLa cells. Second, measurement of correlations between these genes on a single chromosome revealed a cis chromosome-specific transcriptional interaction between two genes spanning 14.3 megabases. Our analysis did not, however, reveal connections between these effects and the three-dimensional conformations of the chromosome. Our findings point to the presence of long-range, chromosome-specific transcriptional regulatory mechanisms that may not depend on chromosome shape

but can be disrupted by alterations to genome structure.

2.2. Results

2.2.1. iceFISH allows per-chromosome transcriptional profiling

Our method utilizes RNA fluorescence *in situ* hybridization[18, 62] (RNA FISH) in a way that allows us to uniquely identify and localize the transcriptional activity of 20 genes from the same chromosome in single cells. For each gene, we wanted to visualize only nascent transcription and ignore mature messenger RNA (mRNA), so we took advantage of the fact that cells transcribe nascent RNAs comprised of exons and introns. The splicing process removes introns and joins exons into mature mRNAs that then leave the nucleus, with introns typically degrading rapidly after being spliced out of the nascent RNA. Labeling the intron thus enables one to measure whether or not the gene is actively transcribing[24], and, if active, the three-dimensional coordinates of that gene[32, 82, 79]. We probed the introns with sets of short, fluorescently labeled nucleic acid probes (Fig. 2.1a,[62]), and using fluorescence microscopy, we detected active sites of transcription in three dimensions without the accompanying mRNAs. Note that even genes considered constitutively active do not always actively transcribe RNA, as transcription occurs in short but intense "bursts" on the order of tens of minutes separated by "off" periods on the order of several hours[30, 10, 60, 72]; researchers believe these bursts arise from random aspects of the transcriptional process[63]. The overall transcription rate is proportional to the probability of finding such a spot for each gene (see Supplementary discussion).

Although targeting a single genes introns provides its 3D position, it does not yield information about the location of the gene relative to the chromosome. To visualize the chromosome, we designed probes targeting the introns of 20 genes along chromosome 19 (Fig. 2.2). Chromosome 19 is an ideal test case for our assay because it is densely populated with highly expressing genes. Applying all of these probes, each labeled with the same fluorophore, on human foreskin fibroblasts, we "painted" chromosome 19 (Fig. 2.1e), allowing one to readily

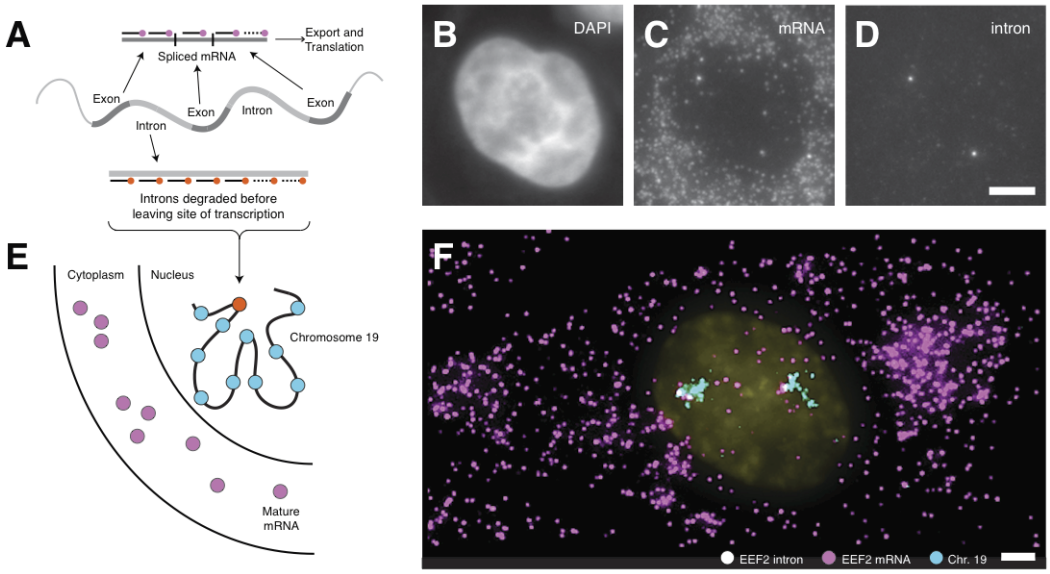

Figure 2.1: a. Depiction of our scheme for labeling the site of transcription by targeting gene introns with a series of labeled oligonucleotide probes. b. DAPI stain of the cells nucleus. c. RNA FISH targeting mRNA from the EEF2 gene (labeled with Alexa594 dye). d. RNA FISH targeting EEF2 introns (labeled with Cy3). e. Depiction of our scheme for labeling chromosome 19 via RNA FISH by labeling several introns simultaneously in a single color. f. 17 gene intron chromosome 19 "paint" (cyan, labeled with Cy3), intron of EEF2 (white, labeled with ATTO 647N), EEF2 mRNA (magenta, labeled with Alexa594), and the nucleus (yellow), labeled with DAPI. All images are maximum z-projections of a three-dimensional z-stack. All scale bars are 5μm long.

visualize the two chromosome copies (Fig. 2.1f), confirming the long known fact that interphase chromosomes are organized into distinct territories[13] (as beautifully demonstrated in Boltzer et al. PLoS Biology 2005[6]). Simultaneously, we applied a differently- colored probe targeting the intron of a particular gene on chromosome 19 (EEF2), and used a probe labeled with a third color to detect EEF2 mRNA (Fig. 2.1f). These chromosomal intron paints demonstrated the ability to simultaneously visualize chromosome structure, gene position, transcriptional activity, and mRNA abundance in a single (Fig. 2.1f) cell, a process which we call iceFISH for intron chromosomal expression FISH. RNase experiments showed that the spots did not result from probes binding to DNA (Fig. 2.3). All spots disappeared within 30 minutes after addition of the transcriptional inhibitor Actinomycin D (Fig. 2.4), and comparing intron and exon probe spot intensities at the site of transcription revealed a strong correspondence between active transcription and the presence of an intron spot (Fig. 2.5). (We also found a correspondence between run-on nascent transcript sequencing data19 and intron spot frequency; Fig. 2.6). Furthermore, colocalization of intron spots with bright exonic transcription sites (Fig.2.7, inset) and with DNA FISH probes targeting the gene locus (Fig. 2.8) show that the intron spot location correctly marks the site of transcription itself.

In order to measure all 20 genes transcriptional status simultaneously, we utilized a color-coding approach that enabled us to uniquely identify each spot in our chromosomal paint. We labeled each genes introns with a particular "pseudocolor", which is a distinct code for each gene consisting of either two or three out of a base palette of five spectrally distinguishable fluorophores (Fig. 2.9A,B; akin to other methods[44, 49]). To assign gene identity, we looked for colocalization of two or three spots in the images we acquired for each fluorescence channel (Fig 2.9C-H) corresponding to the pseudocolor probe labeling scheme we employed, aligning the channels using spots from an mRNA labeled with all 5 dyes. In human foreskin fibroblasts, which are primary cells containing two intact copies of chromosome 19 before DNA replication, we could usually discern two chromosomes clearly separated into individual territories (Fig. 2.9I,J), with roughly 22% of cells having commingled territories

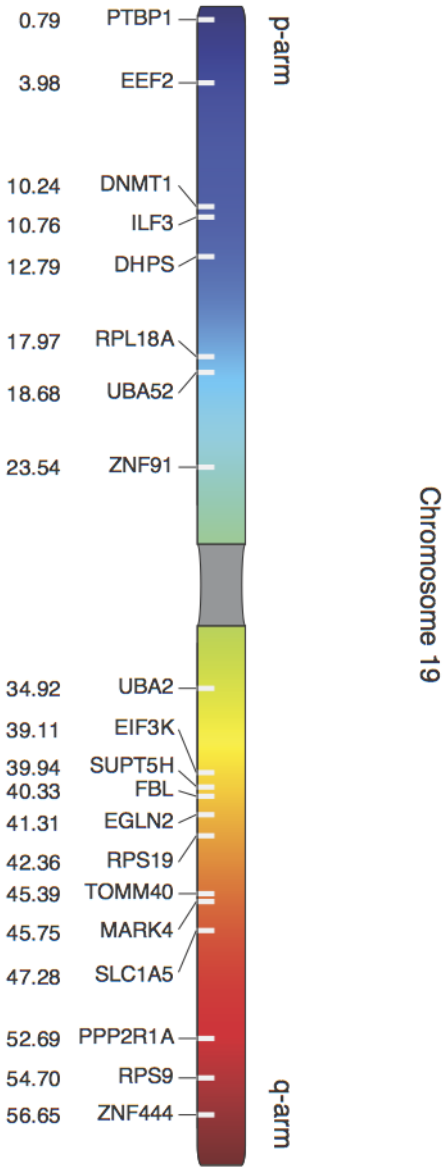

Figure 2.2: We selected many potential genes based on high abundance of their mRNA (determined by RNA-seq and RT-PCR), then narrowing our focus to a subset of those with high intron spot frequencies as measured by RNA FISH. Ultimately, the set of genes we picked had spot frequencies ranging between 10% and 80% of chromosomes exhibiting a spot; see Fig.2.13 for some representative numbers. The smallest genetic distance between loci was 0.36 megabases (between TOMM40 and MARK4)

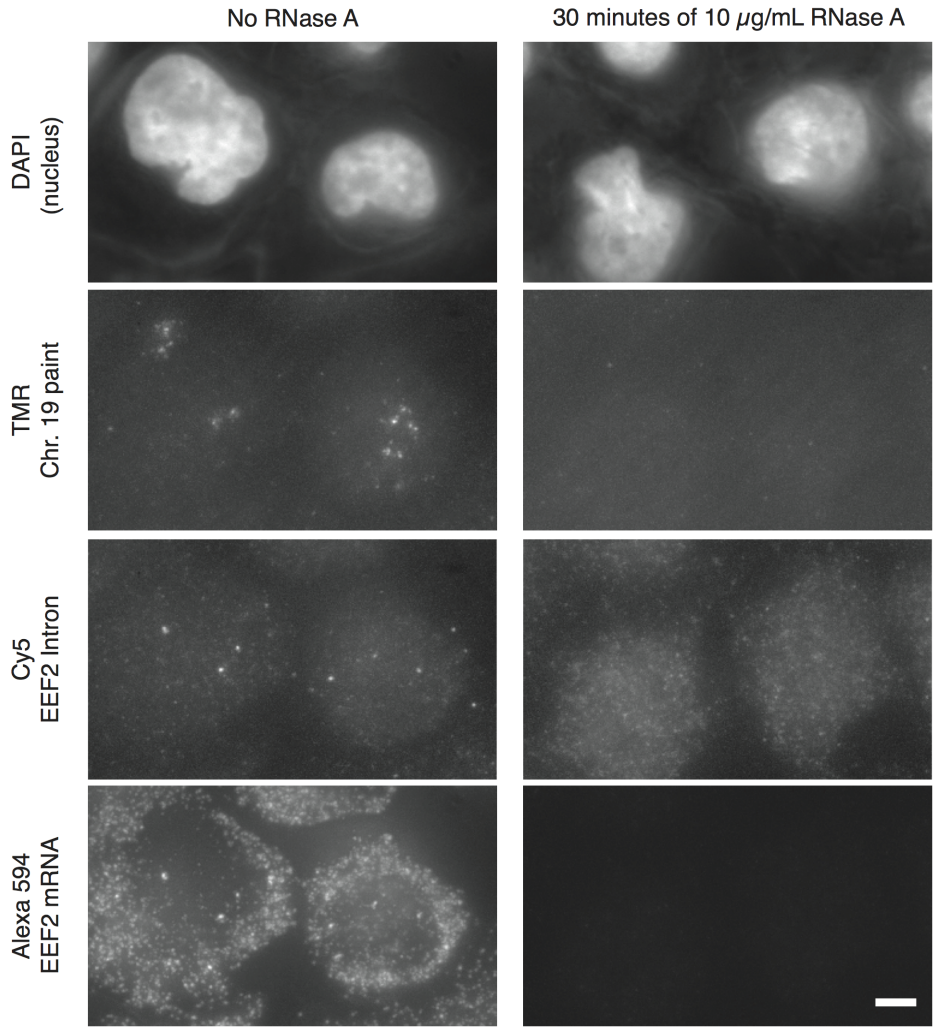

Figure 2.3: We exposed HeLa cells to 10μg/mL of RNase A for 30 minutes after fixation and before hybridization (right panels), with control cells exposed to the same procedure but without the addition of RNase. The top row contains images of the DAPI nuclear stain. The second row contains images in which we labeled the introns of all genes except EEF2, thus painting active genes in chromosome 19, and in the third row we labeled the introns of EEF2. In the fourth row, we labeled EEF2 mRNA. All images are maximum intensity projections of a z-stack of fluorescence images. The scale bar is 5μm long and applies to all images depicted. These results show that RNase A eliminates all the FISH signals we observed in the cells. This shows that the signals we are detecting are due to binding to RNA and not DNA, showing that our probes are detecting RNA from actively transcribing genes and not the DNA of gene itself.

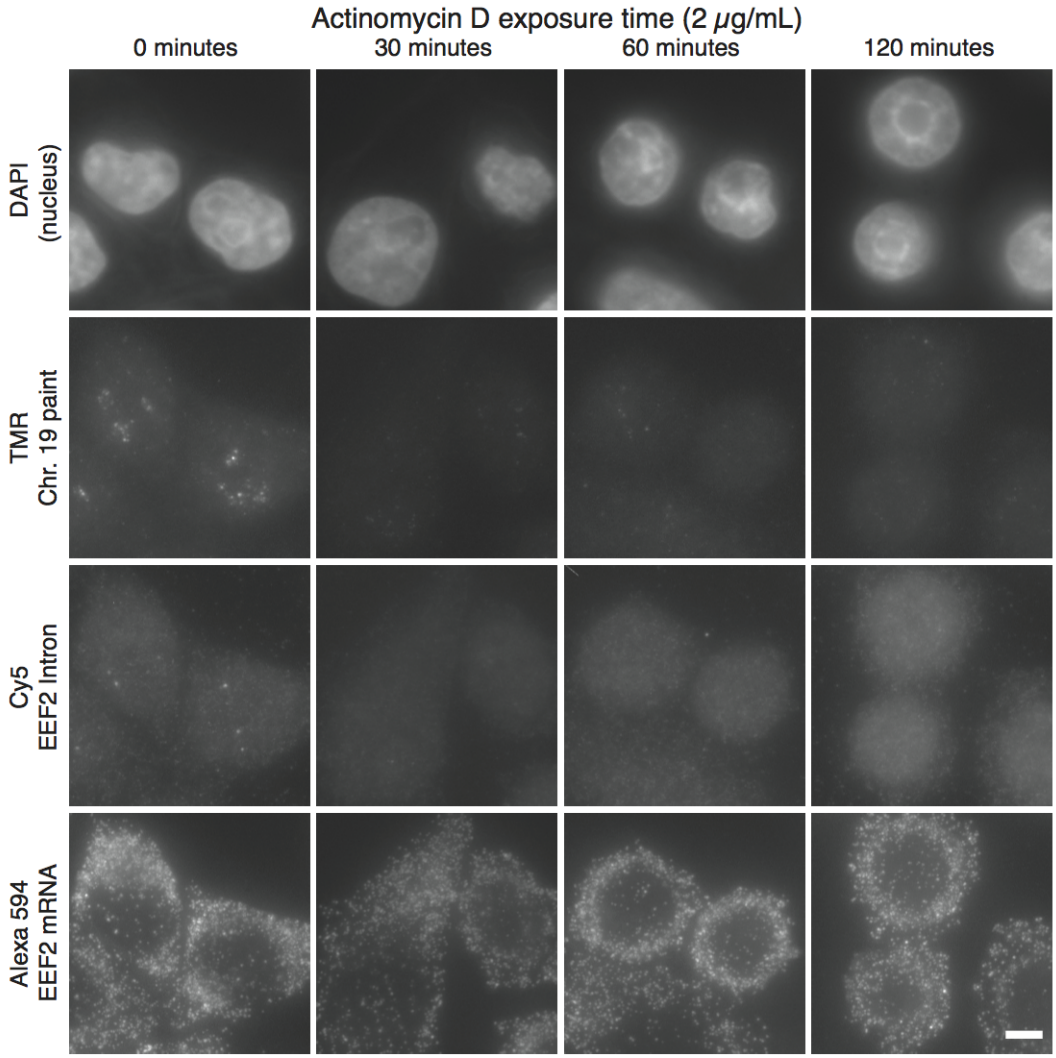

Figure 2.4: We exposed HeLa cells to 2μg/mL of actinomycin D for varying amounts of time as indicated. The top row contains images of the DAPI nuclear stain. The second row contains images in which we labeled the introns of all genes except EEF2, thus painting active genes in chromosome 19, and in the third row we labeled the introns of EEF2. In the fourth row, we labeled EEF2 mRNA. All images are maximum intensity projections of a z-stack of fluorescence images. The scale bar is 5μm long and applies to all images depicted. We found that virtually all the intronic RNA disappeared or greatly diminished in intensity after 30 minutes of Actinomycin D exposure. In the case of EEF2 intron, we found absolutely no spots; in the case of the chromosome paint, we did occasionally see dim spots even at later time points, although they were considerably dimmer. These may be residual RNA still attached to RNA polymerases stalled by Actinomycin D. We observed mature mRNA at all time points, indicating that the cells were still alive and that the treatment did not affect the RNA itself. Altogether, our results show that intronic RNA degrades rapidly; thus, the presence of an intronic RNA spot indicates that the targeted gene is transcriptionally active.

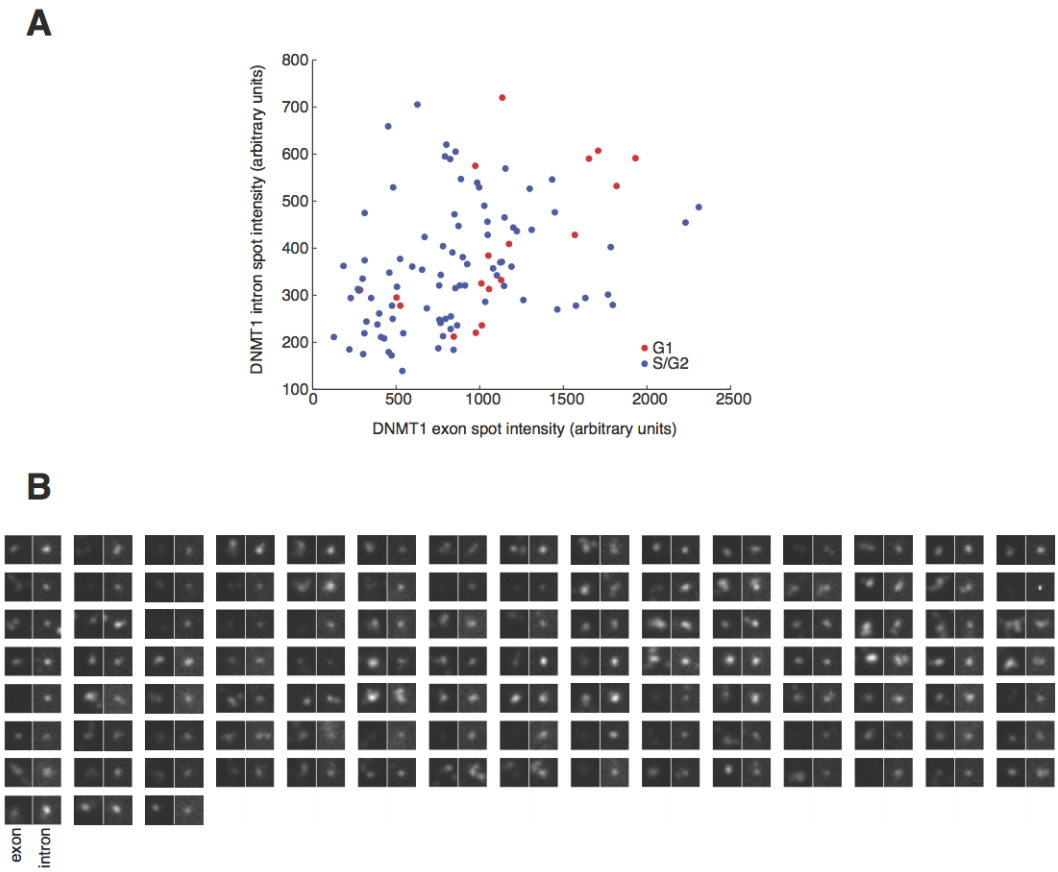

Figure 2.5: A. We show intron vs. exon spot intensity, defined as maximum pixel intensity of the spot minus the nearby background pixel intensity. Blue spots correspond to S/G2 phase cells (as scored by Cyclin A2 mRNA levels); red spots correspond to G1 cells. We analyzed a total of 105 spots. The correlation between intron and exon intensity is 0.378. B. Images depicting all the exon and intron spot pairs we analyzed (exon on left, intron on right). The slight shift between the exon and intron spot results from slight registration shifts in the images between fluorescence channels. We found that only 2 out of 105 intron spots displayed a lack of corresponding exon spot. These results further establish that the presence of an intron spot corresponds to active transcription of the gene. We almost never observe an intron spot without any corresponding exon spot. Moreover, when there is no visible intron spot, we never find a bright transcription site in the exon channel (data not shown). Also, these results provide further evidence that the intron spots are truly located at the site of transcription because the intron spots strongly colocalize with bright exon spots that researchers have shown to represent nascent transcripts emanating from the site of transcription (see Levsky and Singer Science 2002, Vargas et al. PNAS 2005, [10, 60]).

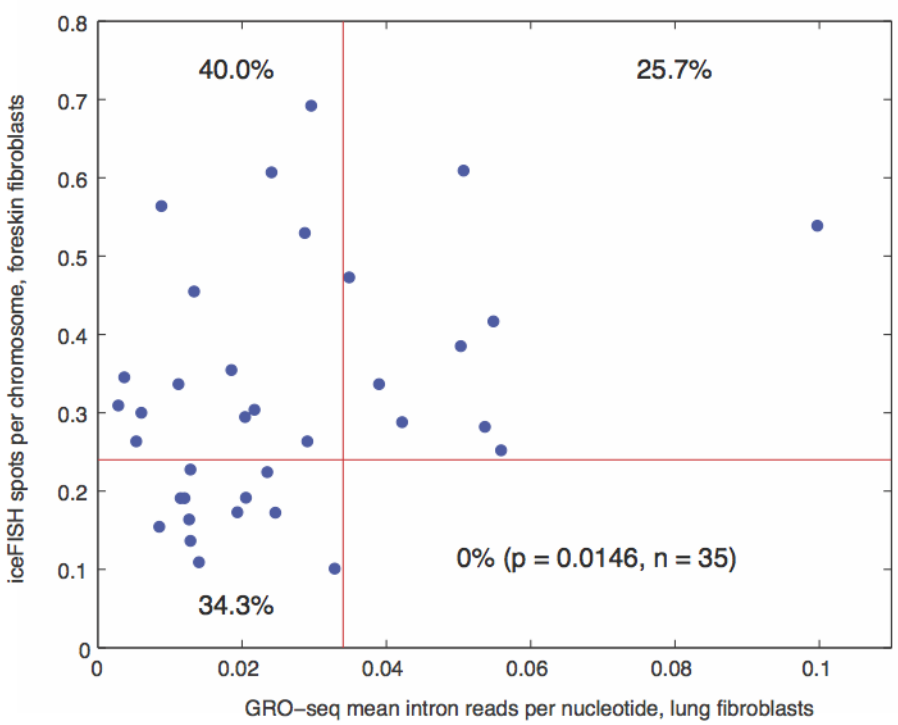

Figure 2.6: In order to compute the GRO-seq mean intron reads[11], we measured the average number of reads within the first 3000 intronic bases in the transcript (which was roughly the same region that we probed by iceFISH in our study). We computed the frequency as discussed at length in the main text. We also included data from a further 15 genes on chromosome 19 not studied in the main paper to provide better sampling in the high GRO-seq regime. Our observation is that above a certain number of GRO-seq reads, the iceFISH spot frequency was always relatively high. We also point out that there is a large set of points that we have not shown that are at the origin of the graph: they have no GRO-seq reads and generally do not show intronic FISH spots. (We have not included these data as we were not interested in intron probes that do not show spots and so did not systematically analyze such probes.) We believe there are many reasons that the correlation we observe is rather noisy: 1. these two sets of data arise from different cell lines and represent completely different treatments and procedures; 2. comparing iceFISH signals from gene to gene is not necessarily valid (nor are those from GRO-seq); and 3. our data are solely from G1 cells, whereas GRO-seq is from a mixed population. As such, we believe that the fact that these data show that high GRO-seq signals correspond to high iceFISH spot frequency is a valid means by which to measure transcriptional activity.

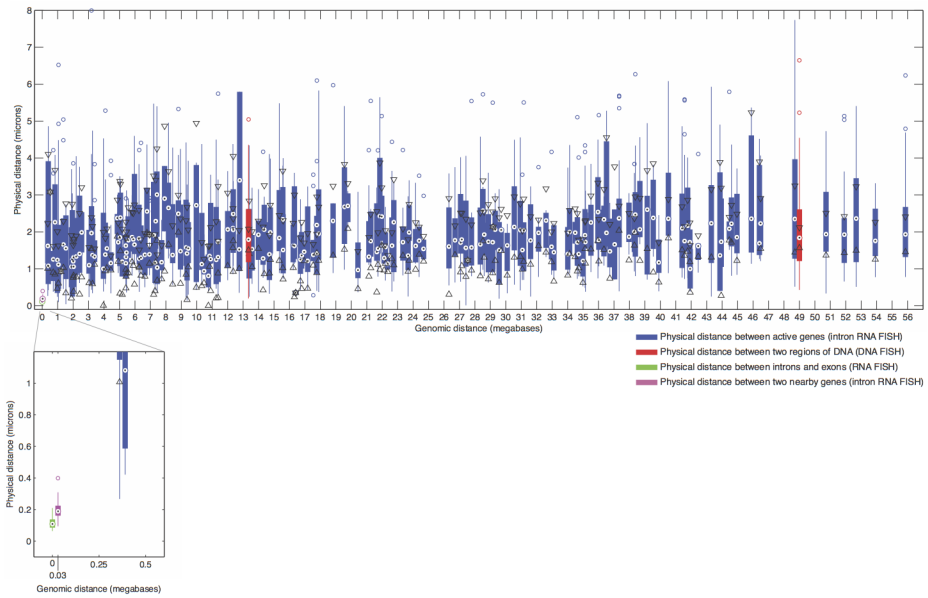

Figure 2.7: Each box plot depicts a statistical analysis of the distances between a particular pair of actively transcribing genes, and we placed the box plot at a location on the x-axis representing the genomic distance between the pair of genes. The spot in the center of the box corresponds to the median distance, the box itself corresponds to 25th and 75th percentiles, and the whiskers reflect the range of the data (with the open circles reflecting data points deemed outliers). The open triangles provide comparison intervals: two medians are different at a 5% significance level if the intervals represented by these triangles do not overlap. The red box plots correspond to distances measured by DNA FISH (the experimental details of which are described in the methods). There is no particular difference between the DNA FISH results and the distances between transcriptionally active loci at the length scales we examined, although our results indicate that genetically proximal transcriptionally active loci (<0.5 megabases) are more spread out physically than pure DNA FISH measurements that do not distinguish between transcriptional status (Mateos-Langerak et al. PNAS 106:3812-3817. 2009). However, a simple model demonstrates that our data are compatible with these previous results (see Supplementary discussion). The inset shows the physical distance between loci separated by 30 kilobases (EIF3K and ACTN4; magenta) and introns and exons of the same gene (DNMT1; green). These data show that intron spots are very close to the site of transcription as measured by exonic probes (we believe the measurements are almost to within our spatial discrimination limit), and that genetically very proximal genes do indeed come very close to each other, showing that the spread out characteristics that we observe are not just an artifact of our method.

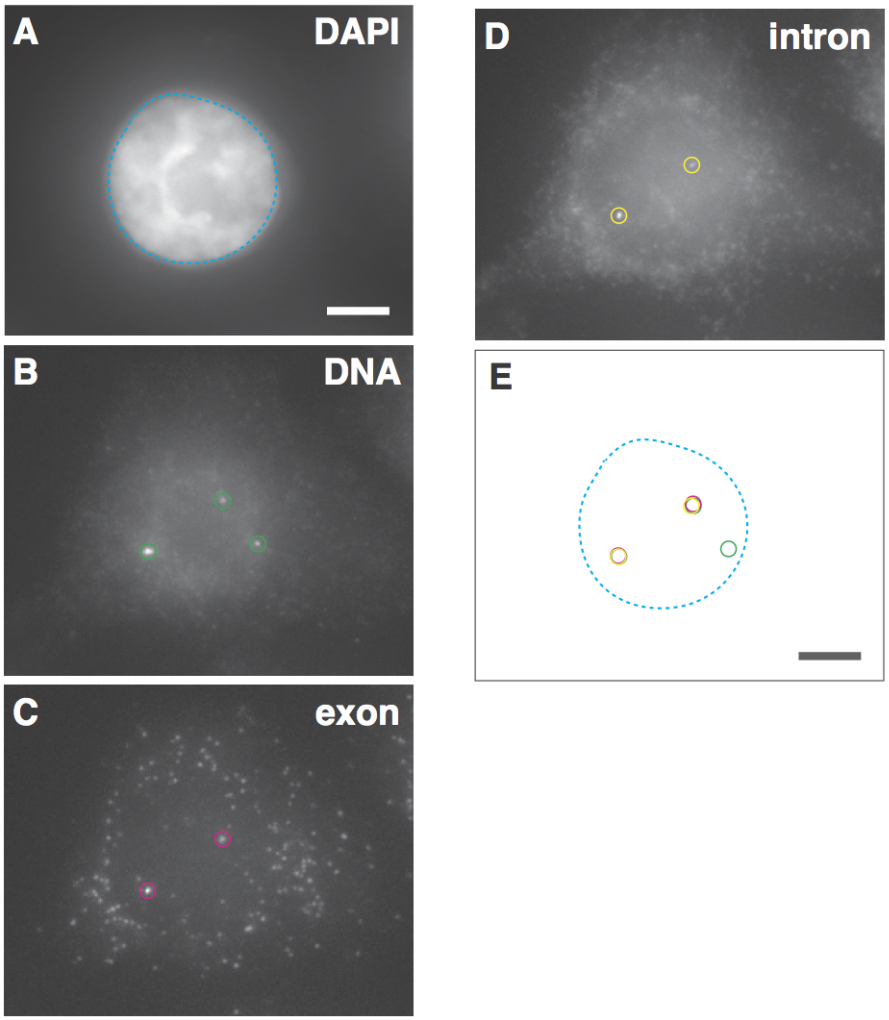

Figure 2.8: As we describe in the methods, we performed a protocol in which we combined DNA and RNA FISH, here using BAC probes targeting the DNA, RNA FISH probes targeting the exons of SLC1A5, and RNA FISH probes targeting the introns of SLC1A5 (b-d). We found that all the exon and intron signals colocalized with the DNA FISH signals (e) to within our detection limits. These results show that the intron RNA FISH signals we observed indeed remain at the site of transcription. Note that we observe three copies of the gene via DNA FISH, but only two of the three copies of the gene are transcriptionally active. This shows that our RNA FISH probes are not inadvertently binding to DNA. Another proof that our probes are not merely targeting DNA is the fact that both the exon and intron signals are brighter than what one would expect from a single RNA molecule. If the probes were bound to the gene's DNA, one would only see fluorescence intensity equivalent to that of a single RNA molecule. The scale bar is 5μm long.

(exclusion of these commingled chromosomes did not alter our results; Fig. 2.10). On average, we found 6 2 expressing genes (out of the 20 labeled) per chromosome. We found that using more probes did not change spot detection efficiency (see Methods), nor does pseudocoloring incur a significant rate of spot misidentification (Fig. 2.11). We ensured that the cells we analyzed were in the G0/G1 stage of the cell cycle by co-labeling Cyclin A2 mRNA and examining only cells with low levels of Cyclin A2, which is abundant during the S, G2, and M phases of the cell cycle (Fig. 2.12, and see [17].

By grouping actively transcribing genes into territories corresponding to each chromosome, we constructed transcriptional profiles showing which of our 20 genes are on and off per chromosome. Taking transcriptional profiles of individual chromosomes in several cells let us 1. directly compare the transcriptional activity of normal and rearranged chromosomes within the same cell (Fig 2.13), and 2. find large-scale interactions between pairs of genes on the same chromosome by measuring correlations or anti-correlations in transcriptional activity (Fig. 2.20). In both cases, our results point to chromosome-specific mechanisms controlling transcription.

2.2.2. Translocations can cause chromosome-wide transcriptional changes

Researchers largely believe that cells regulate transcription of a gene via chromosome-extrinsic trans factors (such as transcription factors) or via local cis factors on the DNA (i.e., sequence elements typically within 1 megabase of the gene itself). Our method enables us to examine the possibility that chromosome-specific mechanisms may also regulate transcription. Translocations, in which large segments of different chromosomes are joined together, provide a means to test this hypothesis: while they disrupt the large-scale structure of a chromosome, the cells trans environment and local cis DNA regulatory code remains unchanged for most genes on the translocated chromosome. Thus, any differences in transcription of a gene(s) between the normal and translocated chromosomes would show that chromosomes possess non-local cis regulatory mechanisms that can be disrupted by translocation.

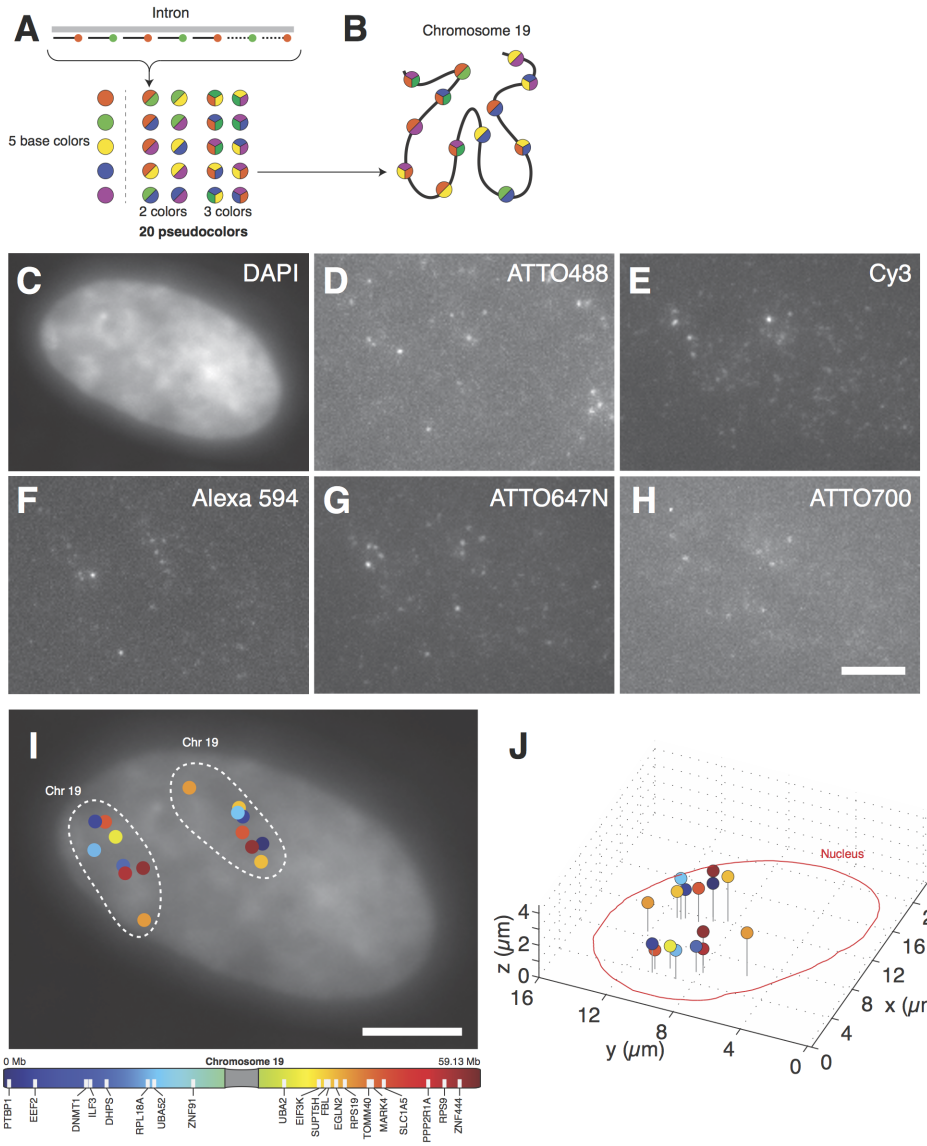

Figure 2.9: a. Our scheme involved labeling each intron with oligonucleotide probes alternately labeled with either two or three different fluorophores, leading to a total of 20 unique pseudocolors. b. Once we identified the pseudocolored transcription sites, we could trace out the chromosomes three- dimensional configuration. c-h. Images for each fluorescence channel from the nucleus (labeled with DAPI in (c)) that we stained with probes labeled with the scheme depicted in (a). Along with probes targeting the introns, we also included probes targeting Cyclin A2 mRNA to determine position in cell cycle and SUZ12 mRNA as a fiducial marker. All images are maximum z-projections of a three-dimensional z-stack. All scale bars are 5μm long. i-j. Computational identification of chromosome 19 gene positions. 3D depiction has axes labeled in μm and the nuclear outline from the DAPI signal is outlined in red.

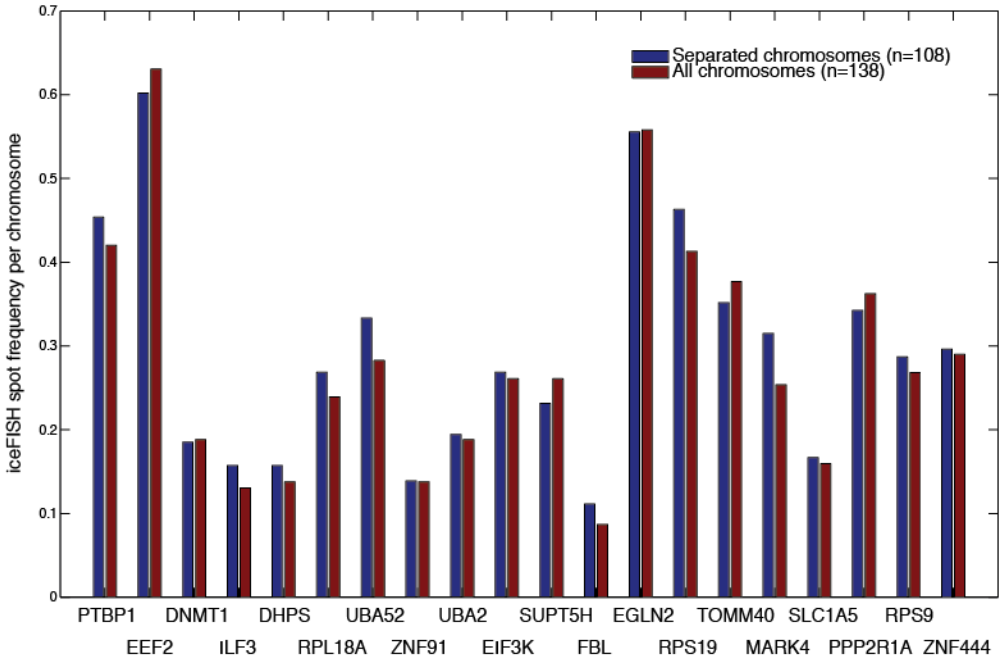

Figure 2.10: In our analysis, we had to exclude chromosome in which both copies of chromosome 19 were too close for us to separate. In order to check whether this exclusion introduced any bias into our measurements, we measured iceFISH spot frequency for all chromosomes (including those which were excluded from transcriptional profiles). We found that the frequency was essentially identical to what we obtained from analyzing just the non-overlapping chromosomes, showing that our exclusion of overlapping chromosomes did not bias our results.

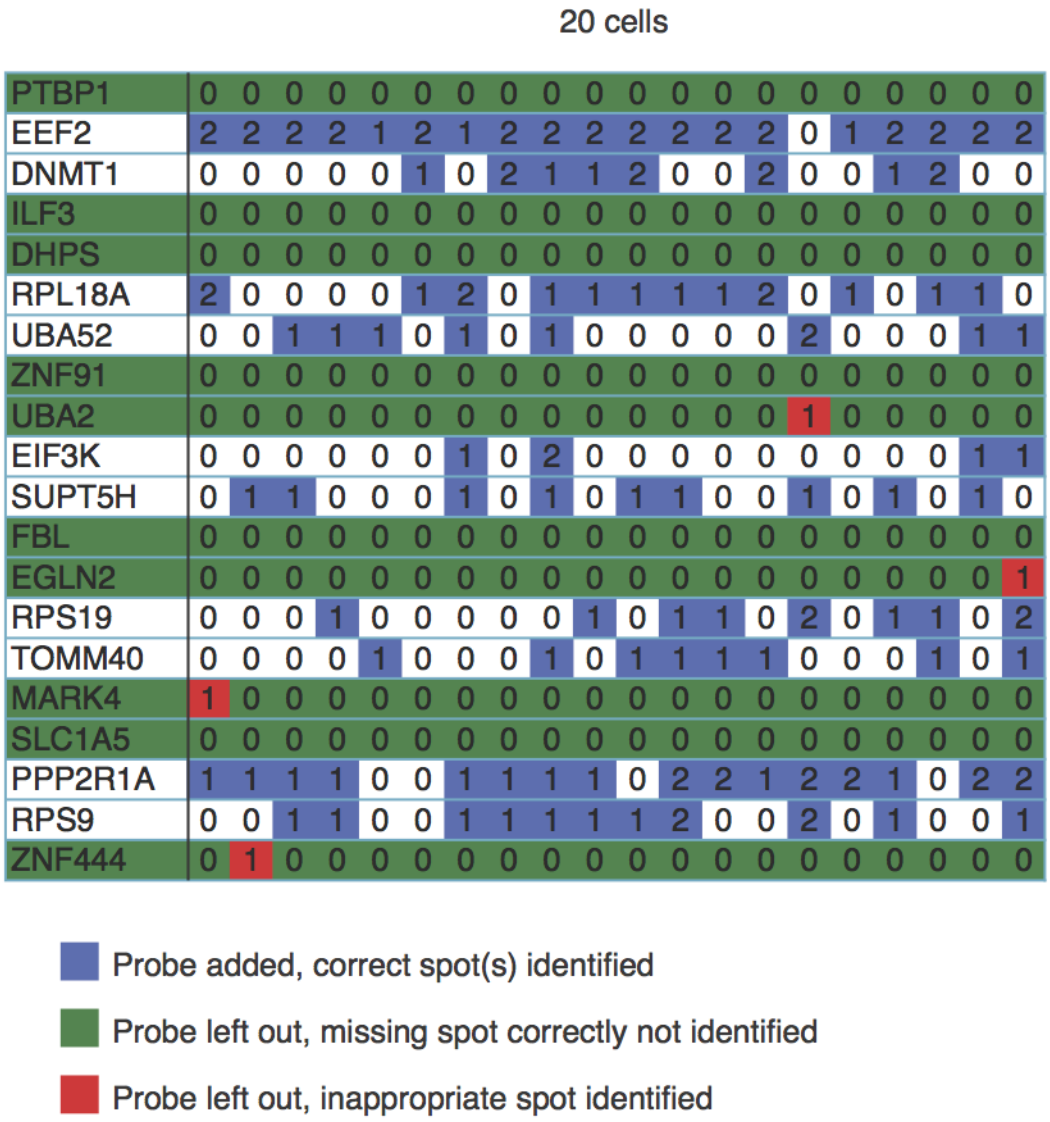

Figure 2.11: In order to assess the degree to which our intron spot color-coding scheme resulted in false positives, we left out 10 of the 20 probes (labeled in green) and ran the resultant images through our normal spot identification pipeline. In blue, we have indicated all the spots that we found that we ended up assigning to genes whose probes had been added to the hybridization (i.e., correct identifications). In red are spots we misidentified in the sense that they correspond to genes that we hadn't added to the hybridization. We found that misidentified spots were relatively rare, with roughly 97% of spots we identified being assigned to genes that we had actually targeted in our hybridization. In order to minimize bias, we had another person in the lab randomly select 10 genes to leave out of the hybridization.

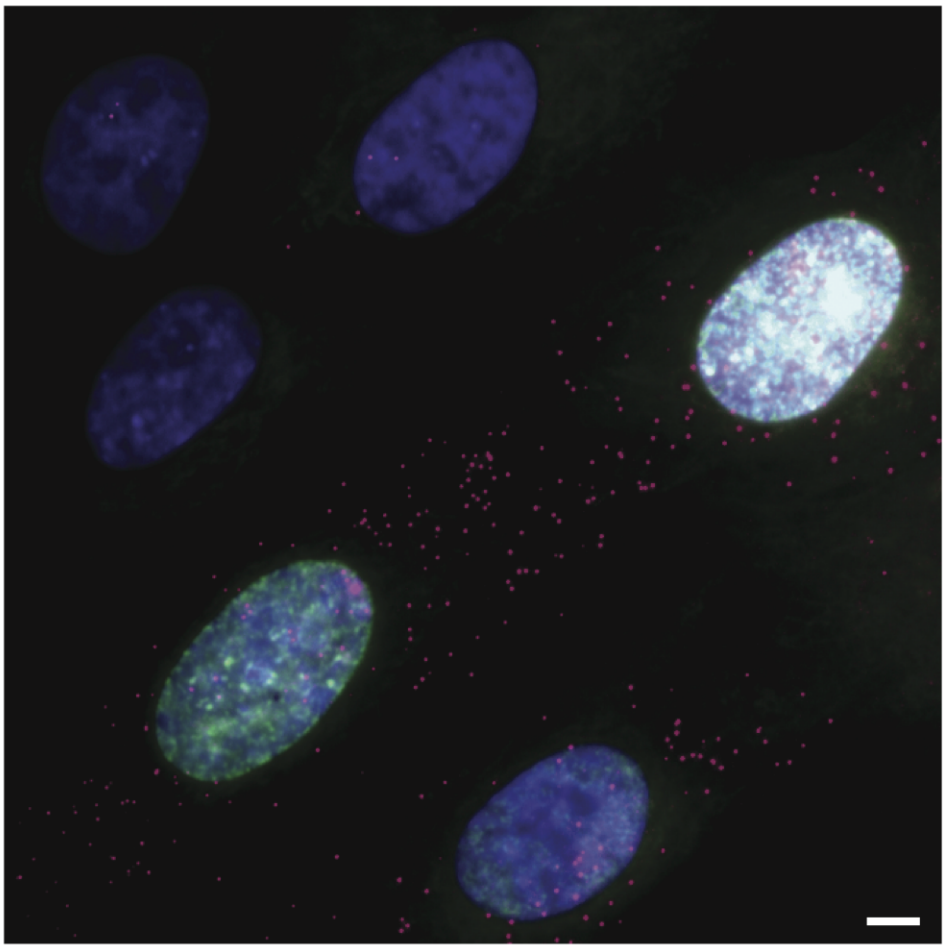

Figure 2.12: In order to isolate cells in G0/G1 for analysis, we performed RNA FISH targeting Cyclin A2 (magenta), which Eward et al. [17] have shown to be present during S, G2, and M phases. We observed highly variegated expression, with some cells having having high levels of expression and some having very few mRNA. In order to demonstrate that the high expressing cells were indeed in S-M phase, we incubated the cells with Click-iT EdU 10μM for 5 minutes before fixation (green); the Click-iT EdU reagent incorporates into polymerizing DNA and produces a signal in cells undergoing DNA replication. We found that every cell display- ing Click-iT EdU signal had high levels of Cyclin A2 mRNA, showing that Cyclin A2 mRNA provides a strong marker for cell-cycle. Sometimes a cell would have high levels of Cyclin A2 but would not be undergoing DNA replication; these cells are in G2, as we observed several double intron spots in these cells, indicating that those cells had already duplicated their DNA. In thus study, we were primarily interested in cells that had not undergone replication, so we selected cells with low levels of Cyclin A2. Note that we could not use the Click-iT EdU kit in combination with iceFISH because we found that incubating cells with the Click-iT reagent resulted in an abolition of transcription. We stained the nuclei with DAPI (purple); the scale bar is 5μm long.

HeLa cells provide a test case for such a study. This widely-used cervical cancer cell line contains two intact copies of chromosome 19 and one copy that is split into two pieces fused to parts of other chromosomes[50]: one, denoted t(6;19), consists of the first 17-20 megabases of chromosome 19 fused to part of chromosome 6, and the other, denoted t(13;19) consists of the remaining 40-43 megabases of chromosome 19 translocated onto a portion of chromosome 13 (Fig. 2.13A; confirmed by G-band karyotyping and DNA FISH, Fig. 2.14). We observed this pattern of genetic rearrangements in our iceFISH data (Fig. 2.13B). We found that most genes on t(13;19) were up to 5 fold more transcriptionally active than those on the normal copies of chromosome 19 (Fig. 2.13C, replicate in Fig. 2.15). This finding is consistent with the existence of chromosome-specific transcriptional regulation that the translocation has disrupted in some way. Intron spot intensities were roughly the same on all the chromosomes we examined (Fig. 2.16), suggesting that transcriptional hyperactivation results from an increased probability of a gene being active rather than an increased rate of transcription when the gene is active (see Supplementary discussion).

We then asked whether the portion of chromosome 13 on t(13;19) also displays heightened transcriptional activity. We found that the transcriptional frequency of two genes from chromosome 13 (DIAPH3 and MTZ1) was roughly 2 fold higher on t(13;19) than on the normal copies of chromosome 13 (Fig. 2.13C, Fig. 2.15), suggesting that this translocation resulted in hyperactivation of all genes on t(13;19) irrespective of origin. Meanwhile, transcription of the chromosome 19 genes on t(6;19) was similar to the normal copies (Fig. 2.13C), suggesting that translocations do not necessarily lead to transcriptional changes. We note also that per-chromosome differences in transcription are difficult to observe using bulk assays that average expression from all chromosomes, which may explain why reports of such effects are not widespread.

We explored whether the hyperactivation of t(13;19) was associated with differences in the chromosomes spatial configuration, examining both the relationship between genomic and physical distance as well as the chromosomes positioning within the nucleus. Previous re-

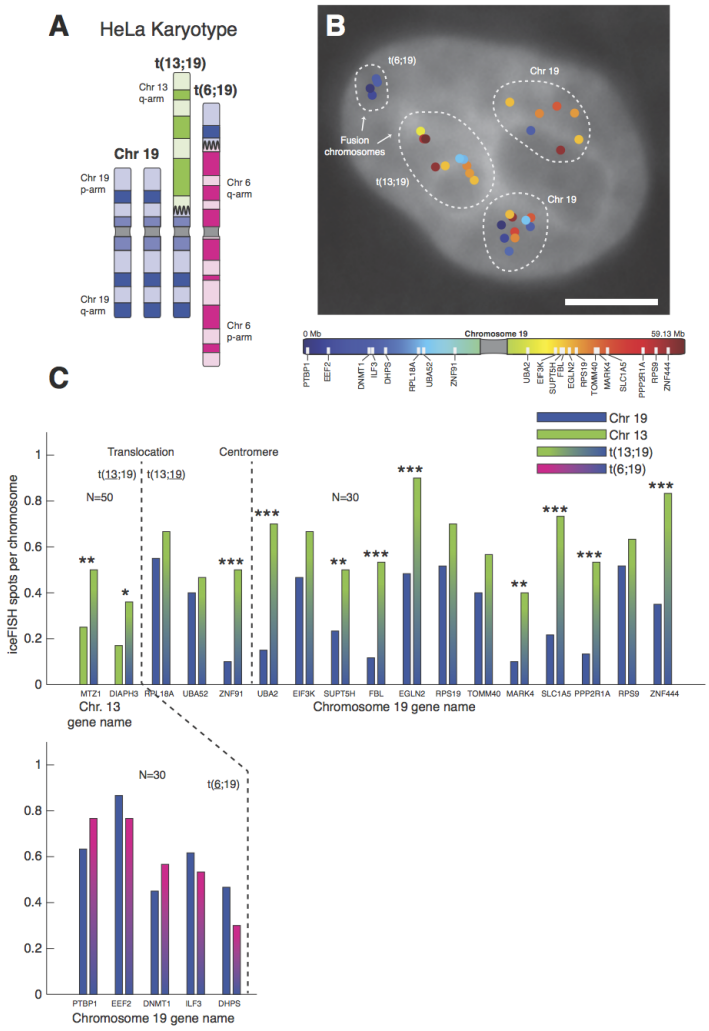

Figure 2.13: a. Schematic showing that our HeLa cells contain two intact copies of chromosome 19. b. Computational identification of actively transcribing genes on chromosome 19 revealed all the 4 chromosomes containing portions of chromosome 19, including the two intact copies and the two translocated pieces. The scale bar is 5μm long. c. Comparison of the transcriptional activity of the genes on chromosome 19 (as measured by frequency of observing a transcription site per chromosome) on the various translocated fragments of chromosome 19 as well as the intact copies of chromosome 19. For the expression of the two genes on chromosome 13 (MTZ1 and DIAPH3), we measured spot frequency as described in Fig. 2.17. We denote p-values for the difference in frequency (as compared to the null hypothesis of no difference) by *** for p <0.001, ** for p <0.01, * for p <0.05.

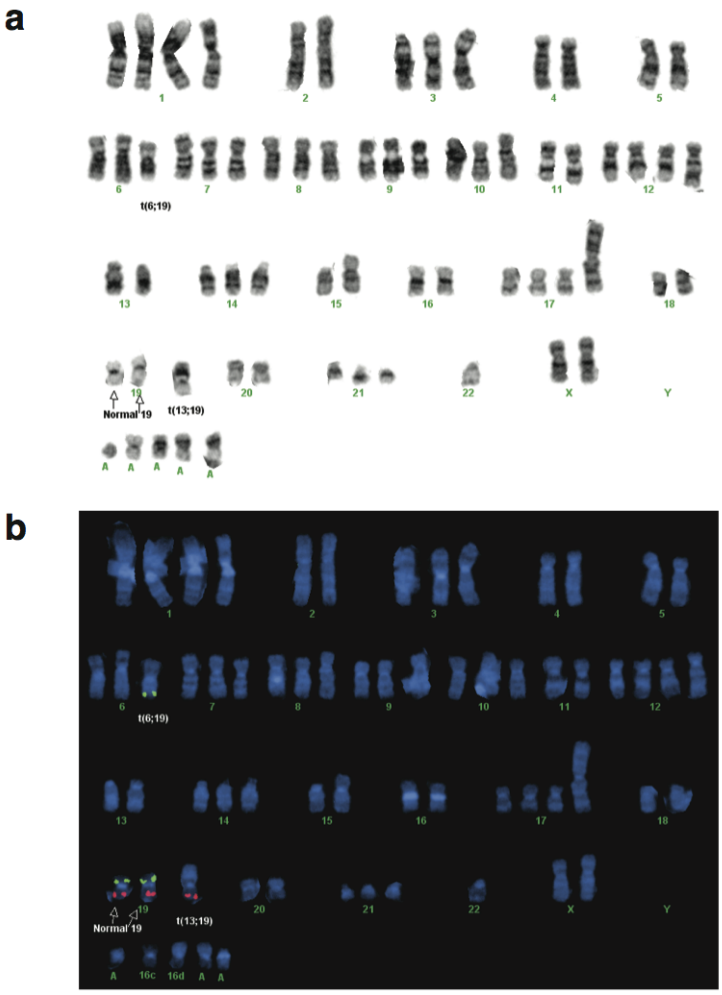

Figure 2.14: A. We performed G-band analysis of a metaphase spread of our chromosomes to identify both the numbers of chromosomes in the cells as well as identify potential translocations. The numbers (and X and Y) correspond to the identities of the intact chromosomes. The specific translocations we were interested in for this study are fusions of portions of chromosome 19 to chromosome 13 and 6, denoted t(13;19) and t(6;19), respectively. B. We wanted to further verify the translocations of chromosome 19 via DNA FISH, which we performed upon the same chromosomes for the G-band analysis in A. We probed the chromosomes with probes against the p-arm of chromosome 19 (green) and the q-arm of chromosome 19 (red). This confirmed the results of our G-band analysis.

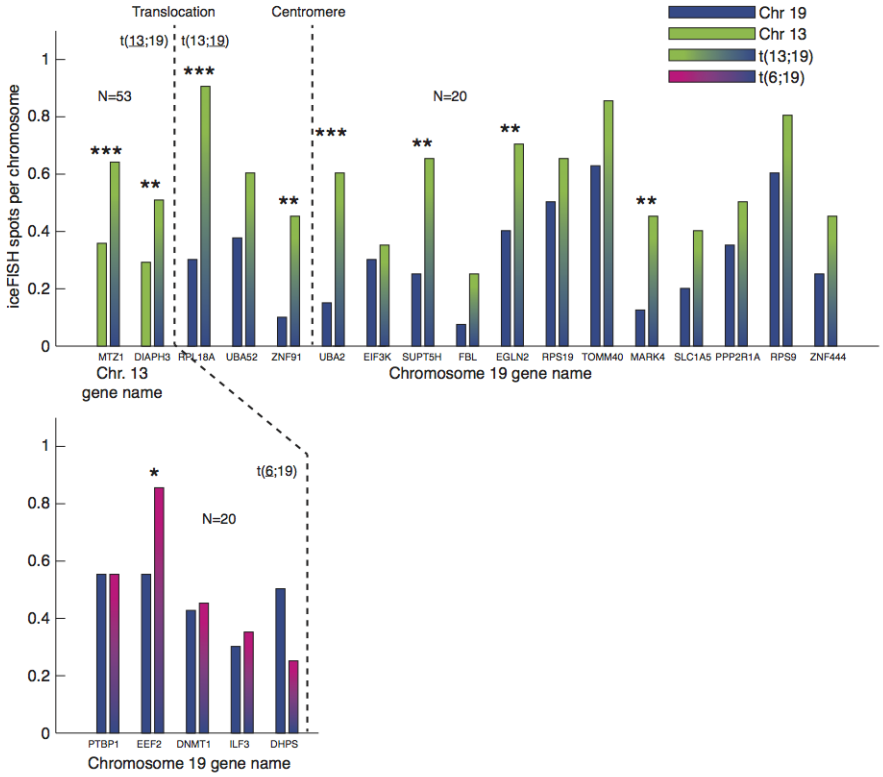

Figure 2.15: Here we present data from a biological replicate of the data presented in Fig. 2.13. We performed the exact same analysis as described in the legend of that figure. In this case, we analyzed 20 cells with the full set of 20 chromosome 19 probes and 53 cells in which we analyzed expression of MTZ1 and DIAPH3 on chromosome 13.

ports often found genes "looped" out towards the periphery of the chromosome territory when active[55] and that active regions of DNA are more physically spread out than inactive ones[23] and these domains may be physically separated[47], raising the possibility that the hyperactivation of t(13;19) may correlate with differences in inter-gene spacing. We compared genomic and physical distance separating pairs of actively transcribing loci (Fig. 2.7), finding that the physical distances between genetically proximal (<5 and especially <1 megabases) genes are considerably larger than those obtained by pure DNA FISH[39, 77] even for "active" DNA[51]. These observations are consistent with a model in which DNA fluctuates between compact and extended configurations depending on transcriptional status[55] (see Supplementary discussion); however, the physical inter-gene spacing on t(13;19) was virtually identical to that of the intact chromosome 19 (Fig. 2.18).

Other reports[73] have found that smaller, gene dense chromosomes are often located at the center of the nucleus, whereas larger chromosomes tend towards the nuclear periphery. Translocations thus may alter the affected chromosomes distance from the nuclear periphery, and one report has suggested that translocation-induced changes in nuclear position may be partly responsible for changes in transcriptional activity[33]. In HeLa cells, we saw that the t(13;19) translocated chromosome was slightly closer to the periphery than the intact chromosome 19s, although this difference was not statistically significant (Fig. 2.19). We did see a significant shift in t(6;19) towards the nuclear periphery, but that chromosome did not show any transcriptional differences. We also found no dependence between transcriptional activity of any of the genes we examined and distance of the chromosome to the nuclear periphery (Fig 2.19). Together, these analyses suggest that spatial configuration may not be responsible for hyperactivation per se, but it is possible that other spatial aspects of chromosomes we are unable to measure cause these effects.

2.2.3. Transcriptional profiling reveals long range cis transcriptional interactions

Chromosome transcriptional profiling demonstrated the potential for differences in transcriptional activity between chromosomes. We next looked for evidence of interactions

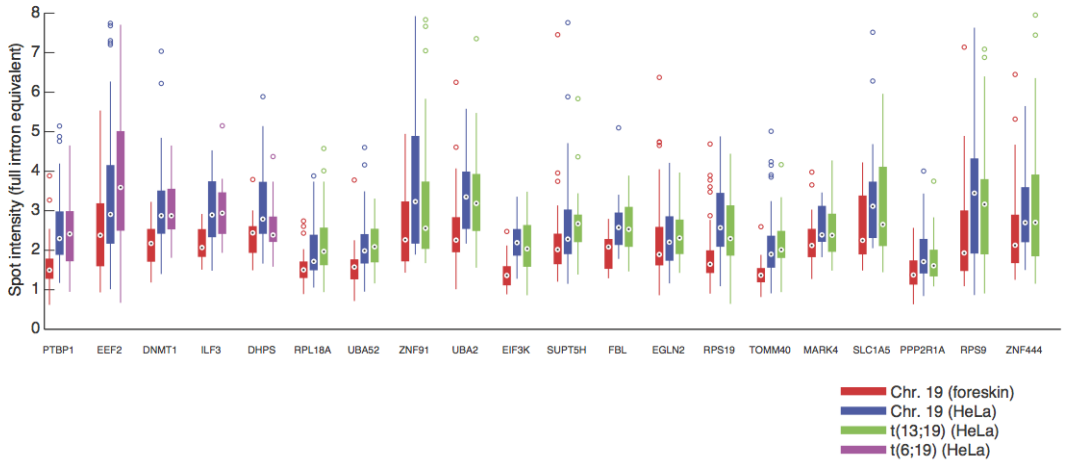

Figure 2.16: We computed spot intensity by finding the maximum pixel intensity in the center of a 3x3 pixel box around the center of the spot and subtracting the median local back- ground pixel intensity around the spot. We then normalized this intensity for each color to number of probes on a cell-by-cell basis by calibrating to the intensities of the SUZ12 mRNA signals, which had 13 oligonucleotide probes in each fluorescence channel. A value of 1 for spot intensity here indicates and intensity equivalent to 1 complete set of intron probes bound. The box plots show spot intensity for each gene on Chr. 19 in foreskin fibroblasts and HeLa cells, as well as the t(6;19) and t(13;19) derivative chromosomes in HeLa cells. Center white spot represents the median, edges correspond to 25th and 75th percentile, and whiskers extend to extrema, with outliers plotted individually as circles. We found that the spot intensities varied from gene to gene, with most of the genes showing a median between 1 and 4 full intron equivalents. The foreskin spot intensities appeared to be somewhat lower than in HeLa cells. In HeLa cells, we found that spot intensities were essentially the same when comparing spots on the intact chromosome 19s and the translocated t(6;19) and t(13;19). This result implies that the changes in transcription of the chromosome 19 genes on t(13;19) arise via changes in transcriptional burst frequency but not changes in burst size. It also highlights the fact that the changes in spot frequency that we observe on t(13;19) are not likely to be the result of a relatively small increase in spot intensity leading to an increase in the number of spots above a putative detection threshold. Were that the case, then we would see many more low intensity spots on t(13;19), which would lower the mean intensity, which is not what we observe.

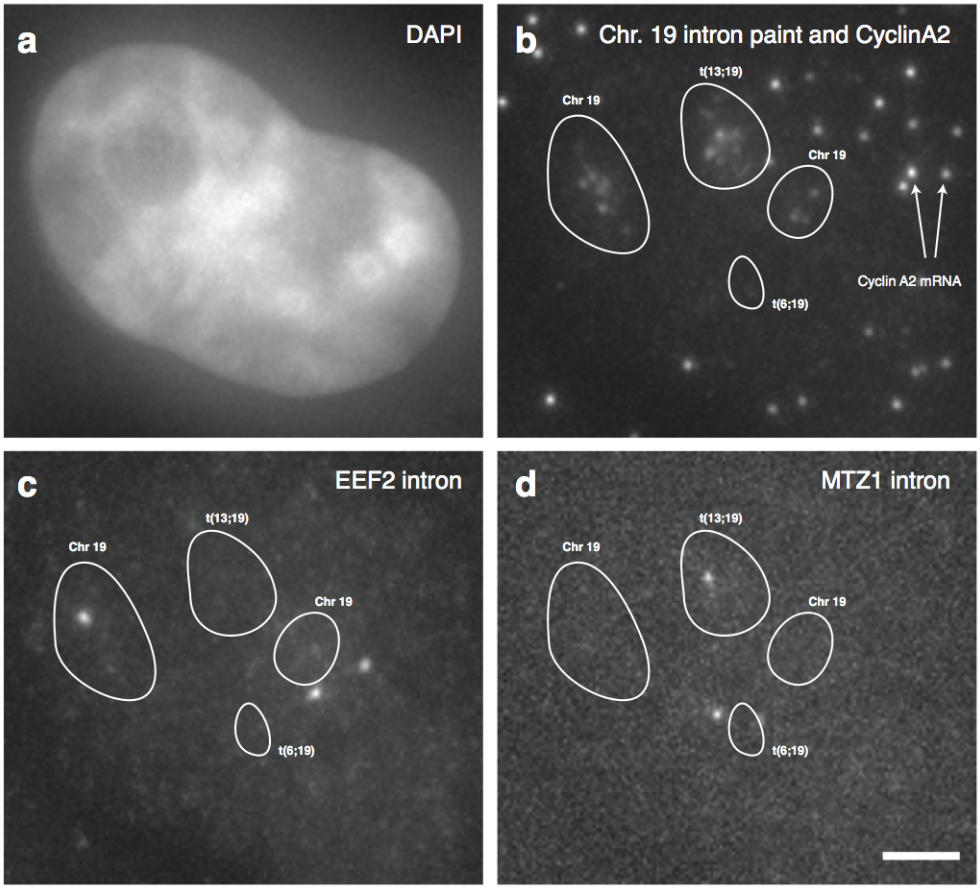

Figure 2.17: To measure the frequency of transcription of genes on the portion of chromosome 13 fused to chromosome 19 (denoted der(19)t(13;19)) in HeLa cells, we used a strategy in which we painted chromosome 19 with probes in one color (Cy3, panel B) to identify the chromosome territory. We also labeled Cyclin A2 mRNA in Cy3 to ensure analyzed cells were in G0/G1 phase of the cell cycle. We also labeled probes targeting EEF2s intron in Atto 647N (C), which is located on the portion of chromosome 19 that is part of der(6)t(6;19) and is not on der(19)t(13;19), thereby helping to identify those portions of chromosome 19 that are not fused to chromosome 13. Finally, we labeled the intron of the gene of interest on chromosome 13 (either MTZ1, in panel D, or DIAPH3, not depicted) with Alexa 594, and if the intron spot appeared in the chromosome territory identified by the chromosome 19 paint, we assigned it to der(19)t(13;19). We assigned spots for these chromosome 13 genes that appeared away from chromosome 19 as coming from one of the two normal copies of chromosome 13 in our HeLa cells. The scale bar is 5μm long.

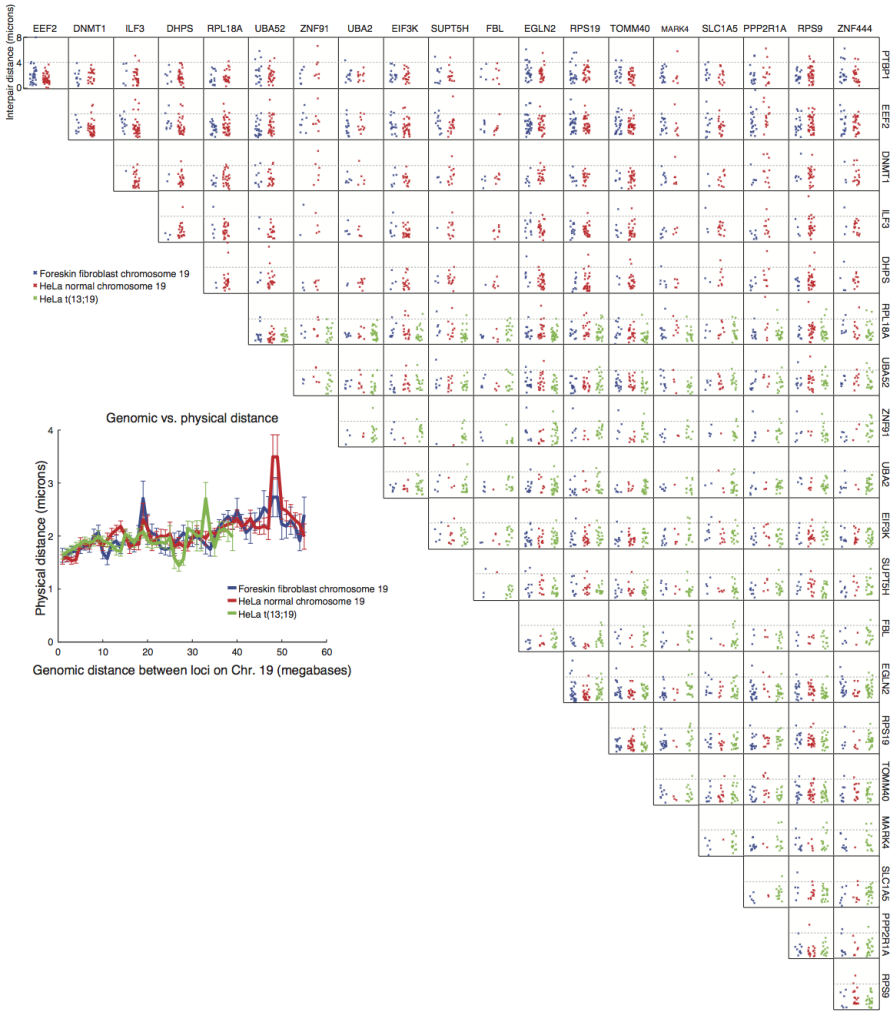

Figure 2.18: We compared the physical distances between all pairs of actively transcribing genes on chromosome 19 or the t(13;19) derivative chromosome from human foreskin fibroblasts and HeLa cells. Our comparison was between all loci on all chromosomes from human foreskin fibroblasts (blue markers), normal chromosome 19s from HeLa cells (red markers), and the t(13;19) derivative chromosome also present in HeLa cells (green markers; only for those genes that are located on this chromo- some). We did not observe any statistically significant differences in average distances for any gene pair when comparing any of these conditions. The inset plot shows a comparison between genomic and physical distance for all active loci, comparing these same classes of chromosomes. The plot contains a rolling average with a 1 megabase window, with the error bars corresponding to the standard error of the mean. We observed no significant differences in the relationship between genomic distance and physical distance between active loci, suggesting that the increased expression on the t(13;19) derivative chromosome does not correlate with an overall increase in the physical distance between active loci.

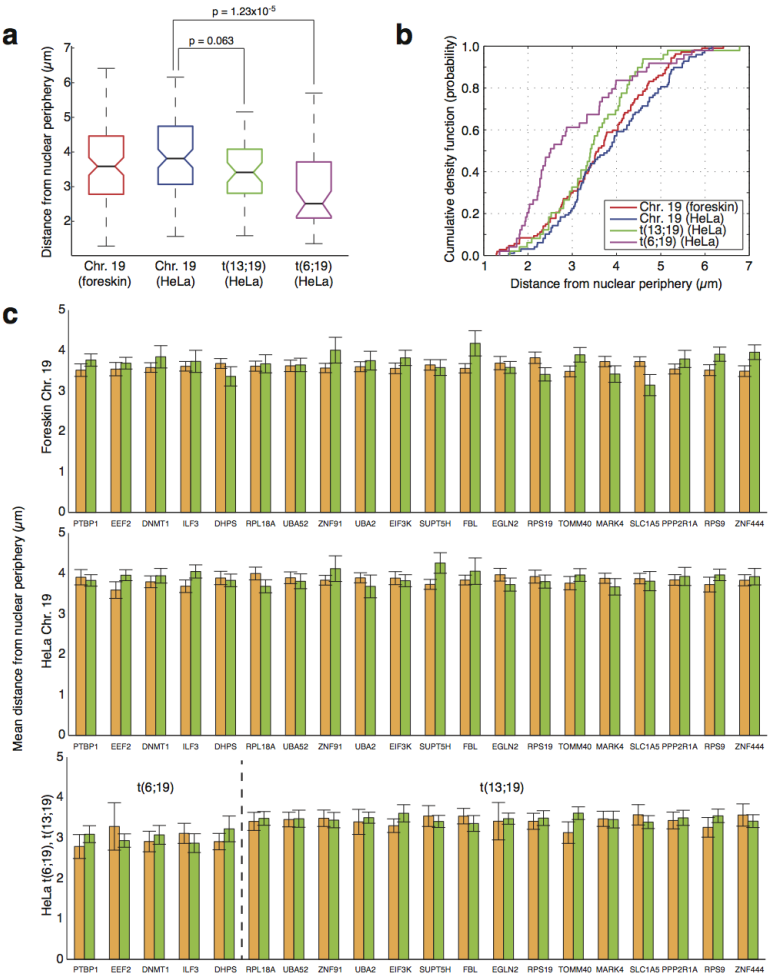

Figure 2.19: A. Box plot showing distance from nuclear periphery for Chr. 19 in foreskin fibroblasts and HeLa cells, as well as the t(6;19) and t(13;19) derivative chromosomes. Center line represents the median, edges correspond to 25th and 75th percentile, and whiskers extend to extrema. Medians are different with a p ¡ 0.05 if the notched regions do not overlap. We computed p-values using the Kolmogorov-Smirnov test. B. The cumulative density function corresponding to A. C. Distance from nuclear periphery for chromosomes in which a given gene is actively transcribing (green) versus inactive (orange). Error bars reflect the standard error of the mean.

governing the transcription of genes within a single chromosome. Specifically, we examined whether the transcriptional status of one gene in our panel (i.e., actively transcribing or transcriptionally inactive) affected the transcriptional status of another gene on the same chromosome. Such an interaction would manifest itself as a deviation from independence, with positive correlations signifying that the two genes A and B would be more likely than chance to be actively transcribing at the same time on the same chromosome, and anti-correlations indicating that the transcriptional statuses of genes A and B would be mutually exclusive.

We found that most pairwise interactions on single chromosomes did not show a significant deviation from independence (Fig.2.20; see Fig.2.22 and 2.22 for individual replicates) (although there may be weak effects that we we have insufficient data to detect). However, one pair of genes, RPS19 and ZNF444 (separated by 14.3 megabases), showed a significant anti-correlation (R = -0.400.08; p=3.99x10-5, Fisher Exact Test). One explanation for this anti-correlation is fluctuations in a potential trans-acting factor, such as a transcription factor, that activated RPS19 and inactivated ZNF444 in some cells while activating ZNF444 and inactivating RPS19 in others. Any such trans factor would, however, affect the copy of the gene on the other chromosome 19 as well[16]. We checked for this potential trans factor by looking for an anti-correlation between RPS19 on one chromosome and ZNF444 on the other copy of chromosome 19 in the same cell. We found that the inter-chromosomal interactions between the genes was qualitatively different, having a mild and less statistically significant positive correlation (R = 0.330.09; p=6.90x10-4, Fisher Exact Test), indicating that the interaction between these genes is not due to a trans factor but rather a cis effect confined to the chromosome itself. The lack of anti-correlation between the chromosome 19 copies (both between the pair of genes and also each gene with itself; Fig. 2.20) also precludes the possibility of genetic imprinting.

To see if these results are cell-type specific or are intrinsic to chromosome 19, we also looked for interactions amongst these same genes on the two intact copies of chromosome

19 in HeLa cells (Fig. 2.13A). Despite the large differences between these two cell types, we found that HeLa cells displayed the same anti-correlation between RPS19 and ZNF444 (Fig. 2.20), and did not display any inter-chromosomal correlations. Moreover, as in the foreskin fibroblasts, none of the other gene pairs displayed any significant interactions. To test whether this behavior depends on large-scale properties of chromosome 19, we checked whether the anti-correlation between RPS19 and ZNF444 persisted on the t(13;19) chromosome in HeLa cells (Fig. 2.20A).

This chromosome includes both of these genes, but we found no anti-correlation between these copies of the two genes (Fig. 2.23). We believe that these data indicate that the long-range regulatory mechanism governing this interaction may require the entire chromosome to be intact in order to function, suggesting that this regulatory interaction may be an intrinsic property of chromosome 19.

To check if chromosome conformation mediates this anti-correlation, we examined the physical distances between all pairs of active genes when RPS19 or ZNF444 were transcriptionally active or inactive (Fig. 2.24). We found no difference in the physical distances between any gene pair, although the anti-correlation prevented us from finding many chromosomes upon which we could locate both genes, and our statistical power was low for several pairs. We also found no dependence between the transcriptional state of these genes and chromosome positioning within the nucleus (Fig. 2.19). Combined with our findings on nuclear positioning of translocated chromosomes (Fig. 2.19), we feel our data suggest that large-scale chromosome conformation does not play a major role in determining the transcriptional status of these genes.

2.3. Discussion

Our method has enabled us to measure transcriptional activity on individual chromosomes in single cells by spatially segregating intron RNA FISH signals to particular chromosome territories. Our results suggest the existence of chromosome-specific regulatory mechanisms

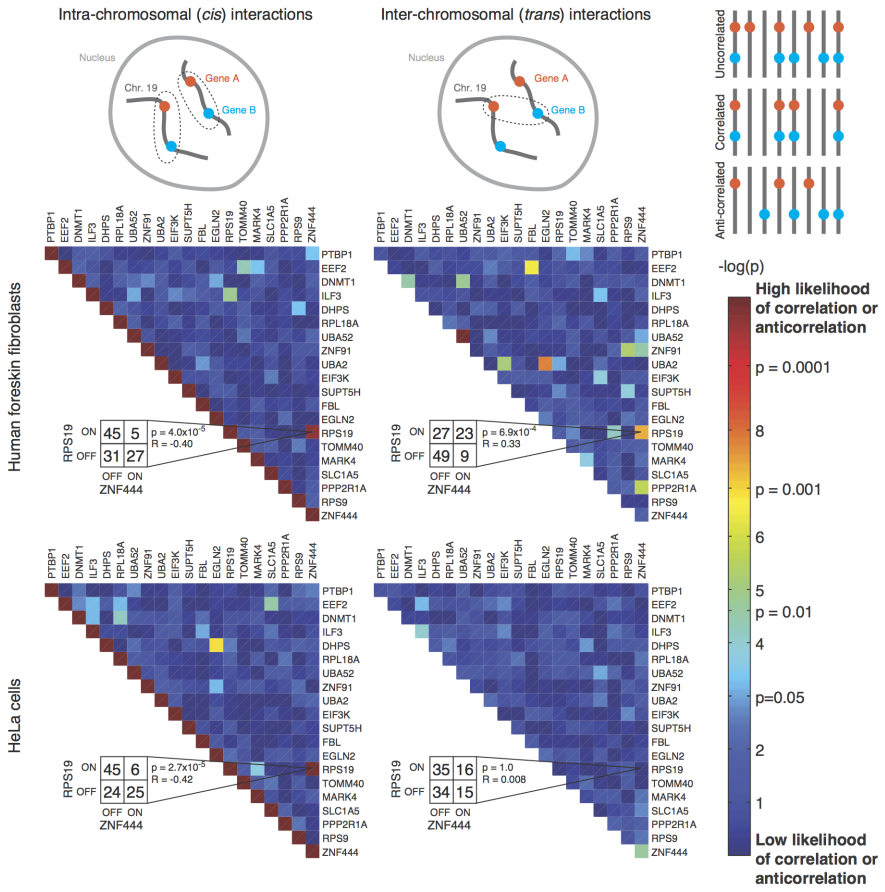

Figure 2.20: a. We reasoned that intrachromosomal transcriptional interactions between two genes would manifest themselves as a correlation or anti-correlation between transcriptional activity of the two genes on the same chromosome. To eliminate the possibility of a trans factor producing the same effect, we also measured the correlation between pairs of genes on opposite chromosomes within the same cell, which would be uncorrelated in the case of a cis (i.e., intra-chromosomal) interaction. b. Heat map showing the deviation from independence of the intrachromosomal transcriptional activity of all pairs of genes we measured in human foreskin fibroblasts, as measured by p-value for obtaining the measurement by random chance under the null-hypothesis that the genes transcribe independently (calculated using the Fisher Exact Test; see methods). A smaller p-value (more red) indicates a more significant deviation from independence. c. Same as (b), but for interchromosomal pairs. Diagonal elements represent interactions between the two copies of the same gene in single cells. d-e., Same as (b,c), but for the two intact copies of chromosome 19 in HeLa cells (identified as described in Fig 2.13). Here, we have presented data combined from two independent biological replicates (see Fig.2.21 and 2.22 for replicates).

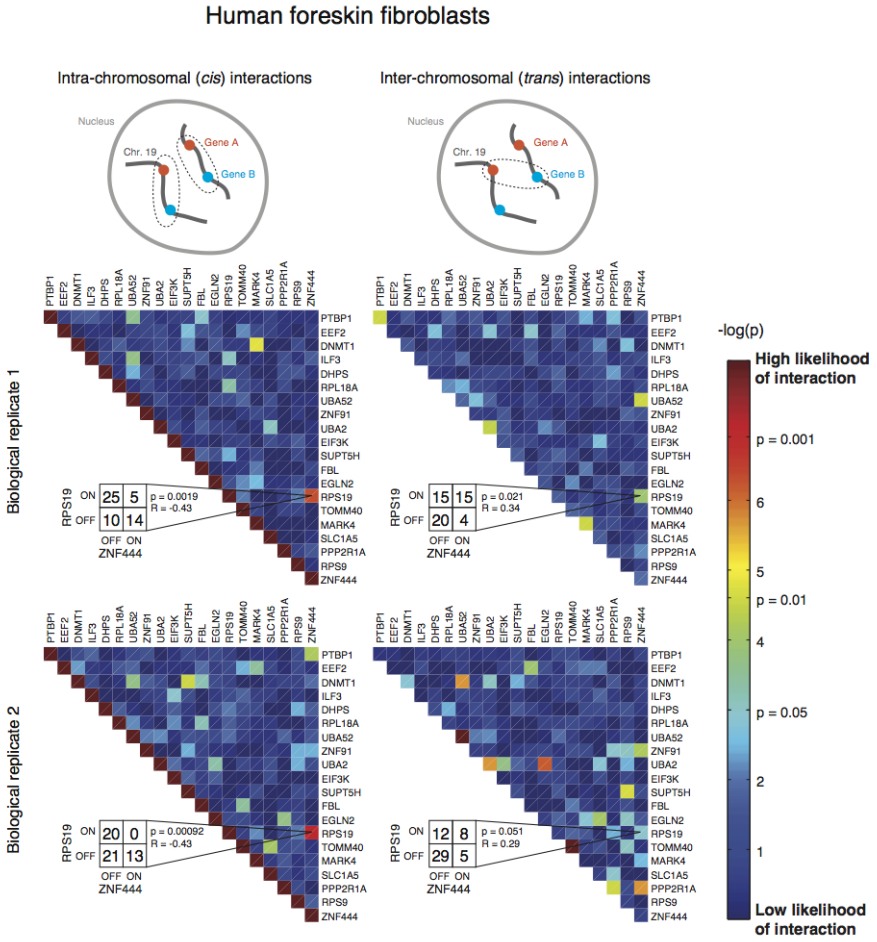

Figure 2.21: The analysis is exactly the same as that performed in 2.20 and as described in the methods.

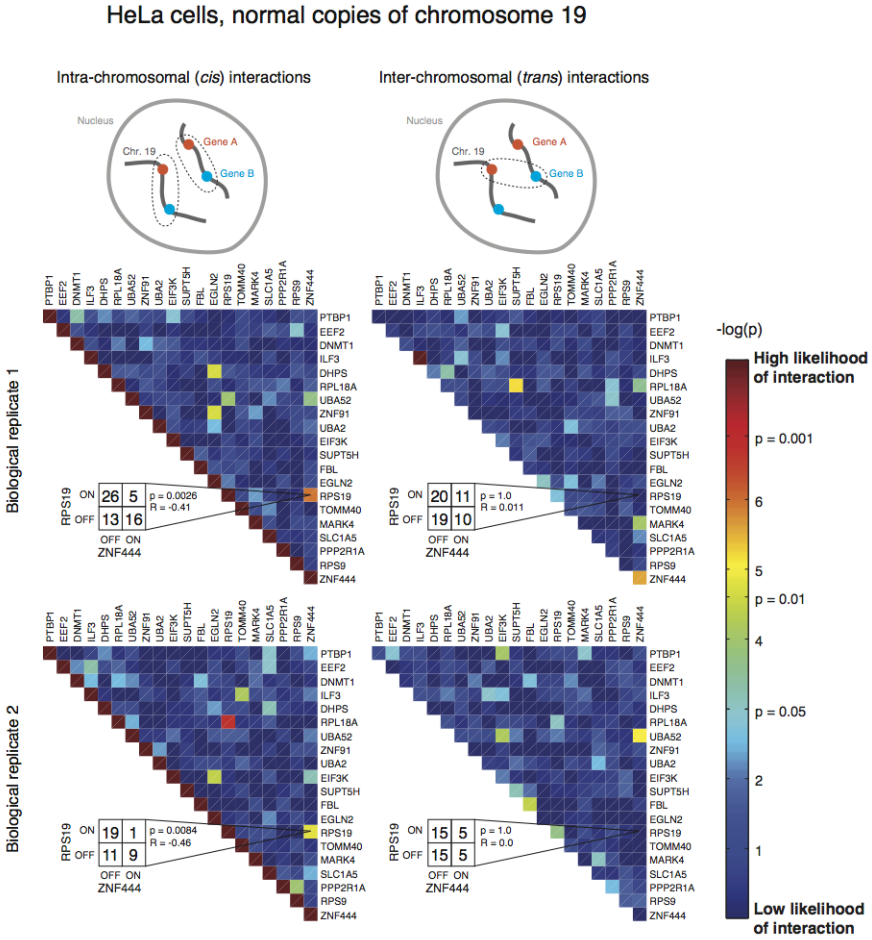

Figure 2.22: The analysis is exactly the same as that performed in 2.20 and as described in the methods.

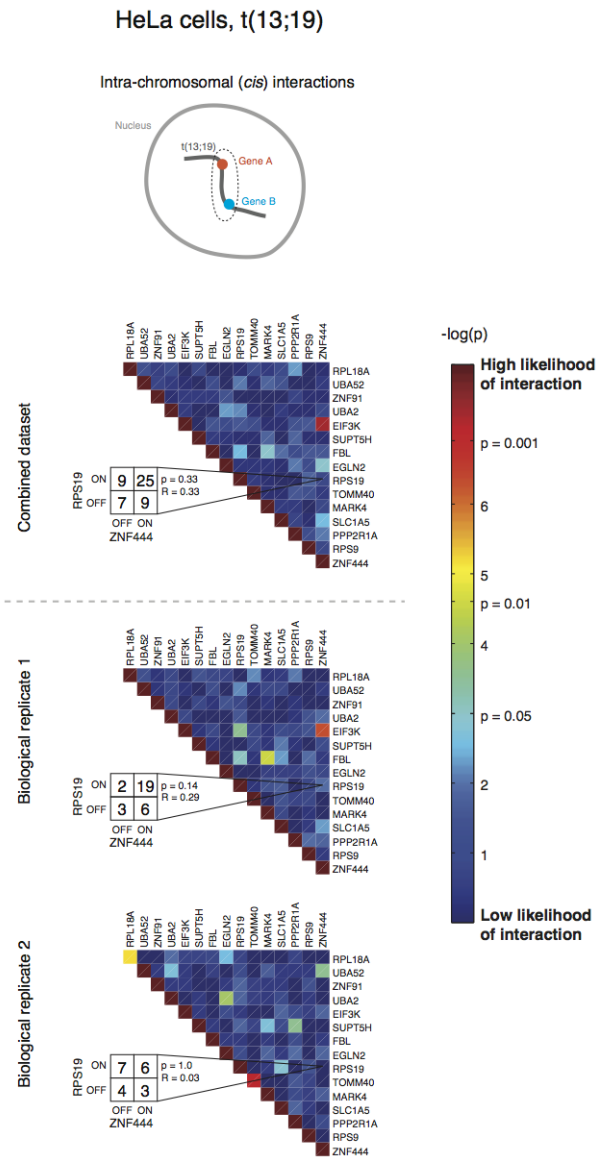

Figure 2.23: The analysis is exactly the same as that performed in Fig.2.20 and as described in the methods. We found no significant correlation between RPS19 and ZNF444, suggesting that the translocation disrupts the whatever the regulatory mechanism is that is responsible for the anti-correlation we observed between this pair of genes on the normal copies of chromosome 19 in both human foreskin fibroblasts and HeLa cells.

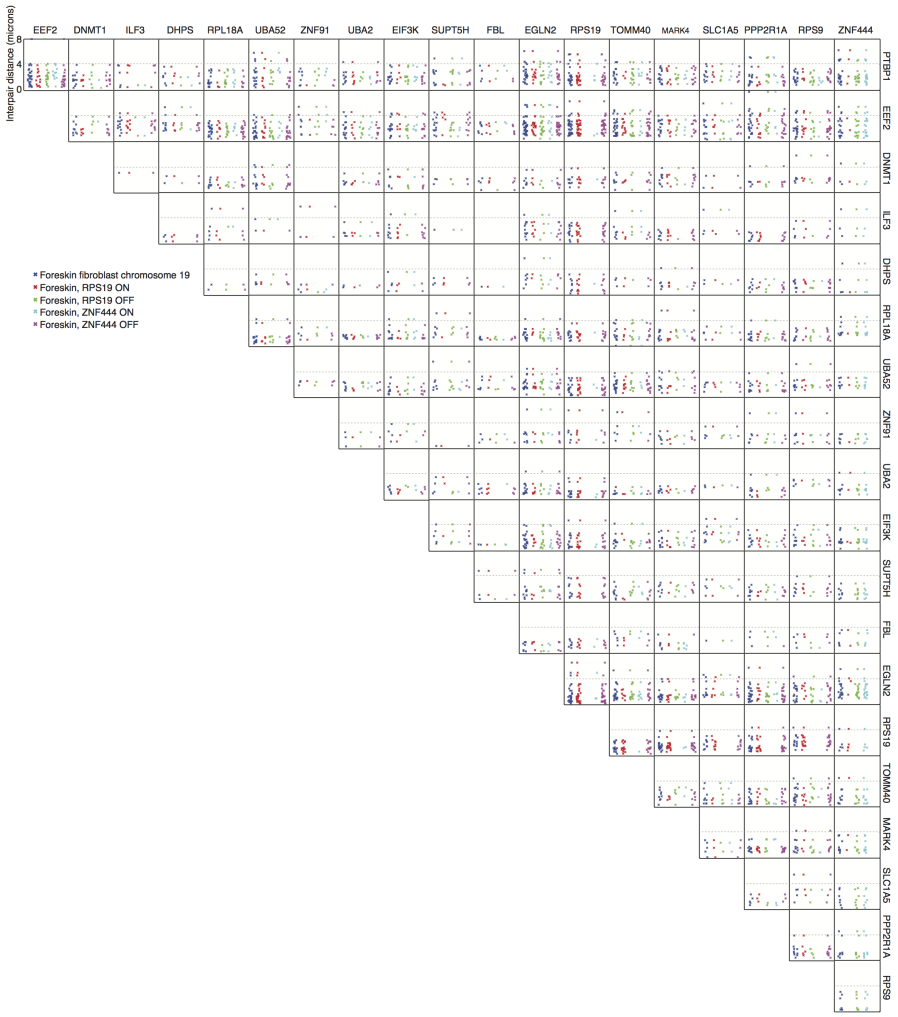

Figure 2.24: We compared the physical distances between all pairs of actively transcribing genes on chromosome 19 in human foreskin fibroblasts. Our comparison was between all loci on all chromosomes (blue markers), chromosomes upon which RPS19 in actively transcribing (red markers), chromosomes upon which RPS19 is transcriptionally inactive (green markers), chromosomes upon which ZNF444 is actively transcribing (cyan) and chromosomes upon which ZNF444 is transcriptionally inactive (magenta markers). We did not observe any statistically significant differences in average distances for any gene pair when com- paring any of these conditions (K-S test, p ¡ 0.05), although we note that our statistical power to resolve these differences is in some cases rather low because of a low number of measurements. We also did not note any particular distance patterns or relationships, although some pairs did appear to display less distance variability than others.

that control the expression of genetically distal genes, and that large-scale chromosome rearrangements may disrupt these mechanisms and influence transcription in a chromosome-specific manner. The key aspect of our method for these measurements is the ability to spatially segregate transcription and thus assign transcription to particular chromosomes, highlighting one of the main advantages of single cell/chromosome imaging over bulk methods that average together signals from multiple sources.

The chromosome-scale transcriptional effects we observed could arise as a consequence of spatial conformational characteristics of the chromosome or from other epigenetic features. Given that our assay also yields the position of active genes, we could assess the former possibility, and we did find that the distance between transcriptionally active loci in general appears to be larger than that measured between loci using DNA FISH, which cannot discriminate between active and inactive loci (see Supplementary discussion). We did not, however, find evidence in our data for particular large-scale conformational features associated with the specific transcriptional phenomena we observed. We note, however, that this does not preclude potential regulatory roles for other structural features beyond the ones examined here. Also, we have not examined short range interactions, nor have we looked at particular categories of genes that may display more gene-specific chromosomal conformation dynamics[69].

There are examples of epigenetic phenomena involving whole-chromosome transcriptional regulation, often related to dosage compensation in cases of aneuploidy[5]. One notable example of chromosome-wide regulation is dosage compensation, in which entire X chromosomes can show increased or decreased transcription or even complete transcriptional inactivation[71]. The molecular mechanisms underlying these epigenetic effects often appear to involve modification of chromatin, and it is possible that such effects may be at play in the effects that we have observed as well. For instance, one could imagine that the transcriptional hyperactivation of t(13;19) we observed in HeLa cells arose because of epigenetic modifications due to double stranded breaks that may progressively spread through the rest

of the chromosome. Testing such hypotheses will require the development of complementary methods to measure, for example, the chromatin status of individual chromosomes. Combined with iceFISH, we believe that such a toolset will allow us to determine the prevalence of these chromosome-level regulatory phenomena and uncover their underlying mechanisms.

2.4. Materials and Methods

2.4.1. Cell culture, fixation, and fluorescent in situ *hybridization*

We grew primary human foreskin fibroblasts (ATCC CRL 2097) or HeLa cells (gift from the lab of Phillip Sharp) in Dulbeccos modified eagles medium with glutamax (DMEM, Life Technologies) supplemented with penicillin/streptomycin and 10% fetal bovine serum. We enriched for G0/G1 phase cells through a double-thymidine block (2mM thymidine in medium) procedure, which arrested cells at the beginning of S phase. We released the cells and let them go through S, G2, M, G1, S, G2, M, and then fixed them when they were in G1. We let the cells go through over one complete cell cycle to minimize any potential transcriptional or structural effects due to the block itself. To fix the cells, we followed the protocol of Raj et al. Nat. Meth. 2008. Briefly, we fixed the cells for 10 minutes at room temperature using 4% formaldehyde/10% formalin in 1x phosphate buffered saline solution (PBS), followed by two rinses in 1x PBS, after which we permeabilized the cells with 70% EtOH and stored at 4C at least overnight.

To perform fluorescence *in situ* hybridization (FISH), we again followed the procedure of Raj et al. Nature Methods 2008 with some minor modifications. We prewashed with a wash buffer containing 10% formamide and 2x saline-sodium citrate (SSC), then hybridized by adding the appropriate amount and type of probe (described later) in a buffer containing 10% formamide, 2x SSC and 10% dextran sulfate (W/V). We empirically determined the optimal concentration of each probe, which in most cases was roughly equivalent to the concentrations used in Raj et al. Nature Methods 2008. We hybridized our samples overnight in a humidified chamber kept at 37C, then washed twice for 30 minutes with wash buffer at

37C (adding DAPI at a concentration of 50 ng/mL in the second wash), and then imaged in 2x SSC as described below.

In the case of the experiments involving Actinomycin D, we incubated HeLa cells in 2 μg/mL of Actinomycin D (Sigma) for 0, 30, 60, and 120 minutes (as described in Fig. 2.4), after which we fixed the cells and performed FISH. We made sure to thoroughly mix the Actinomycin D into the medium before adding it to avoid spatial inhomogeneity in the activity of the drug.

For the RNase experiments, we fixed and permeabilized the cells as just outlined, after which we aspirated the 70% EtOH, washed once with 1x PBS, then added 1x PBS with 10μg/mL of RNase A (Sigma). We incubated the fixed cells at 37C for 30 minutes, washed with 1x PBS, and then proceeded with FISH as outlined above. As a control, we performed the exact same procedure on cells in a neighboring well, but didnt add RNase A to the 1x PBS for the incubation (as described in Fig. 2.3).

2.4.2. Imaging

We imaged all our samples on a Nikon Ti-E inverted fluorescence microscope using a 100x Plan-Apo objective (numerical aperture of 1.43) and a cooled CCD camera (Pixis 1024B from Princeton Instruments). We sequentially acquired three-dimensional stacks of fluorescent images in 6 different fluorescent channels using filter sets for DAPI, Atto 488, Cy3, Alexa 594, Atto 647N, and Atto 700. Our exposure times were roughly 2-3 seconds for most of the dyes except for DAPI (which we exposed for 100ms) and Atto 700 (5 seconds, due to somewhat weaker illumination on our apparatus). The spacing between consecutive planes in our stacks is 0.3μm.

2.4.3. Image analysis

Once we acquired our images, we put them through an image analysis pipeline made up of custom semi-automated spot recognition software we wrote in MATLAB with the following

series of steps:

1. We first identified candidate spots in the three-dimensional image by filtering the image with a Laplacian of Gaussian filter, and taking the top 300 spots as candidates. In some cases, we also chose cells to analyze based on phase in the cell-cycle. In those cases, we chose cells that had little or no Cyclin A2 mRNAs. Our experiments in Fig. 2.12 validate this approach.

2. For each candidate, we then fit the candidate to a Laplacian of Gaussian intensity profile, thereby giving us precise estimates of the center, width, and intensity of the spot.

3. Based on histograms of the intensities and widths, we manually selected a subset of the spots with qualities (uniform width, higher intensity) that were higher than background. This is similar in spirit to the procedure described in Raj et al. Nature Methods 2008, in which the experimenter chose a threshold to separate legitimate RNA spots from background spots. In this case, we erred on the side of including spots that may be background, because our multi- color scheme for spot assignment provided us another means by which to discard background spots.

4. Once we had selected the spots, we then ran software that found the fiducial markers (in this case, probes in all 5 RNA colors targeting SUZ12 mRNA, which are present at an abundance of roughly 20-50 clear cytoplasmic spots per cell). In this manner, we could measure the displacements between different fluorescence channels in each cell individually. We then applied these shifts to align the computationally identified spots between the different fluorescence channels.

5. After alignment, we then ran software that looked for colocalized spots corresponding to the particular pseudocoloring scheme we chose for the introns we targeted. We estimate that our software is roughly 75% accurate in assigning colocalized spots to particular genes at this stage.

6. We then went through a manual correction process in which we corrected mistakes the software made in identifying spots. Common issues were failure to detect dim (but clearly present) signals in one of the fluorescent channels and resolving two spatially close fluorescent spots that the laplacian of gaussian filtering and candidate identification steps had labeled as a single spot.

7. Once we had correctly annotated the introns of the gene loci we had labeled, we then examined cells manually to separate out individual chromosomes. We would discard cells in which the chromosomes overlapped since this made it difficult to assign gene spots to particular chromosomes. In order to determine the distance of the chromosome from the nuclear periphery, we first determined the average position of the spots of the chromosome in x and y and then found the Euclidean distance between this point and the nuclear periphery as outlined by our DAPI stain.

2.4.4. Characterization of error rate

In order to gain some sense of the rate of false positives, we performed a hybridization in foreskin fibroblasts in which we left out 10 of the 20 genes comprising our iceFISH assay (randomly chosen by another member of the lab), and proceeded with our spot identification procedure as usual (Fig. 2.11). We found that our rate of false identification was very low, with the vast majority (97%) of spots we assigned corresponding to genes which we had targeted in our assay.

We also probed a set of 2 genes (RPS19, TOMM40) one at a time with oligonucleotides labeled with a single dye rather than the combination of 2 or 3 dyes used in our pseudocoloring strategy. Our aim was to determine to what extent our pseudocoloring strategy would result in false negatives in spot identification. We found that the spot per chromosome frequencies measured with a singly-colored probe alone were 0.56, 0.27, while the spot frequencies measured by pseudocoloring were 0.57, 0.25, respectively, in a total of 30 cells. Although statistical effects preclude a definitive statement, our results are consistent

with our pseudocoloring strategy correctly identifying virtually all spots detectable by RNA FISH targeting introns.

2.4.5. Probe design

We designed 20 base oligonucleotide probes against introns using custom FISH design software (http://www.biosearchtech.com/stellarisdesigner/). Where possible, we tried to design 16 oligonucleotides targeting the first intron of the gene. We ordered the oligonucleotides from Biosearch Technologies (Novato, CA), who synthesized the oligonucleotides with amine groups attached to the 3 end. We coupled these 3 ends to various organic dyes (including Atto 488 (Atto-Tec), Cy3 (GE), Alexa 594 (Invitrogen), Atto 647N (Atto-Tec), and Atto 700 (Atto-Tec)) as indicated in the text and in ***Supplementary Table 1***. We purified the probes by HPLC as described in Raj et al. Nature Methods 2008.

2.4.6. Karyotyping of HeLa cells

We performed G-band analysis (karyotyping) on metaphase spreads of our HeLa cells following standard procedures. This indicated that our cells contained two intact copies of chromosome 19 and a full third copy of chromosome 19 split into two fragments and fused to other chromosomes (Fig. 2.14). One fragment includes the first half of the chromosome 19 p-arm and is fused to a large portion of chromosome 6. The second fragment is the remaining portion of chromosome 19 (half the p-arm through the centromere and entire q-arm), which is fused to the q-arm of chromosome 13. In order to conclusively demonstrate that chromosome 19 was split in this particular way, we performed a DNA FISH analysis on the same metaphase spreads that we performed the G-band analysis on. We used probes targeting loci within the 19p13 and 19q13 regions on chromosome 19, each labeled with a different fluorophore (Abbott Molecular). The results confirm the results of the G-band analysis. We performed this analysis on 10 cells, each of which showed the same genetic abnormalities, indicating that the cells do not vary much in this particular characteristic from cell to cell.

2.4.7. Click-iT EdU analysis of cell cycle progression

In order to demonstrate that Cyclin A2 mRNA was an accurate marker of position in the cell cycle, we used the Click-iT EdU Alexa Fluor 594 Imaging kit (Invitrogen), which incorporates a targetable chemical into newly replicated DNA. In this case, we incubated foreskin fibroblasts with the 10μM Click-iT EdU reagent for 5 minutes before fixing the cells. We performed our FISH protocol on these cells using a Cyclin A2 mRNA Cy3 probe and after hybridization and wash steps followed the instructions provided with the kit for fluorescently labeling the incorporated EdU. We ultimately did not elect to use the Click-iT EdU kit directly in most of our experiments (and instead opted to use Cyclin A2) because we found that performing the Click-iT EdU procedure interfered with our nascent RNA FISH detection, most likely either due to interference with transcription itself or by making our spot detection less reliable because of additional washing steps associated with the Click-iT procedure.

2.4.8. DNA FISH

We performed DNA FISH with BAC probes from Empire Genomics, using their reference hybridization protocol. In the human foreskin fibroblast cells we applied pairs of fluorescently labeled BAC clones from the human RPCI-11 library targeting human chromosome 19 at positions 2.8-4.5 Mb (268O21), 39.0-39.5 Mb (31D10), or 52.5-52.7 Mb (43N16). We denatured the DNA by immersing the cells in 70% formamide, 2X SSC buffer at 80C for 5 minutes, and then transferred to series of ethanol steps increasing to 70, 85, and then 100% ethanol. We added 10μL of BAC probes to the air dried sample, applied a coverslip, and incubated overnight in humidified slide chamber. The next day we washed the sample with 0.4X SSC at 73C for 2 minutes, removed the coverslip, transferred to room temperature 2X SSC for 1 minute, then to 10μL of 2X SSC with DAPI at 50 ng/mL, and applied a new coverslip. We performed imaging similar to our iceFISH probes with dye pairs Red 5-ROX and TAMRA.

2.4.9. Combined DNA/RNA FISH

We performed a sequential DNA/RNA FISH in HeLa cells by first performing DNA FISH using BAC clones as and then performing RNA FISH, both by following the protocols outlined above. We found that both the bright exonic transcription sites and the intron spots were considerably brighter than single mRNA spots, thus showing that the RNA probes were not simply targeting the DNA directly. We compared the location of SLC1A5 exonic (Alexa 594) and intronic (ATTO647N) RNA to the location of DNA FISH probes using BAC clones RP11-687M15 (TAMRA).

2.4.10. Statistical analysis

In Fig. 2.13, we looked for deviations from independence in the transcriptional frequencies of all pairs of genes we examined. We performed the Fisher Exact Test on all 2x2 tables generated by counting the number of chromosomes where gene A or B was transcriptionally active vs. inactive. We reported the two-sided p-value corresponding to the chance of obtaining a similar deviation from independence via random chance, with a smaller p-value corresponding to a more significant result. In Fig 2.13, we show the results we obtained by analyzing a dataset consisting of the combination of two independent biological replicates; we also performed the analysis on each individual biological replicate, as shown in Fig. 2.21 and 2.22. Note that we have not applied a multiple hypothesis correction in our presentation of the p-values; however, our results would remain statistically significant if we applied the crude correction of just multiplying our p-values by 190, which is the number of pairs of genes we examined. We chose to convey the information in this manner because the number of hypotheses tested depends on the particular question being asked of the data. For instance, if one decides that, based on the human foreskin fibroblast data, one wanted to focus on interactions between RPS19 and ZNF444, then the p-values for the specific hypothesis comparing these two genes in, say, HeLa cells, would not be subjected to this same correction. We leave such interpretative matters to the reader.

We also report the correlation coefficient between RPS19 and ZNF444; although it is a somewhat imperfect measure of the lack of independence for this sort of data, it has the advantage of being familiar to many researchers. We obtained standard errors for the correlation coefficient by bootstrapping.

In Fig. 2.20, we obtained p-values for the difference in transcriptional frequency between the copy of the gene on the t(13;19) (or t(6;19)) chromosome and the copies of the gene on the normal copies of chromosome 19 by rejecting the null hypothesis in which the frequency of transcription was the same for all three copies. We did this by computationally generating the probability density function for the difference in transcriptional frequencies between two sets chosen to match our experimental data in size under the null hypothesis that the frequency is the same for both sets, and then directly calculated the probability of finding our observed difference by chance.

2.4.11. Relationship between intron spot measurements and transcriptional activity

In our measurements, we obtain both the probability of finding an intron spot as well as the intensity of that spot. Here, we present a simple model of intron dynamics that relates transcriptional dynamics to these two measurements.We assume that transcription occurs in bursts, which is supported by several studies in higher eukaryotes as noted in the main text. We assume that the transcription of a gene as a function of time is given by the function $\mu(t) = \mu_0 f(t)$, where μ_0 is a constant is a constant and $f(t)$ is a stochastic process that randomly fluctuates between having value 0 and value 1 (corresponding to the gene being active or inactive, respectively). We do not assume any form for $f(t)$ other than that the time in the active state or the inactive state is on average considerably longer than the time to degrade introns, although several groups model the dwell times in the active or inactive state as being exponentially distributed, and there is experimental support using time-lapse imaging for this view [30, 10, 72]. We assume that the fraction of time the gene is in the active state is given by a. The (continuous) equation governing the intron dynamics is:

$$dI/dt = \mu(t) - \delta I$$

where I is the number of intron molecules and δ is the rate of intron degradation. The steady state of this equation when the $\mu = 0$ or $\mu = 1$ is 0 or μ_0/δ, respectively. The degradation rate, δ, is what determines how rapidly I heads to steady state. Based on our Actinomycin D experiments (Fig. 2.4), we believe the intron half-lives of the genes we examined to be less than 5 minutes. In this case, where δ is considerably larger than the rates of the gene switching on or off, then

$$I(t) \approx \mu(t)/\delta$$

i.e., $I(t)$ is non-zero only when the gene is actively transcribing, and zero when the gene is inactive. The time average of $I(t)$ is then

$$\langle I(t) \rangle = a\mu_0/\delta$$

while the time averaged rate of transcription is given by

$$\langle txn \rangle = a\mu_0$$

By measuring the percentage of the time we observe the gene actively transcribing, we can estimate a, the probability of the gene being active, in absolute terms. When the gene is active, the rate of transcription is μ_0, but we can only measure the intron spot intensity, which is proportional to the rate of transcription. Thus, we cannot measure the rate of transcription when the gene is active up to a constant of proportionality that is $1/\delta$, which in principle may vary from one gene to another. Nevertheless, we can compare

the relative changes in the rate of transcription of the same gene from one chromosome to another by comparing our measurements of both a and μ_0/δ. In our experiments, we found that in virtually all situations, the spot intensity (μ_0/δ) did not change (Fig. 2.5), but we did observe changes in the probability of finding an intron spot (a), which implies a proportional change in the overall time-averaged rate of transcription. We interpret this to mean that whatever causes the changes in transcription on the hyperactivated t(13;19) chromosome in HeLa cells (as compared with the intact chromosome 19s in HeLa cells), it is most likely not something that is changing the rate of transcription when the gene is active, but rather is changing the probability that the gene is active itself. We note that this is not necessarily the same as saying that the transcriptional burst frequency has changed while the transcriptional burst size remains the same: if transcriptional bursts lasted for longer, then both the burst size and the probability of finding a spot would increase, even if the burst frequency remained constant.

2.4.12. Comparison of the distance between active DNA loci to previous experiments

We were somewhat surprised to find that the distance we observed between transcriptionally active loci was quite large even for relatively short genomic separations; for instance, we observed a mean physical displacement of 1.7μm for genes separated by only 0.36 kilobases. We suspected that these large distances were due to the relatively decondensed chromatin thought to accompany actively transcribed genes. To check whether such a hypothesis was consistent with the published literature, we examined the data from the excellent study [51] , in which the authors measured the relationship between the physical and genetic separation of DNA loci. In particular, they examined the distances genes in transcriptionally active regions of DNA (ridges) and transcriptionally inert regions of DNA (anti-ridges), finding that the transcriptionally active regions were considerably more physically spread out than the transcriptionally inert regions.

We posit, given that transcription is fundamentally pulsatile, that the mean physical separation between two loci in a transcriptionally active region fluctuates between a short distance

when the genes are inactive and a long distance distance when the genes are active (consistent with the findings of [76]). From this perspective, the observations by [51] correspond to measuring the mean inactive gene separation (DNA FISH between transcriptionally inert regions) and the weighted average of the mean inactive and active separation (DNA FISH between transcriptionally active regions), weighted by (1-a) and a, respectively, where a is the probability of the gene being active. Our measurements of interpair separations correspond to the mean active gene separation.

We checked for consistency between these different sorts of measurements when comparing our data to that of Fig. 2B, left panel from [51]. At a genetic distance scale of roughly 490 kilobases, Mateos-Langerak report an mean square distance of around 0.23 μm2 for inactive loci and 0.84μm2 for "ridges", the latter of which we believe corresponds to the weighted average of active and inactive loci as described above. Our measurements of a mean square distance of 3.57μm2 between active loci at this genetic distance scale would imply a weighting factor of 0.18, which falls squarely within our observed variation in probabilities of genes transcribing. Thus, we conclude that our data are at least consistent with the previous DNA FISH observations of Mateos-Langerak et al. with this simple model for the distance between active and inactive loci. Further studies may elucidate whether such a model is indeed an accurate description of conformational dynamics.

CHAPTER 3 : Single cell allele-specific expression via single nucleotide variant detection *in situ*

3.1. Background

Advances in single cell imaging have enabled researchers to detect individual RNAs with single molecule resolution[18, 62], more recently in conjunction with single chromosomes[43]. However, such methods typically are unable to distinguish single nucleotide variants in these molecules. Development of such a method with general applicability would be of great utility in fields like genetics and gene regulation, specifically because of its ability to measure allele-specific gene expression at the single cell and single molecule level[29, 31, 19].

Methods also exist to survey the expression levels of RNA species that vary only by single nucleotides on the cell population level through RNA sequencing [67, 45, 88]. The advantage of these techniques is that the measurements are genome-wide. However, despite the 'digital' transcript counts provided by RNA sequencing, the many steps required in upstream procedures to extract total RNA, generate input cDNA by reverse transcription, and library preparation are not fully characterized and may introduce systematic errors that need to be accounted for in subsequent analysis. Overall, the RNA sequencing based methods have the disadvantage of requiring lots of input cellular material that prevent detection of allelic expression behaviors present in a minority of cells and obviously masks cell-to-cell heterogeneity in expression. These methods require amplification of nucleic acid material. Quantification of low copy number RNA species through amplification is susceptible to noise. Direct detection is the best strategy for quantitative comparison with amounts of material spanning several orders of magnitude and between different RNA species.

The few methods available for *in situ* SNV detection tend to be complex and suffer from low efficiency. Specifically, Larsson et al 2010[38] used a rolling circle amplification method that requires a series of enzymatic steps, each of which has issues with reaction efficiencies and technical repeatability. Ideally there would be no enzymatic steps in the assay to

avoid variability in reagent stability over freeze/thaw cycles or different batches. Enzymatic steps can also suffer from different amounts of accessibility throughout a sample used for *in situ*. Larsson et al described experiencing difficulties with assay efficiencies in technical repeats using these methods in the past for direct RNA detection (as little as 1% detection efficiency), thus requiring the use of expensive LNA primers to create cDNA (a DNA complement version of the RNA target) used for detection and subsequent amplification. The most concerning data from their work is the reported beta-actin mRNA molecules per human fibroblast cell with a gaussian distribution ranging from 28-1000 whereas similar experiments performed using tiled singly labeled RNA FISH assay [62] produces consistent beta-actin mRNA counts in no less than 1500 and up to 6000 RNA per human fibroblast cell (data not shown). This discrepancy is most likely due to the variability in enzymatic efficiency.

Here we present a method that discriminates single nucleotide differences in single cells using direct detection and quantification of RNA molecules. Using a probe design strategy employed in DNA directed chemical reactions and other nanotechnology applications [86], we are able to discriminate single nucleotide differences while still using DNA oligonucleotides without exotic chemistries, making them affordable for custom synthesis. By combining this discriminating technology with tiled singly labeled RNA FISH probes[62], we get a high efficiency, quantitative assay for allele-specific gene expression. We validate the capabilities of the method with a panel of melanoma cell lines with different genotypes for the BRAF V600E mutation, apply the method to measure population and single cell allelic imbalance, and discriminate many SNVs at a time to classify whole chromosomes for parental origin.

3.2. Results

3.2.1. The SNP FISH assay

One of the primary difficulties in detecting a single base difference via RNA FISH is that a 20 base oligonucleotide probe will often hybridize to the RNA despite the presence of

a single mismatch. On the other hand, very short oligonucleotide probes, while able to discriminate between single base differences, will often fail to remain bound to the target due to reduced binding energy. Meanwhile, in either case, distinguishing legitimate signals from false positives is a challenge when using just a single probe. We use probe design and high-resolution image analysis to circumvent these issues. Firstly, in order to distinguish between single base mismatches, we used a "toehold probe" strategy in which we hybridize a 28 base single stranded DNA SNV detection oligonucleotide probe to a shorter "mask" oligonucleotide[87, 85, 46] (Fig. 3.1). The remaining single stranded portion of the detection oligonucleotide includes the SNV base and is short enough to confer selectivity based on single base mismatches, but once bound, the mask oligonucleotide dissociates from the detection probe via passive strand displacement, enabling the remainder of the detection probe to bind to the target RNA. This strategy confers specificity while still retaining a sufficient binding energy to prevent the detection probe from rapidly dissociating from the target after hybridization.

The use of a single probe can often lead to a large number of false positive signals, as every off-target binding event is indistinguishable from on-target binding. Typically, one avoids such false positives by relying on the co-localization of multiple probes[62, 61], but that is not possible when one can only use at most a single probe, as is the case in SNV detection. We adopted a strategy in which we used multiple oligonucleotide probes (collectively referred to as the "guide" probe) that bind to the target RNA, thereby robustly identifying the target RNA with a very low rate of false positives and negatives. We then only consider detection probe signals as legitimate if they co-localize with the guide probe signals, thereby clearly distinguishing false positive signals from true positives (Fig. 3.1).

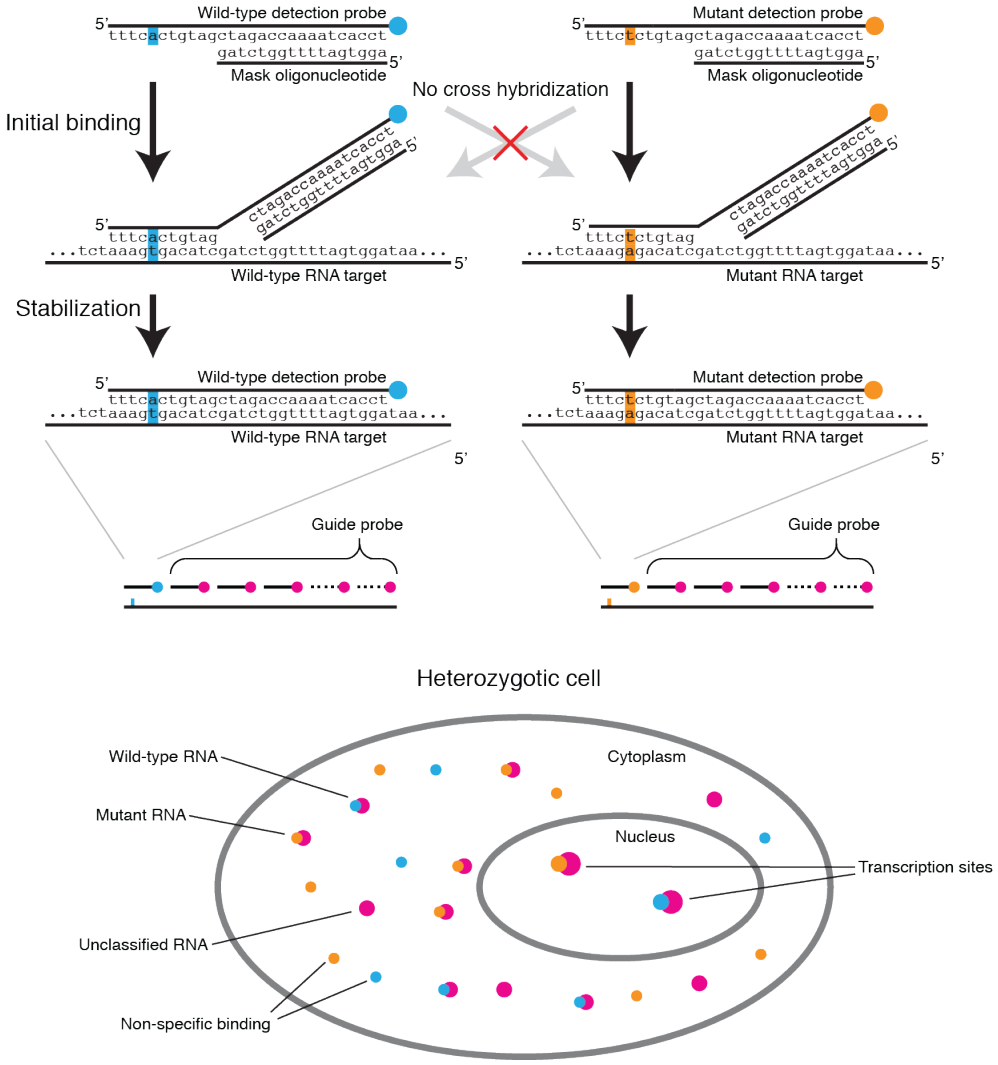

Figure 3.1: Toehold probes enable SNV detection on individual RNA molecules *in situ*. Schematic of the principle behind *in situ* SNV detection, using the T1799A mutation of BRAF as an example.

3.2.2. Validation of the assay using BRAF V600E melanoma cell lines

To demonstrate the efficacy of our method, we utilized a series of melanoma cell lines harboring a well-known mutation in the BRAF oncogene. We used cell lines that were homozygous mutant, heterozygous mutant/wild-type and homozygous wild-type in a mutation of the 1799 position from T to A. We designed two detection probes for this particular SNV, one targeting the mutant and one targeting wild-type transcripts, and utilized a mask oligonucleotide common to both. We found that our scheme performed as expected, clearly revealing both wild-type and mutant transcripts in a heterozygous line (Fig. 3.2a,b; see Fig. 3.3 for homozygous lines). In the homozygous mutant cell line (SK-MEL-28), we found that roughly 56% of the RNA identified by the guide probe co-localized with signals from the mutant detection probe, whereas only 7% of the guide probe signals co-localized with the wild-type detection probe (Fig. 3.4, Fig. 3.5). Conversely, in the homozygous wild-type cell line (WM3918), we found that 58% of guide probe signals co-localized with the wild-type detection probe whereas only 7% of the guide probe signals co-localized with the mutant detection probe. In the heterozygous mutant/wild-type cell line WM9, we found 33% of BRAF transcripts co-localized with the wild-type detection probe while 34% co-localized with the mutant detection probe, indicating that both copies of the gene transcribe equivalently in these cells. In another heterozygous cell line WM983b, we observed 36% and 29% wild-type and mutant mRNA, respectively. Overall, we found that our co-localization efficiency was around 65%, roughly in line with other estimates of efficiency of hybridization of DNA oligonucleotides to RNA[49], and that co-localization itself is not subject to a high rate of false positives (Fig. 3.5). We also found that the presence of the wild-type probe improves specificity of the mutant detection probe and vice-versa (data not shown). The mask oligonucleotide is critical for maintaining this specificity; we observed many false-positive detections when we performed our detection without the mask present (Fig. 3.6a). This approach appears to work for a variety of different target sequence mismatches (Fig. 3.6b). Increasing the toehold length also increases the detection efficiency (Fig. 3.7).

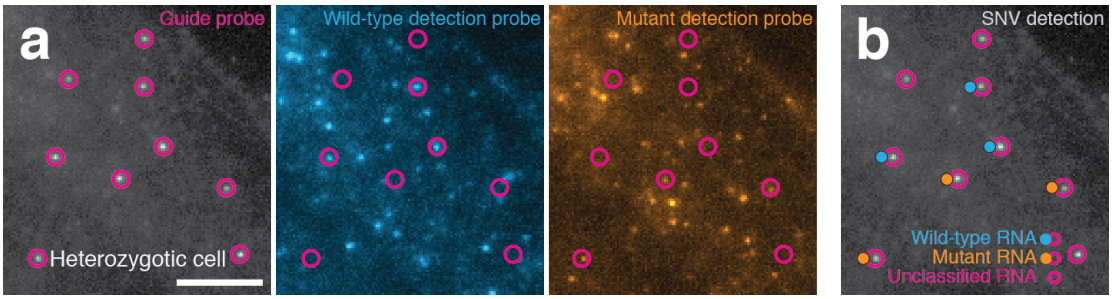

Figure 3.2: a. Visualization of the guide probe detecting BRAF mRNA (ATTO488, left panel) and the wild-type and mutant detection probes (Cy5, Cy3, middle and right panels, respectively). b. Classification of RNA as being either wild-type or mutant using the detection probes. Scale bar is 5μm.

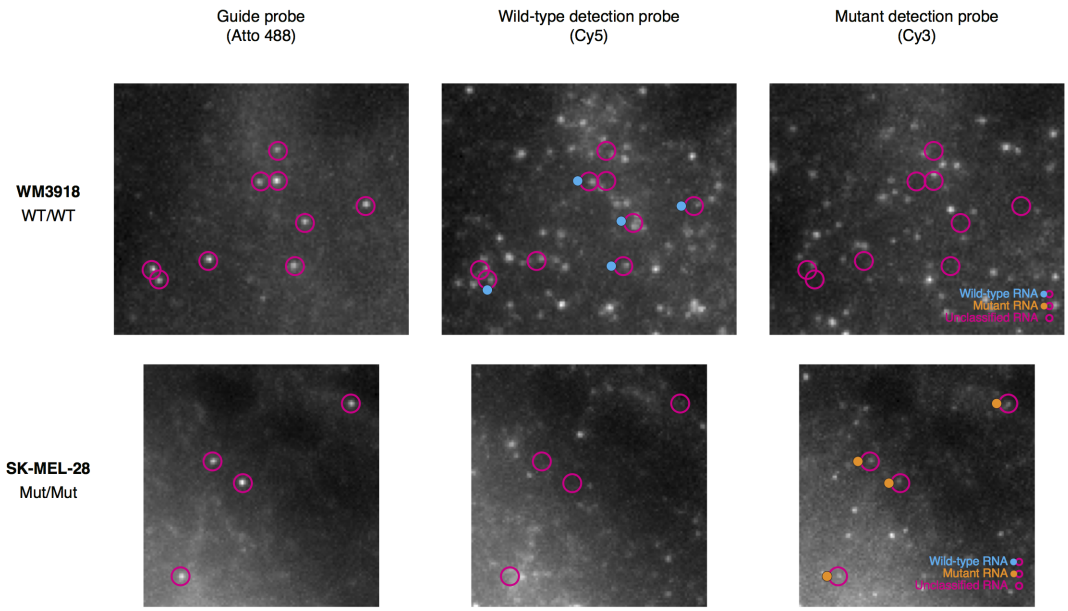

Figure 3.3: Images of SNV detection in homozygous wild-type BRAF cells (top) and homozygous mutant BRAF cells (bottom). The left panels show the guide probe targeting BRAF mRNA, and the middle and right panels show the wild-type and mutant BRAF detection probes, respectively. These are example images corresponding to the data shown in Fig. 3.2. We found little co-localization with the off target probe, as shown quantitatively in Fig. 3.2. Note that each panel shows a z-projection of multiple planes, but that we perform co-localization analysis in three dimensions.

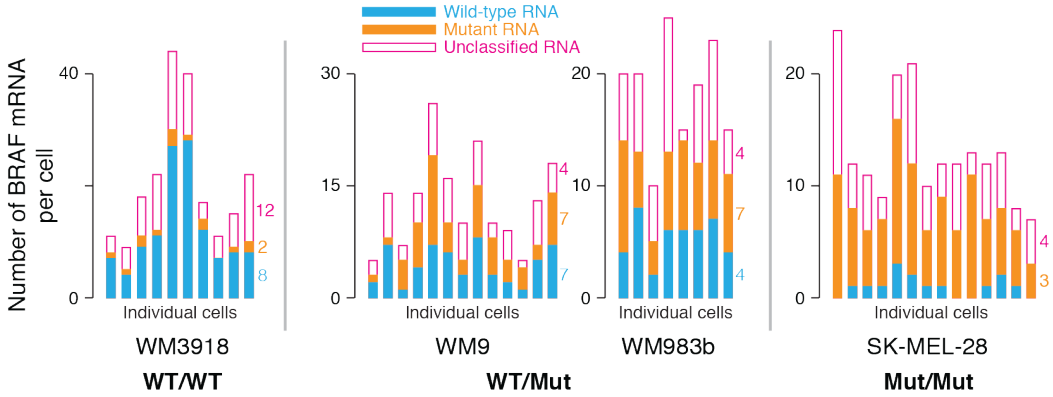

Figure 3.4: Quantification and classification of RNA as wild-type or mutant in a group of single cells. Each sample shown is one of a set of at least two biological replicates. Left: cells with only wild-type BRAF; middle: cells that are heterozygous for BRAF; right: cells that are mutant for BRAF.

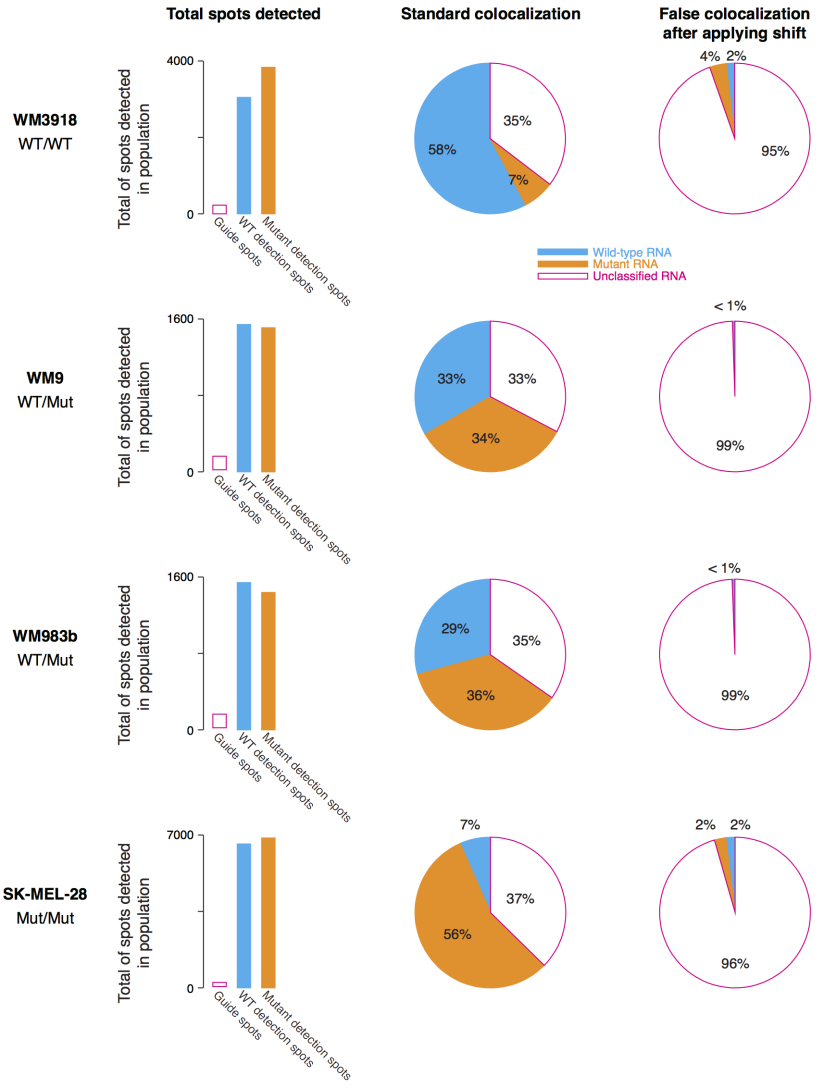

Figure 3.5: Measurement of false positive rates due to random colocalization. Bar graphs show the number of guide spots and detection spots identified in all the cells we analyzed. The pie charts show the degree of colocalization in the original images (left) and after applying a 8 pixel shift in x and y to the detection spots (right). The latter serves as an estimate of how often spots would be likely to colocalize purely by chance. We found that these rates of colocalization were very low, with the vast majority of spots remaining unclassified after applying the shift.

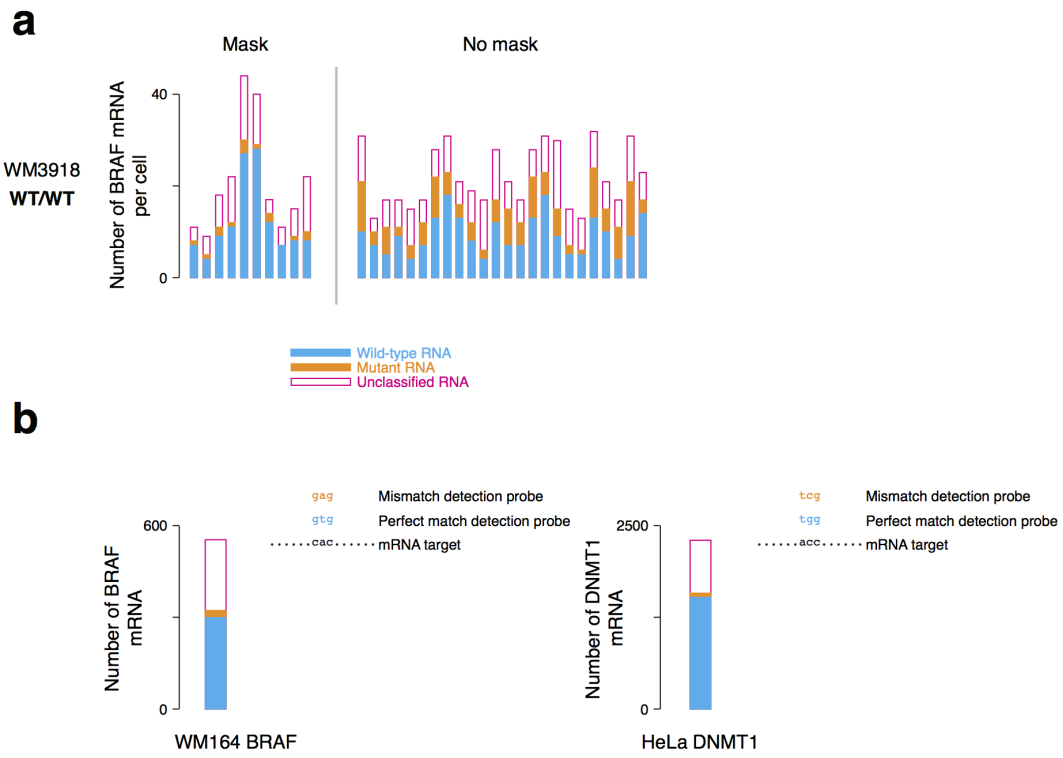

Figure 3.6: a. Addition of mask is required for proper discrimination of single nucleotide variant targets. We probed the WM3918 cell line, which is homozygous for the wild-type allele, with probes targeting both the mutant and wild type allele, either with or without the mask (left, right, respectively). Each bar represents the mRNA counts from a single cell. We found that in the presence of mask, the vast majority of transcripts are wild-type, whereas without mask, a large fraction of the mutant probe spuriously bound to the target. b. Other targets also showed single base mismatch discrimination. We targeted sequences as shown with both perfect match and mismatch detection probes, and found that the perfect match probe was far more likely to bind.

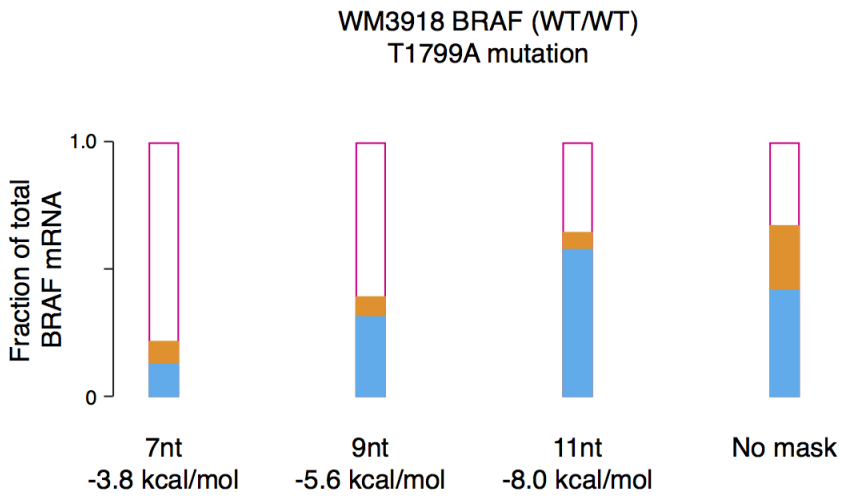

Figure 3.7: Changing the toehold length can change the detection efficiency without dramatically increasing off-target binding. Toehold length is in nucleotides, with the total probe length remaining constant (toehold length changed by changing the mask probe length). With no mask there is dramatically reduced target discrimination and overall detection efficiency saturates around 67%. We computed the free energy change of the toehold binding (given in kcal/mol) using the definition from [87]

3.2.3. Population and single-cell allelic imbalance

Our method for detecting SNVs on RNA molecules enabled us to measure differences in the number of mRNA derived from the maternal vs. paternal copies of a gene, both in the cell population overall and at the single cell level. We explored these possibilities using the GM12878 cell line, for which complete genetic phase information is available[1], making it ideal for studies involving allele-specific expression[27, 67]. We first examined cell population-level imbalances in maternal vs. paternal transcript abundance. We found that the gene DNMT1 displayed no imbalance, whereas EBF1 and SUZ12 had more mRNA from the paternal chromosome (Fig. 3.8; see Fig. 3.9a for number of mRNA one must classify in order to determine that there is an imbalance). Consistent with our findings, a previous study has also found an allelic imbalance in the expression of EBF1 in a similar cell line[29].

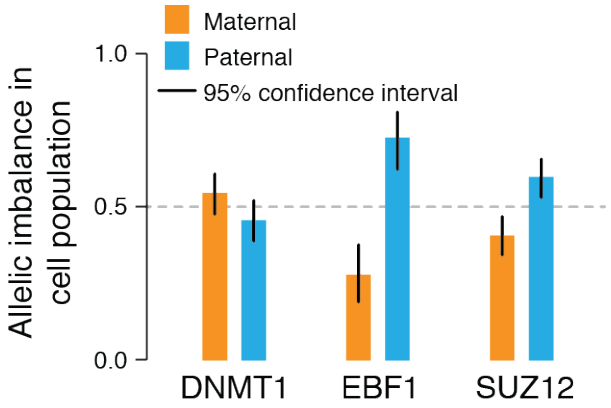

Figure 3.8: We quantified allelic imbalance in the population of the indicated genes by measuring the probability that a transcript comes from either the maternal or paternal allele. Error bars reflect 95% confidence intervals on counting statistics plus an 8 percentage-point differential between maternal and paternal detection efficiency; see methods for details

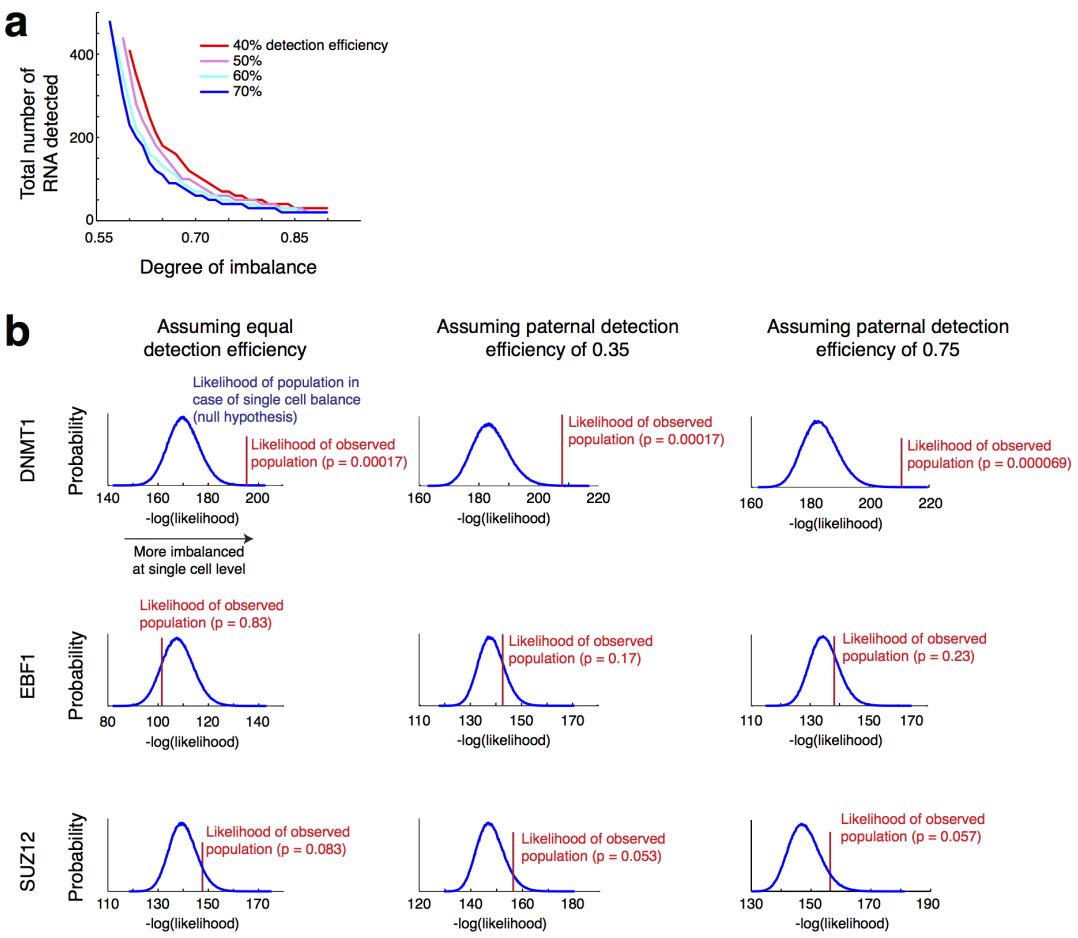

Figure 3.9: a. Using a statistical model, we determined the number of RNAs required to say whether there was an allelic imbalance (for a given actual degree of imbalance). This number is relatively insensitive to the detection efficiency. b. We examined the degree to which changes in the detection efficiency between maternal and paternal detection probes would affect the determination of the presence of single cell imbalance. We found that even very large changes in the detection efficiencies would qualitatively similar conclusions. This is because single cell imbalance manifests as a deviation from the average, thereby making it insensitive to parameters governing the determination of the average itself.

While the cell population average gives us the average imbalance between the maternal and paternal copies of the gene, our method allows us to look for deviations from this average at the single cell level, which would manifest themselves as abnormally large proportions of maternal or paternal transcripts (Fig. 3.10a). In order to quantify the degree of deviation from the average, we took a population of cells and calculated the probability of observing the imbalances detected in that cell population. The null hypothesis is that each transcript in a given cell has a probably of being maternal or paternal equal to that of the cell population average. We found that while DNMT1 displayed allelic balance at the cell population level, a significant number of individual cells deviated from this average ($p = 0.00017$) (Fig. 3.10b). In contrast, while EBF1 and SUZ12 showed imbalance at the cell population level, single cells did not deviate significantly from the average. We note that these imbalances are insensitive to detection efficiency (Fig. 3.9b) and that our analytical method is agnostic as to whether the single cell imbalances are stochastic[16], epigenetic[29] or even genetic in origin.

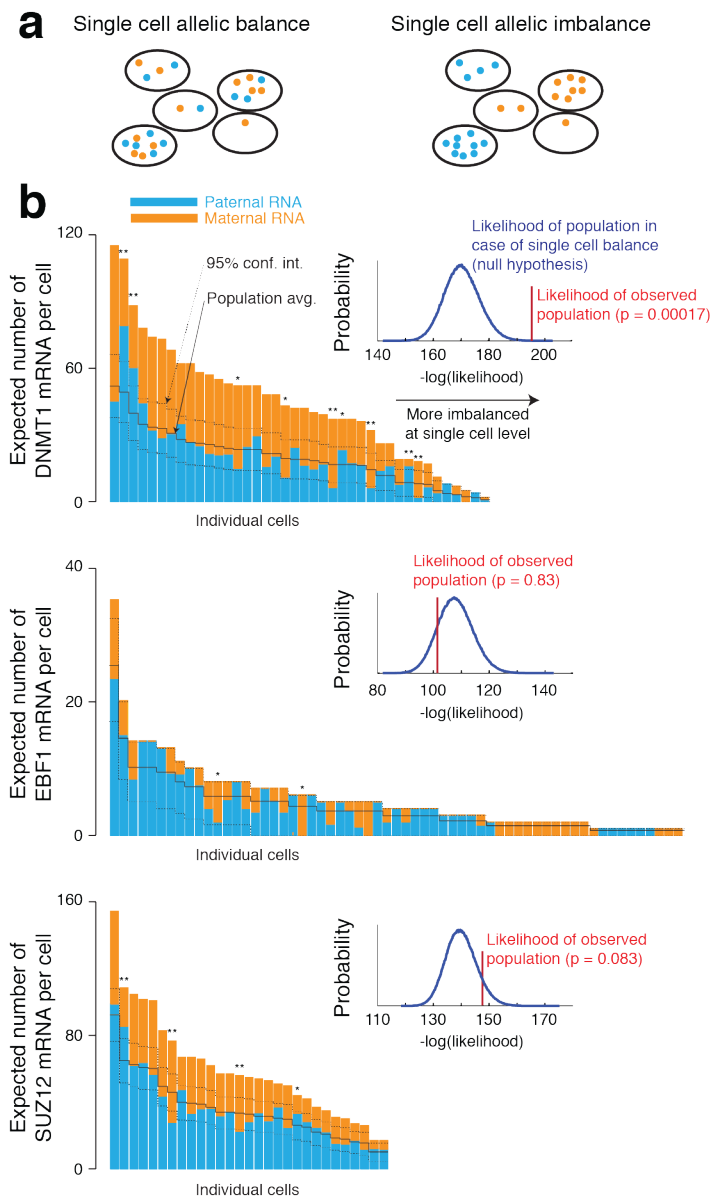

Figure 3.10: a. Diagram of single cell allelic balance and imbalance. b. Allelic imbalance in single cells. The solid black midline represents the average imbalance across cells (from a). The dashed black lines shows the 95% confidence interval on the imbalance for each cell with the null hypothesis that the probability of an RNA being maternal or paternal is independent of which cell it is in. The inset shows the likelihood of the observed population imbalance (red) compared to that of the null model (blue); see methods for details. Note that for EBF1, 90% of cells expressed zero transcripts, so we excluded those cells from the figure. Each sample shown is one of a set of at least two biological replicates. ** represents cells with a p-value below 0.05, and * represents a p-value below 0.10 (p-value defined in methods and Supplementary Note).

3.2.4. Parent-specific chromosome paints

Another application of our method is to distinguish transcription from the maternal vs. paternal chromosomes *in situ*. In previous work[43], we developed a set of probes targeting introns of a set of 31 genes along chromosome 19, yielding an RNA-based chromosome "paint". We used a database of SNVs in GM12878 cells[67] to find SNVs in the introns of these genes and created a set of detection probes designed to label 15 of the introns from the paternal chromosomes in a distinct color. In this manner, we were able to visualize and classify chromosomes as maternal or paternal *in situ* (Fig. 3.11). These results demonstrate that our method is applicable to introns, enabling us to measure allele-specific transcriptional activity directly. Moreover, localization of signals to specific chromosomes can allow one to determine whether a new SNV is on the maternal or paternal copy of the chromosome, or even whether transcription of a gene with no SNV is coming from the maternal or paternal chromosome..

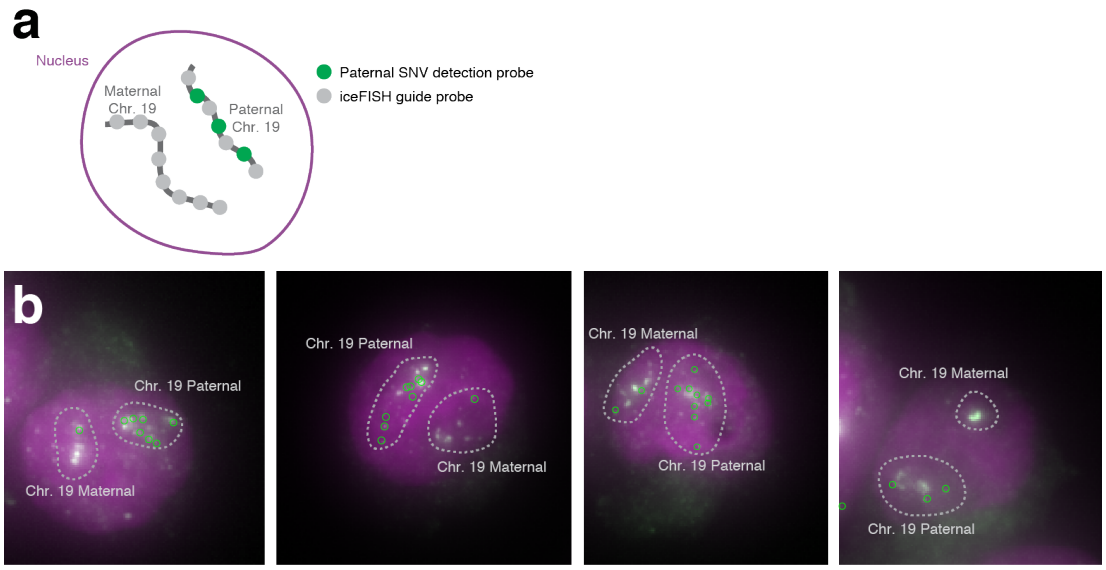

Figure 3.11: a. Illustration of the chromosome detection method. We designed iceFISH probes[43] that target chromosome 19 and SNV detection probes targeting 19 SNPs within 15 of these genes on the paternal chromosome (methods). b. Example images showing the two copies of chromosome 19 (gray dashed regions) with the computationally identified co-localized detection probes labeled with green circles. Each sample shown is one of a set of at least two biological replicates.

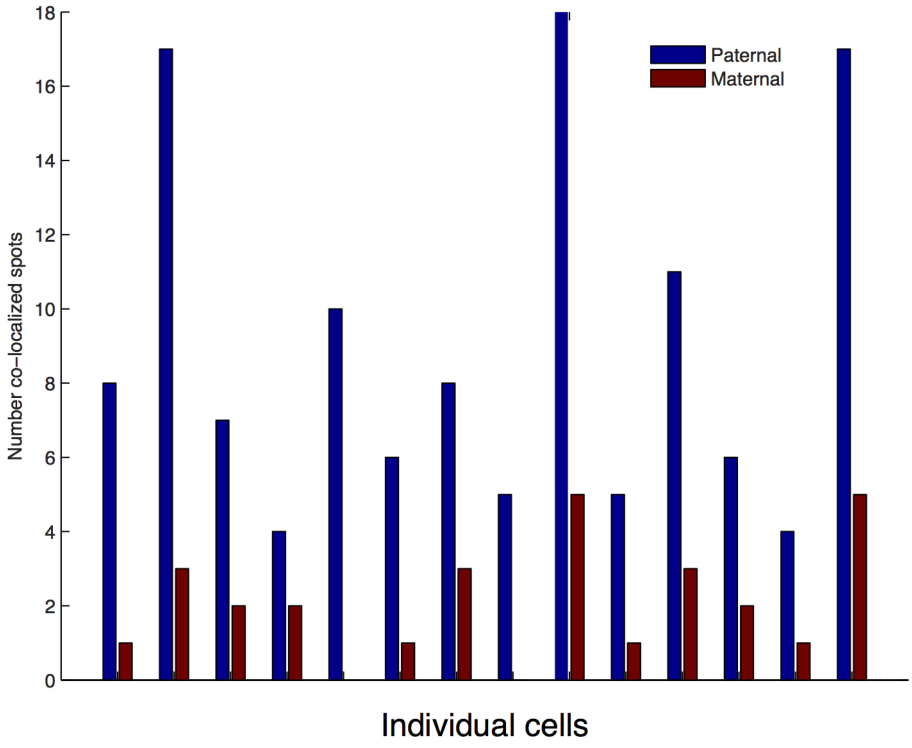

Figure 3.12: Number of co-localized spots in the experiment described in Fig. 3.11. We applied SNV RNA FISH detection probes targeting introns on the paternal copy of chromosome 19; we simultaneously labeled these introns with guide probes targeting several introns on chromosome 19 as described in the methods. We spatially isolated individual chromosomes in individual cells and then counted the number of paternal SNVs detected on each chromosome (by co-localization with the guide intron probes). We designated the chromosome with more paternal SNVs as being the paternal copy.

3.3. Discussion

Here, we have demonstrated the ability to distinguish SNVs with high efficiency and specificity at the level of individual RNA molecules. Our method is simple to implement and uses readily available reagents. It is possible that using different nucleic acid chemistries for the detection probe could help increase the detection efficiency while also reducing off-target binding, which may make application of this method more difficult for more abundant RNA species. Aside from diagnostic applications, particularly in genotyping single cells *in situ*, our method has the potential to reveal new insights into allele-specific effects in gene expression. Classic examples include gene imprinting[2], but genome wide association studies have highlighted the need for tools to quantify the expression of genes in an allele-specific manner to show how disease-associated SNVs affect transcription, and methods like ours will help bridge that gap.

3.4. Materials and Methods

3.4.1. Cell culture and fixation

We grew melanoma cell lines with the BRAF V600E mutation, SK-MEL-28 (Mut/Mut, ATCC cat no HTB-72), WM3918 (WT/WT) and WM398b & WM9 (both WT/Mut) (gifts from the lab of Meenhard Herlyn, Wistar Institute, genotypes verified by the Herlyn lab), using the recommended cell culture guidelines for each line. The SK-MEL-28 cell line is documented as homozygous for the V600E mutation, but our experiments revealed that a subpopulation of the cells was heterozygous (Fig. 3.13), which we excluded from further analysis. We grew the cells on Lab-Tek chambered coverglass (Lab-Tek) and fixed the cells following the protocol in Raj et al. Nat Meth 2008[62]. We obtained GM12878 cells from the Coriell Cell Repositories and grew them according to guidelines. We stored fixed cells in 70% ethanol at 4C for up to 4 weeks before hybridization; the duration of storage did not affect hybridization efficiency. All cells were negative for mycoplasma contamination as verified by DAPI imaging.

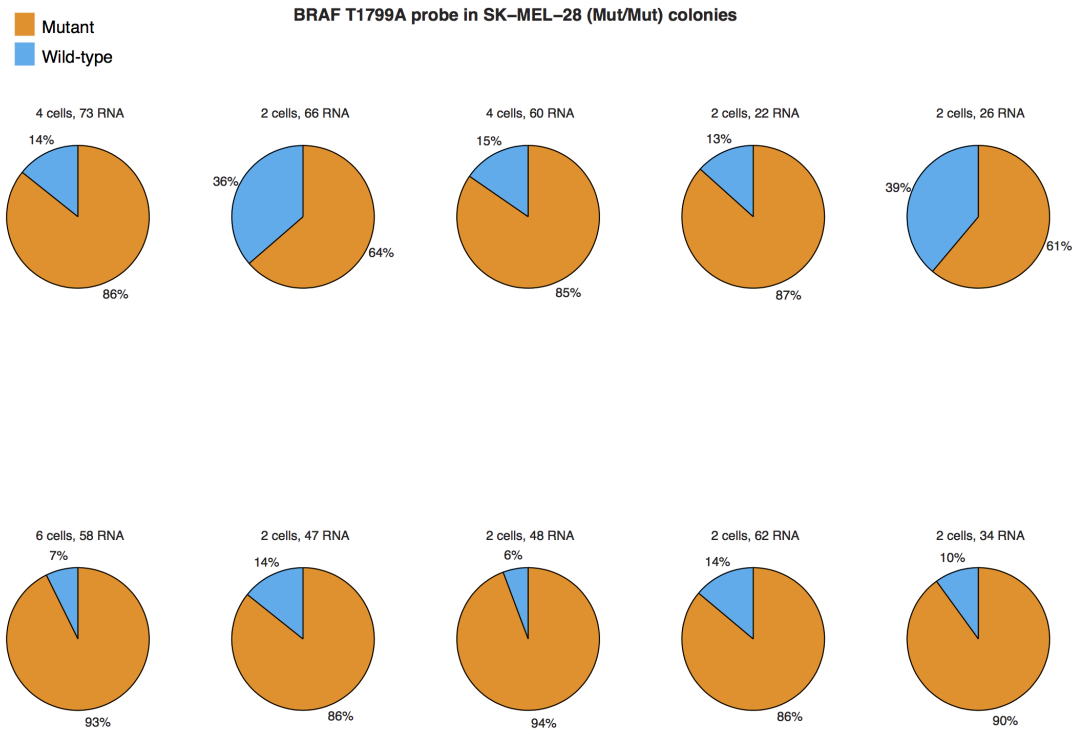

Figure 3.13: We plated out the SK-MEL 28 cells at a low density and grew them until we had groups consisting of 2-6 recently divided cells. Within these groups, the cells can be regarded as genetically identical because they have a common recent ancestor. We found two of the ten groups analyzed had an unusually large number of wild-type transcripts (36% and 39%) compared to the other groups, indicating that those cells were likely to be heterozygous.

3.4.2. Probe design and synthesis

We designed detection probes with the single nucleotide difference located at the 5th base position from their 5 end. We adjusted the total length of the detection oligonucleotide to ensure the hybridization energy with target RNA was similar or greater than that of the guide probe oligonucleotides[87]. We designed mask oligonucleotides complementary to the detection probes that, upon binding to the detection probe, left a 6 to 11 base toehold regions available to target RNAs regions with SNVs. We conjugated guide probe oligonucleotides to ATTO 488 dye (ATTO-TEC) and we interchangeably used Cy3 and Cy5 (GE Healthcare) dyes for the SNV detection probes. We did not observe any changes to detection efficiency when swapping the Cy3/Cy5 dyes. Our choice of dyes was influenced by dye stability after a post-fixation step described below and affinities of some dyes that cause excessive binding to the incorrect target. We listed the detection, mask, and guide probe sequences in the supplementary information.

3.4.3. RNA FISH

We performed RNA fluorescence *in situ* hybridization (FISH) as outlined in Raj et al. Nat Meth 2008[62] with some modifications as outlined presently, most notably a postfixation step after the hybridization to help prevent probe dissociation during imaging. Firstly, our hybridization buffer consisted of 10% dextran sulfate, 2x saline-sodium citrate (SSC) and 10% formamide[49]. We performed the hybridization as before, using final concentrations of 5nM for the guide probe, wild-type and mutant detection probe, and 10nM for the mask, thereby leading to 1:1 mask:detection oligonucleotide ratios. We let the hybridization proceed overnight at 37C. For Lab-Tek chamber samples, we used 50μL hybridization solution with a coverslip and included a moistened paper towel to prevent excessive evaporation in parafilmed culture dish. For suspension cells, we used 50uL hybridization solution in a 1.5mL Eppendorf tube. In the morning, we washed the samples twice with a 2X SSC and 10% formamide wash buffer. Suspension cells included 0.1% Triton-X in the wash buffer. We then performed a postfixation step using 4% formaldehyde in 2X SSC for 30 minutes at

25C to crosslink the detection probes and thereby prevent dissociation during imaging, followed by 2 washes in 2X SSC. We then put the cells into anti-fade buffer with catalase and glucose oxidase[62] to prevent photobleaching of Cy5 during imaging. For the chromosome 19 paints, we used probes against introns of 31 genes with 12-16 oligonucelotides per gene, each at 0.1nM, for the guide probe in Cy3[43]. We added maternal and paternal probes, in Cy3 and Cy5 respectively, for 19 SNV sites within 15 of the chromosome 19 paint genes, added masks, and performed hybridization as described above.

3.4.4. Imaging

We took all our images on a Leica DMI600B automated widefield fluorescence microscope equipped with a 100x Plan Apo objective, a Pixis 1024BR cooled CCD camera and a Prior Lumen 220 light source. We took image stacks in each fluorescence channel consisting of sets of images separated by 0.35μm. Our exposure times were 1500ms and 3500ms for guide and detection probes respectively. We used longer exposure times for the wild-type and mutant detection probes owing to the low signal afforded by single dye molecules relative to the dozens of fluorophores typically used in the guide probes. Step-wise photobleaching traces demonstrated that we were indeed detecting single dyes (Fig. 3.14).

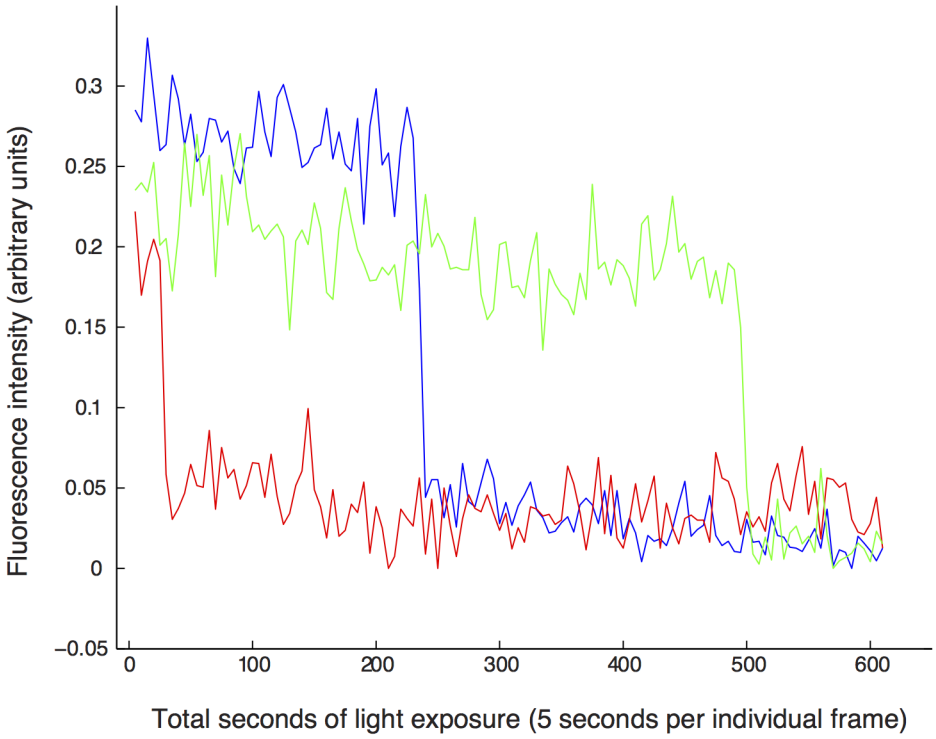

Figure 3.14: Spot fluorescence traces over time show stepwise photobleaching. To show that each of the fluorescence spots we found in the detection probes was indeed from a single fluorescent molecule, we looked for stepwise decrease in fluorescence upon repeated exposure. Here, we show three representative fluorescence traces of spots observed in the fluorescence channel corresponding to the detection probe target- ing SUZ12 mRNA. Each exposure lasted for 5 seconds, and the x-axis shows the total number of seconds of exposure. The y-axis shows the fluorescence intensity in arbitrary units, but note that we did not normalize each trace individually; thus, the fact that the intensities are similar at time zero provides further evidence that the spots correspond to single fluorescent molecules.

3.4.5. Image analysis

Our image analysis consisted of first manually segmenting the cells using custom software written in MATLAB (Mathworks), after which we identified spots using algorithms similar to those we described in Raj et al. Nat Meth 2008. We chose relatively permissive thresholds for spots in the channels for the mutant and wild-type detection probe channels, thereby trying to avoid false negatives due to overly stringent criteria for spot detection. Once we had located the spots, we then denoted spots as colocalized if two spots from different fluorescence channels were within 4 pixels of each other in order to account for a 2 pixel chromatic aberration in portions of the images from the different channels. In the event of a colocalization event in which spots appeared in more than 2 channels or in which more than 2 spots were in the neighborhood of the guide probe, we used colocalized pairs in the rest of the image to correct for shifts between channels, thereby allowing us to tighten the colocalization window.

3.4.6. Bioinformatic analysis of GM12878 to find SNPs

We used the RefSeq gene model to define the genomic coordinates of introns and exons for genes of interest. We queried these regions in the published diploid genome of GM12878 (http://alleleseq.gersteinlab.org) (version Dec 16, 2012) to locate the heterozygous SNPs, and extracted those sequences for probe design.

3.4.7. Statistical analysis of allele-specific expression

We performed a statistical analysis of allele-specific expression in two stages. In the first stage, we combined data from all cells to find evidence for population-level allelic imbalance. Using this data, we computed the mean detection efficiency of the detection probes as well as the average percentage of detected transcripts that originated from the maternal or paternal allele of the gene in question. We computed confidence intervals on these percentages by combining a. the error associated with the number of observations itself (modeled as a multinomial distribution and computed to 95% confidence) and b. the error associated

with uncertainty in the detection efficiency. For the latter, we assumed that the detection efficiency could differ by at most 8% from each other; for example, if the average detection efficiency was 55%, we would compute the imbalance with 59%/51% detection efficiencies, first in favor of maternal and then paternal. Empirically, we have found that our detection efficiencies tend to remain in the 50%-60% range, and so this procedure will ensure that at least one of the detection efficiencies remains in this range. Combining these two sources of error, our error bars likely reflect a greater than 95% confidence interval.

In the next stage, we used the observed detection efficiency and population-level imbalance to ascertain the degree to which single cells displayed allelic imbalance. Our null hypothesis is that each RNA produced at any given period of time would be independently chosen to come from either the maternal or paternal allele at the same frequency as at the population level; in other words, there are no "runs" of maternal or paternal-origin transcripts in single cells. Given this null model, we then computed the probability density of possible observed imbalances for each cell given the population-level imbalance. We used these densities to compute single cell likelihoods for our observed counts and calculated the total likelihood of the population by taking the product of the single cell likelihoods. We then compared the likelihood of our observations to the likelihood one might expect from the null hypothesis by generating 1,000,000 *in silico* counts for each cell based on our multinomial model and computing the likelihood of these observations to generate a distribution of likelihoods corresponding to the null hypothesis. In order to reject the null hypothesis and show that the population of single cells displays cell-to-cell allelic imbalance, we then computed the percentage of the null hypothesis likelihoods that were more extreme than our observation.

CHAPTER 4 : Outlook and applications for the chromosome and allele-specific expression measurements

4.1. Implications of single chromosome profiling and applications of iceFISH

In our study [43] we definitively showed that translocations can lead to chromosome-wide changes to transcriptional activity. Our measurements of transcriptional activity on normal versus translocated copies of chromosomes are impossible to obtain with previous methods. The ability to discriminate the chromosomal source of transcription comes from the use of microscopy and spatially assigning activity to chromosomes within single cells. The physical mechanism that produces the difference in transcriptional activity is still unknown.

The possibility of three-dimensional conformation or nuclear positioning of chromosomes producing these expression changes seems unlikely given our observations (Fig. 2.19). This is in contrast to previous findings that attributed changes in transcription to the positioning of chromosomes and genes within the 3D nucleus[52, 33]. The statistical significance of their findings required high numbers of observations (a larger value of n) and the differences were subtle. Speculating on alternative possibilities that could lead to their conclusions, there may be bias in producing a coordinate system for asymmetrical nuclei and then applying this system across cell types. There are gross structural differences between the nuclei of normal and cancer tissues[90] that could introduce systematic error in comparing nuclear positioning measurements between cell types. One outstanding example is how these coordinate systems developed for nuclear positioning treat the nucleus as a homogenous volume. We feel this is a fundamentally erroneous assumption since the largest substructures of the nucleus, nucleoli, are ignored. Nucleoli, which are the site of ribosomal RNA production and assembly of ribosomal subunits come in different sizes and numbers in each nucleus[34]. Cancer cells can even be identified by their nucleoli that differ from healthy tissue[90]. Overall, these concerns about the measurement methods in published work and the lack of correlation in our studies leads us to dismiss the hypothesis of nuclear positioning regulating transcription.

Instead of 3D positioning, we hypothesize that the translocated portion of chromosome 19 has an epigenetic profile that differs from its intact copy. Translocated chromosomes are created by non-homologous end joining after a set of DNA double-strand breaks. The DNA repair process involves a great deal of chromatin remodeling[66] and it is possible that a translocation event leads to the replacement or modification of histones chromosome-wide. Therefore, we would like to measure the DNA methylation and histone modifications present on the individual copies. Unfortunately, sequence similarity prevents discriminating the chromosome copies since high-throughput sequencing techniques produce short DNA reads that map to a reference genome. To overcome this, we utilized flow assisted cell sorting to separate chromosome species by size[12] (normal 19 is smaller than derivatives t(6;19) and t(13;19)). Once separated, we can probe the material using chromatin immunoprecipitation coupled with DNA sequencing (ChIP-seq). This is an ongoing project in the Raj Lab currently in the sequencing stages. This technique will provide nucleotide resolution of the epigenetic profiles and hopefully provide an answer to what makes translocated versions of chromosomes behave differently than the intact copies.

Using 5-base colors and sophisticated image processing for labeling 20 different genes, the iceFISH assay described in [43] is too complex for rapid diagnostic applications. As a proof of concept, we put together a version of the iceFISH assay that provides simpler interpretation of signals. By labeling two halves of chromosome 19 in two distinct colors, one for each segment split by the site of translocation, we can karyotype interphase cells manually by eye (see Fig. 4.1). This data is simple enough that the analysis could be automated using computational image processing routines that identify overlapping or separated 'blobs' for the two halves of chromosome 19. Since this version of the method uses only a couple fluorescent colors, the remaining third or fourth colors could probe for an RNA biomarker associated with a certain type of cancer or other diseased tissue. Looking forward, an iceFISH assay designed to detect recurring translocations and associated RNA biomarkers has the potential to provide high confidence diagnostics with opportunities for automated analysis.

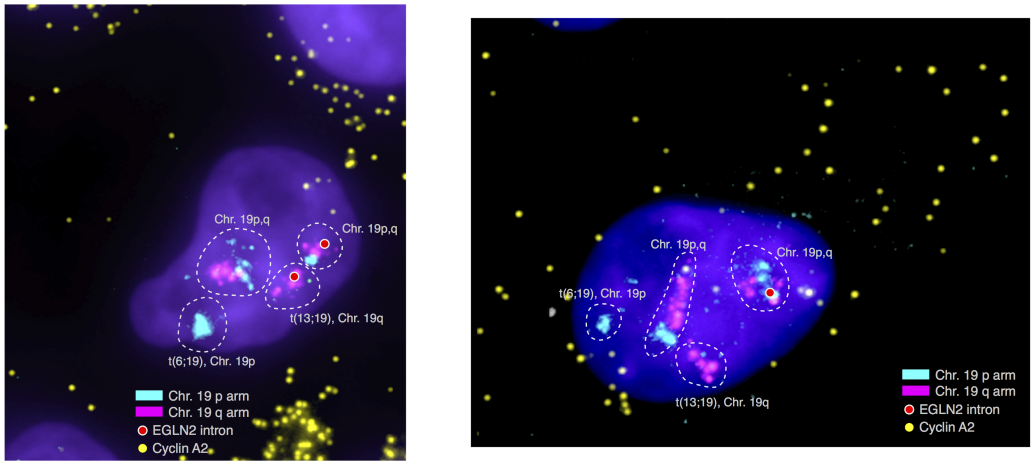

Figure 4.1: Applying the iceFISH assay to specifically detect a recurring, known translocation, karyotyping chromosomal translocations becomes simple to identify by eye. The assay can be combined with RNA expression probes for correlated biomarkers, thus providing an even higher level of confidence in the diagnosis

4.2. Genetic variation within single cells measured by SNP FISH

SNP FISH is a novel assay capable of detecting single nucleotide differences on individual RNA molecules within single cells. We are able to discriminate single nucleotide differences using a masked-probe strategy that leaves only a short region of an oligonucleotide probe accessible in the search for a complementary target. This portion of the probe is called the toehold and the short length maintains specificity when it encounters a destabilizing single-base mismatch. After finding the target, the complementary mask oligonucleotide is displaced and the probe can stably bind the RNA target. By combining this strategy with a tiled probe RNA FISH method, we overcame false positives other single probe strategies suffer from and achieved quantitative, allele-specific measurements of transcription.

SNP FISH opens up many new questions one can study in the field of transcriptional regulation. Without the ability to discriminate between alleles, it is difficult to investigate interactions between genes on the same chromosome. These *cis* interactions could be regulated by control regions of the DNA near the affected genes and would display RNA levels correlated in parental origin. Allelic discrimination would also be useful in studying diseases caused by mutations in the RNA such as many cancers. An open question is how a cancer becomes resistant after exposure to therapies targeting the affected allele. One possibility is that cancer cells increase levels of the problematic allele while maintaining overall RNA levels constant for the affected gene. SNP FISH can quantify the single cell allelic ratio of RNA in cancer cells before and after obtaining resistance to test this hypothesis.

An application of SNP FISH we are currently pursuing is imprinted transcription where one parental allele of a gene is chosen during development as the sole source of RNA. Imprinted genes can lose their parent-specific expression due to a mutation in an imprinting control region. Measurements from the cell population leave open the question of whether the observed biallelic expression is now coming from each cell randomly choosing an allele or all cells transcribing from both alleles. Disruptions in the regulation of imprinted expression have serious disease implications and imprinting is often specific to particular tissues or cell

types[81]. Therefore, the ability to measure gene expression within native tissue with cells identified through RNA biomarkers is essential, making RNA FISH a well-suited strategy to characterize this behavior.

Beyond allelic imbalances caused by deterministic transcriptional regulation, the stochastic nature of transcription is another place to apply SNP FISH. Some genes express with "bursts" of new RNA at infrequent intervals[60]. Provided a short RNA half-life and large enough burst-sizes, a cell could toggle between parental copies of a gene, making it effectively homozygous for each allele at certain times. The SNP FISH assay can quantify the maternal or paternal ratio of RNA on a per cell basis. This data could help develop computational models describing the dynamics of "bursty" gene expression. We already observed examples of this class of genes, in the expression of dual specificity phosphatase 6 (Dusp6) in mouse embryonic fibroblast cells (Fig. 4.2).

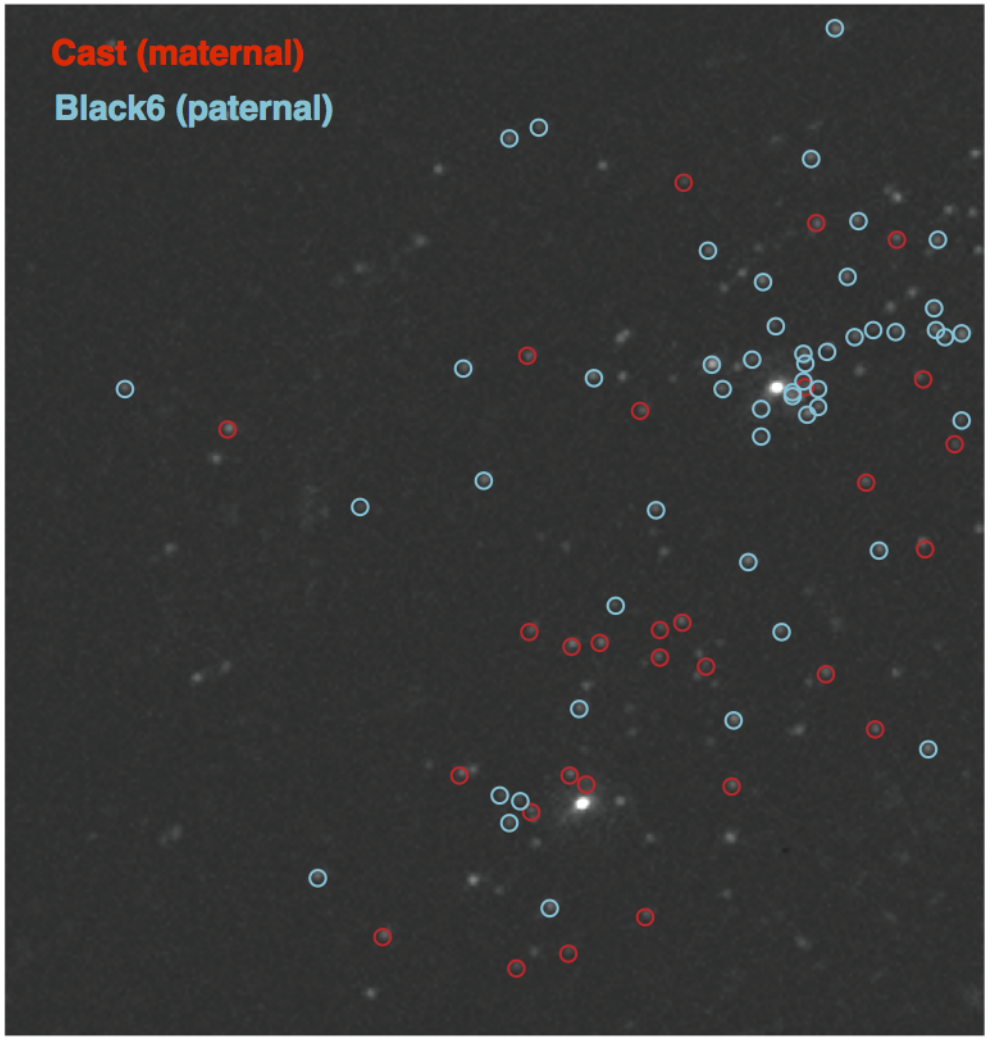

Figure 4.2: Applying SNP FISH in a mouse embryonic fibroblast line derived from a CAST (chromosome 7 and 10) female mouse crossed with a Black6 male, the gene Dusp6 shows transcriptional bursting from transcription sites (bright white spots) with allele-specific labeling of the individual RNA molecules. The Black6 allele shows a cluster of newly transcribed RNA coming from the upper transcription site.

Hopefully, new insights from example applications of SNP FISH described above will produce more interest in how our cells manage transcribing RNA with two copies of each gene.

4.3. Final conclusions

Both the iceFISH and SNP FISH assays provide new details of transcription by discriminating which chromosome copy RNA comes from in single cells. We used iceFISH to describe how genes on a translocated copy of chromosome, a hallmark of cancer, can have drastically different transcriptional behaviors than the copies on intact chromosomes. We did not find any correlation between these transcriptional changes and the spatial positioning of chromosomes or genes within the nucleus. These results force us to rethink how translocations affect transcription and lead to disease. We then built SNP FISH to discern single base differences on RNA *in situ* and used it to quantify allelic levels of RNA in single cells. These tools have the potential to open up new avenues for research in the regulation of gene expression, serve as effective diagnostic tools, and ultimately deepen our understanding of biology to improve medicine.

: Bibliography

[1] Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, October 2010.

[2] Lara K Abramowitz and Marisa S Bartolomei. Genomic imprinting: recognition and marking of imprinted loci. *Current opinion in genetics & development*, 22(2):72–78, 2012.

[3] Kashif Ahmed, Hesam Dehghani, Peter Rugg-Gunn, Eden Fussner, Janet Rossant, and David P Bazett-Jones. Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PloS one*, 5(5):e10531, 2010.

[4] J A Birchler, U Bhadra, M P Bhadra, and D L Auger. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Developmental biology*, 234(2):275–288, 2001.

[5] James A Birchler. Reflections on studies of gene expression in aneuploids. *The Biochemical journal*, 426(2):119–123, 2010.

[6] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, and Thomas Cremer. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, 3(5):e157, 2005.

[7] Miguel R Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS biology*, 4(5):e138, 2006.

[8] Michael Bulger and Mark Groudine. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–39, February 2011.

[9] N P Carter. Cytogenetic analysis by chromosome painting. *Cytometry*, 18(1):2–10, 1994.

[10] Jonathan R Chubb, Tatjana Trcek, Shailesh M Shenoy, and Robert H Singer. Transcriptional pulsing of a developmental gene. *Current biology : CB*, 16(10):1018–25, May 2006.

[11] Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, NY)*, 322(5909):1845–1848, 2008.

[12] L Scott Cram, Carolyn S Bell, and John J Fawcett. Chromosome sorting and genomics. *Methods in cell science : an official journal of the Society for In Vitro Biology*, 24(1-3):27–35, January 2002.

[13] T Cremer and C Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews Genetics*, 2(4):292–301, 2001.

[14] Josée Dostie and Job Dekker. Mapping networks of physical interactions between genomic elements using 5C technology. *Nature protocols*, 2(4):988–1002, 2007.

[15] C Duart-Garcia and M H Braunschweig. The Igf2as transcript is exported into cytoplasm and associated with polysomes. *Biochemical genetics*, 51(1-2):119–130, February 2013.

[16] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science (New York, NY)*, 297(5584):1183–1186, 2002.

[17] K Leigh Eward, Matthew N Van Ert, Maureen Thornton, and Charles E Helmstetter. Cyclin mRNA stability does not vary during the cell cycle. *Cell cycle (Georgetown, Tex.)*, 3(8):1057–1061, 2004.

[18] A M Femino, F S Fay, K Fogarty, and R H Singer. Visualization of single RNA transcripts in situ. *Science (New York, NY)*, 280(5363):585–590, 1998.

[19] Anne C Ferguson-Smith. Genomic imprinting: the emergence of an epigenetic paradigm. *Nature reviews. Genetics*, 12(8):565–75, August 2011.

[20] Carmelo Ferrai, Sheila Q Xie, Paolo Luraghi, Davide Munari, Francisco Ramirez, Miguel R Branco, Ana Pombo, and Massimo P Crippa. Poised transcription factories prime silent uPA gene prior to activation. *PLoS biology*, 8(1):e1000270, 2010.

[21] J E Ferrell and E M Machleder. The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. *Science (New York, NY)*, 280(5365):895–898, 1998.

[22] L E Finlan, D Sproul, I Thomson, S Boyle, E Kerr, P Perry, B Ylstra, J R Chubb, and W A Bickmore. Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS genetics*, 4(3):e1000039, March 2008.

[23] Peter Fraser and Wendy Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143):413–417, 2007.

[24] R. Fremeau, J. Lundblad, D. Pritchett, J. Wilcox, and J. Roberts. Regulation of pro-opiomelanocortin gene transcription in individual cell nuclei. *Science*, 234(4781):1265–1269, December 1986.

[25] Alexandre Gaspar-Maia, Adi Alajem, Eran Meshorer, and Miguel Ramalho-Santos. Open chromatin in pluripotency and reprogramming. *Nature reviews Molecular cell biology*, 12(1):36–47, 2011.

[26] Daniel Gerlich, Joël Beaudouin, Bernd Kalbfuss, Nathalie Daigle, Roland Eils, and Jan Ellenberg. Global chromosome positions are transmitted through mitosis in mammalian cells. *Cell*, 112(6):751–764, 2003.

[27] Jason Gertz, Katherine E Varley, Timothy E Reddy, Kevin M Bowling, Florencia Pauli, Stephanie L Parker, Katerina S Kucera, Huntington F Willard, and Richard M Myers. Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS genetics*, 7(8):e1002228, 2011.

[28] Pamela K Geyer, Michael W Vitalini, and Lori L Wallrath. Nuclear organization: taking a position on gene expression. *Current opinion in cell biology*, 2011.

[29] Alexander Gimelbrant, John N Hutchinson, Benjamin R Thompson, and Andrew Chess. Widespread monoallelic expression on human autosomes. *Science (New York, NY)*, 318(5853):1136–1140, November 2007.

[30] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, December 2005.

[31] Christopher Gregg, Jiangwen Zhang, Brandon Weissbourd, Shujun Luo, Gary P Schroth, David Haig, and Catherine Dulac. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science (New York, N.Y.)*, 329(5992):643–8, August 2010.

[32] J Gribnau, E de Boer, T Trimborn, M Wijgerde, E Milot, F Grosveld, and P Fraser. Chromatin interaction mechanism of transcriptional control in vivo. *The EMBO journal*, 17(20):6020–7, October 1998.

[33] Louise Harewood, Frédéric Schütz, Shelagh Boyle, Paul Perry, Mauro Delorenzi, Wendy A Bickmore, and Alexandre Reymond. The effect of translocation-induced nuclear reorganization on gene expression. *Genome research*, 20(5):554–564, 2010.

[34] AS Henderson, D Warburton, and KC Atwood. Location of ribosomal DNA in the human chromosome complement. *Proceedings of the ...*, 69(11):3394–3398, 1972.

[35] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999–1003, 2012.

[36] Gavin Kelsey and Marisa S Bartolomei. Imprinted Genes . . . and the Number Is? *PLoS genetics*, 8(3):e1002601, 2012.

[37] Katrin Küpper, Alexandra Kölbl, Dorothee Biener, Sandra Dittrich, Johann von Hase, Tobias Thormeyer, Heike Fiegler, Nigel P Carter, Michael R Speicher, Thomas Cremer, and Marion Cremer. Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma*, 116(3):285–306, 2007.

[38] Chatarina Larsson, Ida Grundberg, Ola Söderberg, and Mats Nilsson. In situ detection and genotyping of individual mRNA molecules. *Nature methods*, 7(5):395–397, 2010.

[39] J. Lawrence, R. Singer, and J. McNeil. Interphase and metaphase resolution of different distances within the human dystrophin gene. *Science*, 249(4971):928–932, August 1990.

[40] M. P. Lee, M. R. DeBaun, K. Mitsuya, H. L. Galonek, S. Brandenburg, M. Oshimura, and A. P. Feinberg. Loss of imprinting of a paternally expressed transcript, with antisense orientation to KVLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proceedings of the National Academy of Sciences*, 96(9):5203–5208, April 1999.

[41] Derek Lemons and William McGinnis. Genomic evolution of Hox gene clusters. *Science (New York, N.Y.)*, 313(5795):1918–22, September 2006.

[42] Martin J Lercher, Araxi O Urrutia, Adam Pavlícek, and Laurence D Hurst. A unification of mosaic structures in the human genome. *Human molecular genetics*, 12(19):2411–2415, 2003.

[43] Marshall J Levesque and Arjun Raj. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nature methods*, 10(3):246–248, 2013.

[44] J M Levsky. Single-Cell Gene Expression Profiling. *Science (New York, NY)*, 297(5582):836–840, 2002.

[45] Gang Li, Jae Hoon Bahn, Jae-Hyung Lee, Guangdun Peng, Zugen Chen, Stanley F Nelson, and Xinshu Xiao. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Research*, 40(13):e104, 2012.

[46] Qingge Li, Guoyan Luan, Qiuping Guo, and Jixuan Liang. A new class of homogeneous nucleic acid probes based on specific displacement hybridization. *Nucleic Acids Research*, 30(2):E5, 2002.

[47] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, NY)*, 326(5950):289–293, October 2009.

[48] Stavros Lomvardas, Gilad Barnea, David J Pisapia, Monica Mendelsohn, Jennifer Kirkland, and Richard Axel. Interchromosomal interactions and olfactory receptor choice. *Cell*, 126(2):403–13, July 2006.

[49] Eric Lubeck and Long Cai. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature methods*, 9(7):743–748, 2012.

[50] M Macville, E Schröck, H Padilla-Nash, C Keck, B M Ghadimi, D Zimonjic, N Popescu, and T Ried. Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. *Cancer research*, 59(1):141–150, 1999.

[51] Julio Mateos-Langerak, Manfred Bohn, Wim de Leeuw, Osdilly Giromus, Erik M M Manders, Pernette J Verschure, Mireille H G Indemans, Hinco J Gierman, Dieter W Heermann, Roel van Driel, and Sandra Goetze. Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, 106(10):3812–3817, 2009.

[52] Karen J Meaburn, Prabhakar R Gudla, Sameena Khan, Stephen J Lockett, and Tom Misteli. Disease-specific gene repositioning in breast cancer. *The Journal of Cell Biology*, 187(6):801–812, 2009.

[53] Tom Misteli. Beyond the sequence: cellular organization of genome function. *Cell*, 128(4):787–800, 2007.

[54] Jennifer A Mitchell and Peter Fraser. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes & development*, 22(1):20–25, 2008.

[55] Céline Morey, Clémence Kress, and Wendy A Bickmore. Lack of bystander activation shows that localization exterior to chromosome territories is not sufficient to up-regulate gene expression. *Genome research*, 19(7):1184–1194, July 2009.

[56] Iris Müller, Shelagh Boyle, Robert H Singer, Wendy A Bickmore, and Jonathan R Chubb. Stable morphology, but dynamic internal reorganisation, of interphase human chromosomes in living cells. *PloS one*, 5(7):e11560, 2010.

[57] H Nawata, G Kashino, K Tano, K Daino, Y Shimada, H Kugoh, M Oshimura, and M Watanabe. Dysregulation of gene expression in the artificial human trisomy cells of chromosome 8 associated with transformed cell phenotypes. *PloS one*, 6(9):e25319, January 2011.

[58] Cameron S Osborne, Lyubomira Chakalova, Karen E Brown, David Carter, Alice Horton, Emmanuel Debrand, Beatriz Goyenechea, Jennifer A Mitchell, Susana Lopes, Wolf

Reik, and Peter Fraser. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10):1065–1071, 2004.

[59] Daan Peric-Hupkes, Wouter Meuleman, Ludo Pagie, Sophia W M Bruggeman, Irina Solovei, Wim Brugman, Stefan Gräf, Paul Flicek, Ron M Kerkhoven, Maarten van Lohuizen, Marcel Reinders, Lodewyk Wessels, and Bas van Steensel. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, 38(4):603–13, May 2010.

[60] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS biology*, 4(10):e309, 2006.

[61] Arjun Raj and Sanjay Tyagi. Detection of individual endogenous RNA transcripts in situ using multiple singly labeled probes. *Methods in enzymology*, 472:365–86, January 2010.

[62] Arjun Raj, Patrick van den Bogaard, Scott A Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature methods*, 5(10):877–879, 2008.

[63] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[64] Anita Rauch, Juliane Hoyer, Sabine Guth, Christiane Zweier, Cornelia Kraus, Christian Becker, Martin Zenker, Ulrike Hüffmeier, Christian Thiel, Franz Rüschendorf, Peter Nürnberg, André Reis, and Udo Trautmann. Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. *American Journal of Medical Genetics Part A*, 140A(19):2063–2074, 2006.

[65] Timothy E Reddy, Jason Gertz, Florencia Pauli, Katerina S Kucera, Katherine E Varley, Kimberly M Newberry, Georgi K Marinov, Ali Mortazavi, Brian A Williams, Lingyun Song, Gregory E Crawford, Barbara Wold, Huntington F Willard, and

Richard M Myers. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research*, 22(5):860–869, 2012.

[66] Dorine Rossetto, Andrew W Truman, Stephen J Kron, and Jacques Côté. Epigenetic modifications in double-strand break DNA damage signaling and repair. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 16(18):4543–52, September 2010.

[67] Joel Rozowsky, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, Nitin Bhardwaj, Mark Rubin, Michael Snyder, and Mark Gerstein. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology*, 7:522, January 2011.

[68] Vijay G Sankaran, Jian Xu, and Stuart H Orkin. Advances in the understanding of haemoglobin switching. *British journal of haematology*, 149(2):181–94, April 2010.

[69] Stefan Schoenfelder, Tom Sexton, Lyubomira Chakalova, Nathan F Cope, Alice Horton, Simon Andrews, Sreenivasulu Kurukuti, Jennifer A Mitchell, David Umlauf, Daniela S Dimitrova, Christopher H Eskiw, Yanquan Luo, Chia-Lin Wei, Yijun Ruan, James J Bieker, and Peter Fraser. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics*, 42(1):53–61, 2009.

[70] Jason M Sheltzer and Angelika Amon. The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends in genetics : TIG*, 27(11):446–453, 2011.

[71] Tobias Straub and Peter B Becker. Dosage compensation: the beginning and end of generalization. *Acute kidney injury: diagnosis and classification of AKI: AKIN or RIFLE?*, 8(1):11, 2007.

[72] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science (New York, N.Y.)*, 332(6028):472–4, April 2011.

[73] Takumi Takizawa, Karen J Meaburn, and Tom Misteli. The meaning of gene positioning. *Cell*, 135(1):9–13, 2008.

[74] Renata Taslerová, Stanislav Kozubek, Eva Bártová, Pavla Gajdusková, Roman Kodet, and Michal Kozubek. Localization of genetic elements of intact and derivative chromosome 11 and 22 territories in nuclei of Ewing sarcoma cells. *Journal of structural biology*, 155(3):493–504, 2006.

[75] Bas Tolhuis, Robert Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter de Laat. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, 10(6):1453–1465, 2002.

[76] T Tumbar, G Sudlow, and A S Belmont. Large-scale chromatin unfolding and remodeling induced by VP16 acidic activation domain. *The Journal of Cell Biology*, 145(7):1341–1354, 1999.

[77] G van den Engh, R Sachs, and B. Trask. Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science*, 257(5075):1410–1412, September 1992.

[78] Geert Vandeweyer and R Frank Kooy. Balanced translocations in mental retardation. *Human genetics*, 126(1):133–147, 2009.

[79] Diana Y Vargas, Khyati Shah, Mona Batish, Michael Levandoski, Sourav Sinha, Salvatore A E Marras, Paul Schedl, and Sanjay Tyagi. Single-Molecule Imaging of Transcriptionally Coupled and Uncoupled Splicing. *Cell*, 147(5):1054–1065, 2011.

[80] A E Visser and J A Aten. Chromosomes as well as chromosomal subdomains constitute distinct units in interphase nuclei. *Journal of cell science*, 112 (Pt 1:3353–3360, 1999.

[81] A J Wood and R J Oakey. Genomic imprinting in mammals: emerging themes and established theories. *PLoS genetics*, 2(11):e147, November 2006.

[82] Y Xing, C V Johnson, P R Dobner, and J B Lawrence. Higher level organization of individual gene transcription and RNA splicing. *Science (New York, NY)*, 259(5099):1326–1330, 1993.

[83] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 2011.

[84] Hai Yan, Weishi Yuan, Victor E Velculescu, Bert Vogelstein, and Kenneth W Kinzler. Allelic variation in human gene expression. *Science (New York, NY)*, 297(5584):1143, 2002.

[85] David Yu Zhang, Sherry Xi Chen, and Peng Yin. Optimizing the specificity of nucleic acid hybridization. *Nature Chemistry*, 4(3):208–214, March 2012.

[86] David Yu Zhang and Georg Seelig. Dynamic DNA nanotechnology using strand-displacement reactions. *Nature Chemistry*, 3(2):103–113, 2011.

[87] David Yu Zhang and Erik Winfree. Control of DNA strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society*, 131(47):17303–17314, 2009.

[88] Kun Zhang, Jin Billy Li, Yuan Gao, Dieter Egli, Bin Xie, Jie Deng, Zhe Li, Je-Hyuk Lee, John Aach, Emily M Leproust, Kevin Eggan, and George M Church. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nature methods*, 6(8):613–618, 2009.

[89] Yu Zhang, Rachel Patton McCord, Yu-Jui Ho, Bryan R Lajoie, Dominic G Hildebrand, Aline C Simon, Michael S Becker, Frederick W Alt, and Job Dekker. Spatial Organization of the Mouse Genome and Its Role in Recurrent Chromosomal Translocations. *Cell*, 2012.

[90] Daniele Zink, Andrew H Fischer, and Jeffrey A Nickerson. Nuclear structure in cancer cells. *Nature reviews Cancer*, 4(9):677–687, 2004.