Spatially resolved, highly multiplexed RNA profiling in single cells

A dissertation presented
by

Kok Hao Chen

to

The Department of Chemistry and Chemical Biology

In partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Chemistry

Harvard University
Cambridge, Massachusetts

April 2015

10x Genomics EX2041
Bruker v. Harvard
IPR2024-00526

ProQuest Number: 3738727

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest 3738727

Published by ProQuest LLC (2015). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

©2015 – Kok Hao Chen

All rights reserved.

Dissertation Advisor: Xiaowei Zhuang Kok Hao Chen

Spatially resolved, highly multiplexed RNA profiling in single cells

Abstract

The transcriptome of a cell contains a wealth of information about the cell's current state and its recent history. System-wide analyses of the copy number and localization of RNAs in single cells promise to transform our understanding of many areas of biology, such as the origin and consequences of cellular heterogeneity, functions and mechanisms underlying sub-cellular localization of transcripts, and organization of cell types in tissues. Fluorescent in situ hybridization (FISH) based approaches for measuring single-cell gene expression offer high spatial resolution and single molecule sensitivity, but lack throughput in comparison with sequencing based methods.

In the first part of this thesis, we present multiplexed, error robust FISH (MERFISH) for transcriptome imaging in single cells. We labeled each cellular RNA with multiple complementary oligonucleotide probes, which contain additional readout sequences. Each RNA species is encoded with a unique combination of readout sequences. We then identified these RNAs by using successive rounds of hybridization and imaging, with a different fluorescently labeled readout probe each round. This combinatorial labeling scheme allows the number of detectable RNA species to grow exponentially with the number of hybridization rounds, but the effect of readout errors also increases. By utilizing error-robust encoding schemes similar to

those common in digital electronics, we are able to match the measured sequence of on/off fluorescence signals from individual RNAs to their assigned barcodes with high efficiency and minimal error.

In the second part of this thesis, we optimize the performance of localization based super-resolution microscopy by conducting a systematic characterization of the photo-switching properties of 26 organic dyes. These properties, which include the photons per switching event, on-off duty cycle, photostability, and number of switching cycles, largely dictate the quality of super-resolution images. Our study identified the top performing dyes in each fluorescent channel, allowing us to perform super-resolution imaging in four colors.

We envision that by combining multi-color super resolution microscopy with our multiplexed RNA imaging approach, we can visualize all the RNA content, the transcriptome, of a cell *in situ*. The ability to perform spatially-resolved transcriptomic analysis of cells and tissues will dramatically improve our abilities to understand biology and disease.

Contents

Title…..……i
Copyright……...……ii
Abstract……iii
Table of contents…..……iv
Citations to previously published work……...……vii
Acknowledgments……..……viii
Dedication……..…...x

1 Introduction……1
1.1 Genome and transcriptome are spatially organized in cells…….……1
1.2 Vast utility of a spatially resolved, genome wide expression profiling technique……...3
1.3 Imaging individual transcripts in single cells via fluorescent *in situ* hybridization with single molecule resolution (smFISH)…….....4
1.4 Current approaches to scale up the number of imaged RNA species…….…….5
1.5 Breaking the diffraction barrier with localization-based super-resolution microscopy ……..8
1.6 Outline of this dissertation……..…11

2 Imaging of single-cell transcriptome with multiplexed error-robust FISH (MERFISH)…….....13
2.1 Key Innovations that enabled highly multiplexed measurements……..13
2.1.1 Use of error robust encoding schemes in multiplexed RNA imaging……...13
2.1.2 Sequential labeling scheme with orthogonal readout probes…….…....18
2.1.3 Large scale synthesis of complex oligo probes…….…….22
2.2 Results…….……...24
2.2.1 Initial experiments with chemically synthesized oligos……24
2.2.2 Twelve and thirty two target genes using array synthesized oligos…….…..28
2.2.3 One hundred and forty target genes with modified hamming distance 4 code……..…...39

2.2.3.1 Cell-to-cell variation in gene expression……50
2.2.3.2 Correlation analysis of gene expression……52
2.2.3.3 Spatial distribution patterns of RNA……54
2.2.4 One thousand genes with modified hamming distance 2 code……57
2.3 Discussion……62
2.4 Methods and materials……69

3 Super-resolution imaging……92
3.1 Introduction……92
3.2 Screening of organic dyes for multi-color super-resolution imaging……95
3.2.1 Photo-switching properties and their effect on super-resolution images……95
3.2.2 Effect of buffer composition on switching properties……100
3.2.3 Best performing dyes resolve the hollowness of microtubule filaments……104
3.2.4 Low cross talk four-color super-resolution imaging……107
3.3 Methods and materials……111

4 Conclusion……122
4.1 MERFISH for RNA imaging……122
4.2 Multi-color super-resolution microscopy……124
4.3 Imaging of single cell transcriptome with multi-color super-resolution microscopy……124

Appendix A……127
Appendix B……128
Appendix C……135
Appendix D……138

Bibliography……158

Citations to previously published work

Portions of Chapter 2 have appeared in the following publication:
K. H. Chen*, A. N. Boettiger*, J. R. Moffitt*, S. Wang, X. Zhuang, Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090 (2015) DOI: 10.1126/science.aaa6090. *These authors contributed equally to this work

Portions of Chapter 3 have appeared in the following publication:
G. T. Dempsey*, J. C. Vaughan*, K. H. Chen*, M. Bates, X. Zhuang, Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. Nat. Methods. 8 (2011), pp. 1027–1036. *These authors contributed equally to this work

Acknowledgements

First and foremost, I would like to express my gratitude to my adviser, Prof. Xiaowei Zhuang, for providing the vision on potential projects to embark on. She challenged me to dream big, but at the same time, be strategic on the choice of experiments to pursue. It is also amazing how meticulous and detail-oriented she is. For example, I remembered her catching a mistake in a supporting table with tens of thousands of DNA sequences. I am indebted to her for all the scientific skills she has imparted to me. Her professionalism and passion in research will always be an inspiration to me. I would also like to thank my committee members, Prof. Adam Cohen and Prof. Sunney Xie, whom I also did my rotations with. Thank you for all your guidance.

I have been privileged to work with many phenomenal scientists. In particular, I would like to thank Dr. Alistair N. Boettiger, who provided indispensible guidance on data analyses and experimental techniques, as well as a contagious optimism coupled with endless new ideas. I thank him for being a wonderful and patient mentor, as well as a reliable friend who has always been available to me. I would also like to thank Dr. Jeffrey R. Moffitt who provided invaluable help with experiments and data analyses. I thank him for training me to think critically while at the same time being encouraging and motivating. More importantly, I thank him for being a genuinely caring friend. I would also like to thank Dr. Siyuan Steven Wang for great scientific discussions and experimental help. Additionally, I thank Dr. Graham T. Dempsey, who welcomed me to the Zhuang lab and trained me to do all things STORM-related. I also thank Dr. Joshua C. Vaughan for his mentorship and friendship.

The Zhuang lab has been a stimulating and fun place to work in. I would like to acknowledge our fantastic lab manager, Alec Goodman, for selflessly maintaining the smooth running of the lab. I also thank Matt Kilroy for all his administrative help and for organizing fun

activities for the lab. I thank Dr. Hazen Babcock for technical advice and help with instrumentation. I would also like to acknowledge Bryan Harada for many fruitful discussions and experimental help; Dr. Shu Jia for encouragements and many scientific discussions; Dr. Ethan Garner for thought provoking discussions; Dr. Mark Bates and Dr. John Wu for teaching me many experimental techniques; and the all the rest of the Zhuang lab members who have graciously supported me. These people have truly made my graduate school experience more meaningful.

I acknowledge with gratitude the funding received from the Agency of Science, Technology, and Research from Singapore for both my undergraduate and graduate studies in the United States.

I have been blessed with an extremely supportive family. I thank my parents and in-laws for their love, encouragement, and unwavering support. I am also grateful for the support of my brother-in-law who has always been willing to help in anyway possible despite living far away. Finally, no words can described how grateful I am to my wife for her sacrificial love. The completion of my doctorate would not have been possible without her. To my sons, may this thesis be an inspiration to you.

Philippians 4:13

For my wife and my parents

Chapter 1

Introduction

The completion of the Human Genome Project has given us access to the DNA sequence of our whole genome; this is akin to having a manual filled with instructions needed to make every cells of our body. However, the task of harnessing this manual for the improvement of human health remains a challenge, as it relies on our abilities to decipher its complex language and recognize all its nuances. To do so require the development of new genome-wide tools that can provide a system level understanding of complex biological processes and discover the genetic basis of diseases. Next generation sequencing technologies and state of the art genetic manipulation tools have enabled scientists to investigate the genetic basis of a certain phenotype in a high throughput manner and perform large scale forward and reverse genetic screens to unravel gene function (*1–3*). Single cell sequencing technologies even allowed scientists to probe genomic heterogeneity and transcriptomic diversity of individual cells (*4–6*). However, most of these methods cannot measure the spatial location of the genetic elements of interest or retain the spatial context of the cells (*7*). Therefore, an image based sequencing approach that allows researchers to directly visualize the genome and transcriptome of single cells *in situ* will be invaluable for probing the spatial organization of our genome and transcriptome.

1.1 The human genome and transcriptome are spatially organized in cells

While the central dogma of biology, which describes the linear flow of biochemical information from DNA, to RNA, and then protein, is well understood, we know much less about

how and why the genome and transcriptome are spatially organized in cells. Recent studies have revealed that chromosomal organization and transcripts localization play many fundamental roles in diverse aspects of biology, such as the regulation of transcription and translation, patterning of embryos, cell migration, and neural signaling (*8–12*). For example, the human genome is organized into distinct chromosomal territories with gene-poor chromosomes often found near the nuclear periphery while gene-rich chromosomes located near the center (*12–14*). Likewise, many mRNAs were found expressed in distinct spatial patterns in cells. For example, many neuronal mRNAs are found to be enriched in axons and dendrites compared with the soma (*15*). Spatial localization of many transcripts is also found to be important for their functions. For example, the localization of actin mRNAs at cell's leading edge is correlated with its polarization while the localization of actin mRNAs to focal adhesions is important for cellular migration (*16*, *17*). There is, however, a large host of mRNA whose subcellular localization is not yet understood (*18*).

Because the spatial organization of the transcriptome is an important mechanism for cells to achieve localized translation, it is important that we understand how cells achieve such spatial organization. For example, Actin mRNAs are known to be transported by molecular motors to their destinations through the recognition of a 'zip-code' like sequence found in the 3′ untranslated region of the mRNA by the RNA-binding protein, ZBP1 (*19*). While the regulatory mechanisms underlying the localizations of some RNAs, like actin mRNAs, are rather well studied, the necessary factors as well as the biological roles for the localization of many other mRNAs and non-coding RNAs are unknown. Therefore, a transcriptome-wide technique that can probe both the spatial and genetic information inside a cell will be invaluable for answering these questions pertaining to subcellular organization of RNAs.

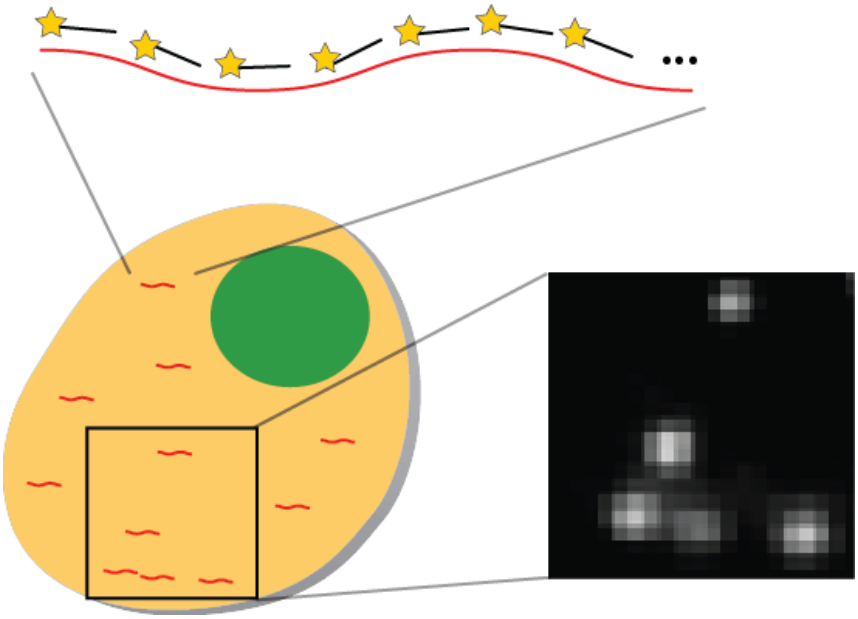
1.2 Vast utility of a spatially resolved, genome wide expression profiling technique

In addition to probing sub-cellular organization of RNA, a transcriptome-wide imaging based expression profiling technique will be able to delineate single cells based on their spatial locations and uncover the underlying distribution of individual cell states. Due to the stochastic nature of gene expression (*20*, *21*), such as transcriptional bursts, and a host of other factors, large amount of cell-cell phenotypic variations are commonly observed in tissues or even seemingly isogenic cells (*22*, *23*). Hence, a single-cell expression profiling technique will not only be useful for understanding the mechanisms of gene regulation under expression noise, but also the origins and biological consequences of cellular heterogeneity.

Since an imaging technique will not require the dissociation and capture of individual cells or marker based fluorescence-activated cell sorting, it can perform direct *in situ* transcriptomic profiling of complex tissues. Expression-based clustering of cells in developing, neuronal, or any other types of tissue based on their transcriptome will allow for the unbiased reconstruction of cell types, potentially enabling the discovery of new cell types or new biomarkers. Furthermore, such spatially resolved gene expression profiling technique may help to illuminate the mechanisms underlying polarity or gradient formation in complex tissues, like the embryos or retina. In addition, it will enable the analysis of diseased cells inside pathological tissues, which maybe hard to isolate or culture separately. Such analysis will facilitate the understanding of disease progression or the design of appropriate treatment plan. Lastly, such single cell technique will allow the probing of transient or rare cell states, which may occur during growth, differentiation, reprogramming or response to external stimuli like pathogens. Image based single cell tracing studies of transcriptomic changes during these cellular processes may help to elucidate the mechanisms that control cell fate or immune response.

1.3 Imaging individual transcripts in single cells via fluorescent in situ hybridization with single molecule resolution (smFISH)

More than a decade ago, Femino et al demonstrated a technique to visualize single RNA transcripts in cells using fluorescently labeled oligonucleotides hybridized *in situ* to target mRNA (*24*). This technique was later modified by Raj et al to use multiple (~50-100) singly labeled probes instead of a few probes with multiple labels (*25*). In these methods, fluorescently labeled oligonucleotides, which are complementary in tandem to a RNA sequence of interest, are hybridized to target RNA molecules in fixed cells; resulting in the appearance of multiple diffraction-limited spots that corresponds to the number of target RNAs (Fig. 1). Detection and counting of these spots using a fluorescent microscope provides the copy number of a particular RNA species in individual cells. This method, termed fluorescent in situ hybridization with single molecule resolution (smFISH), can quantify RNA count without amplification bias (a PCR amplification step is often necessary for other gene expression profiling techniques like quantitative PCR and RNA sequencing) while preserving their native context.



(Continued) Figure 1: Explanation of the fluorescent in situ hybridization with single molecule resolution (smFISH) method. Multiple oligonucleotides with different sequences each with a fluorescent label collectively bind along the same target transcript to produce a punctate signal under light microscope. The number and position of these fluorescent spots correspond to the copy number and spatial localization of the target RNA.

smFISH has proven to be a powerful method for the absolute quantification of RNA, as well as for the study of mechanisms underlying transcript localization, gene expression, and genetic interactions. For example, smFISH was used to quantify the variation in gene expression among genetically identical cells residing in the same environment (*26*), probe the coupling between transcription and splicing (*27*), analyze the localization and abundance of long non-coding RNA (*28*), study the relationship between cellular volume and transcription rate (*29*), and identify genetic mechanisms such as paused promoters (*30*) and shadow enhancers (*31*). In addition, smFISH have also been used to understand the molecular basis of genetic phenomena like incomplete penetrance (*32*), identify new feedback pathways in previously studied genetic circuits (*33*) and infer kinetic properties of gene expression and degradation (*34*).

1.4 Current approaches to scale up the number of imaged RNA species

Despite being a powerful technique, a major limitation of smFISH is the number of genes whose expression level can be measured simultaneously in a single a cell. Its throughput is paltry in comparison with single cell RNA sequencing, which typically measures thousands of genes (*35*). This limits not only the throughput of smFISH to study gene expression, but also the ability to discover novel blocks of co-regulated genes through correlated changes in their expression. Although subsequent work has expanded smFISH using multiple, spectrally distinguishable

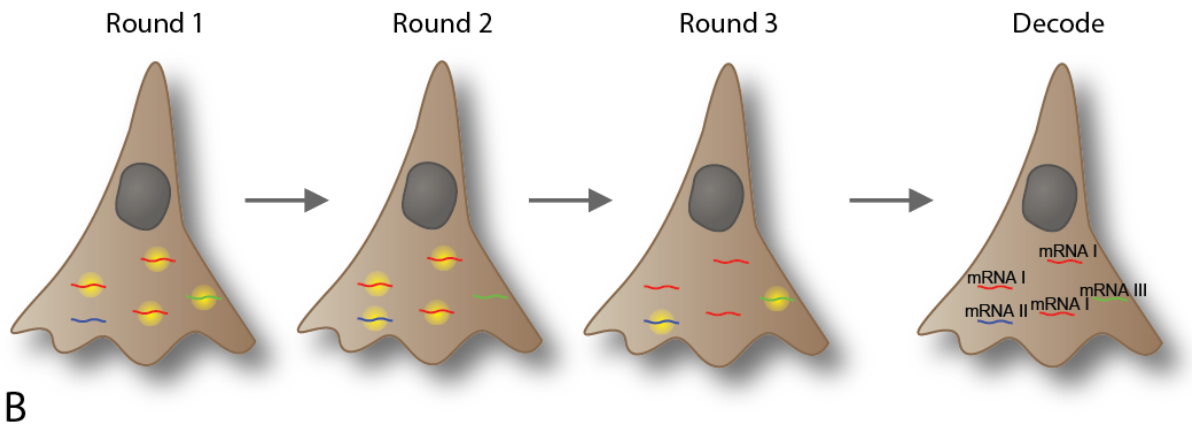
fluorescent dyes to allow the imaging of several different mRNA species simultaneously in the same cell (*32*, *36*), the number of spectrally distinguishable dyes is still substantially smaller than the number of genes active in many common cellular behaviors (let alone the number of genes expressed in a cell).

To overcome this limitation, researchers have introduced the idea of barcoding mRNAs, building off earlier approaches at the chromosomal scale (*38*, *39*). Instead of labeling each RNA species with different spectrally distinguishable dyes, each RNA species was labeled with a unique combination of fluorophores. Levsky et. al. demonstrated barcoding of 10 nascent RNA transcripts by using 4 different dyes: each transcript was identified by a unique combination of 2 or 3 of the 4 dyes (*37*). Levesque et. al. recently used chromatic barcoding to label 20 different genes' introns by assigning 2 or 3 (out of five) fluorophores to each target (*40*). They were then able to probe which 20 genes were actively transcribing in the same cell. Lubeck et. al. have also adapted this approach to identify 32 different mRNA transcripts (*41*). Their labeling scheme barcoded each transcripts with combinations of 3 out of 7 different 'colors'. In order to expand the number of spectrally distinguishable dyes and to resolve all these transcripts, they used a multi-activator STORM (*42*) based imaging assay. Using this assay, they were able to identify co-regulated genes and different clusters of yeast cells based on their response to calcium stress.

Further extending the chromatic barcoding scheme to the transcriptome scale is challenging because the total number of possible barcodes is limited by the number of spectrally distinguishable dyes. We sought to develop a sequential imaging approach to overcome this restriction. Analogous to the different combination of colors, RNAs can be barcoded using combination of hybridization rounds. In other words, instead of labeling mRNAs with different colors in a single hybridization step, groups of mRNAs are labeled with the same dye

sequentially in multiple hybridization steps, so that co-localization of fluorescent signals from different combinations of imaging rounds can be used to identify the transcripts (Figure 2a). Figure 2b illustrates an example of how the same three transcripts can be barcoded with either color or sequence combinations. Each gene in this scheme can be represented by the same sequence of binary letters, or bits, in both the chromatic and sequential barcoding approaches. Binary representation is particularly useful when we consider readout errors and the application of coding theory in Section 2.1.3.

A



B

Color Combinations	Binary Code			Sequential Imaging		
	Green	Red	Blue	Round 1	Round 2	Round 3
mRNA I	1	1	0	mRNA I	mRNA I	mRNA I
mRNA II	0	1	1	mRNA II	mRNA II	mRNA II
mRNA III	1	0	1	mRNA III	mRNA III	mRNA III

Figure 2: A) Cartoon illustration of how combinatorial labeling can be achieved with sequential imaging. B) Color combinations vs. sequential imaging as comparable barcoding methods. For example, to barcode three genes with colors, one may designate mRNA I as red and green, mRNA II as red and blue, and mRNA III as blue and green. This same barcode may be achieved

using sequential imaging with a single fluorophore in which mRNA I is imaged in round 1 and 2, (Continued) mRNA II is imaged in round 2 and 3, and mRNA III is imaged in round 1 and 3. Both method barcode the mRNAs with the same set of binary codewords: mRNA I as [1 1 0], mRNA II as [0 1 1], and mRNA III as [1 0 1].

Given N available bits, a maximum of 2^N -1 RNA species can be barcoded if all possible combination of '0's and '1's are used (1 is subtracted because the codeword with all '0's cannot be used). Thus, the number of RNA species that can be imaged simultaneously in individual cells grow exponentially with the number of imaging rounds. With N = 16, over 65,000 RNA species, which should cover the whole human transcriptome, could be identified. However, we believe that extending the sequential hybridization scheme to the transcriptome scale will face two major roadblocks: 1) Inefficiency in RNA labeling and signal removal during each rounds of hybridization will lead to an drop in detection efficiency and the corresponding increase in misidentification errors. 2) Labeling of many RNA targets will require long experimental time and large amount of fluorescent probes, which are costly to synthesize. Therefore, to enable combinatorial labeling of RNAs at the transcriptome scale with sequential hybridizations, we addressed these challenges by 1) developing a novel sequential labeling method with both high detection efficiency and high removal rate, 2) designing modified error robust encoding schemes to barcode RNAs, and 3) adapting a large scale complex oligopool synthesis protocol.

1.5 Breaking the diffraction barrier with localization-based super-resolution microscopy

As we continue to scale up the number of RNA species imaged in a cell, we will encounter another imaging roadblock - the optical diffraction limit. The fluorescence emitted

from a point source becomes blurred when imaged through a lens based system due to optical diffraction. The blurred spot has a width of ~200-300nm in the lateral dimension and ~500-600 in the axial dimension for visible light. The width of the blurred spot sets the minimum separable distance at which two point emitters can no longer be distinguished from each other. This spatial resolution limit has constrained the application of light microscope to resolve cellular structures that are more than several hundred nanometers.

This resolution limit can be overcome by patterned illumination or stochastically activating the fluorescence from a sparse subset of single molecules (*43*). Because the centroid position of a single fluorescent emitter can be determined with very high precision, this process of localizing a subset of single molecules can be repeated many times to yield a high-resolution image from the centroid positions of all the molecules (figure 3). Such localization based super resolution microscopy, named Stochastic Optical Reconstruction Microscopy (STORM) or (Fluorescence) Photoactivated Localization Microscopy ((F)PALM), was first demonstrated in 2006 (*44–46*). Since then, super resolution microscopy has proven to be extremely useful in nano-scale imaging of biological samples (even in live cells), such as measuring distances between different types of synapse (*47*), mapping the spatial architecture of focal adhesion complex (*48*), probing the morphology of clarithin coated pits during endocytosis (*49*), revealing the periodic structure of actin cytoskeleton in neurons (*50*), and etc.

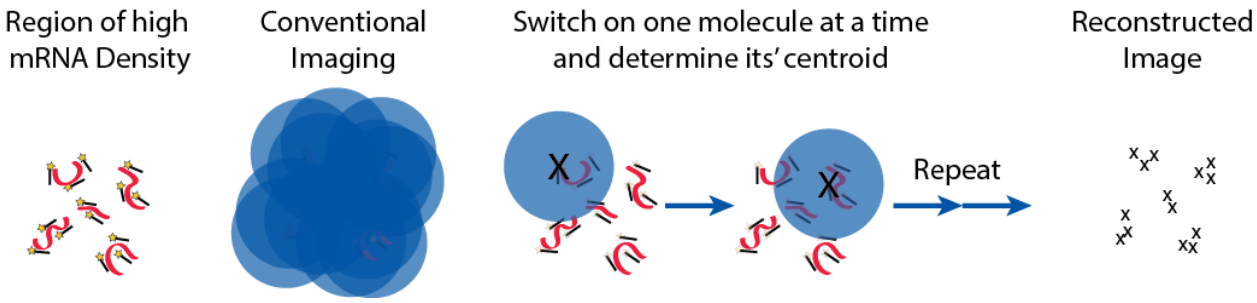


Figure 3: Explanation of the Stochastic Optical Reconstruction Microscopy (STORM) method. In densely labeled region of a cell, individual mRNAs may not be optically resolvable with conventional imaging due to the diffraction limit. However, if individual flurophores can be switched on one at a time, its centroid position can be determined with high accuracy, allowing the reconstruction of a high resolution image of these mRNAs.

Localization based super-resolution imaging methods rely on the ability of organic fluorophores or fluorescent proteins to photo-switch (or 'blink') or photo-activate. Since the uncertainty in the determination of the centroid position of a point emitter scales inversely with the square root of the number of collected photons, fluorophores that emit more photons when it blinks can provide higher STORM resolution. In addition, STORM resolution also depends on the Nyquist criterion, which specifies that the sampling interval must be smaller than half the desired resolution. Therefore, the switching duty cycle, the fraction of time a fluorophore spends in the on state, of a fluorophore is also a crucial property, as it will constrain the maximum density of molecules allowed in a diffraction-limited spot. Other factors such as the number of switching cycles and photo-bleaching rate also contributes to STORM image quality.

Many organic fluorophores have been used for STORM imaging, but very few had been systematically characterized for these important photo-switching properties, namely their photon output, duty cycle, switching cycle, and photo-bleaching rate. These properties can significantly

impact a probe's imaging performance. To address this need, we evaluated 26 commercially available organic dyes for their suitability for STORM imaging and correlated their photo-switching properties with the quality of the super-resolution images. We identified the best performing dyes for four different spectral channels and demonstrated four color super-resolution imaging.

1.6 Outline of this dissertation

The ability to perform spatially-resolved transcriptomic analysis of RNA molecules *in situ* promises to dramatically improve our abilities to understand biology and disease. The ability to reveal the transcriptomes of different cell types in complex tissues will allow us to probe the underlying mechanism of how cells are organized in tissues and how biological complexity can arise from cells having essentially the same genome.

For these reasons, we developed a highly multiplexed approach for the *in situ* imaging of hundreds to thousands of different RNA species. In Chapter 2 of this thesis, I describe the key innovations, as well as the design and development process of this method. With our method, we demonstrated simultaneous imaging of 140 transcripts using an error correcting encoding scheme or 1001 transcripts using an error detection scheme. Both encoding schemes could be scaled up to image even more number of transcripts. We envision that this method will have broad applications in both academic research as well as clinical settings.

In Chapter 3 of this thesis, I describe our efforts to understand and characterize the photo-switching properties of 26 commercially available organic fluorophores. These fluorophores have fluorescence wavelengths spanning from ~500nm (blue) to ~800nm (near infrared). Our work not only provides a resource for scientists to select the best performing dyes in different

spectral channels for multi-color super resolution imaging, but also provides a guideline for evaluating other fluorophores.

In Chapter 4 of this thesis, I discuss how our method for highly multiplexed RNA imaging can be scaled up to the transcriptome scale, and how it can be combined with super resolution imaging to visualize the whole transcriptome at nano-scale resolution.

Chapter 2

Imaging of single-cell transcriptome with multiplexed error-robust FISH (MERFISH)

This work was done in collaboration with Dr. Alistair Boettiger, Dr. Jeff Moffitt, and Dr. Steven Wang under the supervision of Prof. Xiaowei Zhuang. Most of the results from this work are published in (*51*).

2.1 Key innovations that enabled highly multiplexed measurements

2.1.1 Use of error robust encoding schemes in multiplexed RNA imaging

Combinatorial labeling that identifies each RNA species by multiple distinct signals offers a route to rapidly increase the number of RNA species that can be probed simultaneously in individual cells. With N distinct signals, a maximum of 2^N-1 barcodes, i.e. RNA species, can be probed. With just 16 distinct signals, over 65,000 RNA species, which should cover the entire human transcriptome (mRNA and non coding RNA), could be identified (*52*). In the chromatic barcoding approach, where fluorophores with different fluorescence wavelengths are used as distinguishable signals, one barcodes more mRNAs by introducing more colors. However, development of new spectrally distinguishable dyes and imaging of tens of colors are challenging. Thus, the number of chromatic barcode does not scale easily to the transcriptome.

In contrast, with a sequential labeling scheme (Figure 4), where each RNA species is encoded with a N-bit binary word and the sample is probed with N corresponding rounds of hybridization, each round targeting only the subset of RNAs that should read '1' in the

corresponding bit, the number of barcodes is limited only by the number of hybridization rounds that can be performed. As shown later in this chapter, 16 rounds of hybridization can be performed with almost no sample degradation. Consequently, the sequential imaging scheme scales more favorably to the whole transcriptome. Indeed, barcoding via sequential hybridization provides the first feasible route to transcriptome-scale imaging of RNA.

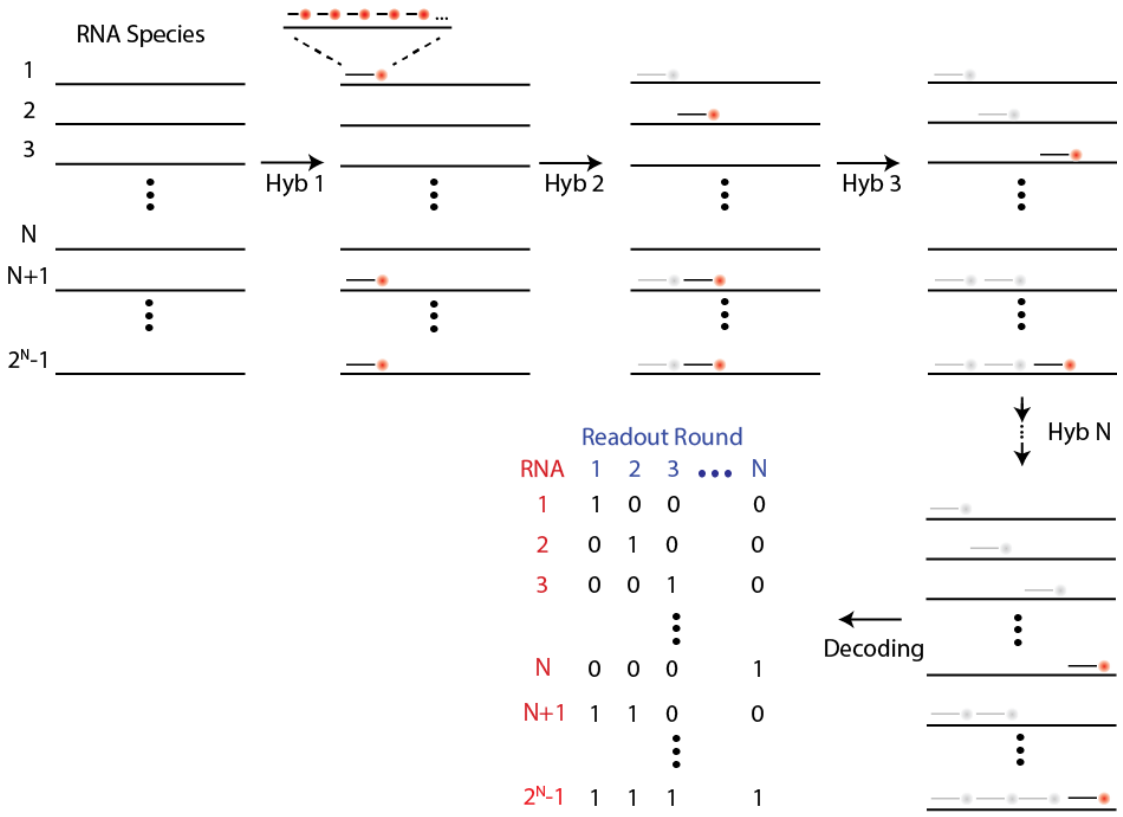


Figure 4: Combinatorial labeling with sequential imaging: 2^N-1 different RNA species can be uniquely encoded with all N-bit binary words. In each hybridization round, FISH probes that are targeted to all RNA species that have a '1' in the corresponding bit are included. To increase the ability to discriminate RNA spots from background, each RNA is addressed with multiple FISH probes per hybridization round. Signal from the bound probes is extinguished before the next round of hybridization. This process continues for all N hybridization rounds (Hyb 1, Hyb 2, and etc), and all 2^N-1 RNA species can be identified by the unique on-off pattern of co-localized fluorescence signals in each hybridization round.

However, combinatorial labeling approaches face a significant challenge because not only does the number of addressable RNA species increases exponentially with N (Figure 5A; black symbols), but the error rates also grow with N (Figure 5). Misidentification or missed detection errors can arise from false positives (0→1 conversion) and false negatives (1→0 conversion) in smFISH experiments. False positives can be due to background or off-target hybridization that produces a spurious fluorescent signal. On the other hand, false negatives can be due to some RNA molecules not being labeled with a sufficient number of probes to be detected above the background noise. In a combinatorial labeling scheme, such errors will result in some RNA species being not properly counted (reduced calling rate) and some RNA species being misidentified as another species (increased misidentification rate). Therefore, it is important to keep these error rates low in a combinatorial labeling experiment. Although error rates per hybridization round are low (~10% false negative rate and ~4% false positive rate) in our measurements [~90% detection efficiency per hybridization is comparable to what was reported in the literature (*41*, *53*)], the probability of error-free detection of an RNA after *N* rounds of hybridization still diminishes rapidly as *N* increases. In other words, as N increases, the fraction of RNAs properly counted (the calling rate) would decrease and, more troublingly, the fraction of RNAs that are identified as incorrect species (the misidentification rate) would rapidly increase (Figure 5C, D; black symbols). After 16 rounds of hybridizations, the majority of RNA molecules would be misidentified! See Section 2.4.6 for details concerning the calculation of error rates.

As we were developing our combinatorial labeling and sequential imaging approach, Lubeck et. al. described a sequential hybridization scheme using DNASE 1 digestion as a way to remove probes in between hybridization rounds (*54*). However, they only demonstrated the

simultaneous imaging of 12 genes (with 88.5% probe removal efficiency (per hyb) and 77.9% re-staining efficiency). It is likely that the rapid increase of error rates with the number of hybridization rounds has also been a major challenge in this approach.

A

Transcripts	Assigned codeword	Measured codeword	Types of error
RNA I	0 1	0 1	Not affected
RNA II	1 0	0 0	Missed RNA II
RNA III	1 1	0 1	Misidentified as RNA I

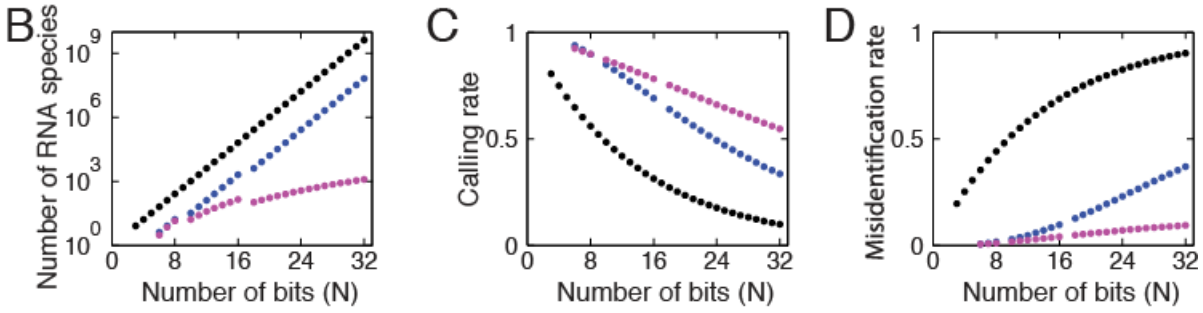


Figure 5: (A) Effect of false negative error (1→0 conversion, red) on three example transcripts: RNA I is not affected, RNA II is not properly counted (reduced calling rate) while RNA III is misidentified as RNA I (increased misidentification rate). The number of addressable RNA species (B), the rate at which these RNAs are properly identified – calling rate (C), and the rate at which RNAs are incorrectly identified as a different RNA species – misidentification rate (D) plotted as a function of the number of bits (N) in the binary words encoding RNA. Black, a simple binary code that includes all 2^N-1 possible binary words. Blue, the HD4 code where the Hamming distance separating words is 4. Magenta, the modified HD4 (MHD4) code where the number of '1' bits are kept at four. The calling and misidentification rates are calculated with per bit error rates of 10% for the 1→0 error and 4% for the 0→1 error. 10% and 4% are close to the experimentally measured per bit error rate.

To address the challenge associated with the increase in error rates with the number of imaging rounds, we designed error-robust encoding schemes, in which only a subset of the 2^N-1 possible codewords, separated by a certain Hamming distance, was used to encode RNAs. In a codebook where the Hamming distance between all the codewords is 4 (HD4 code), at least four bits must be read incorrectly to change one code word into another (Figure 6A). As a result, every single-bit error produces a word that is uniquely close to a single code word, allowing such errors to be detected and corrected (Figure 6B). Double-bit errors produce words with an equal Hamming distance of 2 from multiple code words and, thus, can be detected but not corrected (Figure 6C). Such a code should substantially increase the calling rate and reduce the misidentification rate (Figure 5C, D, blue symbols). To further account for the fact that it is more likely to miss a hybridization event (an 1→0 error) than to misidentify a background spot as an RNA (an 0→1 error) in smFISH measurements, we designed a modified HD4 (MHD4) code, in which the number of '1' bits were kept relatively low—only four per word—to reduce error. In this MHD4 code, we also kept the number of '1' bits constant to avoid biased detection. If the number of '1' bits in all the codewords were not kept constant, biased detection can happen because RNAs encoded with more number of '1' bits will suffer from lower calling rate than the RNAs encoded with less '1' bits. This MHD4 code should further increase the calling rate and reduce the misidentification rate (Figure 5C, D, magenta symbols).

If more accurate measurements are desired, the use of encoding schemes with a Hamming distance greater than 4, such as the Golay or Reed-Muller code, will further enhance the error detection and correction capabilities. An encoding scheme that has all its codewords separated by a minimum Hamming distance of d, can detect d-1 errors or correct (d-1)/2 errors. In general, the choice of encoding scheme depends on the number of barcodes desired as well as

the amount of errors that can be tolerated. For example, the codes that are more efficient (more codewords for a given number of bits used) can be used for high-throughput screens to generate hypotheses for many gene regulatory networks and pathways, whereas the codes better at error detection and correction are more appropriate for higher accuracy investigations of specific networks and pathways.

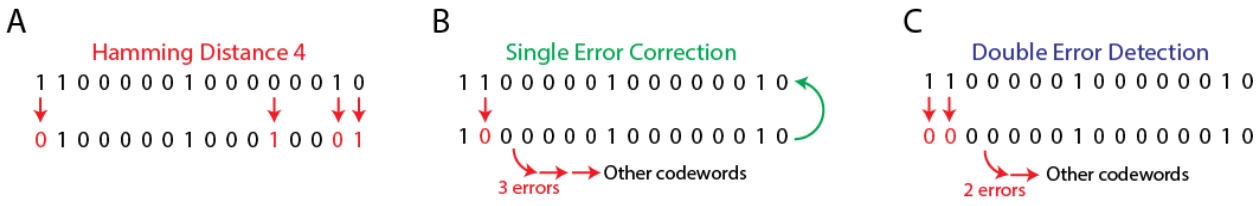
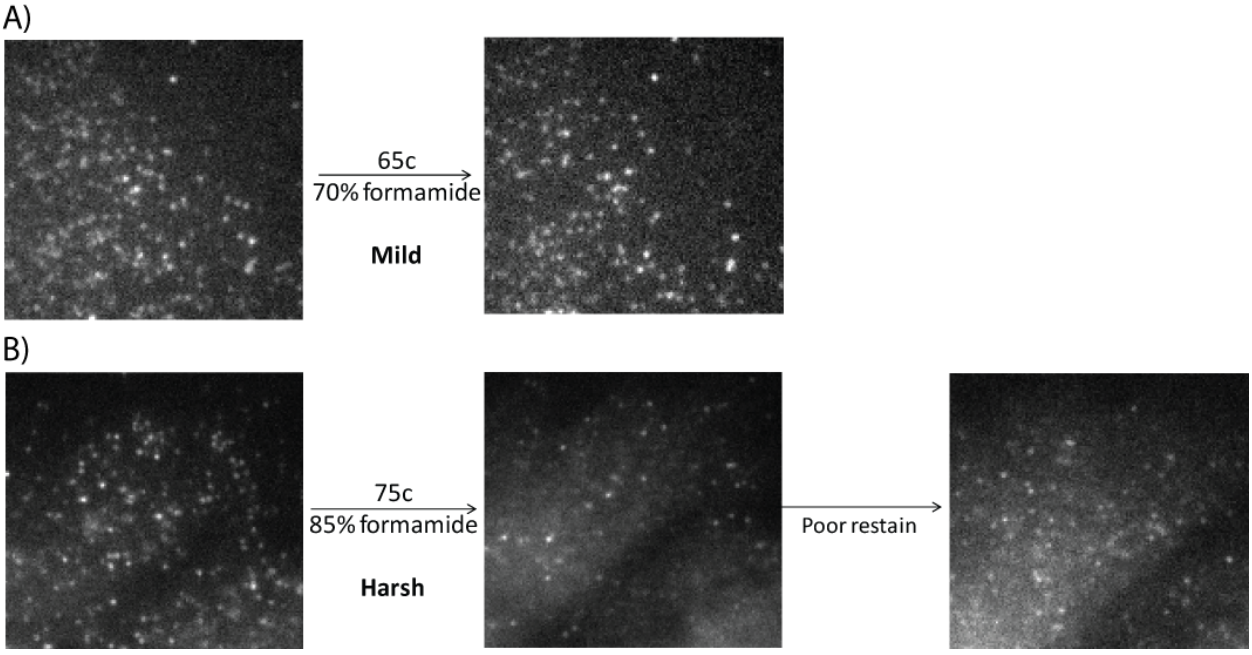


Figure 6: Schematic descriptions of Hamming distance and its use in the identification and correction of errors. (**A**) Schematic representation of a Hamming distance of 4 for 16-bits codewords. (**B-C**) Schematic showing the ability of an encoding scheme with Hamming distance 4 to correct single-bit errors (B) or detect but not correct double-bit errors (C). Red arrows highlight bits at which the indicated words differ. Two code words are separated by a Hamming distance of 4 if one of the words has to flip four bits from '1' to '0' or '0' to '1' to convert into the other word. Single-bit error correction is possible because if a measured word differs from a legitimate code word by only one bit, it is most likely an error that arises from misreading this code word, since the code words of all the other RNA species will differ from the measured word by at least three bits. In this case, we can correct the measured word to the code word that differs by only one bit. If a measured word differs from a legitimate code word by two bits, this measured word can still be identified as an error, but correction is no longer possible since more than one legitimate code word differs from this measured word by two bits.

2.1.2 Highly multiplexed RNA imaging via sequential readout of orthogonal sequences

The sequential imaging technique requires that fluorescent labels be added, imaged, removed, and re-added to the same mRNAs (Figure 2). We explored multiple ways to implement

the sequential hybridization scheme. We tried to wash out the FISH probes after each imaging round and re-stain mRNAs using the same set of FISH probes. We tested this method with 20 and 40-nt long probes for a range of wash-out conditions involving different concentrations of formamide (from 0-90%), different temperature incubations (from 22°C to 85°C), and different amount of time (10mins to 24hrs). We found that probe removal is ineffective for milder wash conditions (Figure 7A), whereas the harsher wash-out conditions remove a fraction of the target mRNAs themselves, jeopardizing subsequent relabeling and decoding (Figure 7B). Under a balanced compromise of wash-out stringency and transcript preservation, we still observe a substantial decrease in labeling efficiency over the course of 4 rounds of staining. Overall, we think that the wash out method has high error and is unlikely to extend to tens of hybridization rounds. However, it is possible that washout of probes could be greatly facilitated by using toe-hold displacement probes to facilitate the removal of the fluorescent probes from the previous stain under gentler conditions (*55*). This method has not been tested and could be returned to at a later date.



(Continued) Figure 7: Challenge of using a washout method to implement sequential hybridization. A) Under mild conditions, such as 65°C and 70% formamide, large fraction of probes remain bound to mRNAs. B) Under slightly harsher conditions, such as 75°C and 85% formamide, majority of signals are removed after washing, but re-staining of mRNAs using the same probe set is inefficient, most likely because the harsher wash-out conditions remove a fraction of the target mRNAs themselves, preventing subsequent relabeling.

To achieve high efficiency of both RNA detection and probe removal, we developed a two-steps labeling approach (Figure 8). In our approach, the entire probable mRNA sequence is first annealed to a tile of 'encoding-FISH probes'; each probe comprising a RNA targeting sequence and two flanking readout sequences that are not complimentary to the transcriptome. Subsequent rounds of staining then detect the same mRNAs by hybridizing to these readout sequences using a set of 'readout probes', which have orthogonal sequences to eliminate cross talk between hybridization rounds. For each RNA species, its encoding probes only contain readout sequences that correspond to the hybridization rounds where this RNA should read "1". This step converts each RNA species into a set of readout sequences that uniquely corresponds to the codeword of the RNA. Co-localization of fluorescent spots originating from different combination of imaging rounds can then be used to identify the RNA species. By using readout sequences (which need not be limited to two per encoding probe), this two-steps labeling approach allows the maximal use of RNA space for targeting sequences while enabling many rounds of sequential readout.

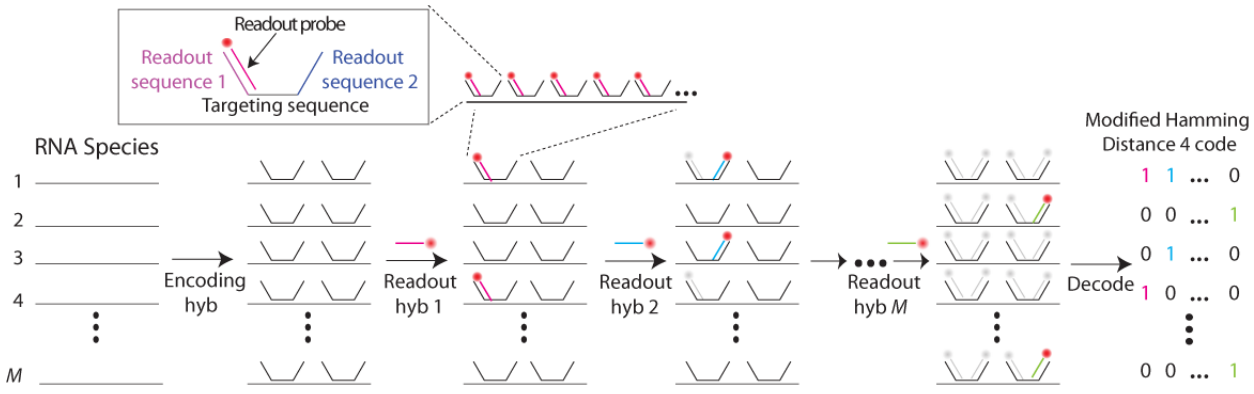


Figure 8: Schematic diagram of the implementation of a MHD4 code for RNA identification. Each RNA species is first labeled with a tile of encoding probes that convert the RNA into a unique combination of readout sequences (Encoding hyb). These primary probes each contain a central RNA targeting region flanked by two readout sequences, drawn from a pool of M different sequences, each associated with a specific hybridization round. M subsequent rounds of hybridization with the fluorescent secondary probes are used to probe the readout sequences (hyb 1, hyb 2, …, hyb M). The bound probes are inactivated by photobleaching between successive rounds of hybridization. For clarity only one possible pairing of the readout sequences is depicted; however all possible pairs of the four readout sequences are used at the same frequency and distributed randomly along each cellular RNA.

In addition to the high detection efficiency enabled by using large tile of encoding probes and high fluorophore removal efficiency made possible by complete photo-bleaching of fluorescent dyes, our labeling scheme have additional practical advantages. We substantially reduced probe synthesis cost by only labeling RNA once during the primary hybridization step and we also significantly diminished the total hybridization time for an experiment: we found that efficient hybridization to the readout sequences took only 15 minutes whereas efficient direct hybridization to cellular RNA required more than 10 hours. This is due to the fact that we have designed our readout sequences to contain no secondary structure and, unlike cellular RNAs, they are not occluded by RNA binding proteins.

2.1.3 Large scale synthesis of complex oligo probes

Although we have reduced the amount of probes needed by designing a two-step labeling method, the quantity and complexity of probes needed to stain a thousand RNA species is still extremely high. For example, to stain a thousand RNA targets with a tile of 50 probes per RNA, we will need 50,000 total unique sequences and approximately 200μM concentration of probes. With traditional chemical oligo synthesis method, this will amount to approximately $540,000 per experiment (90bps oligo synthesized by Integrated DNA Technologies, IDT), which is prohibitively costly. An Oligopaint approach has been previously developed to generate a large number of oligo probes from array-derived oligonucleotide pools containing tens of thousands of custom sequences to label chromosome DNA (*56*). We adapted this approach to enzymatically amplify probes from oligo-pools, which can contain up to ~100,000 unique sequences, for our experiments (Figure 9). We further increased the yield of this protocol by using in vitro transcription followed by reverse transcription (*57*). Details of the synthesis protocol can be found in section 2.4.2.

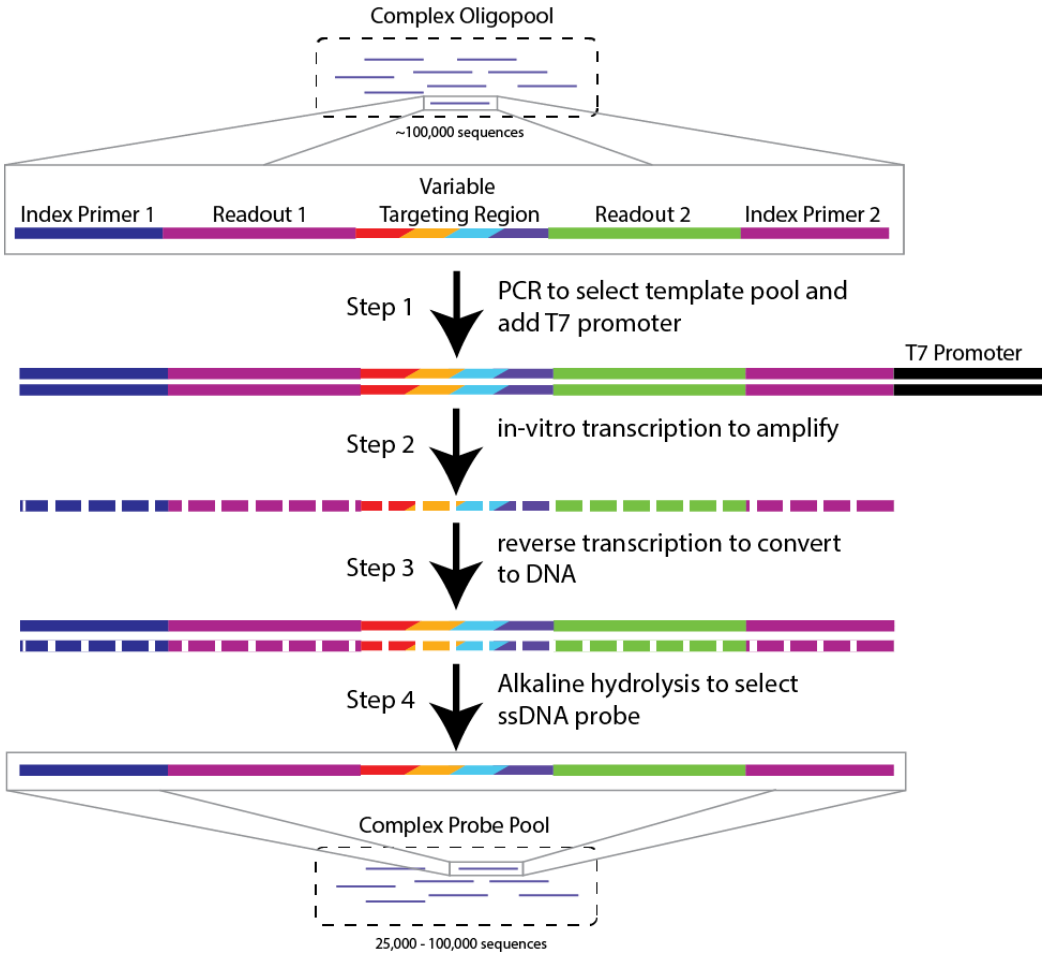


Figure 9: Production of the encoding probes. An array-synthesized complex oligopool, containing ~100k sequences, is used as a template for the enzymatic amplification of the encoding probes for different experiments. Each template sequence in the oligopool contains a central target region that can bind to a cellular RNA, two flanking readout sequences, and two flanking index primers. In the first step, the required template molecules for a specific experiment are selected and amplified with an indexed PCR reaction. To allow amplification via in vitro transcription, a T7 promotor is added to the PCR products during this step. In the second step, RNA is amplified from these template molecules via in vitro transcription. In the third step, this RNA is reverse transcribed back into DNA. In the final step, the template RNA is removed via alkaline hydrolysis, leaving only the desired ssDNA probes. This protocol produces ~2 nmol of complex pools of encoding probes containing ~20,000 different sequences for the 140-gene experiments or ~100,000 different sequences for the 1001-gene experiments.

2.2 Results

In order to optimize the staining protocol, estimate hybridization efficiency, and test the feasibility of a two-step labeling scheme, we first designed and performed photo-bleaching, co-localization, and sequential readout experiments using synthetic oligos obtained from IDT. Then, we tested the feasibility of using array-derived oligos in MERFISH by performing sequential readout experiments of twelve genes using an eight-bits Hamming distance four codebook. We then scaled up to thirty-two genes and one hundred and forty genes using the twelve bits and sixteen bits modified Hamming distance four codebooks respectively. Finally, we measured a thousand genes simultaneously in single cells by using a fourteen bits modified Hamming distance two codebook.

2.2.1 Initial experiments with chemically synthesized oligos

To optimize our smFISH protocol, we used the degree of co-localization of two tiles of probes targeting the same transcript as the metric. We designed ten probes for β-actin (*ACTB*) mRNA and tagged half of the probes with Cy5 and the other half with Cy3 (Figure 10A). We fixed and labeled BS-C-1 cells with these probes (synthesized by IDT). Detailed staining protocol is in 2.4.3 and sequences of the probes are in appendix A. If the amount of background spots are low and the hybridization efficiency is high, spots we observed from two different color channels should show a high degree of co-localization (Figure 10B-C). We varied hybridization stringency by changing the concentration of formamide used in the hybridization buffer. The concentration of probes also affects the binding efficiency and the amount of background. When the amount of formamide is too high or the concentration of probes is too low, probes are unable to hybridize to their targets (Figure 10D-E). If the concentration of formamide is too low or the

concentration of probes is too high, the amount of off-target hybridization (background) is high, leading to smaller degree of co-localization. We found that using 20-50% formamide (v/v) and 0.5-5nM (concentration per sequence) of probes are optimum. The maximum degree of co-localization observed is ~80%, similar to what was reported in the literature (*53*).

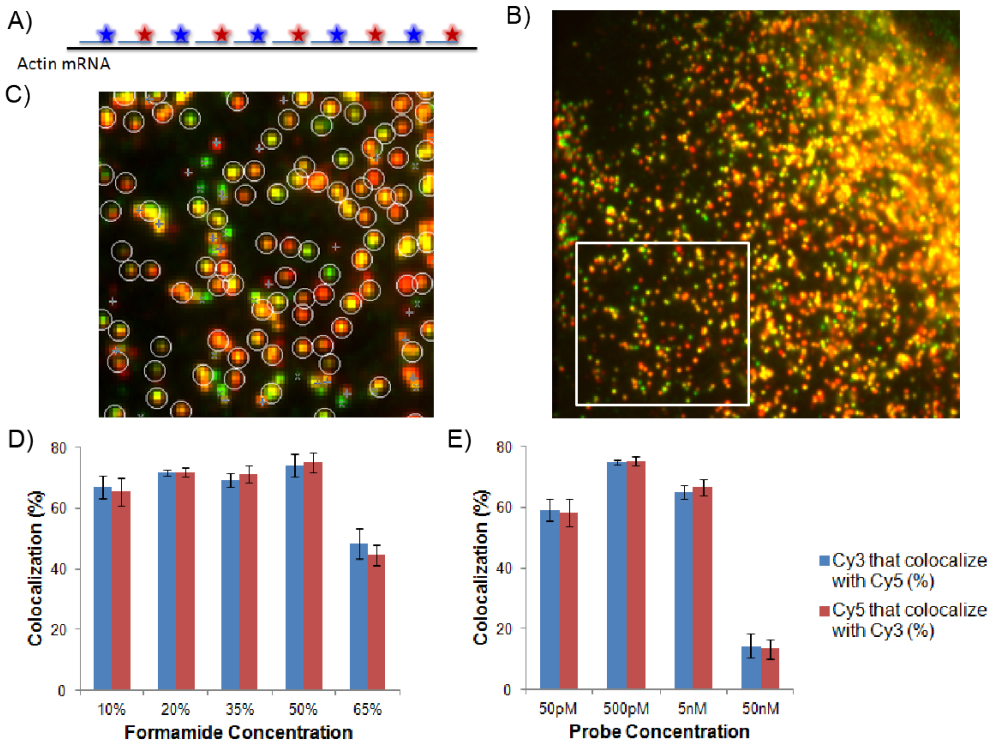


Figure 10: Optimization of smFISH protocol: A) Labeling of *ACTB* mRNA using two tiles of probes. Each tile consists of five probes. One of the tiles is tagged with Cy3 while the other is tagged with Cy5. B) Fluorescence image showing co-localization of Cy3 (green) and Cy5 (red) spots. Sample shown was stained at 50% formamide and 500pM (per probe concentration). White box indicates a zoomed in region displayed in C. C) Quantification of the degree of co-localization. Circles indicate co-localized spots; crosses indicate the position of a Cy3 spot without a corresponding Cy5 spot; plusses indicate the position of a Cy5 spot without a corresponding Cy3 spot. D) Fraction of co-localization under different formamide and E) probe concentration. The samples were stained at 37°C and washed at 47°C.

To estimate the hybridization efficiency per probe, five probes were used to label β-actin mRNAs in BSC-1 cells at 50% formamide and 2.5nM total concentration. The samples were illuminated for approximately two minutes, so that all the fluorescent molecules in the field of view were photo-bleached. The number of photo-bleaching steps for each fluorescent spot, from a sparse region containing no overlapping mRNAs, was counted from the photo-bleaching movies. Binomial fitting of the distribution of number of photo-bleaching steps gave an estimated hybridization efficiency per probe of ~40% (this value is similar to what was reported in the literature (*24*, *41*, *58*)). Therefore, the detection efficiency of mRNA with a tile of 5 probes is ~90%. This estimated number is higher than the value measured from the co-localization assay likely because the co-localization assay is more sensitive to background (background spots due to dust or cellular auto-fluorescence do not photo-bleach but will lower the degree of co-localization).

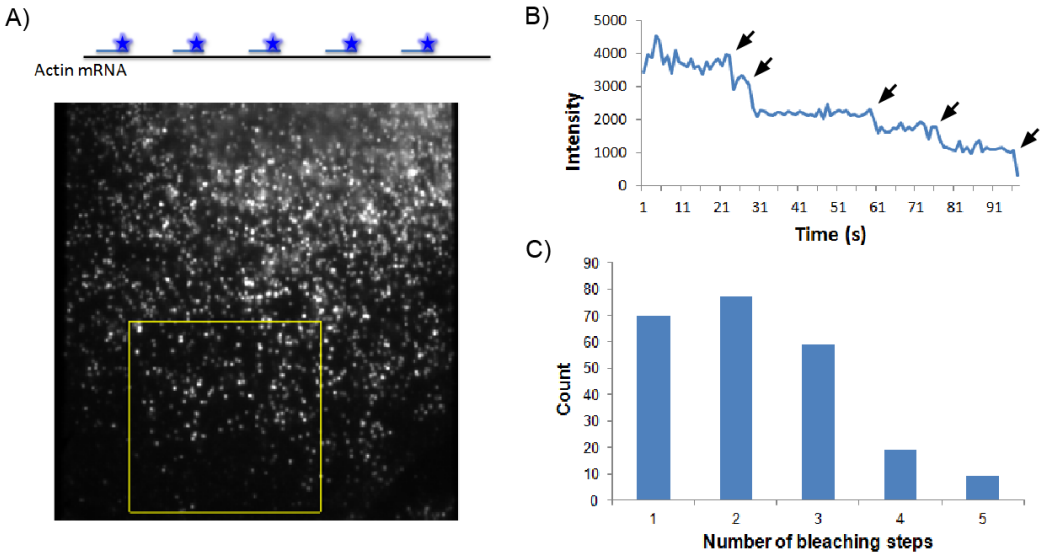


Figure 11: Estimation of hybridization efficiency. A) Labeling of β-Actin mRNA using a tile of 5 probes: Yellow box showed a sparse region of fluorescently labeled mRNAs in the cytoplasm of a cell that was used for photo-bleaching analysis. B) An example fluorescence time trace.

Each (Continued) arrow indicates a photo-bleaching step. C) Histogram of the number of photo-bleaching steps. Binomial fit to the distribution gave an estimated hybridization efficiency of 42%.

To test the feasibility of the two-step labeling approach, we tried to label one target transcript three times through the use of readout probes. We first labeled GAPDH mRNAs in BSC-1 cells with fifteen encoding probes split into three different tiles (Figure 12A). Then, we hybridized three different fluorescently labeled readout probes to the overhang of the encoding probes sequentially. Sequences for the encoding probes (purchased from IDT) are in appendix A. After each round of hybridization, we imaged the cells using STORM (Figure 12B) and completely photo-bleached all the signals before the next round of staining. We aligned the images from different rounds using beads as fiducial markers. We measured ~90% co-localizations of mRNA clusters between any two stains, indicating that the two-step protocol can stain and re-stain mRNAs with high fidelity and the sample was stable through 3 rounds of staining. This experiment also indicated that the readout probes can bind to readout sequences with high efficiency. No STORM localizations were observed after photo-bleaching or in control experiment where hybridization was performed with non-complementary readout probes.

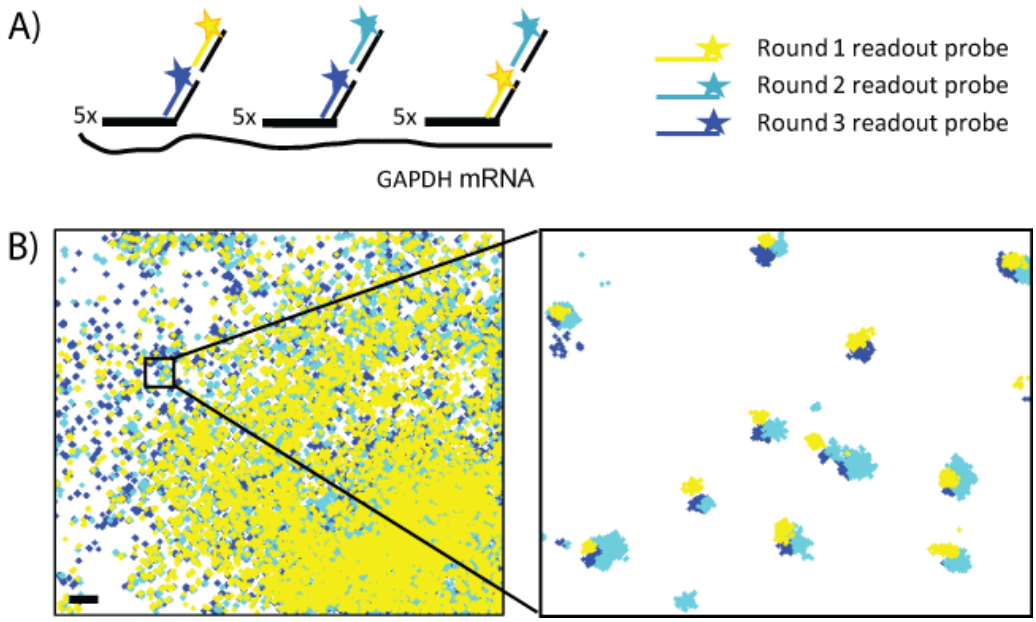


Figure 12: Two-step labeling of one gene using IDT synthesized oligos A) Schematic of experimental design. 3 tiles of encoding probes (5 probes in each tile) are first hybridized to target mRNAs. Then, 3 different fluorescently labeled readout probes are annealed to the overhangs of these encoding probes one at a time, staining 10 target sites on *GAPDH* mRNA each time. Between each round of hybridization, the cells were imaged with STORM and all the signals were photo-bleached. B) Results of experiment. STORM image of a labeled cell false colored based on the readout round. Each dot indicates a STORM localization. Zoom in shows localization clusters of putative *GAPDH* mRNAs. 90% of clusters from this cell contained signals from all three hybridization rounds.

2.2.2 Twelve and thirty two target genes using array derived oligos

To test the feasibility of using array derived oligos for MERFISH experiments, we designed encoding probes for twelve target transcripts (*THBS1*, *FLNA*, *TNC*, *DYNC1H1*, *SPTBN1*, *SPTAN1*, *PCNX*, *ASCC3, PRKCA, GTF3C4, POLQ,* and *STARD9*) with the eight bits extended Hamming code. Codebook assignment is in Appendix B. This codebook allowed a maximum of fourteen encoded words – we used twelve of them to encode transcripts while two of them were left as misidentification controls. We designed two types of misidentification

controls: blank words that were not represented with encoding probes and no-target words that had encoding probes identical to those designed for measured RNAs but had no cellular RNA target. The targeting regions of these probes were composed of random sequences that were screened against the human transcriptome to ensure that they contain no significant homology to any human RNA.

We designed ~192 encoding probes for each mRNAs to achieve a high signal to background ratio as well as a small error rate. As described previously, each encoding probe consisted of a targeting region and two flanking readout sequences that allow combinatorial read out of mRNA identities. To reduce the length of encoding probes, we designed each encoding probe to contain two of the four readout sequences associated with that RNA (Figure 13). Therefore, a maximum of ~96 readout probes can bind to each cellular RNA per hybridization round. These probes were custom ordered from Custom Array, Inc and enzymatically amplified to reach sufficient quantity for in situ hybridization experiments (Figure 9). We also switched the cell line to IMR90 (human fibroblast cells) because this cell line has a sequenced transcriptome and the cells are flat, allowing us to image almost the whole cell without z-sectioning. Lastly, we also implemented a computer controlled flow system for automated sequential labeling and imaging. Details of the flow and imaging setup can be found in section 2.4.4.

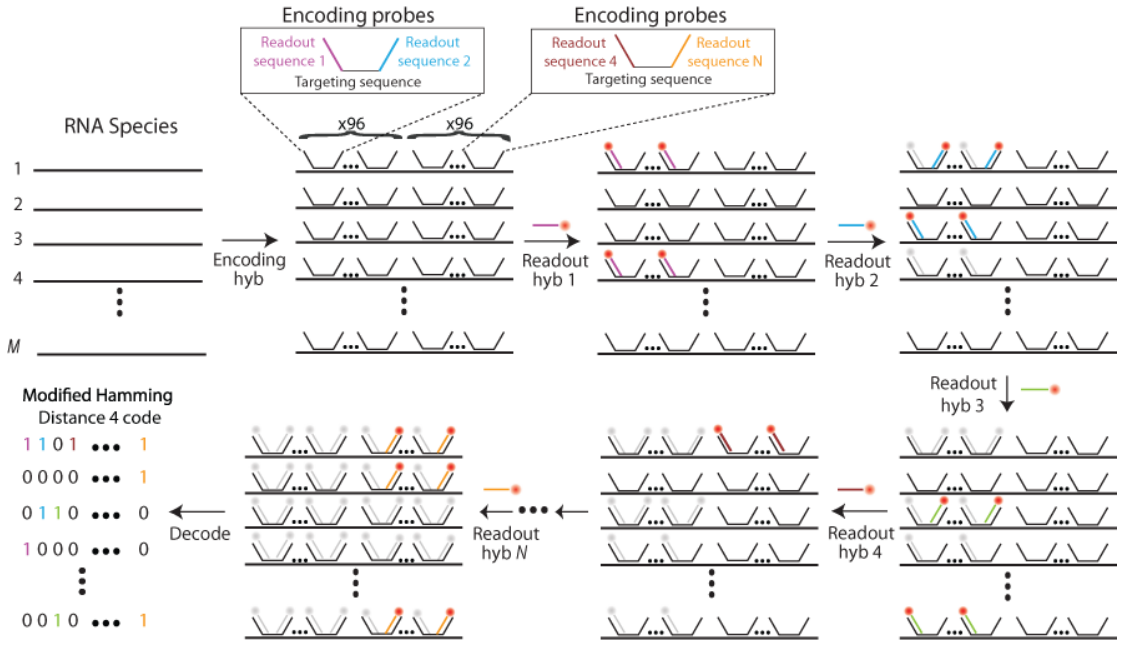


Figure 13: Schematic diagram of the implementation of a MHD4 code for MERFISH: Each RNA species is first labeled with ~192 encoding probes that convert the RNA into a unique combination of readout sequences. To reduce the length of encoding probes, we designed each encoding probe to contain two of the four readout sequences associated with that RNA. Encoding probes for a specific RNA species contain a unique combination of four of the N readout sequences, which correspond to the four hybridization rounds where this RNA should read '1'. For clarity only one possible pairing of the readout sequences is depicted; however all possible pairs of the four readout sequences are used at the same frequency and distributed randomly along each cellular RNA.

We labeled IMR90 cells with encoding probes derived from array-synthesized oligo-pool. We wash the cells to remove unbound or non-specifically bound probes and mount the sample in a flow chamber. We performed 8 additional readout hybridizations on the sample, taking an image each time before labeling with the next readout probe using conventional diffraction-limited wide-field imaging with oblique-incidence illumination geometry. After each round of hybridization with the fluorescent readout probes, fluorescent spots corresponding to individual RNAs were clearly detected (Figure 14A-B). We then constructed binary words from

the observed fluorescent spots based on their on-off patterns across the 8 hybridization rounds. By combining measurements from twenty cells, we counted 4840 clusters that matched our codebook and 2860 clusters that differed by one bit (Figure 14C).

Assuming that the probability of 1→0 error is equal and independent between hybridizations, we can use the ratio of perfectly matched words to one-bit error words to estimate the per bit false negative (1→0) and false positive (0→1) rate in our experiment (Equation i). Since all the encoded transcripts have four '1's in their code words, the probability of observing a perfectly matched word is P^4, where P is the per bit staining efficiency. The factor of ¼ is to account for the four different ways in which a RNA can lose or gain a bit. Therefore, the calculated per bit false negative and false positive rate for this experiment were 12% and 0.3% respectively.

$$\frac{\text{Perfectly matched clusters}}{\text{Clusters with one-bit error}} = \frac{P^4}{P^3 4(1-P)} \qquad \text{(i)}$$

We found that error correction increased the number of detected mRNA by ~2 fold, and this improvement is independent of the transcript's abundance. We detected only 1 word associated with our false positive controls and all the transcripts have higher count than the misidentification controls. To judge the accuracy of our measurement, we compared our measured abundances of these 12 mRNAs to the abundances measured from bulk RNA sequencing. Remarkably, we found that there is excellent agreement between these two measurements; the transcripts count measured using the multiplexed FISH approach correlated with gene expression measured using RNA-seq with a ρ of 0.96 (Figure 14D).

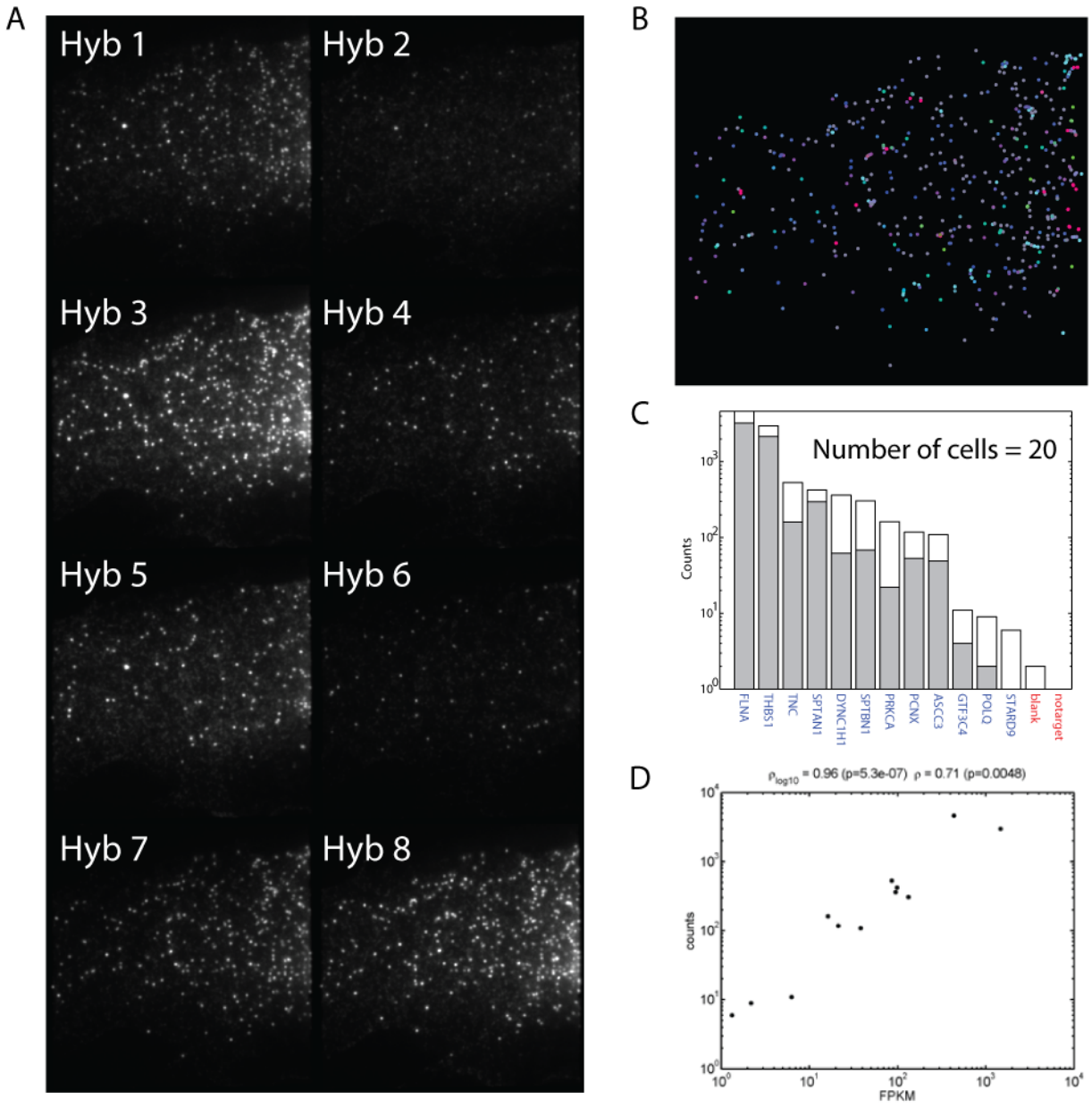


Figure 14: MERFISH measurement of 12 target transcripts. A) Images of RNA molecules in an IMR90 cell after each hybridization round (Hyb 1 – Hyb 8). B) The localizations of all detected single molecules in the cell colored based on their measured binary words. C) Histogram of counts before (grey) and after (white) error correction. D) Scatter plot of the measured mRNAs counts versus the abundance determined by bulk sequencing in fragments per kilobase per million reads (FPKM).

To test the feasibility of scaling up the throughput, we encoded thirty-two target transcripts (*CACNA1A, AFF4, THBS1, FLNA, TNC, DYNC1H1, SPTBN1, SPTAN1, PCNX, ASCC3, PRKCA, GTF3C4, POLQ, STARD9, TLN1, FBN1, MALAT1, NUP98, KIAA1199, GTF3C1, NOTCH2, IQGAP1, DOCK7, FLNB, TEAD1, TPR, SRRM2, SSH1, DNAJC13, KIAA1147, FRY,* and *BRCA2*) with the modified sixteen bits Hamming distance 4 codes, but truncated the codewords to twelve bits. This codebook allowed a maximum of thirty eight encoded words – we used thirty two of them to encode transcripts while six of them were left as misidentification controls. We performed MERFISH experiment as described above, but did twelve instead of eight readout hybridization rounds. Clear fluorescent spots were observed in all the hybridization rounds (Figure 15A-C). From 101 cells measured in one experiment, we counted 66,469 transcripts– 29,213 perfectly matched clusters and 37,256 clusters with one-bit error (Figure 15D). Error correction increased the number of detected transcripts by more than two fold, but it also increased the median count of misidentification controls from 0.01 per cell to 0.3 per cell, which is still at a very low level. Thirty transcripts were detected with counts above this value. Using the ratio of perfectly matched clusters to the clusters that gained or lost one bit, we estimated that the per bit false negative rate (1→0) was 19% and the per bit false positive rate (0→1) was 3% in this experiment. These low error rates suggest that the RNA calling rate is high and misidentification rate is low.

To test for potential technical bias in our MERFISH measurements, we probed the same 32 mRNAs with a different codebook by shuffling the code words among different mRNA species and changing the encoding probe sequences. Measurement with these two codebooks showed good agreement with a Pearson correlation coefficient of 0.97 (Figure 15E), again confirming that our misidentification rate is low. In addition, the measured abundances of these

32 mRNAs correlated with the abundances measured from bulk RNA sequencing with a ρ of 0.89 (Figure 15F), indicating that our MERFISH measurement was accurate.

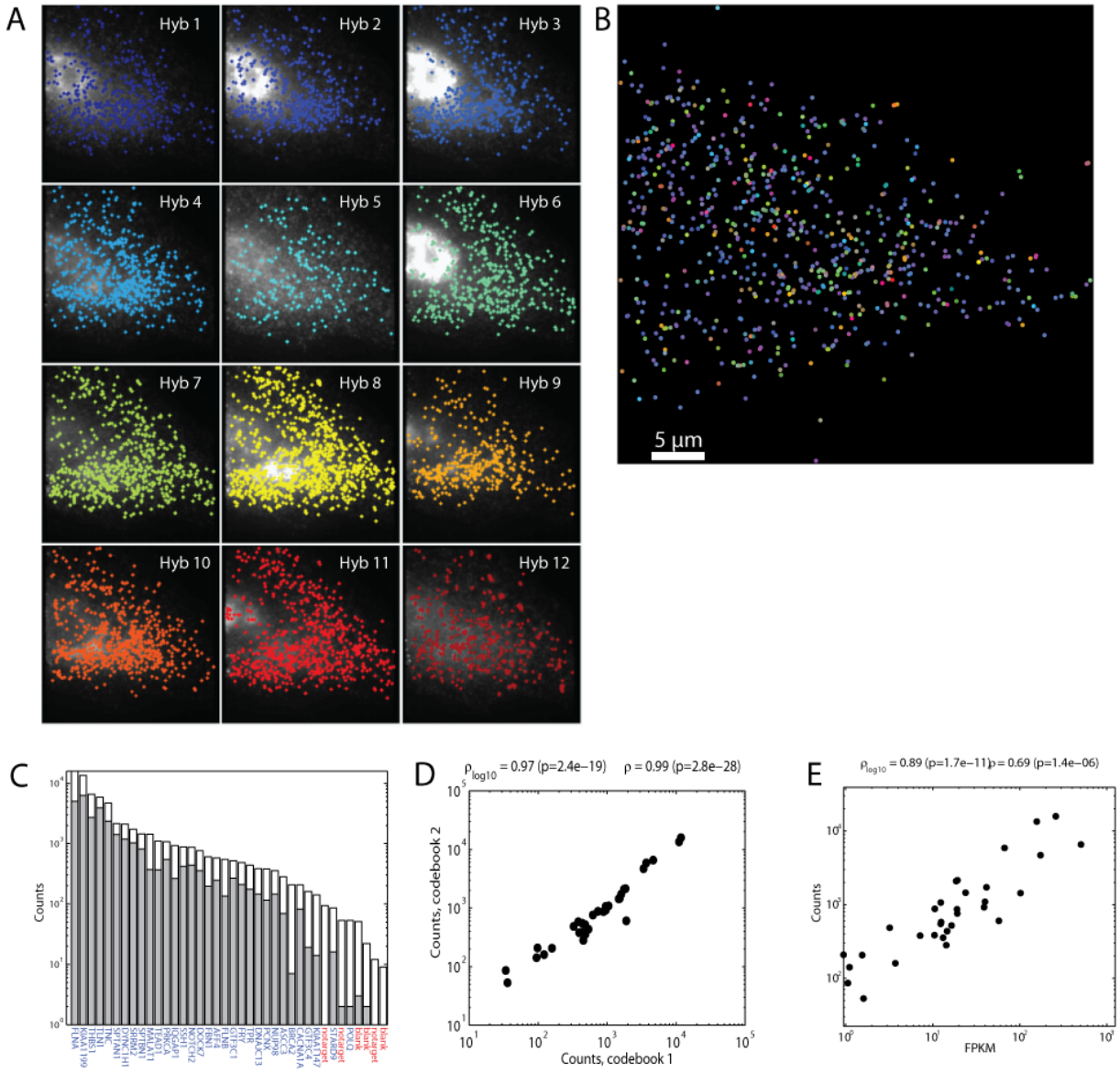


Figure 15: MERFISH measurement of 32 target transcripts: A) Images of RNA molecules in an IMR90 cell after each hybridization round (Hyb 1 – Hyb 12). False colored spots indicate identified molecules. B) The localizations of all detected single molecules in the cell colored based on their measured binary words. C) Scatter plot of the mRNAs abundances measured with

(Continued) two shuffled codebooks. D) Scatter plot of the measured mRNAs counts versus the abundance determined by bulk sequencing in fragments per kilobase per million reads (FPKM).

In addition, we performed the following optimization experiments to reduce the error rates in the MERFISH experiments. These optimization experiments were done with a shortened protocol; instead of performing twelve additional hybridizations after the first encoding hybridization, all the readout probes were added simultaneously in a single readout hybridization step. We then measured the brightness of mRNA spots under different formamide concentration, encoding probe concentration, and probe tile size in order to find the best labeling condition and probe design for the encoding probes.

First, we titrated the amount of formamide used in the encoding hybridization buffer to adjust the stringency of encoding hybridization (Figure 16 A-J). We did the titration experiment under both low (5μM) and high (50μM) concentrations of encoding probes because their concentration also affects their binding efficiency. Under both concentrations, if the amount of formamide used is too high, the encoding probes were unable to hybridize to the target transcripts, resulting in low fluorescence signal (Figure 16D and 16I). On the other hand, if the amount of formamide used is too low, there is high background (Figure 16E) most likely due to off target hybridization or poor signal (Figure 16A) potentially due to the formation of probe's secondary structure. Out of all the conditions tested, 30% formamide at 50uM encoding probe concentration appeared to give the brightness mRNA signal.

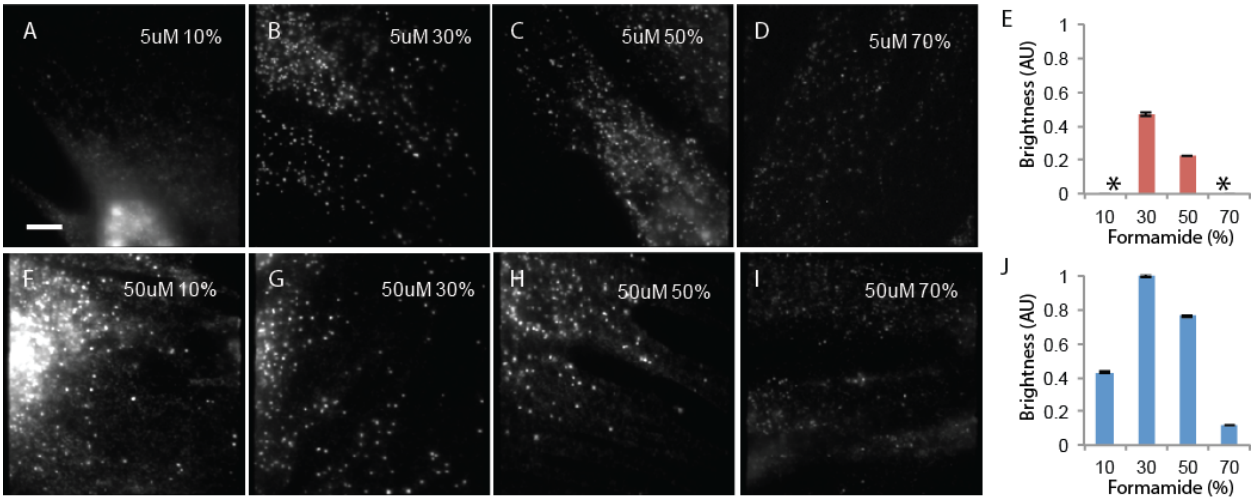


Figure 16: Optimization of hybridization stringency. Concentration of formamide used in the encoding hybridization buffer was varied from A, F) 10%, B, G) 30%, C, H) 50%, D, I) to 70%. Scale bars are 10μm. Relative brightness of mRNA spots for (E) 5μM and (J) 50μM probe concentrations. Spot brightness was calculated by setting an intensity threshold and calculating the area under Gaussian fitting of the spots. Stars indicate staining conditions where there were very few spots with brightness above threshold. Error bars indicate standard error of the mean.

We also varied the amount of encoding probe used during the encoding hybridization. We hybridized between 0 (negative control), 2, 8, 30, and 120μM of encoding probes in 30% formamide encoding buffer (Figure 17). We found that the brightness of mRNA spots increased with higher concentration of probes, but the amount of background also increased (Figure 17A-E). Since the fluorescence signals from mRNA spots are still much higher than the background fluorescence from the off target bindings of individual probe, we favor using greater than 30μM concentration of probes. We also observed no fluorescent spots when no encoding probes were added (Figure 17A), indicating that the readout probes had little off target background.

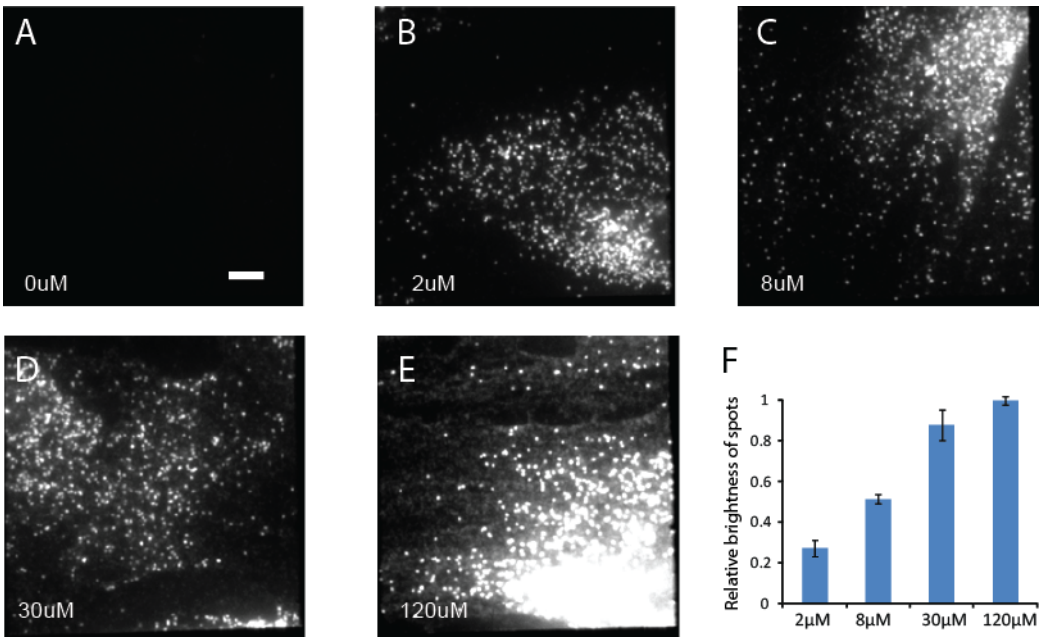


Figure 17: Optimization of probe concentration: Cells we stained with A) 0μM, B) 2μM, C) 8μM, D) 30μM, or E) 120μM of encoding probe. Scale bars are 10μm. F) Relative brightness of mRNA spots. Error bars indicate standard error of the mean.

Additionally, we designed an experiment to test the effect of probe tile size on the labeling efficiency of RNAs. For the same thirty-two target transcripts, we designed four different sets of experiments with 200, 100, 50 or 12 encoding probes for each RNA species. We found that the brightness of RNAs increases proportionally with the size of the tile used (Figure 18). This result suggests that using higher tile size will give better detection efficiency of RNAs. Furthermore, we explored a few alternative design of encoding probes: Encoding probes using two 30bps readout sequences (Figure 19A), one 30bps readout sequence (Figure 19B), or one 20bps readout sequence (Figure 19C) per probe were tested. We found that encoding probes with two 30bps readout sequences yielded similar spot brightness as encoding probes with one 30bps readout sequence, despite us using only half the tile size; indicating that doubling the amount of binding sites indeed resulted in twice the amount of readout out probes bound. Additionally, we

also observed that using 20bps readout probes resulted in similar spot brightness as 30bps readout probes, suggesting that they have similar binding efficiency.

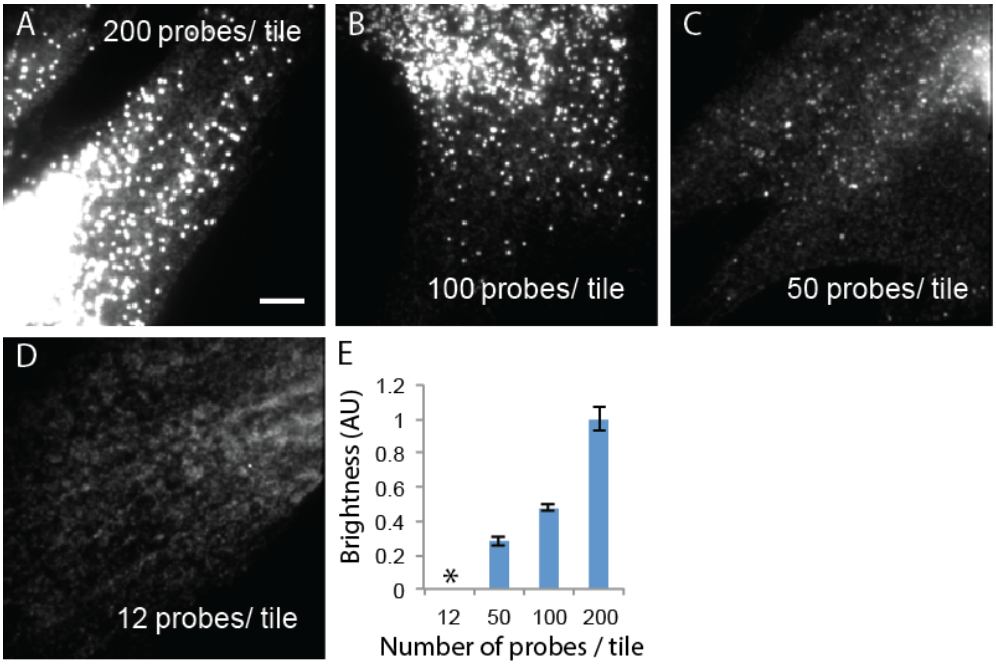


Figure 18: Optimization of tile size: Cells we stained with a tile of A) 200, B) 100, C) 50, or D) 12 encoding probes per RNA. Scale bars are 10μm. E) Relative brightness of mRNA spots. Error bars indicate standard error of the mean.

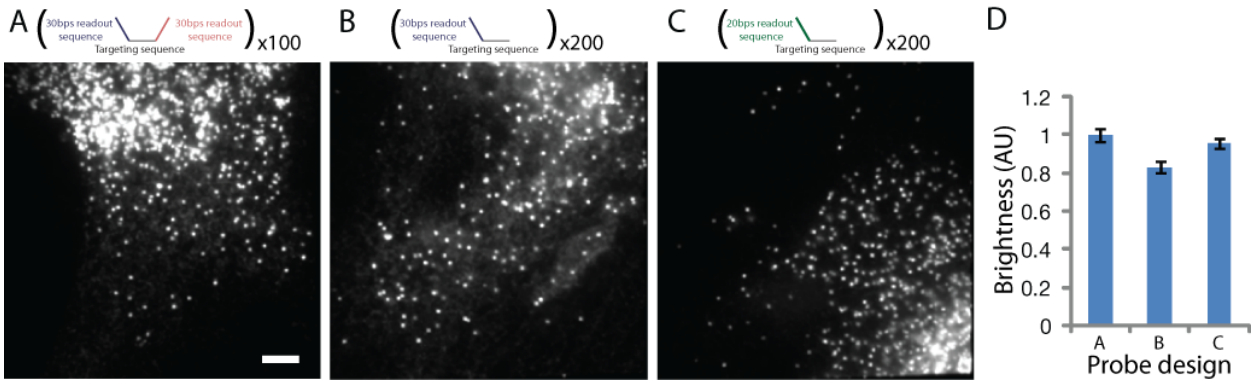


Figure 19: Optimization of probe design: Cells we stained with encoding probes containing A) two 30bps readout sequences but 100 probes per tile, B) one 30bps readout sequences and 200

(Continued) probes per tile, C) one 20bps readout sequences and 200 probes per tile. Scale bars are 10μm. D) Relative brightness of mRNA spots. Error bars indicate standard error of the mean.

2.2.3 One hundred and forty target genes with modified hamming distance 4 code

To test the feasibility of scaling up this error-robust, multiplexed imaging approach, we performed a 140-gene measurement on human fibroblast cells (IMR90) using a 16-bit MHD4 code to encode 130 RNA species while leaving 10 code words as misidentification controls (Appendix B). We used the same probe design and labeling scheme as the 32-gene experiment described above. Fluorescent spots corresponding to individual RNAs were clearly detected in all 16 rounds of hybridization and imaging (Fig. 20). After aligning the fluorescent spots observed in a cell from all imaging rounds, we constructed binary words from the fluorescent spots based on their on-off patterns across the 16 hybridization rounds (Fig. 21A). If the word exactly matched one of the 140 MHD4 code words (exact matches) or differed by only one bit (error-correctable matches), we assigned it to the corresponding RNA species (Fig. 21B-C). Within a cell, more than 1500 RNA molecules corresponding to 87% of the 130 encoded RNA species were detected after error correction (Fig. 21D). Similar observations were made in ~400 cells from 7 independent experiments. On average, ~4 times as many RNA molecules and ~2 times as many RNA species were detected per cell after error correction as compared with the values obtained before error correction.

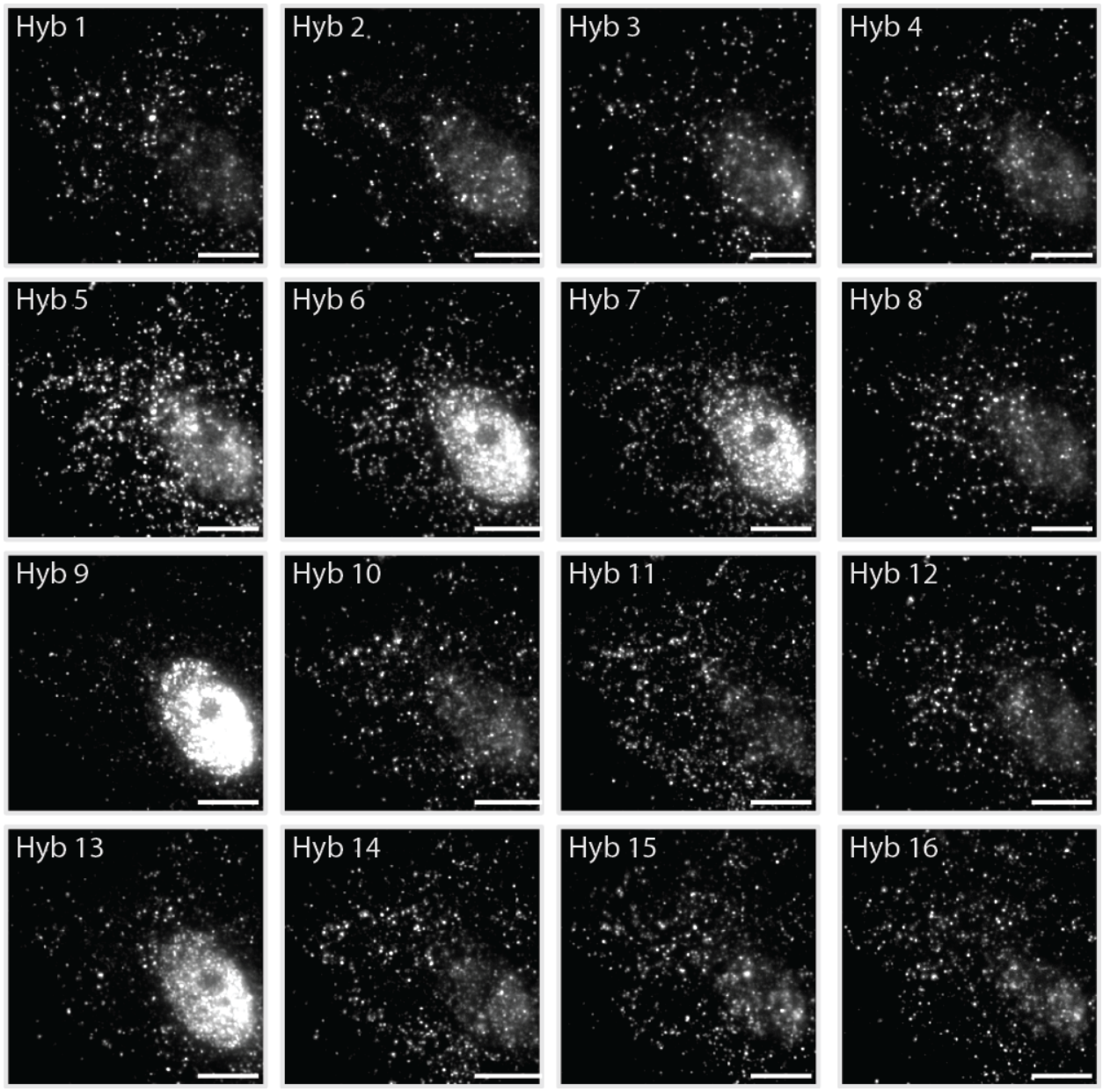


Figure 20: MERFISH measurement of 140 RNA species with a 16-bit MHD4 code: Images of RNA molecules in an example IMR90 cell after each hybridization round (Hyb 1 – Hyb 16).

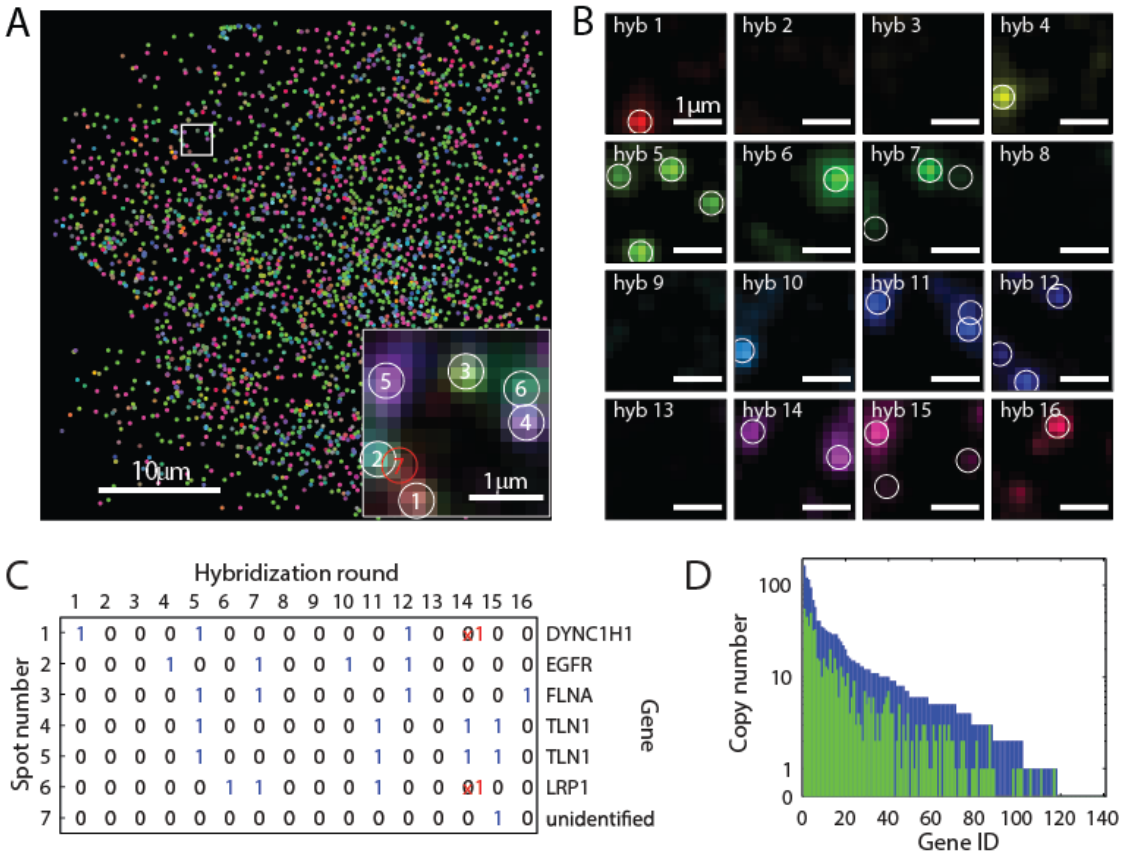


Figure 21: Decoding of 140 RNA species in single cells: A) The localizations of all detected molecules in a cell colored based on their measured binary words B) Fluorescent images from each round of hybridization for the boxed sub-region in A) with circles indicating potential RNA molecules. C) Corresponding code words for the spots identified in (B). Red indices represent the corrected bits. D) The RNA copy number for each gene observed without (green) or with (blue) error correction in this cell.

Two types of errors can occur in the copy number measurement of these RNAs: 1) Some molecules from other RNA species are misidentified as this RNA species, and 2) some molecules of this RNA species are not detected, leading to a drop in calling rate. To assess the extent of misidentification, we utilized the 10 misidentification control words — code words that were not associated with any cellular RNA. Although matches to these control words were observed, they occurred far less frequently than the real RNA-encoding words: 95% of the 130

RNA-encoding words were counted more frequently than these control words (observed at a median count of 0.9/cell with error correction, 0.04/cell without correction). Moreover, we typically found the ratio of the number of exact matches to the number of matches with one-bit errors for a real RNA-encoding word to be substantially higher than the same ratios observed for the misidentification controls (Fig. 22A, B). Using this ratio as a measure of the confidence in RNA identification, we found that 91% of the 130 RNA species had a confidence ratio greater than the maximum confidence ratio observed for the misidentification controls (Fig. 22C), demonstrating a high accuracy of RNA identification. Using the number of perfectly matched words and words with one-bit error, we computed the average per bit false negative (1→0) and false positive (0→1) rate to be 10% and 4% respectively. From these error rates, we estimated that our average calling rate after error correction to be ~80%, i.e. 80% of the fluorescent spots corresponding to a RNA species were decoded correctly. We note that although the remaining 20% of spots contributed to a loss in detection efficiency, most of them did not cause species misidentification because they were decoded as double-bit error words and discarded.

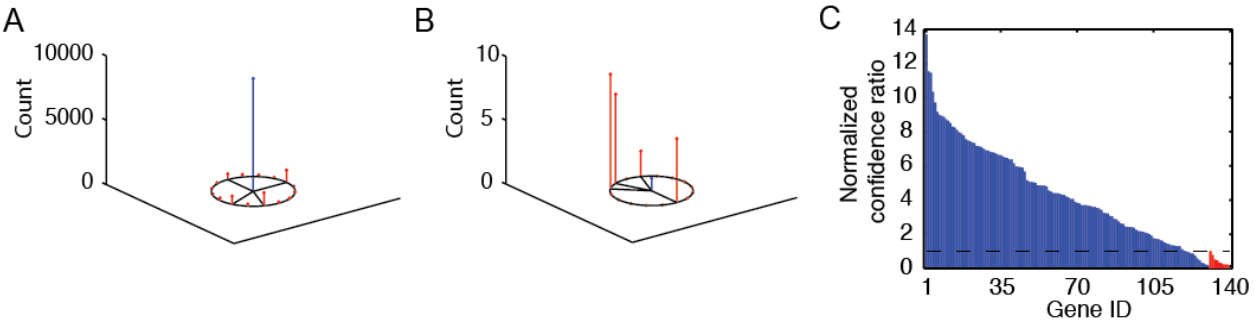


Figure 22: A) The number of measured words exactly matching the code word corresponding to FLNA, represented by the blue bar in the center of the circle, and the number of measured words with one-bit error compared to the code word of FLNA, represented by the 16 red bars on the circle. (B) The same as (A) but for a code word that was not assigned to any RNA — a misidentification control word. The solid lines connect the exact match to one-bit error words

(Continued) that are generated by 1→0 errors. Based on the observation that the ratio of the number of exact matches to the number of error-correctable matches for a real RNA-encoding word was typically substantially higher than the same ratios observed for the misidentification controls, we defined this ratio as a confidence ratio for RNA identification. C) The confidence ratio measured for all 130 RNA species (blue) and 10 misidentification control words not assigned to any RNA (red) using this 16-bit MHD4 code.

To measure the precision of our method, we repeated the 130 genes experiment and found that the measured RNA copy number agrees very well between different experiments (Fig. 23A). To further test for potential technical bias in our measurements, we probed the same 130 RNAs species with a different MHD4 codebook by shuffling the code words among different RNA species (Appendix B) and changing the encoding probe sequences. The copy numbers of individual RNA species per cell measured with these two codebooks (a total of 4 experiments with codebook 1 and 3 experiments with codebook 2) showed excellent agreement with a Pearson correlation coefficient of 0.94 (Fig. 23B), indicating that the choice of encoding scheme did not bias the measured counts. To further probe the difference between these two codebooks, we utilized the error-correction ability of the MHD4 code to determine the average 1→0 error rates and 0→1 error rates for each hybridization round for the two codebooks (Fig. 23C-F). Noticeably, a few hybridization rounds consistently produced higher error rates than the others, likely due to the different binding efficiencies of these readout probes to the encoding probes. Nonetheless, usage of two different codebooks yielded similar 1→0 error rates (10% on average) and 0→1 error rates (4% on average). As a result, measurements with different codes gave similar misidentification and calling rates (Fig. 24A-D). More importantly, because of the ability of MHD4 code to rectify bias from individual readout probe, the calling rates are largely independent of the genes' abundance.

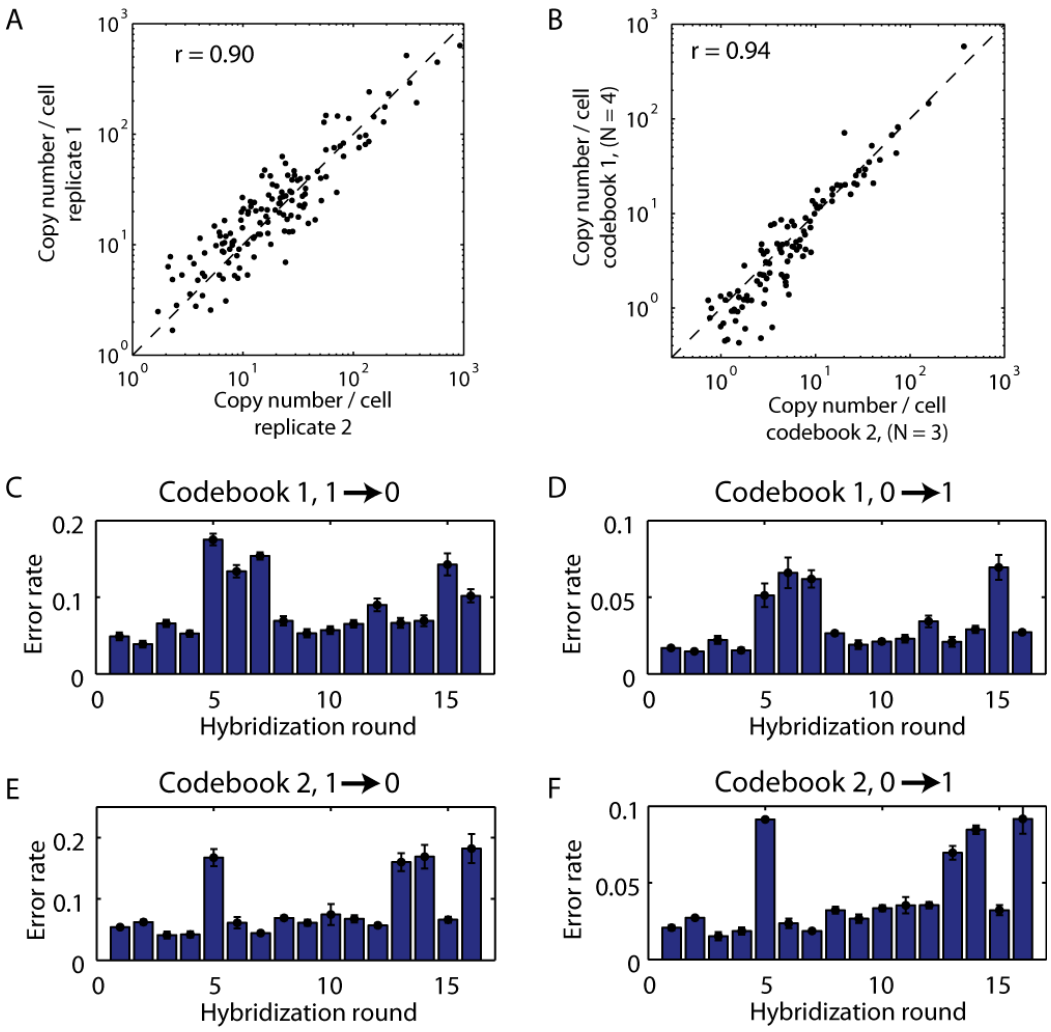


Figure 23: Reproducibility of MERFISH measurements: A) Scatter plot of the average copy number of each RNA species per cell measured in two replicated experiments. The Pearson correlation coefficient is 0.90 with a p-value of $1x10^{-49}$. The dashed line corresponds to the y = x line. B) Scatter plot of the average copy number of each RNA species per cell measured with two different shuffled MHD4 codebooks from a total of 4 experiments with codebook 1 and 3 experiments with codebook 2. The Pearson correlation coefficient is 0.94 with a p-value of $1x10^{-53}$. C - F) The average error rates for the 1→0 error (C, E) and 0→1 error (D, F) for each hybridization round for two different shuffled codebooks.

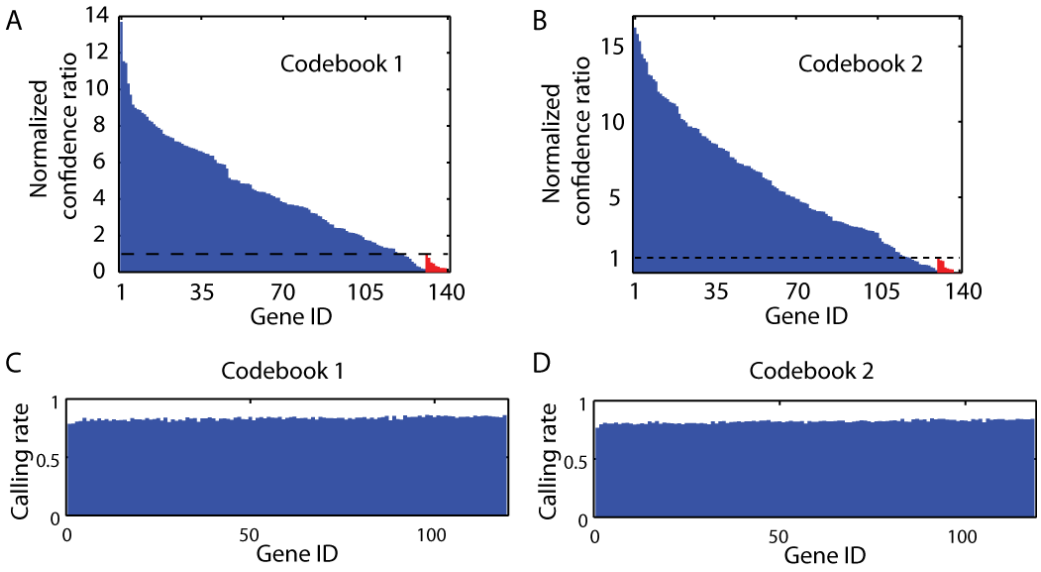
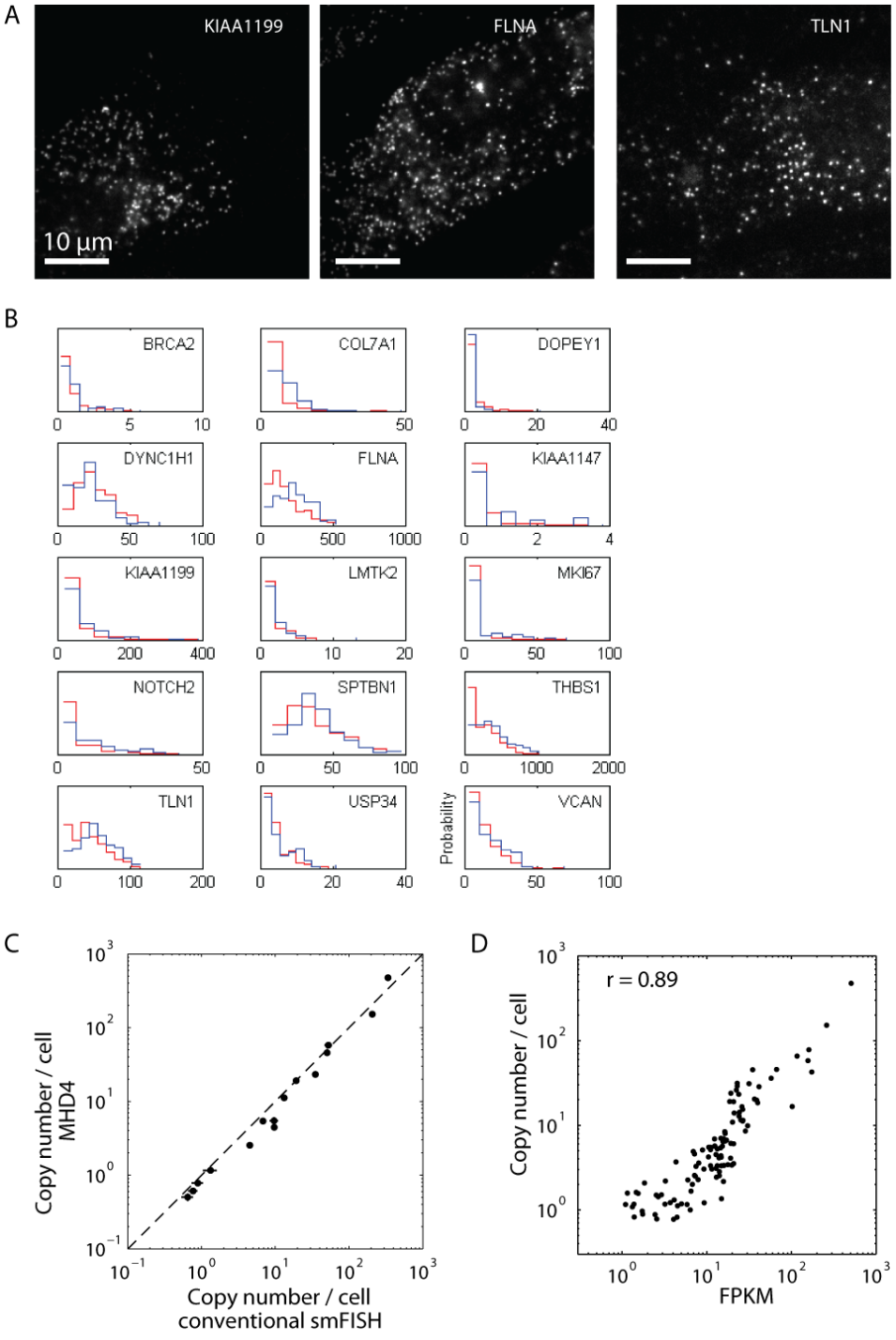


Figure 24: Performance of MERFISH with two different MHD4 encoding schemes: A - B) The average normalized confidence ratio measured for the 130 RNA species (blue) and the 10 misidentification control words not assigned to any RNA (red) for codebook 1 and 2. C - D) The calling rate for each RNA species measured using codebook 1 and 2 estimated from the 1→0 and 0→1 error rates in figure 25.

To measure the accuracy of our method, we compared the copy number derived from MERFISH measurements with the copy number measured with conventional smFISH and the gene expression measured with bulk RNA sequencing. We selected 15 out of the 130 transcripts, spanning the full measured abundance range, and performed smFISH measurements of these transcripts in IMR90 cells (Fig. 25A). For each of these genes, both the average copy number and the copy number distribution across many cells agreed quantitatively between our MERFISH and conventional smFISH measurements (Fig. 25B, C). The ratio of the copy numbers determined by these two approaches was 0.82 ± 0.06 (mean ± SEM across the 15 measured RNA species, Fig. 25C), which agrees with the estimated 80% calling rate for our MERFISH approach. The quantitative match between this ratio and our estimated calling rate over the full

measured abundance range additionally supports our assessment that the misidentification error was low. Given that the agreement between the MERFISH and conventional smFISH results extended to the genes at the lowest measured abundance, we estimate that our measurement sensitivity was better than 1 copy per cell.



(Continued) Figure 25: Accuracy of MERFISH measurements: A) Example smFISH images of three different transcripts (KIAA1199, FLNA, and TLN1). B) The distributions of RNA copy numbers in single cells for 15 genes. Red stairs: distributions constructed from ~400 cells in the 140-gene measurements using the MHD4 codes. Blue stairs: distributions constructed from ~100 cells in the conventional smFISH measurements. C) Comparison of the average RNA copy numbers per cell measured in the 140-gene experiments using the MHD4 codes to those determined by conventional smFISH for 15 genes. The average ratio of the copy number measured using the MHD4 measurements to that measured using conventional smFISH is 0.82 ± 0.06 (mean ± SEM across 15 genes). The dashed line corresponds to the y = x line. D) Scatter plot of the average copy number of each RNA species per cell versus the abundance determined by bulk sequencing in fragments per kilobase per million reads (FPKM). The Pearson correlation coefficient between the logarithmic abundances of the two measurements was 0.89 with a p-value of 3×10^{-39}.

As a final validation, we compared the abundance of each RNA species averaged over 400 cells from 7 independent measurements to those obtained from a bulk RNA sequencing measurement that we performed on the same cell line. Our imaging results correlated remarkably well with bulk sequencing results with a Pearson correlation coefficient of 0.89 (Fig. 25D).

In addition, the sample was stable throughout the 16 rounds of iterative labeling and imaging: The change in the number of fluorescent spots from round to round matched the predicted change based on the relative abundances of RNA species targeted in each round derived from bulk sequencing, and we did not observe a systematic decreasing trend with increasing number of hybridization rounds (Fig. 26A). The average brightness of the spots varied from round to round with a standard deviation of 40%, likely due to different binding efficiencies of the readout probes to the different readout sequences on the encoding probes (Fig. 26B). We observed only a small, systematic decreasing trend in the spot brightness with increasing

hybridization rounds, which was on average 4% per round (Fig. 26B). In addition, these fluorescent spots were efficiently extinguished with a brief photobleaching treatment (Fig. 26B). The remaining average brightness of the spots after photobleaching was similar to that of the autofluorescence of the cell. After photobleaching, the average number of fluorescent spots detected above threshold was 0.02% of the number of spots prior to photobleaching.

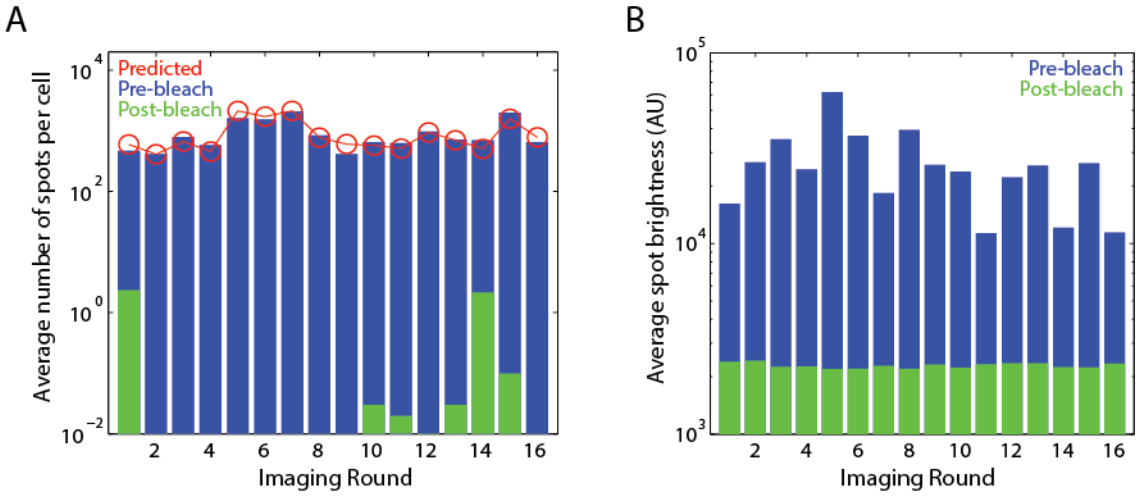


Figure 26: The number and average brightness of the fluorescent spots detected in the 16 rounds of hybridization before and after photobleaching averaged across all measurements with one of the 16-bit MHD4 code: A) The number of fluorescent spots observed per cell before (blue) and after (green) photobleaching as a function of hybridization round. Photobleaching reduces the number of fluorescent spots by two or more orders of magnitude. Hybridization rounds without green bars represent rounds in which no molecules were observed after bleaching. Also depicted is the expected change in the number of fluorescent spots from round to round (red circles) predicted based on the relative abundances of the RNA species targeted in each hybridization round derived from bulk RNA sequencing. The average discrepancy between the observed and predicted number of spots for each hybridization is only 15% of the mean number of spots. This discrepancy does not systematically increase with the number of hybridization rounds. (B) The average brightness of the identified fluorescent spots in each hybridization round averaged across all measurements with the first 16-bit MHD4 code both before (blue) and after (green) photobleaching. Brightness varies by 40% (standard deviation) across different hybridization

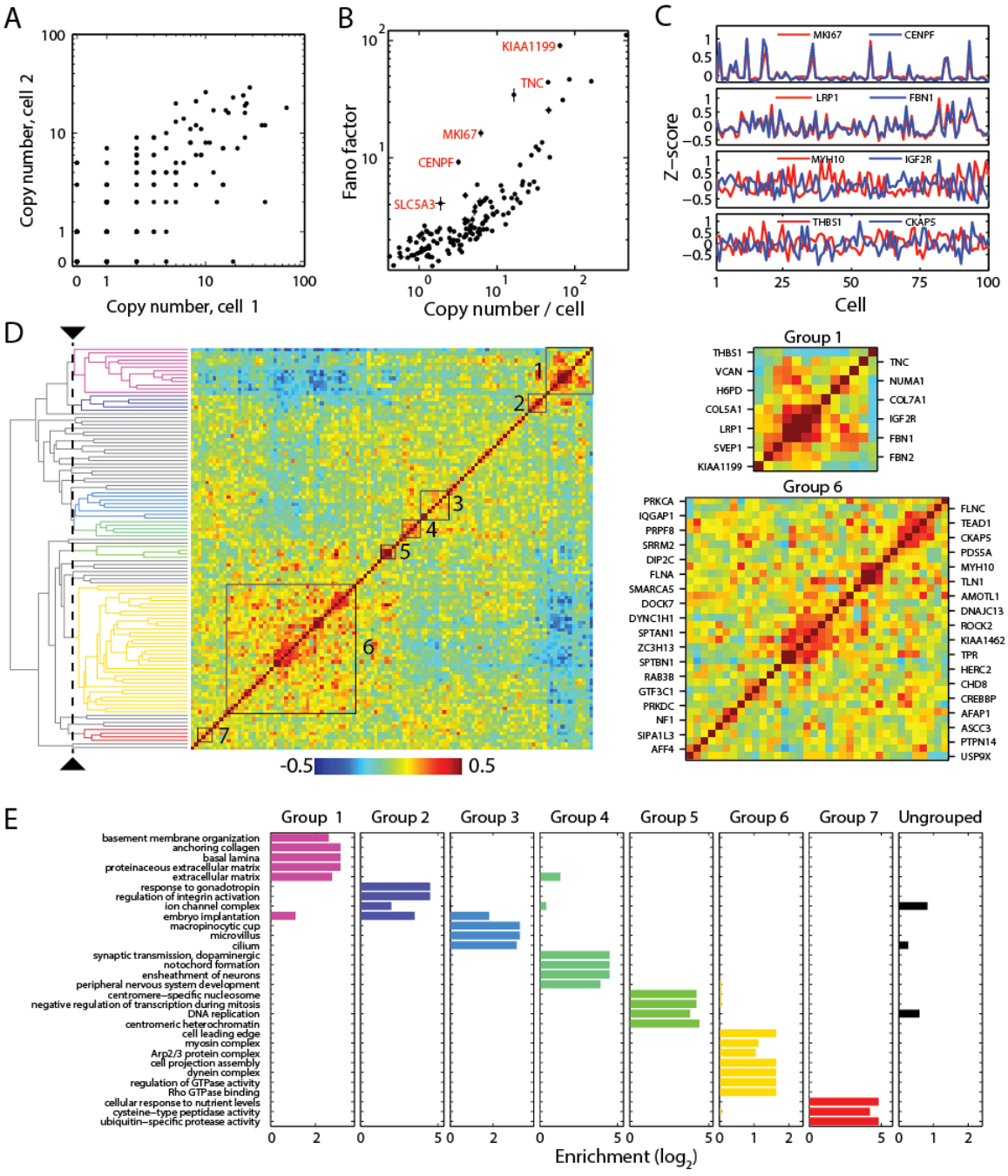
(Continued) rounds. The variation pattern is reproducible between experiments with the same code, likely due to differences in the binding efficiency of the readout probes to the different readout sequences. There is a small systematic trend of decrease in the brightness with increasing hybridization rounds, which is on average 4% per round.

2.2.3.1 Cell-to-cell variation in gene expression

The MERFISH approach allows parallelization of measurements of many individual RNA species and co-variation analysis between different RNA species. We first illustrated the parallelization aspect by examining the cell-to-cell variation in the expression level of each of the measured genes (Fig. 27A). To quantify the measured variation, we computed the Fano factors, defined as the ratio of the variance to the mean RNA copy number, for all measured RNA species. The Fano factors substantially deviated from 1, the value expected for a simple Poisson process, for many genes and exhibited an increasing trend with the mean RNA abundance (Fig. 27B), consistent with a previous observation for other cell types (*20*). A simple model for promoter regulation — the promoter stochastically switches between on and off states with global constraints on the kinetic rates — has been previously suggested to rationalize such a trend (*20*, *59*). Based on this model, this trend of increasing Fano factors with mean RNA abundance can be explained by changes in the transcription rate and/or promoter off-switching rates but not by changes in the promoter on-switching rate.

Moreover, we identified several RNA species with substantially larger Fano factors than this average trend. For example, we found that *SLC5A3, CENPF, MKI67, TNC* and *KIAA1199* displayed Fano factor values substantially higher than those of the other genes expressed at similar abundance levels. The high variability of some of these genes can be explained by their association with the cell cycle. For example, two of these particularly 'noisy' genes *MKI67* and

CENPF are both annotated as cell-cycle related genes (*60*), and based on their bimodal expression (Fig. 27C), we propose that their transcription is strongly regulated by the cell cycle. Other high-variability genes did not show the same bimodal expression patterns and are not known to be associated with the cell cycle. Understanding the origin and implications of noisy gene expression is an active topic of current research.



(Continued) Figure 27: Cell-to-cell variations and pairwise correlations for the RNA species determined from the 140 genes MERFISH measurements. A) Comparison of gene expression levels in two individual cells. B) Fano factors for individual genes. Error bars represent standard error of the mean determined from 7 independent data sets. C) Z-scores of the expression variations of four example pairs of genes showing correlated (top two) or anti-correlated (bottom two) variation for 100 randomly selected cells. Z-score is defined as the difference from the mean normalized by the standard deviation. D) Matrix of the pairwise correlation coefficients of the cell-to-cell variation in expression for the measured genes, shown together with the hierarchical clustering tree. The seven groups identified by a specific threshold on the cluster tree (dashed line) are indicated by the black boxes in the matrix and colored lines on the tree, with grey lines on the tree indicating ungrouped genes. Different threshold choices on the cluster tree could be made to select either smaller subgroups with tighter correlations or larger super-groups containing more weakly coupled subgroups. Two of the seven groups are enlarged on the right. E) Enrichment of 30 selected, statistically significantly enriched GO terms in the seven groups. Enrichment refers to the ratio of the fraction of genes within a group that have the specific GO term to the fraction of all measured genes having that term. Top 10 statistically significantly enriched GO terms for each of the seven groups are shown in Appendix C. Not all of the GO terms presented here are in the top 10 list.

2.2.3.2 Correlation analysis of gene expression

Analysis of co-variations in the expression levels of different genes can reveal which genes are co-regulated and elucidate gene regulatory pathways. At the population level, such analysis often requires the application of external stimuli to drive gene expression variation; hence, correlated expression changes can be observed among genes that share common regulatory elements influenced by the stimuli. At the single-cell level, one can take advantage of the natural stochastic fluctuations in gene expression for such analysis and can thus study multiple regulatory networks without having to stimulate each of them individually. Such co-

variation analysis can constrain regulatory networks, suggest new regulatory pathways, and predict function for unannotated genes based on associations with co-varying genes (*22*, *61*).

We applied this approach to the 140-gene measurements and examined the ~10,000 pairwise correlation coefficients describing how the expression levels of each pair of genes co-varied from cell to cell. Many of the highly variable genes showed tightly correlated or anti-correlated variations (Fig. 27C). To better understand the correlations for all gene pairs, we adopted a hierarchical clustering approach, commonly used in the analysis of both bulk and single-cell expression data (*62*, *63*), to organize these genes based on their correlation coefficients (Fig. 27D). Many genes that share a common regulatory pathway, such as *FBN1* and *FBN2* or *FLNA* and *FLNC*, are often grouped close together. Additionally, from the cluster tree structure, we identified seven groups of genes with substantially correlated expression patterns (Fig. 27D and Appendix C). Within each of the seven groups, every gene showed significantly stronger average correlation with other members of the group than with genes outside the group (Appendix C). To further validate and understand these groups, we identified gene ontology (GO) terms (*64*) enriched in each of these seven groups. Notably, the enriched GO terms within each group shared similar functions and were largely unique to each group (Fig. 27E and Appendix C), validating the notion that the observed co-variation in expression reflects some commonalities in the regulation of these genes.

Here, we describe two of these groups as illustrative examples. The predominant GO terms associated with Group 1 were terms associated with the extracellular matrix (ECM) (Fig. 27D, E and Appendix C). Notable members of this group included ECM components, such as FBN1, FBN2, COL5A, COL7A and TNC, and glycoproteins linking the ECM and cell membranes, such as VCAN and THBS1. The group also included an unannotated gene,

KIAA1199, which we would predict to play a role in ECM metabolism based on its association with this cluster. Indeed, this gene has recently been identified as an enzyme involved in the regulation of hyaluronan, a major sugar component of the ECM (*65*).

Group 6 contained many genes that encode vesicle transport proteins and proteins associated with cell motility (Fig. 27D, E and Appendix C). The vesicle transport genes included microtubule motors and related genes *DYNC1H*, *CKAP1*, and factors associated with vesicle formation and trafficking, like DNAJC13 and RAB3B. Again, we found an unannotated gene, *KIAA1462*, within this cluster. Based on its strong correlation with DYNC1H1 and DNAJC13, we predict that this gene may be involved in vesicle transport. The cell motility genes in this group included actin-binding proteins encoding genes like *AFAP1, SPTAN1, SPTBN1*, and *MYH10*, and genes involved in the formation of adhesion complexes, like *FLNA* and *FLNC*. Several GTPase-associated factors involved in the regulation of cell motility, attachment and contraction also fell into this group, including DOCK7, ROCK2, IQGAP1, PRKCA, and AMOTL1. The observation that some cell motility genes correlated with vesicle transport genes is consistent with the role of vesicle transport in cell migration (*66*). An additional interesting feature of group 6 is that a subset of these genes, in particular those related to cell motility, was anti-correlated with members of the ECM group discussed above (Fig. 27D). This anti-correlation may reflect regulatory interactions that mediate switching of cells between adherent and migratory states.

2.2.3.3 Spatial distribution patterns of RNA

As an imaging based approach, our method also allowed us to investigate the spatial distributions of many RNA species simultaneously. Several patterns emerged from the visual

inspection of individual genes, with some RNA transcripts enriched in the perinuclear region, some enriched in the cell periphery, and some scattered throughout the cell (Fig. 28A). To identify genes with similar spatial distributions, we determined the correlation coefficients for the spatial density profiles of all pairs of RNA species and organized these RNAs based on the pairwise correlations again using the hierarchical clustering approach. The correlation coefficient matrix showed groups of genes with correlated spatial organizations, and the two most notable groups with the strongest correlations are indicated in Fig. 28B. Group I RNAs appeared enriched in the perinuclear region whereas group II RNAs appeared enriched near the cell peripheral region (Fig. 28C). Quantitative analysis of the distances between each RNA molecule and the cell nucleus or the cell periphery indeed confirmed this visual impression (Fig. 28D).

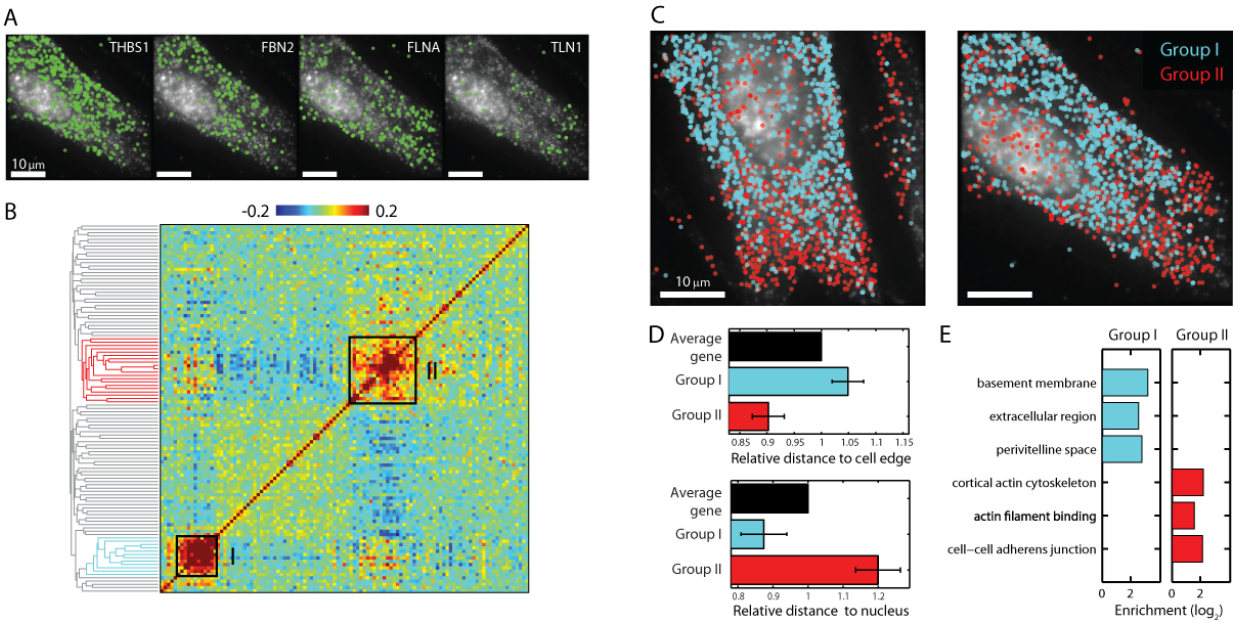


Figure 28: Distinct spatial distributions of RNAs observed in the 140-gene measurements: A) Examples of the spatial distributions observed for four different RNA species in a cell. B) Matrix of the pairwise correlation coefficients describing the degree with which the spatial distributions of each gene pair are correlated, shown together with the hierarchical clustering tree. Two strongly correlating groups are indicated by the black boxes on the matrix and color on the tree.

(Continued) C) The spatial distributions of all RNAs in the two groups in two example cells. Cyan symbols: group I genes; Red symbols: group II genes. D) Average distances for genes in group I and genes in group II to the cell edge or the nucleus normalized to the average distances for all genes. Error bars represent SEM across 7 data sets. E) Enrichment of GO terms in each of the two groups.

Group I contained genes encoding extracellular proteins such as FBN1, FBN2 and THSB1, secreted proteins such as PAPPA, and integral membrane proteins such as LRP1 and GPR107. These proteins have no obvious commonalities in function. Rather a GO analysis showed significant enrichment for location terms, such as extracellular region, basement membrane, or perivitelline space (Fig. 28E). To reach these locations, proteins must pass through the secretion pathway, which often requires translation of mRNA at the endoplasmic reticulum (ER) (*67*, *68*). Thus, we propose that the spatial pattern that we observed for these mRNAs reflects their co-translational enrichment at the ER. The enrichment of these mRNAs in the perinuclear region (Fig. 28C, D, cyan), where the rough ER resides, supports this conclusion.

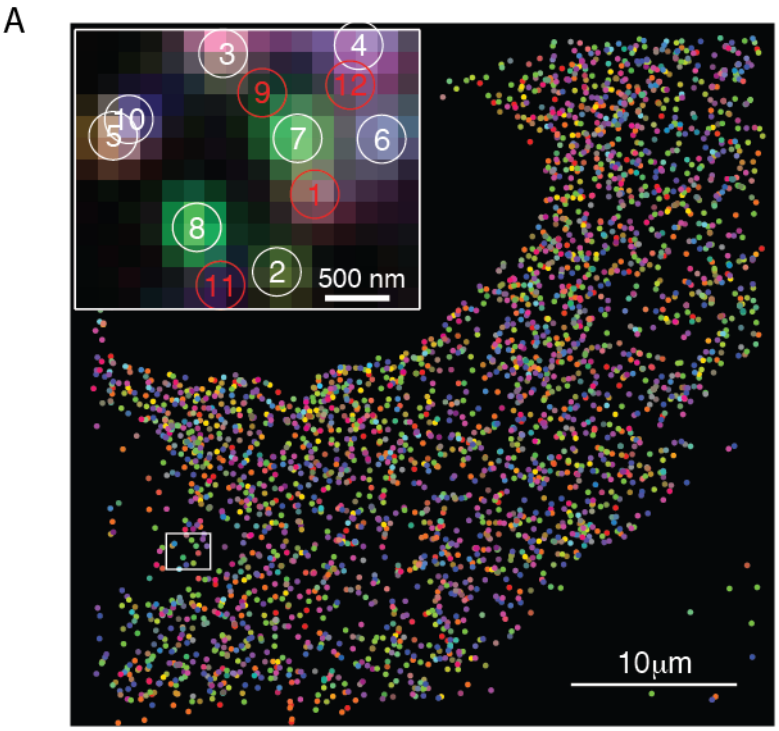
Group II contained genes encoding the actin-binding proteins, including filamins FLNA and FLNC, talin TLN1, and spectrins SPTAN1 and SPTBN1; the microtubule-binding protein CKAP5; and the motor proteins MYH10 and DYNC1H1. This group was enriched with GO terms such as cortical actin cytoskeleton, actin filament binding, and cell-cell adherens junction (Fig. 28E). It has been shown previously that β-actin mRNA is enriched near the cell periphery in fibroblasts as are mRNAs that encode members of the actin-binding Arp2/3 complex (*69*, *70*). The enrichment of group II mRNAs in the peripheral region of the cells (Fig. 28C, D) suggests that the spatial distribution of the Group II genes might be related to the distribution of actin cytoskeleton mRNAs.

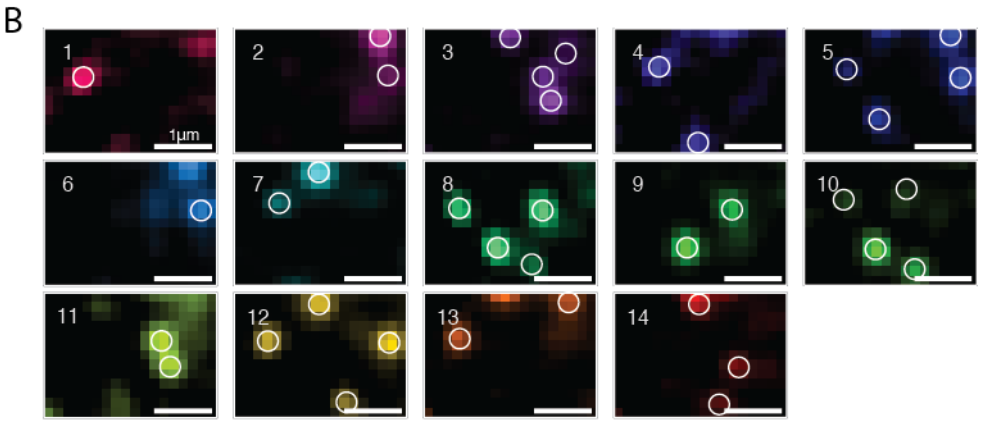
2.2.4 One thousand genes with modified hamming distance 2 code

Finally, we sought to further increase the throughput to ~1000 RNA species. This increase could be achieved with our MHD4 code by increasing the number of bits per code word to 32 while maintaining the number of '1' bits per word at four (Fig. 5B). While the stability of our samples across many hybridization rounds (Fig. 26) suggests that such an extension is potentially feasible, we pursued an alternative approach that did not require an increase in the number of hybridizations by relaxing the error correction requirement but keeping the error detection capability. For example, by reducing the Hamming distance from 4 to 2, we could use all 14-bit words that contain four '1' bits to encode 1001 genes and probe these RNAs with only 14 rounds of hybridization. However, because a single error can produce a word equally close to two different code words, error correction is no longer possible for this modified Hamming-distance-2 (MHD2) code. Hence we expect the calling rate to be lower and the misidentification rate to be higher with this encoding scheme.

To evaluate the performance of this 14-bit MHD2 code, we set aside 16 of the 1001 possible code words as misidentification controls and used the remaining 985 words to encode cellular RNAs (Appendix B). Among these 985 RNAs, we included 107 RNA species probed in the 140-gene experiments as an additional control. We performed the 1001-gene experiments in IMR90 cells using a similar procedure as described above. To allow all encoding probes to be synthesized from a single 100,000-member oligopool, we reduced the number of encoding probes per RNA species to ~94. Fluorescent spots corresponding to individual RNA molecules were again clearly detected in each round of hybridization with the readout probes (Fig. 29A). Based on their on-off patterns, these spots were decoded into RNA (Fig. 29A-C). 581 RNA

species were detected in the cell shown in Fig. 29A, and similar results were obtained in ~200 imaged cells in 3 independent experiments.





C

Hybridization round

Spot number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	unidentified
2	0	0	0	0	0	0	0	1	0	1	0	1	0	1	FLNC
3	0	0	1	0	0	0	1	0	0	0	0	1	0	1	ZFPM2
4	0	1	0	0	1	1	0	0	0	0	0	0	1	0	LTBP2
5	1	0	0	0	0	0	0	1	0	0	0	1	1	0	TP53BP1
6	0	1	0	0	1	1	0	0	0	0	0	1	0	0	INTS1
7	0	0	1	0	0	0	0	1	1	0	1	0	0	0	PINK1–AS
8	0	0	0	0	1	0	0	1	1	1	0	0	0	0	MLLT4
9	0	0	0	0	0	0	0	0	0	1	0	0	0	0	unidentified
10	0	0	0	1	1	0	1	0	0	1	0	0	0	0	TRIM14
11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	unidentified
12	0	0	1	0	0	0	0	0	0	0	0	0	0	0	unidentified

(Continued) Figure 29: Simultaneous measurements of 1001 genes in single cells using MERFISH with a 14-bit MHD2 code: A) The localizations of all detected single molecules in a cell colored based on their measured binary words. Inset: the composite, false-colored fluorescent image of the 14 hybridization rounds for the boxed sub-region with numbered circles indicating potential RNA molecules. Red circles indicate unidentifiable molecules, the binary words of which do not match any of the 14-bit MHD2 code words. B) Images of individual hybridization round for the insert from (A). C) The corresponding binary word for each of the spots identified in (B) with the RNA species to which it is decoded. 'unidentified' implies that the measured binary word does not match any of the 1001 code words.

77% of all real RNA words were detected more frequently than the median count for the misidentification controls (~0.3/cell) instead of the 95% value observed in the MHD4 measurements. Using the same confidence ratio analysis as described above, we found that 73% (instead of 91% for the MHD4 measurements) of the 985 RNA species were measured with a confidence ratio larger than the maximum value observed for the misidentification controls (Fig. 30A). RNA copy numbers measured from these 73% RNA species showed excellent correlation with our bulk RNA sequencing results (Pearson correlation coefficient r = 0.76; Fig. 30B, black). It is worth noting that the remaining 27% of the genes still exhibit good, albeit lower, correlation with the bulk RNA sequencing data (r = 0.65; Fig. 30B, red), but we took the conservative measure of excluding them from further analysis. The lack of an error correction capability also decreased the calling rate of each RNA species: When comparing the 107 RNA species common in both the 1001-gene and 140-gene measurements, we found that the copy numbers per cell of these RNA species were lower in the 1001-gene measurements (Fig. 30C). The total count of these RNAs per cell was ~1/3 of that observed in the 140-gene measurements. Thus the lack of error correction in the MHD2 code produced a ~3-fold decrease in the calling rate, which is

consistent with the 3-4 fold decrease in calling rate observed for the MHD4 code when error correction was not applied. Using the ratio of perfectly matched words to words with single-bit lost or single-bit gain, we estimated that the per bit false negative rate to be 19% and positive rate to be 4%, which is similar to the values measured from the 140-genes experiments, further confirming that the ~3-fold decrease in calling rate is mainly due to the switch in encoding scheme. Despite the expected reduction in calling rate, the good correlations found between the copy numbers observed in the 1001-gene measurements and those observed in the 140-gene measurements, as well as in bulk RNA sequencing measurements, indicates that the relative abundance of these RNAs can be quantified with the MHD2 encoding scheme.

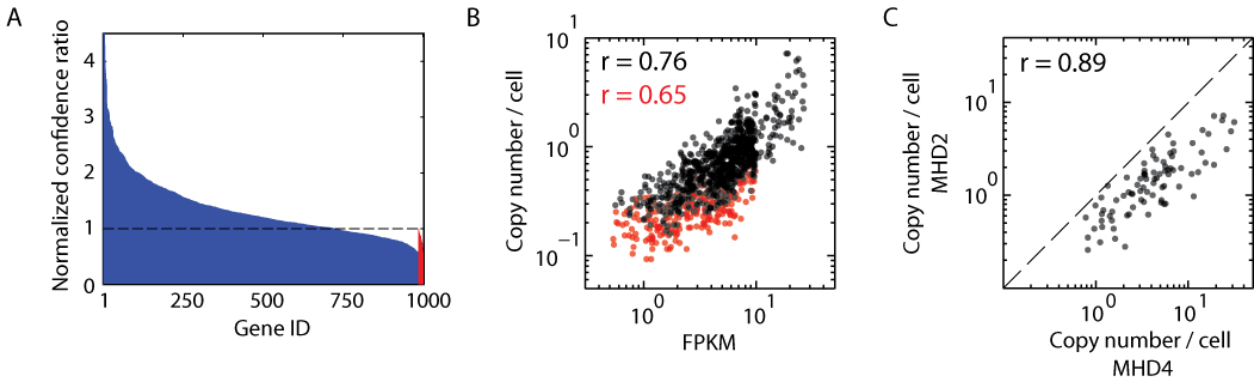


Figure 30: Precision and accuracy of the 1001 genes MERFISH measurement: A) The normalized confidence ratios measured for the 985 RNA species (blue) and the 16 misidentification control words not targeted to any RNA (red). B) Scatter plot of the average copy number per cell measured in the 1001-gene experiments versus the abundance measured via bulk sequencing. The black symbols are for the 73% of genes detected with confidence ratios higher than the maximum ratio observed for the misidentification controls. The Pearson correlation coefficient is 0.76 with a p-value of 3x10-133. The red symbols are for the remaining 27% of genes. The Pearson correlation coefficient is 0.65 with a p-value of 3x10-33. C) Scatter plot of the average copy number for the 107 genes shared in both the 1001-gene measurement with the MHD2 code and the 140-gene measurement with the MHD4 code. The Pearson

(Continued) correlation coefficient is 0.89 with a p-value of 9x10-30. The dashed line is correspond to the y = x line.

Simultaneously imaging ~1000 genes in individual cells substantially expanded our ability to detect co-regulated genes. Fig. 31A shows the matrix of pairwise correlation coefficients determined from the cell-to-cell variations in the expression levels of these genes. Using the same hierarchical clustering analysis as described above, we identified ~100 groups of genes with correlated expression (Appendix D). Remarkably, nearly all of these ~100 groups showed statistically significant enrichment of functionally related GO terms (Fig. 31B; Appendix D). These included some of the groups identified in the 140-gene measurements, such as the group associated with cell replication genes and the group associated with cell motility genes (Fig. 31A, B, groups 7 and 102), as well as many new groups. The groups identified here included 46 RNA species lacking any previous GO annotations, for which we can now hypothesize function based on their group association (Appendix D). For example, *KIAA1462* is part of the cell motility group, as also shown in the 140-gene experiments, suggesting a potential role of this gene in cell motility (Fig. 31A, group 102). Likewise, *KIAA0355* is part of a new group enriched in genes associated with heart development (Fig. 31A, group 79), and *C17orf70* is part of a group associated with ribosomal RNA processing (Fig. 31A, group 22). Using these groupings, we can also hypothesize cellular functions for 61 transcription factors and other partially annotated proteins of unknown functions. For example, the transcription factors Z3CH13 and CHD8 are both members of the cell motility group, suggesting their potential role in the transcriptional regulation of cell motility genes. While these predicted functions based on gene-association analysis require further validation, our co-variation data provide a resource for generating hypotheses on gene function and regulation.

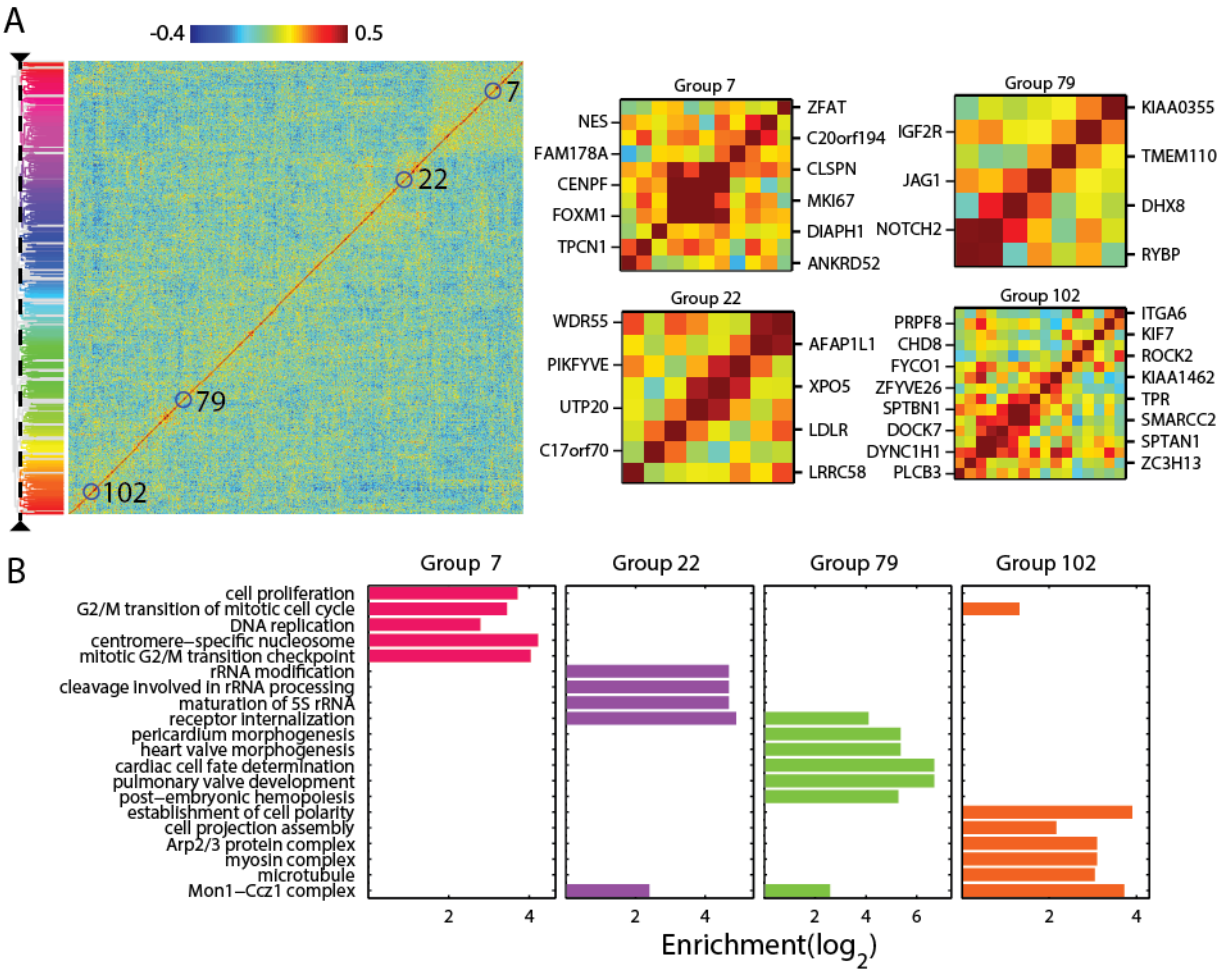


Fig. 31: Co-variation analysis of the RNA species measured in the 1001 genes MERFISH measurements. A) Matrix of all pairwise correlation coefficients of the cell-to-cell variation in expression for the measured genes shown with the hierarchical clustering tree. The ~100 identified groups of correlated genes are indicated by color on the tree. Zoom in of four of the groups described in the text are shown on the right. B) Enrichment of 20 selected, statistically significantly enriched GO terms in the four groups. The statistically most significantly enriched GO terms (maximum 10) for each of the ~100 groups are shown in Appendix D.

2.5 Discussion

In summary, we have developed a highly multiplexed, error robust FISH (MERFISH) scheme for transcriptome scale RNA imaging in single cells. Using combinatorial labeling,

sequential hybridization and imaging, and two different error-robust encoding schemes, we simultaneously imaged either 140 or 1001 genes in hundreds of individual human fibroblast cells. Of the two encoding schemes presented here, the MHD4 code is capable of both error detection and error correction, and hence can provide a higher calling rate and a lower misidentification rate than the MHD2 code, which instead can only detect but cannot correct errors. MHD2, on the other hand, provides a faster scaling of the degree of multiplexing with the number of bits than MHD4. Other error-robust encoding schemes can also be used for such multiplexed imaging, and experimenters can set the balance between detection accuracy and ease of multiplexing based on the specific requirements of the experiments.

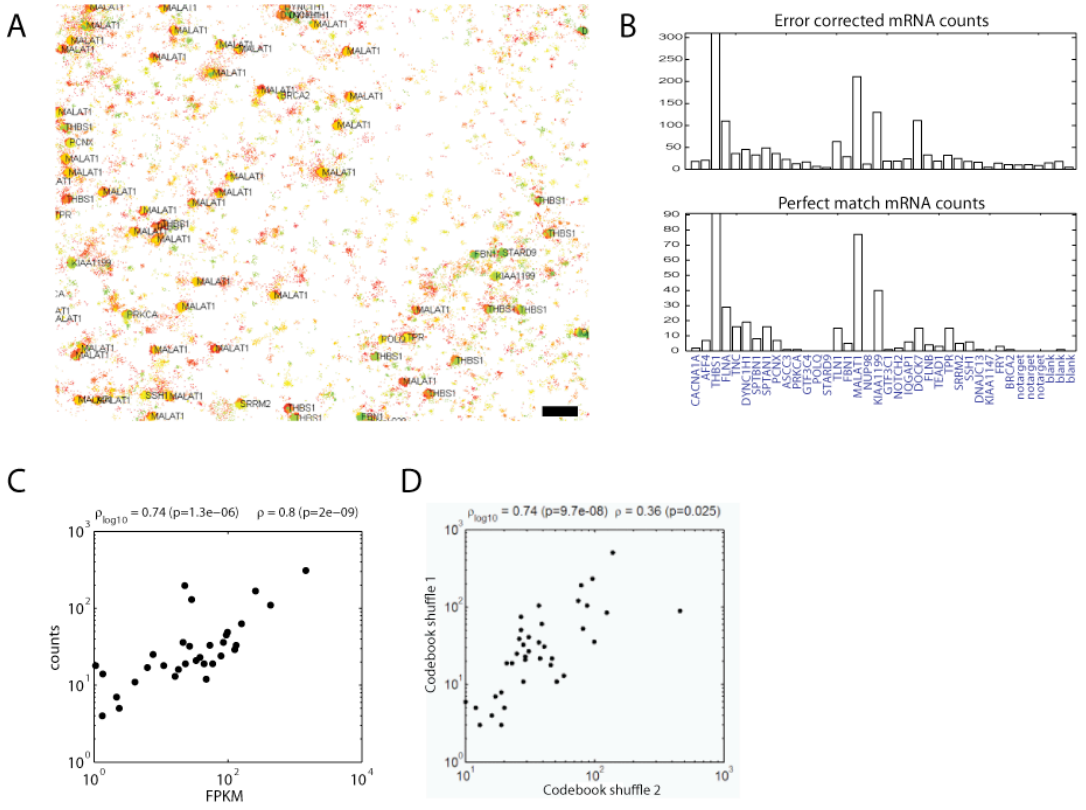
By increasing the number of bits in the code words, it should be possible to further increase the number of detectable RNA species with either a MHD4 or MHD2 code. For example, using the MHD4 code with 32 total bits and four or six '1' bits would increase the number of addressable RNA species to 1,240 or 27,776, respectively — the latter is the approximate scale of the human transcriptome. The predicted misidentification and calling rates are still reasonable for the 32-bit MHD4 code (shown in Fig. 5C, D, magenta for the MHD4 code with four '1' bits and similar rates were calculated for the MHD4 code with six '1' bits). If more accurate measurements are desired, an additional increase in the number of bits would allow the use of encoding schemes with a Hamming distance greater than 4, further enhancing the error detection and correction capability. While an increase in the number of bits by adding more hybridization rounds would increase the data collection time and potentially lead to sample degradation, these problems could be mitigated by utilizing multiple colors to readout multiple bits in each round of hybridization.

As the degree of multiplexing is increased, it is important to consider the potential increase in the density of RNAs that need to be resolved in each round of imaging. Based on our imaging and sequencing results, we estimate that including the whole transcriptome of the IMR90 cells would lead to a total RNA density of ~200 molecules/μm^3. Using our current imaging and analysis methods, we could resolve 2-3 molecules/μm^3 per hybridization round (*71*), which would reach a total RNA density of ~20 molecules/μm^3 after 32 rounds of hybridization. This density should allow all but the top 10% most expressed genes to be imaged simultaneously or a subset of genes with even higher expression levels to be included. By utilizing more advanced image analysis algorithms to better resolve overlapping images of individual molecules, such as compressed sensing (*72*, *73*), it would be possible to extend the resolvable density by ~4-fold and thus allow all but the top 2% most expressed genes to be imaged all together. Finally, theoretical predictions (*41*) indicate that the use of super-resolution imaging (*43*, *74*) could increase the resolvable density to ~10^5 molecules/μm^3, which should be ample to address the entire transcriptome even in cell types with RNA densities substantially higher than that of IMR90. However, it is important to note that RNAs in densely packed structures, such as p-bodies and stress granules, may still elude measurement.

To test the feasibility of using super-resolution imaging for MERFISH measurements, we performed a preliminary STORM-MERFISH experiments to probe 32 genes (Figure 32A). Clusters of localizations in different hybridization rounds were connected into a single string of measured words, corresponding to a potential RNA molecule, if the distance between clusters was smaller than 80 nm. Combining counts from four cells, we observed 397 clusters that matched our codebook and 1227 clusters that differed by one bit (Figure 32B). We also observed

a median count of 1.7 misidentification controls per cell. Twenty nine out of the thirty two transcripts were detected with counts above this value.

Using the ratio of perfectly matched clusters to the clusters that gained or lost one bit, we estimated that the per bit false negative rate (1→0) was 40% and the per bit false positive rate (0→1) was 11%. The measured abundances of these 32 mRNAs correlated with the abundances measured from bulk RNA sequencing with a ρ of 0.74 (Figure 32C), indicating that the measurements were still accurate. Measurement with two shuffled codebooks also showed good agreement with a Pearson correlation coefficient of 0.74 (Figure 32D). These preliminary MERFISH experiments using STORM imaging gave higher per bit error rates that what we obtained with conventional imaging. We are currently optimizing the labeling and imaging conditions to reduce these error rates.



(Continued) Figure 32: MERFISH measurements of thirty-two target transcripts using STORM Imaging. A) Aligned STORM image of mRNAs, localizations are false-colored based on the hybridization round of which they were imaged. Clusters of putative mRNAs are defined with localizations density greater than a hundred and twenty per pixel. Localizations in a cluster are enlarged for clarity. B) Histogram of counts for the transcripts and controls from four fields of view before and after applying error correction. C) Scatter plot of the measured mRNAs counts versus the abundance determined by bulk sequencing in fragments per kilobase per million reads (FPKM). D) Scatter plot of the mRNAs counts measured with two shuffled codebooks.

Next, we compare Our MERFISH approach with a recently described Fluorescence *In Situ* Sequencing (FISSEQ) approach for sequencing transcripts inside cells via rolling circle amplification and ligation (*75*). Although our method does not measure as many RNA species, offer single base pair resolution, or ability for *de novo* transcript discovery, it offers much higher detection efficiency and counting accuracy (by using multiple probes to stain RNA molecules directly), better spatial resolution (rolling circle amplicons are much bigger than RNA molecules), is faster (18 hours versus 10 days (*76*)) and is amplification free. Perhaps the most significant advantage of our approach is the detection efficiency. With the current implementation of FISSEQ, almost all the transcripts, including those of the relatively high abundance genes, are measured on average less than once per cell and any pair of genes is rarely detected in the same cell, which makes single-cell analysis challenging.

In addition, our transcriptome scale MERFISH method also complements existing single cell RNA sequencing methods (*4*, *77*) by providing spatial information and cellular context of identified transcripts. Substantial effort has currently been invested into methods to isolate and process individual cells for sequencing, including fluorescence activated cell sorting (FACs) (*78*), micromanipulation (*79*), microfluidics (*80*, *81*), and laser dissection (*82*). By isolating and

lysing cells, the spatial context of a cell relative to the tissue and to other cells is mostly lost, as is information about cell size and morphology - information that often contains critical cues about cell behavior. In contrast, FISH does not require the isolation of cells, and has been a leading technique in understanding the spatial organization of transcripts.

Moreover, our MERFISH approach does not suffer from amplification noise or RNA extraction variability. Thus, this method can potentially probe very low abundance transcripts that are typically excluded from single cell RNA analysis due to technical variability (*83*, *84*). Furthermore, FISH based techniques are known to have very high detection efficiency. A recent comparison of single-cell sequencing measurements to RNA FISH for individual genes find only 5-20% the number of transcripts by sequencing as counted by FISH (*84*). Therefore, our MERFISH method is particularly suited for applications that depend on absolute molecule counts, such as analysis of expression noise and comparison to Poisson models or inference of kinetic properties of transcription (*30*, *59*, *85*).

In conclusion, by performing combinatorial labeling with sequential imaging instead of color-based barcodes, we substantially increased the number of RNAs that can be detected simultaneously in single cells. Through the use of error correcting codes and encoding/sequential readout hybridization scheme, we have circumvented the rapid growth of molecule identification errors from combinatorial labeling schemes, and demonstrated highly precise and accurate multiplexed measurement of a thousand RNA species. We have also illustrated the utility of the data derived from highly multiplexed RNA imaging by using co-variation and correlation analysis to reveal distinct sub-cellular distribution patterns of RNAs, to constrain gene regulatory networks, and to predict functions for many previously unannotated or partially annotated genes with unknown functions. We anticipate that many more quantitative analyses could be applied to

such data sets that include the spatial localization and copy number information of many RNA species in individual cells. Given its ability to quantify RNAs across a wide range of abundances without amplification bias while preserving native context, we envision that our MERFISH approach will enable many applications of in situ transcriptomic analyses of individual cells in culture or complex tissues.

2.4 Methods and materials

2.4.1 Probe design

Each RNA species in our target set was randomly assigned a binary code word either from all 14 possible codewords of the 8-bit MHD4 code, 38 possible codewords of the 12-bit MHD4 code, 140 possible code words of the 16-bit MHD4 code or from all 1001 possible code words of the 14-bit MHD2 code. The encoding schemes are provided in Appendix B.

We used array-synthesized oligopools as templates to make the encoding probes (*57*, *86*). The template molecule for each encoding probe contains three components: i) a central targeting sequence for *in situ* hybridization to the target RNA, ii) two flanking readout sequences designed to hybridize each of two distinct readout probes, and iii) two flanking primer sequences to allow enzymatic amplification of the probes (Fig. 9). The readout sequences were taken from the 16 possible readout sequences each corresponding to one hybridization round. The readout sequences were assigned to the encoding probes such that for any RNA species each of the 4 readout sequences were distributed uniformly along the length of the target RNA and appeared at the same frequency. Template molecules for the 12-gene, 32-gene, and 140-gene library also included a common 20-nucleotide (nt) priming region between the first PCR primer and the first readout sequence. This priming sequence was used for the reverse transcription step described below.

We embedded multiple experiments in a single array-synthesized oligopool, and used PCR to selectively amplify only the oligos required for a specific experiment. Primer sequences for this indexed PCR reaction were generated from a set of orthogonal 25-nt sequences (*87*). These sequences were trimmed to 20 nt and selected for i) a narrow melting temperature range (70 - 80°C), ii) the absence of consecutive repeats of 3 or more identical nucleotides, and iii) the

presence of a GC clamp, i.e. one of the two 3' terminal bases must be G or C. To further improve specificity, these sequences were then screened against the human transcriptome using BLAST+ (*88*), and primers with 14 or more contiguous bases of homology were eliminated. Finally, BLAST+ was again used to identify and exclude primers that had an 11-nt homology region at the 3' end of any other primer or a 5-nt homology region at the 3' end of the T7 promoter. The forward primer sequences (Primer 1) were determined as described above, whereas the reverse primers each contain a 20-nt sequence as described above plus a 20-nt T7 promoter sequence to facilitate amplification via *in vitro* transcription (Primer 2). The primer sequences used in the 140-gene and 1001-gene experiments are listed below:

Experiment Name	Primer 1 Sequence (Index Primer 1)	Primer 2 Sequence (T7 promoter plus the reverse complement of Index Primer 2)
12-gene	TTGGACGGCGCGCGTAAGAC	TAATACGACTCACTATAGGGTTCGACAAGCGGAAGCC
32-gene Codebook 1	TCATCCCGCAACGGCACATG	TAATACGACTCACTATAGGGTACTGCCGATCGCCCACCAC
32-gene Codebook 2	TCACGCGCGGCATACACCAG	TAATACGACTCACTATAGGGTTTGGACGTCGTGTCACCCG
140-gene Codebook 1	GTTGGTCGGCACTTGGGTGC	TAATACGACTCACTATAGGGAAAGCCGGTTCATCCGGTGG
140-gene Codebook 2	CGATGCGCCAATTCCGGTTC	TAATACGACTCACTATAGGGTGATCATCGCTCGCGGGTTG
1001-gene	CGCGGGCTATATGCGAACCG	TAATACGACTCACTATAGGGCGTGGAGGGCATACAACGC

30-nt-long readout sequences were created by concatenating fragments of the same orthogonal primer set generated above by combining one 20-nt primer with a 10-nt fragment of another. These readout sequences were then screened, using BLAST+, for orthogonality with the

index primer sequences and other readout sequences (no more than 11 nt of homology) and for potential off-target binding sites in the human genome (no more than 14 nt of homology). Fluorescently labeled readout probes with sequences complementary to the readout sequences were used to probe these readout sequences, one in each hybridization round. All used readout probes sequences are listed below:

Bit	Readout probes
1	CGCAACGCTTGGGACGGTTCCAATCGGATC/3Cy5Sp/
2	CGAATGCTCTGGCCTCGAACGAACGATAGC/3Cy5Sp/
3	ACAAATCCGACCAGATCGGACGATCATGGG/3Cy5Sp/
4	CAAGTATGCAGCGCGATTGACCGTCTCGTT/3Cy5Sp/
5	GCGGGAAGCACGTGGATTAGGGCATCGACC/3Cy5Sp/
6	AAGTCGTACGCCGATGCGCAGCAATTCACT/3Cy5Sp/
7	CGAAACATCGGCCACGGTCCCGTTGAACTT/3Cy5Sp/
8	ACGAATCCACCGTCCAGCGCGTCAAACAGA/3Cy5Sp/
9	CGCGAAATCCCCGTAACGAGCGTCCCTTGC/3Cy5Sp/
10	GCATGAGTTGCCTGGCGTTGCGACGACTAA/3Cy5Sp/
11	CCGTCGTCTCCGGTCCACCGTTGCGCTTAC/3Cy5Sp/
12	GGCCAATGGCCCAGGTCCGTCACGCAATTT/3Cy5Sp/
13	TTGATCGAATCGGAGCGTAGCGGAATCTGC/3Cy5Sp/
14	CGCGCGGATCCGCTTGTCGGGAACGGATAC/3Cy5Sp/
15	GCCTCGATTACGACGGATGTAATTCGGCCG/3Cy5Sp/
16	GCCCGTATTCCCGCTTGCGAGTAGGGCAAT/3Cy5Sp/

The readout probes used for the 12-gene libraries were probes 1 through 8. The readout probes used for the 32-gene libraries were probes 1 through 12. The readout probes used for the 140-

gene libraries were probes 1 through 16. The readout probes used for the 1001-gene experiment were probes 1 through 14. /3Cy5Sp/ indicates a 3' Cy5 modification.

To design the central targeting sequences of the encoding probes, we first compiled the abundance of different transcripts in IMR90 cells using Cufflinks v2.1 (*89*), total RNA data from the ENCODE project (*90*), and human genome annotations from Gencode v18 (*52*). Probes were designed from gene models corresponding to the most abundant isoform using OligoArray2.1 (*91*) with the following constraints: the target sequence region is 30-nt long; the melting temperatures of the hybridized region of the probe and cellular RNA target is greater than 70 °C; there is no cross hybridization targets with melting temperatures greater than 72 °C; there is no predicted internal secondary structures with melting temperatures greater than 76 °C; and there is no contiguous repeats of 6 or more identical nucleotides. Melting temperatures were adjusted to optimize the specificity of these probes and minimize secondary structure while still producing sufficient numbers of probes for our libraries. To decrease computational cost, isoforms were divided into 1-kb regions for probe design. Using BLAST+, all potential probes that mapped to more than one cellular RNA species were rejected. Probes with multiple targets on the same RNA were kept.

For each gene in the 12, 32, and 140-gene experiments, we generated 198 putative encoding probe sequences by concatenating the appropriate index primers, readout sequences, and targeting regions as shown in Fig 9. To address the possibility that concatenation of these sequences introduced new regions of homology to off-target RNAs, we used BLAST+ to screen these putative sequences against all human rRNA and tRNA sequences as well as highly expressed genes (genes with FPKM > 10,000). Probes with greater than 14 nt of homology to rRNAs or tRNAs or greater than 17 nt of homology to highly expressed genes were removed.

After these cuts, we had ~192 probes per gene for both MHD4 codebooks used in the 140-gene experiments. We followed the same protocol for the 1001-gene experiments: Starting with 96 putative targeting sequences per gene, we obtained ~94 encoding probes per gene after these additional homology cuts. We decreased the number of encoding probes per RNA for the 1001-gene experiments so that these probes could be synthesized from a single 100,000-member oligopool as opposed to two separate pools. We designed each encoding probe to contain two of the four readout sequences associated with each code word, hence only half of the bound encoding probes can bind readout probe during any given hybridization round. We used ~192 or ~94 encoding probes per RNA to obtain high signal-to-background ratios for individual RNA molecules. The number of encoding probes per RNA could be substantially reduced but still allow single RNA molecules to be identified (*41*, *92*, *93*). In addition, increasing the number of readout sequences per encoding probe or using optical sectioning methods to reduce the fluorescence background may allow further reduction in the number of the encoding probes per RNA.

We designed two types of misidentification controls. The first control — blank words — were not represented with encoding probes. The second type of control — no-target words — had encoding probes that were not targeting any cellular RNA. The targeting regions of these probes were composed of random nucleotide sequences subject to the same constraints used to design the RNA targeting sequences described above. Moreover, these random sequences were screened against the human transcriptome to ensure that they contain no significant homology (>14-nt) to any human RNA. The 12-gene measurements contained 1 blank words and 1 no-target words. The 32-gene measurements contained 3 blank words and 3 no-target words. The 140-

gene measurements contained 5 blank words and 5 no-target words. The 1001-gene measurements contained 11 blank words and 5 no-target words.

2.4.2 Probe synthesis

The encoding probes were synthesized using the following four steps, and this synthesis protocol is illustrated in Fig. 9.

Step 1: The template oligopool (CustomArray) was amplified via limited-cycle PCR on a Bio-Rad CFX96 using primer sequences specific to the desired probe set. To facilitate subsequent amplification via in vitro transcription, the reverse primer contained the T7 promoter. All primers were synthesized by IDT. This reaction was column purified (Zymo DNA Clean and Concentrator; D4003).

Step 2: The purified PCR products were then further amplified ~200-fold and converted into RNA via a high yield in vitro transcription according to the manufacturer's instructions (New England Biolabs, E2040S). Each 20 μL reaction contained ~1 μg of template DNA from above, 10 mM of each NTP, 1x reaction buffer, 1x RNase inhibitor (Promega RNasin, N2611) and 2 μL of the T7 polymerase. This reaction was incubated at 37 °C for 4 hours to maximize yield. This reaction was not purified before the following steps.

Step 3: The RNA products from the above in vitro transcription reaction were then converted back into DNA via a reverse transcription reaction. Each 50 μL reaction contained the unpurified RNA produce from Step 2 supplemented with 1.6 mM of each dNTP, 2 nmol of a reverse transcription primer, 300 units of Maxima H- reverse transcriptase (Thermo Scientific, EP0751), 60 units of RNasin, and a final 1x concentration of the Maxima RT buffer. This reaction was incubated at 50 °C for 45 minutes, and the reverse transcriptase was inactivated at

85°C for 5 minutes. The templates for the 140-gene libraries contain a common priming region for this reverse transcription step; thus, a single primer was used for this step when creating these probes. Its sequence was CGGGTTTAGCGCCGGAAATG. A common priming region was not included for the 1001-gene library; thus, the reverse transcription was conducted with the forward primer: CGCGGGCTATATGCGAACCG.

Step 4: To remove the template RNA, 20 μL of 0.25 M EDTA and 0.5 N NaOH was added to the above reaction to selectively hydrolyze RNA, and the sample was incubated at 95 °C for 10 minutes. This reaction was then immediately purified by column purification using a 100-μg-capacity column (Zymo Research, D4030) and the Zymo Oligo Clean and Concentrator protocol. The final probes were eluted in 100 μL of RNase-free deionized water, evaporated in a vacuum concentrator, and then resuspended in 10 μL of encoding hybridization buffer (recipe below). Probes were stored at -20 °C. Denaturing poly-acrylamide gel electrophoresis and absorption spectroscopy were used to confirm the quality of the probes and revealed that this probe synthesis protocol converts 90-100% of the reverse-transcription primer into full length probe and of the probe that is constructed, 70-80% is recovered during the purification step. This protocol is similar to another recently published protocol (23) but provides a substantially larger yield.

Fluorescently labeled readout probes have sequences complementary to the readout sequences described above and a Cy5 dye attached at the 3' end. These probes were obtained from IDT and HPLC purified.

2.4.3 Sample labeling

Human primary fibroblasts (American Type Culture Collection, IMR90) and grivet's kidney epithelial cells (American Type Culture Collection, BS-C-1) were used in this work. These cells are relatively large and flat, facilitating wide-field imaging without the need for optical sectioning. Cells were cultured with Eagle's Minimum Essential Medium. Cells were plated on 22-mm, #1.5 coverslips (Bioptechs, 0420-0323-2) at 350,000 cells/coverslip and incubated at 37 °C with 5% CO2 for 48-96 hours within petri dishes. Cells were fixed for 20 minutes in 4% paraformaldehyde (Electron Microscopy Sciences, 15714) in 1x phosphate buffered saline (PBS; Ambion, AM9625) at room temperature, reduced for 5 minutes with 0.1% w/v sodium borohydride (Sigma, 480886) in water to reduce background fluorescence, washed three times with ice-cold 1x PBS, permeabilized for 2 minutes with 0.5% v/v Triton (Sigma, T8787) in 1x PBS at room temperature or overnight in 70% Ethanol at 4 °C, and washed three times with 1x PBS.

Cells were incubated for 5 minutes in encoding wash buffer comprising 2x saline-sodium citrate buffer (SSC) (Ambion, AM9763), 30% v/v formamide (Ambion, AM9342), and 2 mM vanadyl ribonucleoside complex (NEB, S1402S). 10 µL of 10 µM (12-gene experiments), 10 µL of 25 µM (32-gene experiments), 10 µL of 100 µM (140-gene experiments) or 200 µM (1001-gene experiments) encoding probes in encoding hybridization buffer was added to the cell-containing coverslip and spread uniformly by placing another coverslip on top of the sample. Samples were then incubated in a humid chamber inside a 37 °C-hybridization oven for 18-36 hours. Encoding hybridization buffer is composed of encoding wash buffer supplemented with 1 mg/mL yeast tRNA (Life technologies, 15401-011) and 10% w/v dextran sulfate (Sigma, D8906-50G).

Cells were then washed with primary encoding wash buffer, incubated at 47 °C for 10 minutes, and this wash was repeated for a total of three times. A 1:1000 dilution of 0.2-µm-diameter carboxylate-modified orange fluorescent beads (Life Technologies, F-8809) in 2xSSC was sonicated for 3 minutes and then incubated with the sample for 5 minutes. The beads were used as fiducial markers to align images obtained from multiple successive rounds of hybridization, as described below. The sample was washed once with 2xSSC, and then post-fixed with 4% v/v paraformaldehyde in 2xSSC at room temperature for 30 minutes. The sample was then washed three times with 2xSSC and either imaged immediately or stored for no longer than 12 hours at 4 °C prior to imaging. All solutions were prepared as RNase-free.

For conventional smFISH measurements, pools of 48 fluorescently-labeled (Quasar 670) oligonucleotide probes per RNA were purchased from Biosearch Technologies. 30-nt probe sequences were taken directly from a random subset of the targeting regions used for the multiplexed measurements. Cells were fixed and permeabilized as described above. 10 µL of 250 nM oligonucleotide probes in encoding hybridization buffer (described above) was added to the cell-containing coverslip and spread uniformly by placing another coverslip on top of the sample. Samples were then incubated in a humid chamber inside a 37 °C-hybridization oven for 18 hours. Cells were then washed with encoding wash buffer (described above) at 37 °C for 10 minutes, and this wash was repeated for a total of three times. The sample was then washed three times with 2xSSC and imaged in imaging.

2.4.4 Successive rounds of hybridizations and imaging

The sample coverslip was assembled into a Bioptech's FCS2 flow chamber, and the flow through this chamber was controlled via a home-built fluidics system composed of three

computer-controlled 8-way valves (Hamilton, MVP and HVXM 8-5) and a computer-controlled peristaltic pump (Rainin, Dynamax RP-1) (Fig. 33A-B). The readout probes were routed through the three daisy-chained valves, and the wash buffer, bleach buffer, and imaging buffer were connected to Valve 1 to ensure that all the tubes were washed prior to the next round of hybridization. Liquid flow was enabled by the peristaltic pump and a flow rate of 0.5mL/min was used for all the steps. The sample was imaged on a home-built microscope constructed around an Olympus IX-71 body and a 1.45 NA, 100x oil immersion objective and configured for oblique incidence excitation. We note that, from an imaging perspective, multiplexed FISH is simply a series of smFISH experiments. Thus, any optical sectioning technique could be employed to allow the imaging of RNAs in thicker cells or tissues.

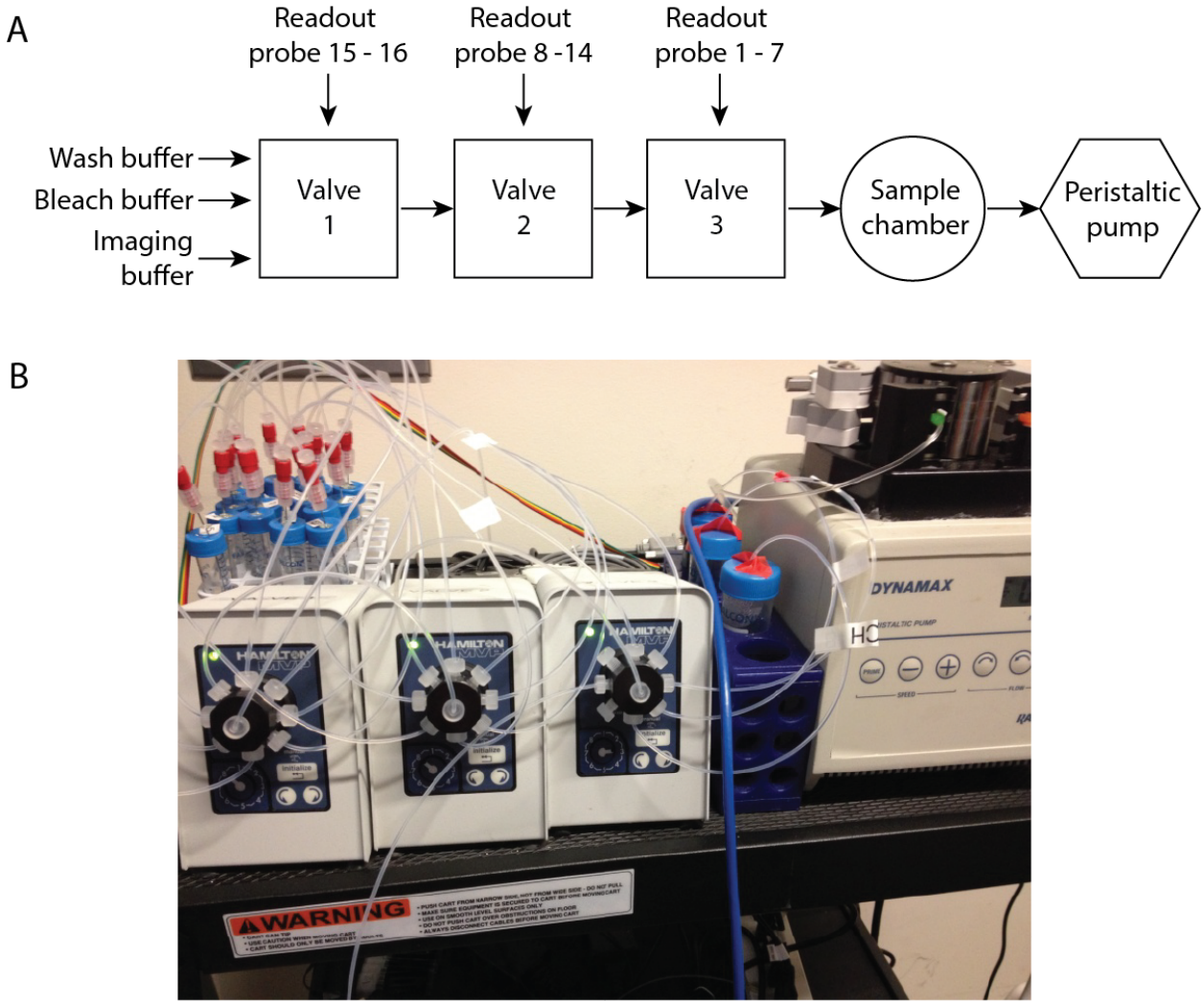


Figure 33: Design of automated platform for multiplexed FISH experiments. A) Schematic of the different inputs and outputs through the sample chamber. B) Photograph of the fluidic setup. On the left are three valves connected to the 16 different readout probes in hybridization buffer, wash buffer, bleach buffer, and imaging buffer according to schematic from A. The output from the first valve is connected to the sample chamber. All the liquids are pulled through the chamber via the peristaltic pump (on the right).

The objective was heated to 37 °C with a Bioptechs objective heater. Constant focus was maintained throughout the imaging process with a home-built, auto-focusing system.

Illumination was provided at 641 nm, 561 nm, and 405 nm using solid state lasers (MPB communications, VFL-P500-642; Coherent, 561-200CWCDRH; and Coherent, 1069413/AT) for excitation of our Cy5-labeled readout probes, the fiducial beads, and nuclear counterstains, respectively. These lines were combined with a custom dichroic (Chroma, zy405/488/561/647/752RP-UF1) and the emission was filtered with a custom dichroic (Chroma, ZET405/488/561/647-656/752m). Fluorescence was separated with a QuadView (Photometrics) using the dichroics T560lpxr, T650lpxr, 750dcxxr (Chroma) and the emission filters ET525/50m, WT59550m-2f, ET700/75m, HQ770lp (Chroma) and imaged with an EMCCD camera (Andor, iXon-897). The camera was configured so that a pixel corresponds to 167 nm in the sample plane. The entire system was fully automated, so that imaging and fluid handling were performed for the entire experiment without user intervention.

Sequential hybridization, imaging, and bleaching proceeded as follows. 1 mL of 10 nM of the appropriate fluorescently labeled readout probe in readout hybridization buffer (2xSSC; 10% v/v formamide; 10% w/v dextran sulfate, and 2 mM vanadyl ribonucleoside complex) was flown across the sample, flow was stopped, and the sample was incubated for 15 minutes. Then 2 mL of readout wash buffer (2xSSC, 20% v/v formamide; and 2 mM vanadyl ribonucleoside complex) was flown across the sample, flow was stopped, and the sample was incubated for 3 minutes. 2 mL of imaging buffer comprising 2xSSC, 50 mM TrisHCl pH 8, 10 % w/v glucose, 2 mM Trolox (Sigma-Aldrich, 238813), 0.5 mg/mL glucose oxidase (Sigma-Aldrich, G2133), and 40 μg/mL catalase (Sigma-Aldrich, C30) was flown across the sample. Flow was then stopped, and then approximately 75 to 100 regions were exposed to ~25 mW 642-nm and 1 mW of 561-nm light and imaged. Each region was 40 μm by 40 μm. The laser powers were measured at the microscope backport. Because the imaging buffer is sensitive to oxygen (51), the ~50 mL of

imaging buffer used for a single experiment was made fresh at the beginning of the experiment and then stored under a layer of mineral oil throughout the measurement. Buffer stored in this fashion was stable for more than 24 hours.

After imaging, the fluorescence of the readout probes was extinguished via photobleaching. The sample was washed with 2 mL of photobleaching buffer (2xSSC and 2 mM vanadyl ribonucleoside complex), and each imaged region of the sample was exposed to 200 mW of 641-nm light for 3 s. To confirm the efficacy of this photobleaching treatment, imaging buffer was reintroduced, and the sample was imaged as described above.

The above hybridization, imaging, and photobleaching process was repeated either 8 times for the 12-gene measurements, 12 times for the 32-gene measurements, 16 times for the 140-gene measurements or 14 times for the 1001-gene measurements. An entire experiment was typically completed in <18 hours.

Following completion of imaging, 2 mL of a 1:1000 dilution of Hoescht (ENZ-52401) in 2xSSC was flown through the chamber to label the nuclei of the cells. The sample was then washed immediately with 2 mL of 2xSSC followed by 2 mL of imaging buffer. Each region of the sample was then imaged once again with ~1 mW of 405-nm light.

2.4.5 Image analysis

Fluorescent spots were identified and localized in each image using a multi-Gaussian-fitting algorithm (38) assuming a Gaussian with a uniform width of 167 nm. This algorithm was used to allow partially overlapping spots to be distinguished and individually fit. RNA spots were distinguished from background signal, i.e. signal arising from probes bound non-

specifically, by setting the intensity threshold required to fit a spot with this software. Due to variation in the brightness of spots between rounds of hybridization, this threshold was adjusted appropriately for each hybridization round to minimize the combined average of the 1→0 and 0→1 error rates across all hybridization rounds (140-gene measurements) or to maximize the ratio of the number of measured words with four '1' bits to those with three or five '1' bits (1001-gene measurements). The location of the fiducial beads was identified in each frame using a faster single-Gaussian fitting algorithm.

Images of the same sample region in different rounds of hybridization were registered by rotating and translating the image to align the two fiducial beads within the same image that were most similar in location after a coarse initial alignment via image correlation. All images were aligned to a coordinate system established by the images collected in the first round of hybridization. The quality of this alignment was determined from the residual distance between five additional fiducial beads, and alignment error was typically ~20 nm.

Fluorescence spots in different hybridization rounds were connected into a single string, corresponding to a potential RNA molecule, if the distance between spots was smaller than 1 pixel. For each string of spots, the on-off sequence of fluorescent signals in all hybridization rounds were used to assign a binary word to the potential RNA molecule, in which '1' was assigned to the hybridization rounds that contained a fluorescent signal above threshold and '0' was assigned to the other hybridization rounds. Measured words were then decoded into RNA species using the MHD4 code or the MHD2 code discussed in the main text. In the case of the MHD4 code, if the measured binary word matched the code word of a specific RNA perfectly or differed from the code word by one single bit, it was assigned to that RNA. In the case of the MHD2 code, only if the measured binary word matched the code word of a specific RNA

perfectly, was it assigned to that RNA. To determine the copy number per cell, the number of each RNA species was counted in individual cells within each 40 μm by 40 μm imaging area. We note that this number accounts for the majority but not all RNA molecules within a cell because a fraction of the cell could be outside the imaging area or focal depth. Tiling images of adjacent areas and adjacent focal planes could be employed to improve the counting accuracy.

In the 140-gene experiments, some regions of the cell nucleus occasionally contained too much fluorescence signal to properly identify individual RNA spots. In the 1001-gene experiments, the cell nucleus generally contained too much fluorescent signal to allow identification of individual RNA molecules. These bright regions were excluded from all subsequent analysis. Employment of super-resolution imaging and/or optical sectioning could potentially allow individual molecules in these dense nucleus regions to be identified, which will be particularly useful for probing those non-coding RNAs that are enriched in the nucleus.

2.4.6 Bulk RNA sequencing

Total RNA was extracted from IMR90 cells cultured as above using the Zymo Quick RNA MiniPrep kit (R1054) according to the manufacturer's instructions. polyA RNA was then selected (NEB; E7490), and a sequencing library was constructed using the NEBNext Ultra RNA library preparation kit (NEB; E7530), amplified with custom oligonucleotides, and 150-bp reads were obtained from on a MiSeq. These sequences were aligned to the human genome (Gencode v18) and isoform abundance was computed with cufflinks (45).

2.4.7 Error calculation and analysis

Analytic expressions were derived for the dependence of the number of possible code words, the calling rate, and the misidentification rate on *N*. The calling rate is defined as the fraction of RNA molecules that are properly identified. The misidentification rate is defined as the fraction of RNA molecules that are misidentified as a wrong RNA species. For encoding schemes with an error-detection capability, the calling rate and misidentification rate does not add up to 1 because a fraction of the molecules not called properly can be detected as errors and discarded and, hence, not misidentified as a wrong species. These calculations assume that the probability of misreading bits is constant for all hybridization rounds but differs for the 1→0 and 0→1 errors. Experimentally measured average 1→0 and 0→1 error rates (10% and 4% respectively) were used for the estimates shown in Fig. 5B-D. For simplicity, the word corresponding to all '0's was not removed from calculations.

For the simple binary encoding scheme in which all possible *N*-bit binary words are assigned to unique RNA species, the number of possible code words is 2^N. The number of words that could be used to encode RNA is actually $2^N - 1$ because the code word '00…0' does not contain detectable fluorescence in any hybridization round, but for simplicity the word corresponding to all '0's was not removed from subsequent calculations. The error introduced by this approximation is negligible. For any given word with *m* '1's and *N*−*m* '0's the probability of measuring that word without error — the fraction of RNAs that is properly called — is

$$(1-p_1)^m(1-p_0)^{N-m}, \qquad \text{(iii)}$$

where p_1 is 1→0 error rate and p_0 is 0→1 error rate per bit. Because different words in this simple binary encoding scheme can have different numbers of '1' bits, the calling rate for different words will differ if $p_1 \neq p_0$. The average calling rate, reported in Fig. 5C, was

determined from the weighted average of the value of Eq. (iii) for all words. This weighted average is

$$\frac{1}{2^N}\sum_{m=0}^{N}\binom{N}{m}(1-p_1)^m(1-p_0)^{N-m}, \qquad \text{(iv)}$$

where $\binom{N}{m}$ is the binomial coefficient and corresponds to the number of words with *m* '1' bits in this encoding scheme. Since in this encoding scheme every error produces a binary word that encodes a different RNA, the average misidentification rate for this encoding scheme, reported in Fig. 5D, follows directly from (2):

$$1-\frac{1}{2^N}\sum_{m=0}^{N}\binom{N}{m}(1-p_1)^m(1-p_0)^{N-m}. \qquad \text{(v)}$$

To calculate the scaling and error properties of the extended Hamming distance 4 (HD4) code, we first created the generator matrix for the desired number of data bits using standard methods (*94*). The generator matrix determines the specific words that are present in a given encoding scheme and was used to directly determine the number of encoded words as a function of the number of bits. In this encoding scheme, the calling rate corresponds to the fraction of words measured without error as well as the fraction of words measured with a single-bit error. For code words with *m* '1' bits, this fraction is determined by the following expression:

$$(1-p_1)^m(1-p_0)^{N-m}+m\,p_1^{1}(1-p_1)^{m-1}(1-p_0)^{N-m}+(N-m)p_0^{1}(1-p_1)^m(1-p_0)^{N-m-1} \qquad \text{(vi)}$$

where the first term is the probability of not making any errors, the second term corresponds to the total probability of making one 1→0 error at any of the *m* '1' bits without making any other 0→1 errors, and the final term corresponds to the total probability of making one 0→1 error at any of the *N-m* '0' bits without making any 1→0 errors. Because the number of '1' bits can differ between words in this encoding scheme, the average calling rate reported in Fig. 5C was

computed from a weighted average over Eq. (v) for different values of *m*. The weight for each term was determined from the number of words that contain *m* '1' bits as determined from the generator matrix described above.

Because RNA-encoding words are separated by a minimum Hamming distance of 4, at least 4 errors are required to switch one word into another. If error correction is applied, then 3 or 5 errors could also convert one RNA into another. Thus, we estimate the misidentification rate from all possible combinations of 3-bit, 4-bit and 5-bit errors for code words with *m* '1' bits. Technically, >5-bit errors could also convert one RNA into another, but the probability of making such errors is negligible because of the small per-bit error rate. We approximate this expression with

$$\begin{aligned}
&\sum_{i=0}^{4}\binom{m}{i}\binom{N-m}{4-i}p_1^i p_0^{4-i}(1-p_1)^{m-i}(1-p_0)^{N-m-(4-i)} + \\
&\sum_{i=0}^{3}\binom{m}{i}\binom{N-m}{3-i}p_1^i p_0^{3-i}(1-p_1)^{m-i}(1-p_0)^{N-m-(3-i)} + \\
&\sum_{i=0}^{5}\binom{m}{i}\binom{N-m}{5-i}p_1^i p_0^{5-i}(1-p_1)^{m-i}(1-p_0)^{N-m-(5-i)} +
\end{aligned} \tag{vi}$$

The first sum corresponds to all of the ways in which exactly four mistakes can be made. Similarly, the second and third sums correspond to all of the ways in which exactly three or five mistakes can be made. Eq. (vi) provides an upper bound for the misidentification rate because not all three, four, or five bit errors produce a word that matches or would be corrected to another legitimate word. Again because the number of '1' bits can differ between words, the average misidentification rate reported in Fig. 5D is calculated as a weighted average of Eq. (vi) over the number of words that have *m* '1' bits.

To generate our MHD4 code where the number of '1' bits for each code word is set to 4, we first generated the HD4 codes as described above, and then removed all code words that did

not contain four '1's. The calling rate of this code, reported in Fig. 5C, was directly calculated from Eq. (v) but with m = 4 because all code words in this code have four '1' bits. The misidentification rate of this code, reported in Fig. 5D, was calculated by modifying Eq. (vi) with the following considerations: (i) the number of '1' bits, m, was set to 4 and (ii) errors that produce words that do not contain three, four, or five '1' bits were excluded. Thus, the expression in Eq. (5) was simplified to

$$\begin{aligned}
&\binom{4}{2}\binom{N-4}{2}p_1^2 p_0^2(1-p_1)^2(1-p_0)^{N-6} + \\
&\binom{4}{1}\binom{N-4}{2}p_1^1 p_0^2(1-p_1)^3(1-p_0)^{N-6} + \binom{4}{2}\binom{N-4}{1}p_1^2 p_0^1(1-p_1)^2(1-p_0)^{N-5} \\
&\binom{4}{2}\binom{N-4}{3}p_1^2 p_0^3(1-p_1)^2(1-p_0)^{N-7} + \binom{4}{3}\binom{N-4}{2}p_1^3 p_0^2(1-p_1)^1(1-p_0)^{N-6}
\end{aligned} \qquad \text{(vii)}$$

Again, this expression is an upper bound on the actual misidentification rate because not all words with four '1's are valid code words.

Estimates of the 1→0 and 0→1 error rates for each hybridization round

To compute the probability of misreading a bit at a given hybridization round, we used the error correcting properties of the MHD4 code. Briefly, the probabilities of 1→0 or 0→1 errors were derived in the following way. Let the probability of making an error at the *i*th bit, i.e. *i*th hybridization round, be p_i and the actual number of RNA molecules of the given species be A, then the number of exact matches for this RNA will be $W_E = A\prod_{i=1}^{16}(1-p_i)$ and the number of one-bit error corrected matches for this RNA corresponding to errors at the *i*th bit will be $W_i = A\frac{p_i}{(1-p_i)}\prod_{j=1}^{16}(1-p_j)$. The p_i can be directly derived from the ratio: $W_i/W_E = \frac{p_i}{(1-p_i)}$. This ratio assumes that the one-bit error-corrected counts were only generated from single-bit errors from

the correct word and that multi-error contamination from other RNA words is negligible. Given that our error rate per hybridization round is small and that it takes at least three errors to convert one RNA-encoding word into a word that would be misidentified as another RNA, the above approximation should be a good one.

To compute the average 1→0 or 0→1 error probabilities for each of the 16 hybridization rounds, we use the above approach to calculate the per-bit error rates for each bit of every gene, sort these errors based on whether they correspond to a 1→0 or a 0→1 error, and then take the average of these errors for each bit weighted by the number of counts observed for the corresponding gene.

Estimates of the calling rate for individual RNA species from actual imaging data

With the estimates of the 1→0 or 0→1 error probabilities for each round of hybridization as determined above, it is possible to estimate the calling rate for each RNA based on the specific word used to encoded it. Specifically, the fraction of an RNA species that is called correctly is determined by

$$\prod_{i=1}^{N}(1-p_i)+\sum_{j=1}^{N}\frac{p_j}{(1-p_j)}\prod_{i=1}^{N}(1-p_i), \qquad \text{(viii)}$$

where the first term represent the probability of observing an exact match of the code word and the second term represent the probability of observing an error-corrected match (i.e. with one-bit error). The values of the per-bit error rate p_i for each RNA species are determined by the specific code word for that RNA and the measured 1→0 or 0→1 error rates for each round of hybridization. If the code word of the RNA contains a '1' in the *i*th bit, then p_i is determined from the 1→0 error rate for the *i*th hybridization round; if the word contains a '0' in the *i*th bit, p_i is determined from the 0→1 error rate for the *i*th hybridization round.

2.4.8 Biological Analysis

Hierarchical clustering analysis of the co-variation in RNA abundance

Hierarchical clustering of the co-variation in gene expression for both the 140-gene and 1001-gene experiments was conducted as follows. First, the distance between every pair of genes was determined as 1 minus the Pearson correlation coefficient of the cell-to-cell variation of the measured copy numbers of these two RNA species, both normalized by the total RNA counted in the cell. Thus, highly correlated genes are 'closer' to one another and highly anti-correlated genes are 'further' apart. An agglomerative hierarchical cluster tree was then constructed from these distances using the Unweighted Pair Group Method with Arithmetic mean (UPGMA). Specifically, starting with individual genes, we constructed hierarchical clusters by identifying the two clusters (or individual genes) that are closest to one another according to the arithmetic mean of the distances between all inter-cluster gene pairs. The pairs of clusters (or individual genes) with the smallest distance are then grouped together and the process is repeated. The matrix of pairwise correlations was then sorted based on the order of the genes within these trees.

Groups of genes with substantial co-variations were identified by selecting a threshold on the hierarchical cluster tree (indicated by the dashed lines in Figs. 29D and 32A) that produced approximately 10 groups of genes each of which contains at least 4 members for the 140-gene experiments or approximately 100 groups each of which contains at least 3 members for the 1001-gene experiments. We note that one can change the threshold in order to identify either more tightly coupled smaller groups or larger groups with relatively loose coupling.

A probability value for the confidence that a gene belongs to a specific group was determined by computing the difference between the average correlation coefficient between that gene and all other members of that group and the average correlation coefficient between that

gene and all other measured genes outside that group. The significance (p-value) of this difference was determined with the student's t-test and is provided in Tables S2 and S4.

Because hierarchical clustering is inherently a one-dimensional analysis, i.e. any given genes can only be a member of a single group, this analysis does not allow all correlated gene groups to be identified. Higher dimension analysis, such as principal component analysis or k-means clustering, could be used to identify more co-varying gene clusters (*63*).

Analysis of RNA spatial distributions

To identify genes that have similar spatial distributions, we subdivided each of the measured cells into 2x2 regions and calculated the fraction of each RNA species present in each of these bins. To control for the fact that some regions of the cell naturally contain more RNA than others, we calculate the enrichment for each gene — the ratio of the observed fraction in a given region for a given RNA species to the average fraction observed for all genes in that same region. For each pair of RNA species, we then determined the Pearson correlation coefficient of the region-to-region variation in enrichment of these two RNA species for each cell and averaged the correlation coefficients over ~400 cells imaged in 7 independent data sets. We then clustered RNA species based on these average correlation coefficients using the same hierarchical clustering algorithm described above. Because of the large number of cells used for the analysis, we found that the coarse spatial binning (2x2 regions per cell) was sufficient to capture the spatial correlation between genes and finer binning did not produce more significantly correlated groups.

To measure the distances of genes from the nuclei and from the cell edge, we first used brightness thresholds on our cell images to segment the nuclei and identify the cell edge. We

then measured the distance from every RNA molecule to the nearest part of the nucleus and nearest part of the cell edge. For each data set, we computed the average distance for each RNA species averaged over all the cells measured. We then averaged these distances for the group I genes, group II genes or all genes. Only those RNA species with at least 10 counts per cell were used in this analysis to minimize statistical error on the distance values.

Gene ontology (GO) analysis

Groups of genes were selected from the hierarchical trees as discussed above. A collection of GO terms (*64*) was determined for all measured RNA species as well as the RNA species associated with each group from the most recent human GO annotations (http://geneontology.org/page/download-annotations) using both the annotated GO terms and terms immediately upstream or downstream of the found annotations. The enrichment of these annotations was calculated from the ratio of the fraction of genes within each group that have this term to the fraction of all measured genes that have this term and the p-value for this enrichment was calculated via the hypergeometric function. Only statistically significantly enriched GO terms with a p-value less than 0.05 were considered.

Chapter 3

Super-resolution imaging

This work was done in collaboration with Dr. Graham T. Dempsey, Dr. Joshua C. Vaughan, and Dr. Mark Bates under the supervision of Prof. Xiaowei Zhuang. Most of the results from this work are published in (*95*).

3.1 Introduction

Super-resolution microscopy is a rapidly developing technique that has already seen many applications in various fields of biological sciences, such as cell biology, microbiology and neurobiology (*43*, *74*). Indeed, super-resolution microscopy promises to revolutionize biology and medicine by allowing researchers to directly image dynamic and complex bio-molecular processes at nano-scale resolution.

There are two major approaches to break the diffraction limit. The first approach use patterned illumination to saturate stimulated emission from an ensemble of fluorophores within a sub-diffraction region. This category includes stimulated emission depletion (STED) microscopy (*96*)and saturated structured illumination microscopy (*97*). The second approach relies on mechanisms to activate a subset of individual flurophores within a diffraction-limited region at different times. Because the position of these emitting point sources can be measured with a precision that is much higher than that allowed by the diffraction limit, super-resolution image can be reconstructed from the localizations of all the individual fluorophores.

STORM uses stochastic activation and precise localization of photoswitchable fluorophores to achieve super-resolution imaging (45). Many fluorescent proteins or organic fluorophores can reversibly photo-switch or ‘blink’ under continuous wave illumination with a

single laser wavelength. The molecular nature of these 'dark' states could be the triplet states, formation of thiol adduct, long-lived radicals, or etc. The fraction of time that the fluorophore spends in its 'on' state versus the time it remains 'off', i.e. its duty cycle, can constrain the maximum number of molecules that can be localized within a diffraction limited spot. The maximum fluorophore density in turn limits the image resolution according to the Nyquist sampling theory. For a N-dimensional image having spatial feature of size r, the minimum required molecular density of fluorophores is:

$$\text{Nyquist Molecular Density} \approx (2/r)^N \quad \text{(ix)}$$

Therefore, it is beneficial to have a low duty cycle for super-resolution imaging. Additionally, the number of photon emitted during the 'on' state is also critical for super-resolution imaging because the precision of localizing the centroid position of an activated fluorophore depends on the number of photons collected, N (*98*):

$$\sigma \approx \frac{s}{\sqrt{N}} \quad \text{(ix)}$$

where s is the standard deviation of the imaging point spread function and σ is the uncertainty in the fitted position. Thus, a fluorophore with high photon number and low duty cycle should yield the sharpest STORM image, while a fluorophore with low photon number or high duty cycle should give a blurry or spotty STORM image (Fig. 34).

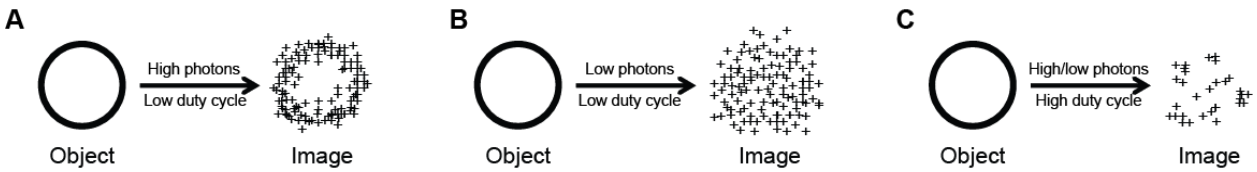


Figure 34: The effect of photon number and duty cycle (fraction of time in the on state) on STORM image quality for an example structure (a ring-like object). A) A fluorophore with high photon number and low duty cycle produces a hollow, ring-like image with high localization

(Continued) precision and sufficient density. B) A fluorophore with low photon number and low duty cycle maintains a large number of localizations but suffers reduced localization accuracy, blurring the ring-like structure C) A fluorophore with high on-off duty cycle requires reduction in the density of fluorescent probes to allow single-molecule localization, which in turn reduces the number of localizations and adversely affects the overall resolution.

Although many 'blinking' fluorophores can be utilized for STORM, only a few of them have been characterized in terms of photon number and duty cycle. In addition, photo-stability and number of switching cycles are also important factors that influence STORM image quality, but are poorly characterized under STORM imaging conditions. Consequently, incomplete information on the photoswitching properties of these fluorophores makes it difficult to choose the optimal fluorophores and imaging conditions for attaining high-quality super-resolution images. In this Chapter, I describe our efforts to evaluate 26 commercially available fluorophores and their suitability for STORM imaging under different buffer conditions. We also identified several high-quality switchable dyes in different spectral ranges, which we used to demonstrate low cross-talk, four-color super-resolution imaging.

3.2 Screening of organic dyes for multi-color super-resolution imaging

3.2.1 Photo-switching properties and their effect on super-resolution images

To characterize the photo-switching properties of fluorophores, we performed single-molecule characterization using dye-labeled proteins that were adsorbed to coverglass. We illuminated the fluorophores with light at 488 nm, 561 nm, 647 nm or 752 nm, to excite blue, yellow, red or NIR dyes, respectively. Under these conditions, many fluorophores reversibly transition between on and off states with a characteristic duty cycle and number of emitted photons per switching event (Fig. 35 A-C). For each dye, we determined four switching

properties: (i) number of photons detected per switching event; (ii) on/off duty cycle; (iii) survival fraction (described below); and (iv) number of switching cycles. Because the switching performance and photostability of many dyes are enhanced by the presence of a primary thiol and/or a low oxygen environment, we performed the experiments in the presence of an enzymatic oxygen scavenging system and a primary thiol.

We recorded the number of photons detected in each switching event for many molecules, from which we constructed a photon number histogram and determined the mean value (Fig. 35 D, F, and H). The on/off duty cycle was measured as the fraction of time spent in the on state averaged over many molecules. Because the molecules were originally in the on state and a certain illumination time is required for them to reach a quasi-equilibrium between the on and off states, we used a sliding window of 100 sec to monitor changes in the on/off duty cycle as a function of exposure time (Fig. 35 E, G, and I, red). The duty cycle value during the time window 400-600 sec was used to represent the characteristic, equilibrium duty cycle because most dyes achieved quasi-equilibrium by this point. Since some dye molecules only switched for a limited number of cycles before photobleaching, a substantial fraction of molecules may be photobleached by the time the on/off equilibrium is reached. We therefore also characterized the survival fraction of unbleached molecules as a function of illumination time (Fig. 35 E, G, and I, blue). This quantity can be used together with the time-dependent duty cycle value to determine what degree of photobleaching has occurred prior to achieving a sufficiently low density to resolve single molecules. Finally, we also determined for each molecule the number of switching cycles during the data acquisition time, which represents a lower bound value given that some molecules did not photobleach before the end of the observation period.

We correlated the above properties with the quality of STORM images for each dye by recording images of clathrin-coated pits (CCPs), which have a spherical morphology (Fig. 35 J-R). CCPs in BS-C-1 cells were immunostained using anti-clathrin heavy chain monoclonal antibody and dye labeled secondary antibody. CCPs provided an excellent metric for assessing a dye's performance because of their small size (~100-200 nm diameter) and the requirement for high spatial resolution to reveal their hollow, spherical shell-like structure. The 3D STORM images were taken using astigmatism imaging. When imaging Alexa 647-labeled CCPs, the hollow shape of the coated pit was clearly resolved (Fig. 35 J-L) due to the high photon yield of individual switching events (~5000 photons, Fig. 35D) and low on/off duty cycle (~0.001, Fig. 35E). In contrast, the low photon yield of Atto 655 (~660 photons, Fig. 35F), and the resulting low localization precision, blurred the CCP image substantially such that its hollowness was no longer apparent (Fig. 35 M-O), despite the fact that the duty cycle of Atto 655 was comparable to that of Alexa 647. In a different manner, the high duty cycle of Cy5.5 (~0.007, Fig. 35 I) also negatively impacted the quality of the CCP images (Fig. 35 P-R). Because of the large fraction of time each dye spending in the fluorescent state, a considerable reduction in labeling density (or photobleaching of molecules prior to imaging) was necessary to ensure single-molecule detection and localization. As a result, the sparse localization pattern no longer resolved the hollow morphology of the CCP as expected from the Nyquist criterion, despite the high photon yield of Cy5.5 (~6000 photons, Fig. 35 H). These examples demonstrate that both high photon number and low duty cycle are essential to achieving high-resolution STORM image. We performed the same set of experiments on 23 other dyes and their photon number and duty cycle values are summarized in Table 1.

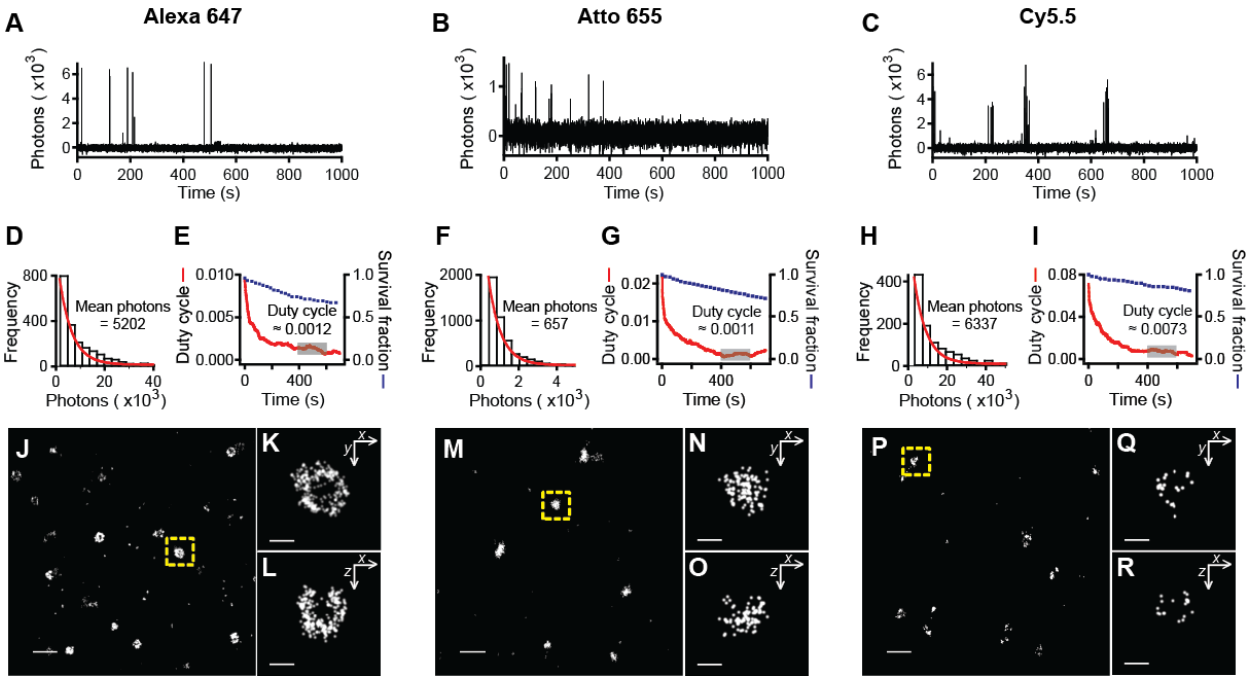


Figure 35: Methods of dye characterization, using Alexa 647, Atto 655, and Cy5.5 as representative examples: A-C) Example single-molecule fluorescence time traces. D, F and H) Histograms of photon number distributions. The indicated mean value was derived from the single exponential fit of the distribution (red curves). E, G, I) and duty cycle plots along with the corresponding survival fractions. The on-off duty cycle value was calculated for each dye and plotted versus time (red curve) to show how each value begins high when most molecules are in the fluorescent state and reaches a quasi-equilibrium later. Reported values are the average duty cycle measured between 400–600 s (gray box). The fraction of molecules that survived photobleaching was plotted together with the duty cycle (blue squares). The dyes exhibited three distinct switching behaviors: switching events with high photon yield and low duty cycle (Alexa 647); with low photon yield and low duty cycle (Atto 655); and with high photon yield and high duty cycle (Cy5.5). J-R) Images of immunolabeled clathrin-coated pits in three dimensions using Alexa Fluor 647, Atto 655 and Cy5.5.

Dye	Extinction (M^{-1} cm^{-1})	Photons per switching cycle		Equilibrium Duty Cycle		Survival Fraction		Number of Cycles (mean)	
		BME	MEA	BME	MEA	BME	MEA	BME	MEA
Cy2	150000	4583	6241	0.0005	0.00012	0.19	0.12	0.7	0.4
FITC	70000	1321	639	0.00031	0.00041	0.9	0.75	16	17
Fluorescein	70000	966	1493	0.00034	0.00032	0.83	0.51	15	4
Alexa 488	73000	427	1193	0.0017	0.00055	1	0.87	139	17
Atto 488	90000	1110	1341	0.0022	0.00065	0.99	0.98	49	11
Atto 520	110000	868	1231	0.00061	0.0015	0.86	0.92	17	9
TAMRA	90430	2025	4884	0.0049	0.0017	0.99	0.85	59	10
Cy3	150000	8158	11022	0.0003	0.0001	0.55	0.17	1.6	0.5
Cy3B	130000	2057	1365	0.0004	0.0003	0.89	1	5	8
Atto 565	120000	8825	12082	0.00041	0.00061	0.52	0.66	7	10
Alexa 568	88000	1686	2826	0.0028	0.00058	0.99	0.58	52	7
Cy3.5	150000	8028	5291	0.0005	0.0017	0.61	0.89	3.3	5.7
Atto 647N	150000	4433	3254	0.0035	0.0012	0.65	0.24	39	9
Atto 647	150000	944	1526	0.0016	0.0021	0.84	0.46	24	10
Alexa 647	270000	5202	4640	0.0013	0.00065	0.73	0.66	26	16
Dyomics 654	220000	3090	4130	0.0018	0.0011	0.64	0.79	19	20
Atto 655	125000	657	1105	0.0011	0.0006	0.78	0.65	22	17
Cy5.5	250000	6337	5831	0.0073	0.0069	0.85	0.87	25	16
Atto 680	125000	987	1656	0.0024	0.0019	0.91	0.65	41	8
Atto 740	120000	463	1126	0.0014	0.00047	0.96	0.31	14	3
Cy7	200000	997	852	0.0004	0.0003	0.49	0.48	4.6	5
Dy 750	220000	749	792	0.0002	0.0006	0.58	0.55	5	6
IR 800 CW	240000	2540	2753	0.038	0.0018	1	0.6	127	3
Alexa 790	260000	740	591	0.0014	0.00049	0.62	0.54	5	5

Table 1: Excitation wavelength, dichroic mirrors and emission filters used for characterization and imaging for each spectral range were 488 nm, T495LP (Chroma) and ET535/50m (Chroma) for blue-absorbing dyes; 561 nm, Di01-R561 (Semrock) and FF01-617/73-25 (Semrock) for yellow-absorbing dyes; 647 nm, Z660DCXRU (Chroma) and ET700/75m (Chroma) for red-absorbing dyes; 752 nm, Q770DCXR (Chroma) and HQ800/60m (Chroma) for NIR-absorbing dyes, respectively. Dye-switching properties are reported in the presence of GLOX and 10 mM MEA as well as GLOX and 140 mM βME.

The dyes also exhibited large variations in their photostability and number of switching cycles (Table 1). Although the three dyes in Figure 35 showed comparably high survival fractions after the molecules reached an on/off equilibrium state, some other dyes photobleached much more rapidly. For example, Cy2 showed little recovery to the on state after the initial switching off event, and by the time the duty cycle reached equilibrium, only a very small fraction of molecules survived. Consequently, it was essentially impossible to reconstruct a STORM image using Cy2 (Fig. 36A). Related to the survival fraction is the number of switching cycles afforded by a fluorophore. The detection of multiple switching events from the same fluorophore reduces the stochasticity of the localization error. This effect directly impact the STORM image quality in that a dye with a low number of switching cycles often results in an image with more ill-defined spatial features due to localization errors, whereas a dye with a large number of switching cycles results in a more continuous image due to repetitive sampling of the same structure and thus lower noise in the final image. For instance, dyes with relatively few switching cycles, such as Cy3, produced STORM images with sparse localization densities and overall poor quality, despite its high photon number and low on/off duty cycle (Fig. 36B). In contrast, fluorophores that exhibited a high number of switching cycles, such as Atto 647N, resulted in more continuous and better-defined CCP structures (Fig. 36C).

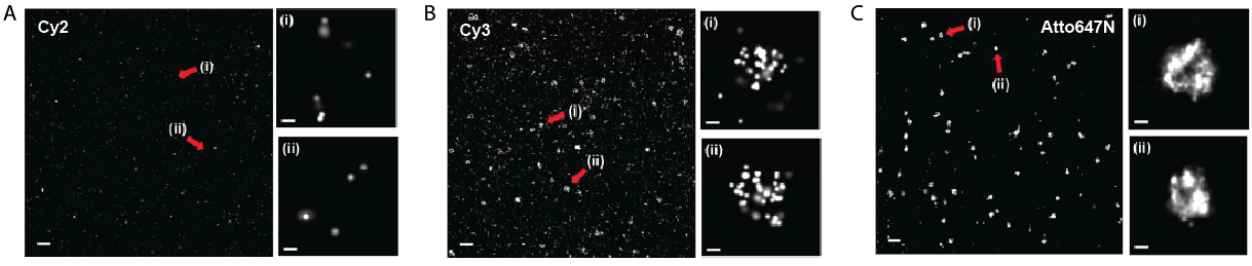


Figure 36: Images of immunolabeled clathrin-coated pits (CCPs) using A) Cy2, B) Cy3, and C) Atto647N labeled antibodies. Cy2 and Cy3 labeled antibodies yielded low quality STORM

(Continued) images due to poor photostability and low number of switching cycles respectively. In contrast, Atto647N is able to reveal the hollow structure of CCP.

3.2.2 Effect of buffer composition on switching properties

Additionally, the switching properties of dyes depend strongly on buffer composition. We therefore tested the switching performance of all dyes under four different buffer conditions: (i) our standard buffers containing both an oxygen scavenging system (glucose oxidase with catalase, or GLOX) and a primary thiol (βME or MEA) — "GLOX + thiol"; (ii) a buffer with GLOX but no thiol — "GLOX only"; (iii) a buffer with a thiol (MEA) but no GLOX — "Thiol only", and (iv) a buffer with neither GLOX nor thiol —"No GLOX or thiol" (Table 2). We found that nearly all of the tested dyes photobleached rapidly under the "No GLOX or thiol" or "Thiol only" conditions, while several dyes switched reasonably well in the "GLOX only" condition.

	No Glox or Thiol	Thiol only	Glox only	Glox and Thiol
Cy2	-	-	-	-
FITC	-	+	+	+
Fluorescein	-	+	+	+
Alexa Fluor 488	-	+	++	++
Atto 488	-	+	++	++
Atto 520	-	+	++	++
TAMRA	-	-	+	+
Cy3	-	-	+	+
Cy3B	-	+	++	++
Atto 565	-	-	+	+
Alexa Fluor 568	-	-	+	++
Cy3.5	-	-	+	+
Atto 647N	+	+	-	++
Atto 647	+	+	+	++
Alexa Fluor 647	-	+	-	++
Dyomics 654	-	+	+	++
Atto 655	-	+	++	++
Cy5.5	-	+	-	++
Atto 680	-	+	++	++
Atto 740	-	+	+	+
Cy7	-	+	-	++
Dy 750	-	+	-	++
IRDye 800CW	-	-	-	+
Alexa Fluor 790	-	+	-	+

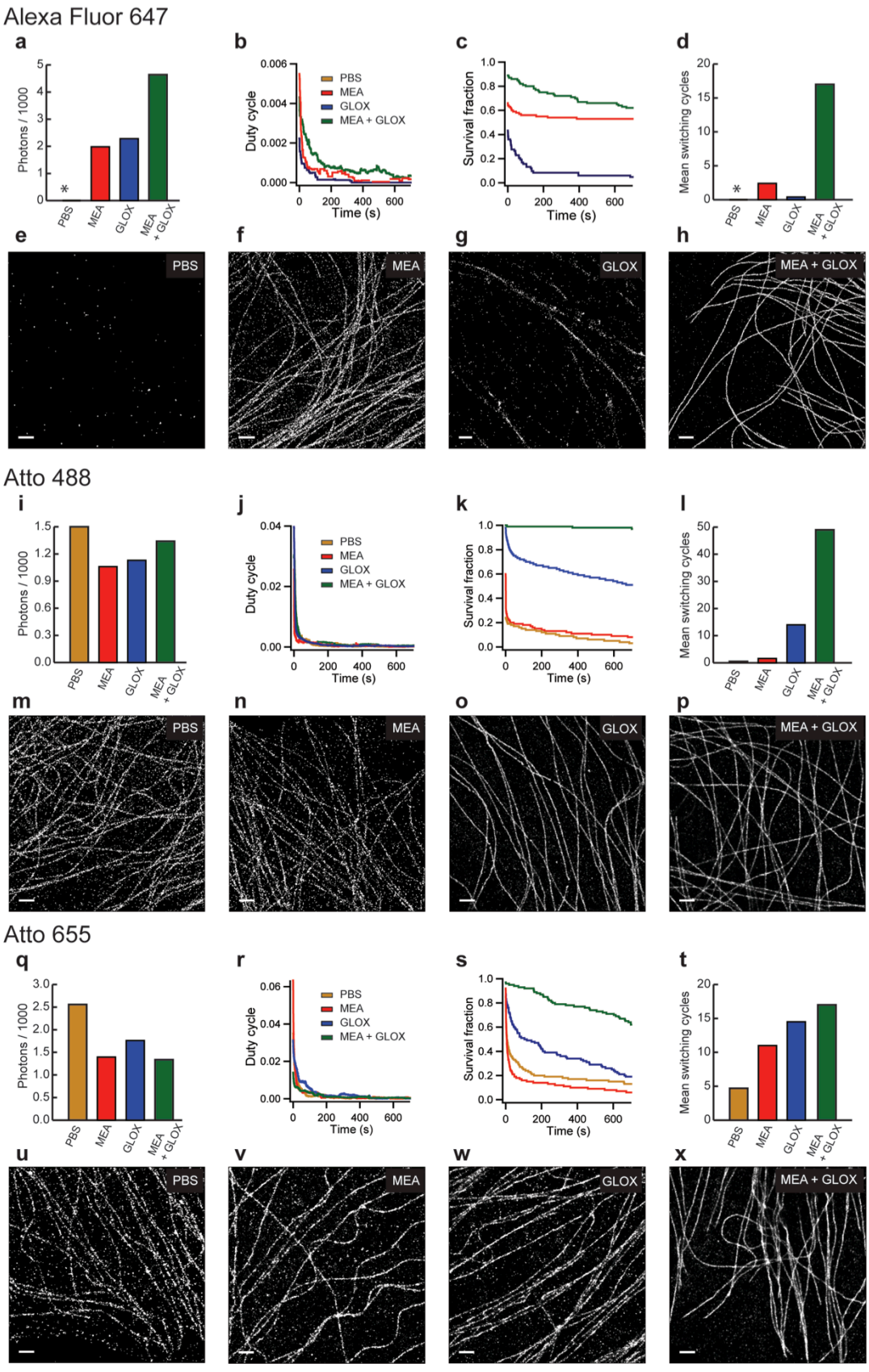
Table 2: Cells were immunostained for microtubules using each of the dyes. The thiol used was MEA. Dyes were assigned a "–", "+" or "++" if rapid bleaching (no image), low to moderate photoswitching (low to moderate quality image) or robust photoswitching (high quality image) was observed, respectively. Whereas lower image quality (+) was observed for Atto 647N and Alexa Fluor 750 in the 'GLOX and thiol' condition, reasonable image quality (++) was observed for these two dyes when βME was used instead of MEA

To more concretely illustrate the effect of thiol or oxygen removal on dye switching, we quantitatively characterized the switching properties of Alexa 647, Atto 488, and Atto 655 under all four buffer conditions (Fig. 38). Consistent with previous observations, we found that Alexa

647 required thiol for photoswitching and that the photon yield per switching cycle, photostability, and number of switching cycles further increased when oxygen was removed by GLOX. In contrast, Atto 488 switched well in the absence of thiol, but performed best in buffer containing both thiol and GLOX. In particular, the photostability of Atto 488 increased in the order of “No GLOX or thiol” or "Thiol only" < "GLOX only" < "GLOX + thiol", and this order correlated well with STORM image quality. In the case of Atto 655, we observed substantially better performance in the "GLOX + thiol" condition and found that even the "GLOX only" condition yielded better photostability and image quality than the "Thiol only" condition.

Furthermore, we found that the switching behavior was rather sensitive to the thiol concentration, as illustrated by more detailed investigations of Atto 488, Cy3B, Alexa 647, Atto655, and DyLight 750. We found that while the addition of thiol enhanced the dye’s switching performance and the STORM image quality, high concentrations (100 mM) of MEA diminished either the number of switching cycles (Cy3B), or number of photons per switching cycle (Atto 488 and Alexa 647), or both (DyLight 750 and Atto 655) (Fig. 39). Optimal performance was achieved under relatively low MEA concentrations for these five dyes.

These results emphasize the importance of buffer formulation in STORM imaging. We therefore recommend that users optimize buffer conditions for a given probe. Overall, we observed large variability in the switching properties of the different dyes. The number of detected photons per switching event ranged from a few hundred to several thousand (Table 1). This wide range of photon numbers led to wide variability in the localization precision and hence STORM image resolution. An even broader range of duty cycles was observed, spanning two orders of magnitude from 0.0001 to 0.04 (Table 1).

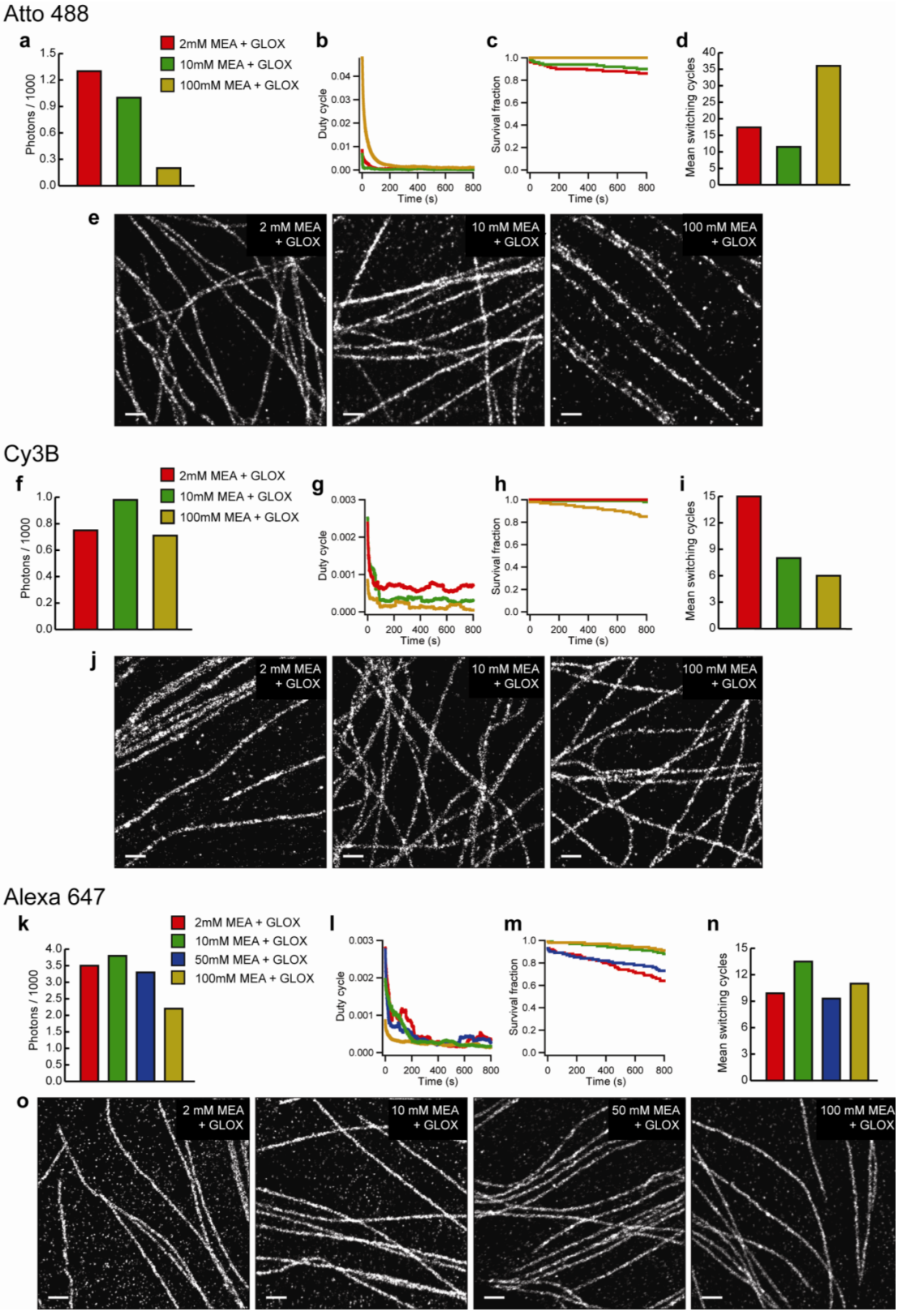

Alexa Fluor 647
a
Photons / 1000
PBS
MEA
GLOX
MEA + GLOX
b
Duty cycle
Time (s)
c
Survival fraction
Time (s)
d
Mean switching cycles
e
f
g
h
Atto 488
i
j
k
l
m
n
o
p
Atto 655
q
r
s
t
u
v
w
x

(Continued) Figure 38: Effect of buffer conditions on the switching properties of Alexa 647, Atto 488, and Atto 655 dyes and the corresponding STORM image quality. All 3 dyes performed best in buffers containing both GLOX and thiol (MEA). Reasonably good photoswitching performance was also observed for Atto 488 and Atto 655 in buffer containing GLOX only (o,w), while the switching in all other conditions was relatively poor (e,f,g,m,n,u,v). Quantitative analysis from single molecule measurements of photon numbers per switching cycle, duty cycle, survival fraction, and number of switching cycles ((a-d) for Alexa 647, (i-l) for Atto 488, and (q-t) for Atto 655) corroborates these findings.

Accordingly, the STORM image qualities afforded by these dyes also showed large variability. Dyes with high photon numbers, low duty cycles, high survival fractions, and many switching cycles in general performed well and produced high quality images, while dyes with low photon numbers, high duty cycles, low survival fractions and/or few switching cycles did not (e.g., fluoresein, FITC, Cy2, Cy3, Atto 565, Cy5.5, Atto 740, Alexa 790, IRDye 800CW).

3.2.3 Best performing dyes resolve the hollowness of microtubule filaments

Several dyes within our screen showed good overall properties, resulting in high-quality super-resolution images. Among these dyes, Alexa 647 (similarly its structural analog Cy5) and Dyomics 654 emerged as top choices for their overall excellent properties. When used to image microtubules, we observed that Alexa 647 and Dyomics 654 were able to resolve the hollowness of immunostained microtubule filaments (Fig. 40). Figure 40C-D show the transverse profile of localizations with two resolved peaks separated by 36-38 nm, as expected for the projection of a 25 nm diameter cylinder that has been broadened by primary antibodies (~10 nm in size) prior to staining with dye-labeled secondary antibodies.


Atto 488
a
2mM MEA + GLOX
10mM MEA + GLOX
100mM MEA + GLOX
Photons / 1000
b
Duty cycle
Time (s)
c
Survival fraction
Time (s)
d
Mean switching cycles
e
2 mM MEA + GLOX
10 mM MEA + GLOX
100 mM MEA + GLOX
Cy3B
f
2mM MEA + GLOX
10mM MEA + GLOX
100mM MEA + GLOX
Photons / 1000
g
Duty cycle
Time (s)
h
Survival fraction
Time (s)
i
Mean switching cycles
j
2 mM MEA + GLOX
10 mM MEA + GLOX
100 mM MEA + GLOX
Alexa 647
k
2mM MEA + GLOX
10mM MEA + GLOX
50mM MEA + GLOX
100mM MEA + GLOX
Photons / 1000
l
Duty cycle
Time (s)
m
Survival fraction
Time (s)
n
Mean switching cycles
o
2 mM MEA + GLOX
10 mM MEA + GLOX
50 mM MEA + GLOX
100 mM MEA + GLOX

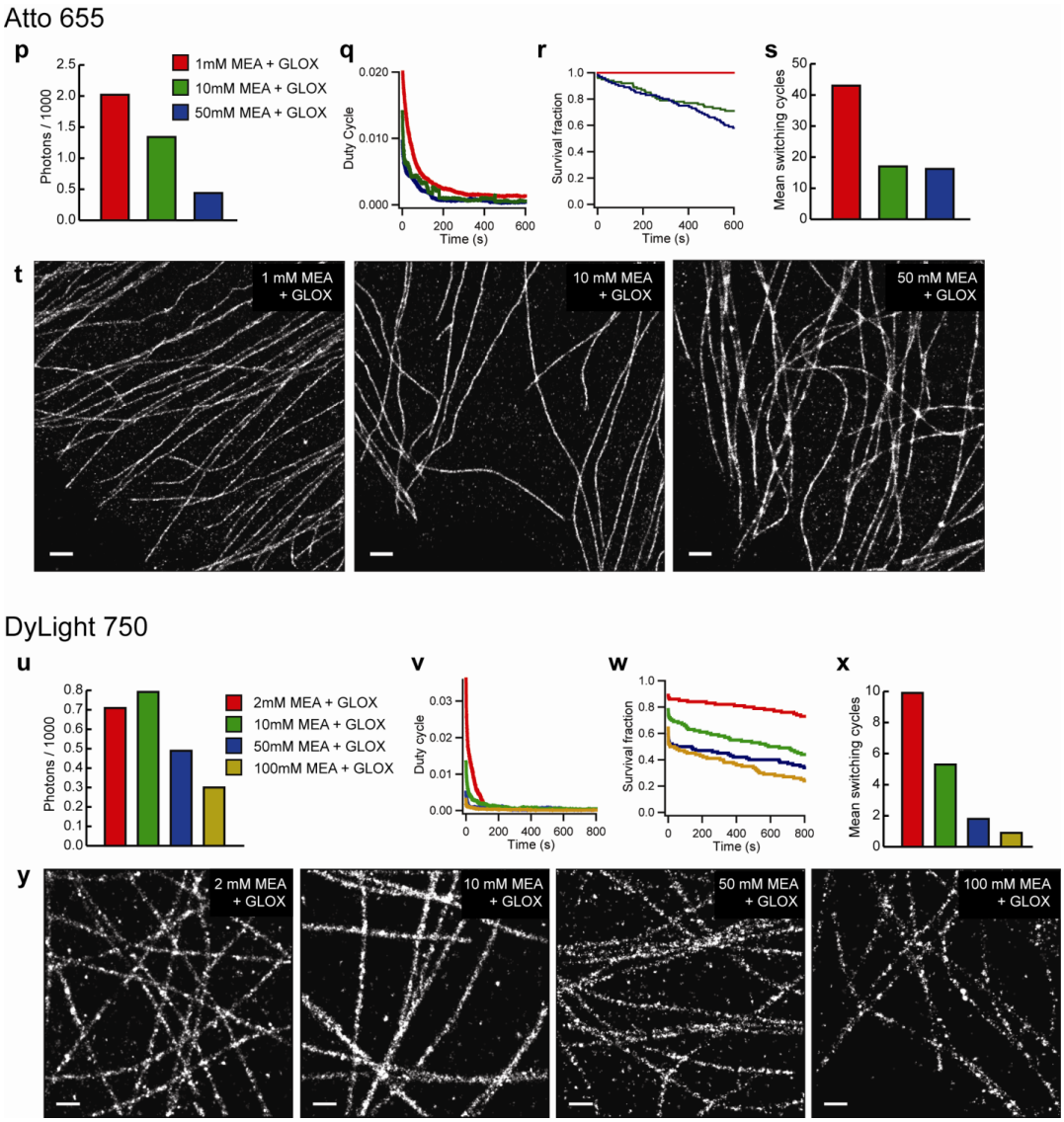


Figure 39: Dependence of the switching properties of Atto 488, Cy3B, Alexa 647, Atto 655, and DyLight 750 on thiol concentration and the corresponding STORM image quality. (a-d, f-i, k-n, p-s, u-x) Single-molecule measurements of photon number per switching cycle, duty cycle, survival fraction, and number of switching cycles of dyes under different MEA concentrations ((a-d) for Atto 488, (f-i) for Cy3B, (k-n) for Alexa 647, (p-s) for Atto 655, and (u-x) for DyLight 750). (e,j,o,t,y) STORM images of microtubules labeled with Atto 488 (e), Cy3B (j), Alexa 647 (o), Atto 655 (t), and DyLight 750 (y) under different MEA concentrations. In all cases, we observed a reduction in the photon number at high (50-100 mM) thiol concentrations and in some cases a reduction in the number of switching cycles as well. The optimal performance was achieved at low to moderate (2-10 mM) thiol concentrations. Scale bars: 500 nm for (e,j,o,y) and

(Continued) 1 µm for (t). In all measurements, GLOX was used, and only the concentration of MEA was varied.

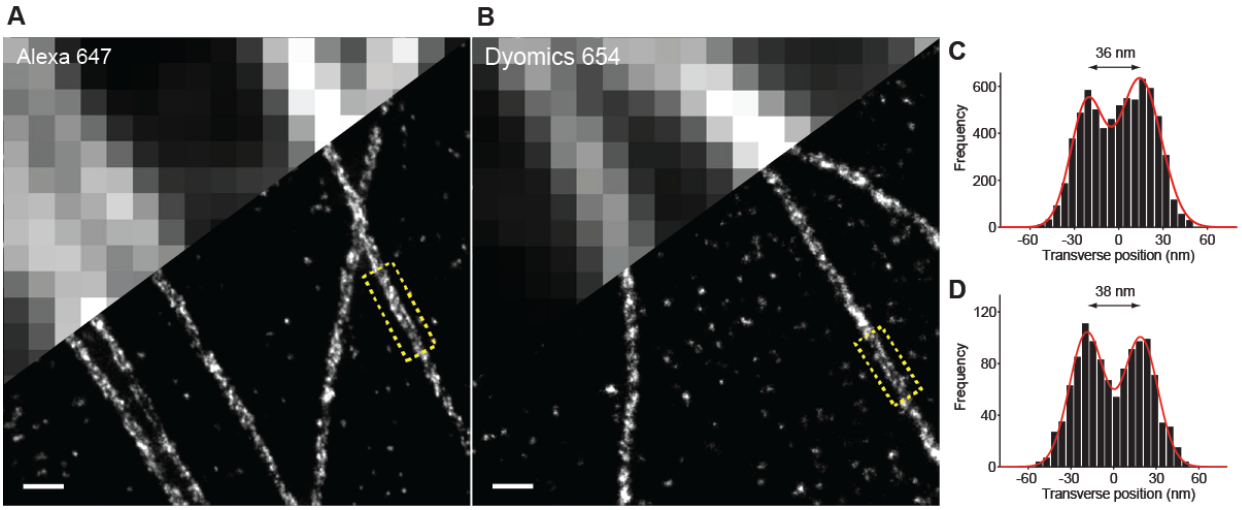


Figure 40: Alexa Fluor 647 and Dyomics 654 resolve the hollow structure of immunostained microtubules. (A, B) STORM images of microtubules immunostained with Alexa Fluor 647 (A) and Dyomics 654 (B) and partially overlaid conventional fluorescence images (top left). (C, D) Transverse profiles of localizations corresponding to regions boxed in yellow in A and B. Fitting of the profile by two Gaussian functions (red lines) gave the expected distances between the two peaks. Scale bars, 250 nm

3.2.4 Low cross-talk four-color super-resolution imaging

From our screen, we identified dyes with good performance in each of four distinct spectral ranges (blue: 480–540 nm, yellow: 545–600 nm, red: 640–700 nm, and NIR: 740–805 nm) to demonstrate four-color STORM imaging. These dyes were Atto 488 in the blue range, Cy3B in the yellow range, Alexa 647, Cy5, and Dyomics 654 in the red range, and DyLight 750, Cy7, and Alexa 750 in the NIR range. Several of these dyes (Atto 488, Dyomics 654, DyLight 750, and Cy7) were used here for the first time in STORM imaging. It should be noted that the red dyes Alexa 647, Cy5, and Dyomics 654 performed considerably better than even the best-performing dyes in other spectral regions. In comparison, Atto 488, Cy3B, DyLight 750, Cy7

and Alexa 750 were substantially dimmer and Cy3B additionally exhibited a lower number of switching cycles and a wider distribution of photons per switching cycle. Nevertheless, when compared to other dyes within each corresponding spectral band, these dyes exhibited the best overall performance in terms of photon output, duty cycle, survival fraction, number of switching cycles, and quality of super-resolution images.

To demonstrate the four-color imaging capability, we imaged a model sample of in vitro prepared microtubules with Atto 488, Cy3B, Alexa 647 and DyLight 750 (Fig. 41A). The microtubules, which were labeled separately with each of these dyes, were mixed and immobilized on coverglass prior to imaging. The excellent spectral separation afforded by the probes (Fig. 41B) allowed us to achieve low crosstalk between the four channels (Fig. 41C-F). We observed 0-3% crosstalk between different channels, except for Cy3B which exhibited higher crosstalk (~8%) into the Atto 488 and Alexa 647 channels (Fig. 41G).

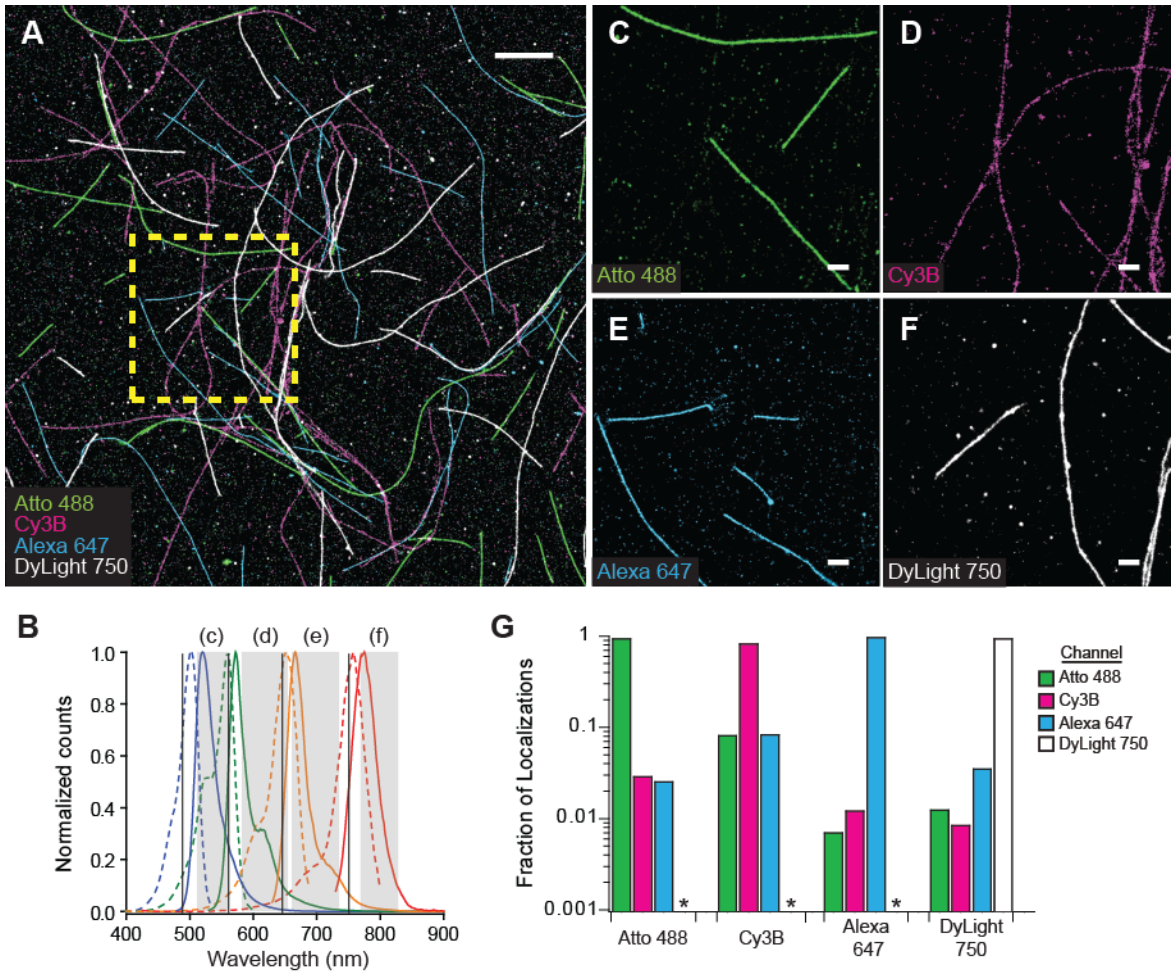


Figure 41: Four-color STORM imaging of in vitro assembled microtubule filaments and cross-talk analysis. (A) Four-color STORM image of in vitro assembled microtubules labeled with each of the four dyes, Atto 488 (green), Cy3B (magenta), Alexa Fluor 647 (cyan) and DyLight 750 (white). (B) Spectral separation of the four dyes, with the black vertical lines representing the excitation wavelength used and the gray regions highlighting the emission filter range for each of the dyes. (B–F) STORM images in each of the spectral regions for the boxed region in A. (G) Cross-talk between channels measured from control microtubule samples (undetectable in channels marked by an asterisk). Scale bars, 2μm (A) and 500 nm (C–F).

Finally, we demonstrated the utility of these probes in multicolor imaging of four different cellular targets. For this, we immunostained cells for tubulin, Tom20 (marker for mitochondrial outer membrane), ATL1 (marker for endoplasmic reticulum (ER)), and acetylated

tubulin using Atto 488, Cy3B, Alexa 647, and DyLight 750, respectively. Figure 42A-D shows STORM images for each of the four individual channels over the same field of view within a single cell. The excellent spectral separation allowed for easy color distinction of each cellular structure. Figure 5E-F shows the magnified, overlaid images of the ER with mitochondria and microtubules with acetylated tubulin, respectively.

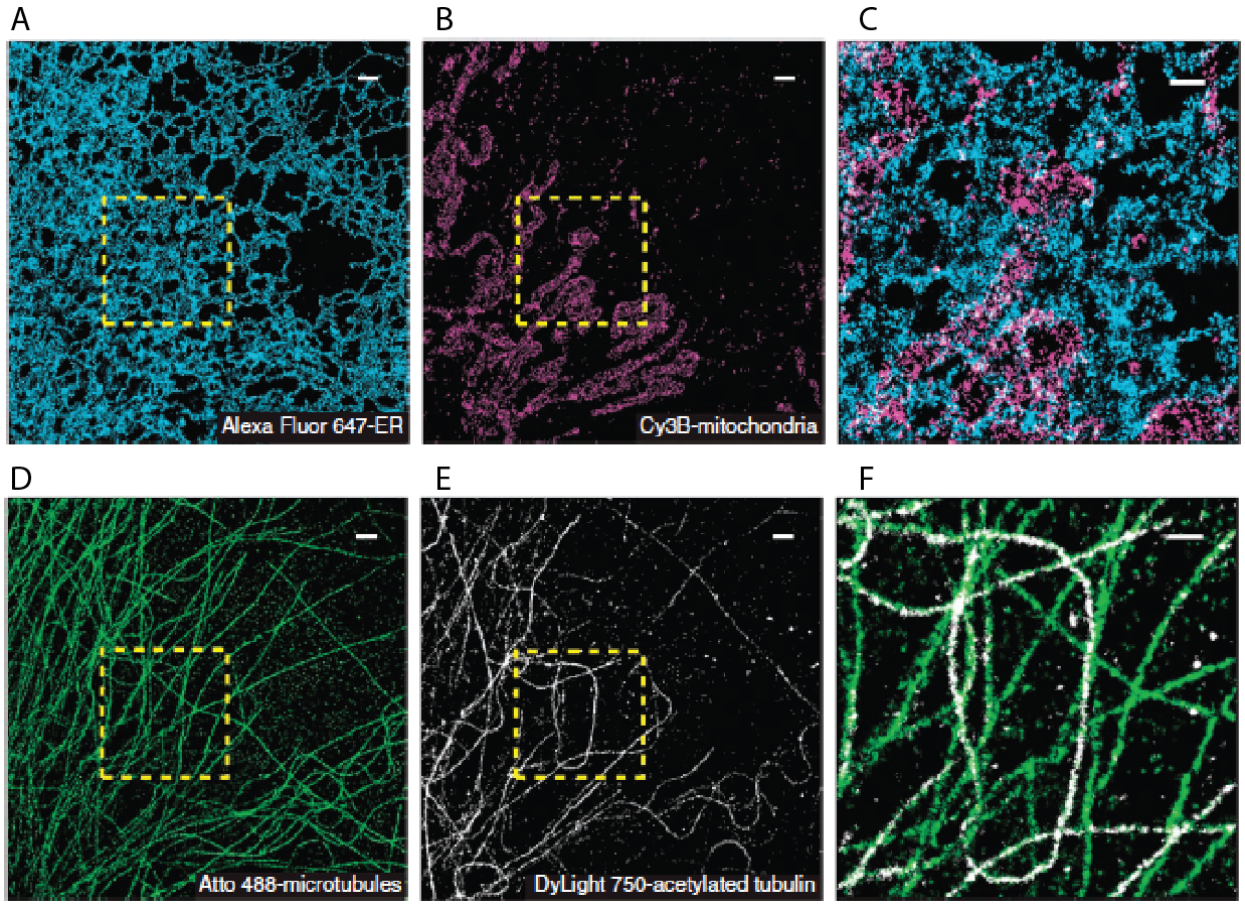


Figure 42: Four-color STORM imaging of cellular structures. (A–D) Individual channels of a four-color image of Alexa Fluor 647–labeled ER (cyan), Cy3B-labeled mitochondria (magenta), Atto 488–labeled microtubules (green) and DyLight 750–labeled acetylated tubulin (white) in a single fixed cell. (E) Magnification of the Cy3B and Alexa Fluor 647 channels of the boxed region in A–D showing extensive contact between mitochondria and the ER. (F) Magnification of the Atto 488 and DyLight 750 channels of the boxed regions in A–D showing overlap of

(Continued) acetylated tubulin with a subset of microtubule filaments. Scale bars, 1 μm (A–D) and 500 nm (E,F).

3.3 Methods and materials

3.3.1 Choice of dyes and antibodies

Dyes were obtained from Amersham (Cy2, Cy3, Cy3B, Cy3.5, Cy5, Cy5.5 and Cy7), Invitrogen (Alexa Fluors (here referred to as Alexa) 488, 568, 647, 750 and 790), ATTO-TEC (Atto 520, Atto 565, Atto 647, Atto 655, Atto 680 and Atto 740), Sigma (Atto 488 and Atto 647N), Anaspec (5(6)-TAMRA), Pierce (FITC, fluorescein and DyLight 750), Dyomics (Dyomics 654) and LI-COR (IRDye 800CW). All purchased dyes contained an N-hydroxysuccinimidyl ester (NHS) group for conjugation to protein, with the exception of FITC, which contains an isothiocyanate moiety for protein coupling. Primary antibodies were purchased from Abcam (mouse antibody to clathrin, ab2731; rat antibody to tubulin, ab6160; mouse antibody to acetylated tubulin, ab24610; goat antibody to myc, ab9132), and Santa Cruz Biotechnology (rabbit antibody to TOM20, sc-11415). Unconjugated secondary antibodies were purchased from Jackson ImmunoResearch Laboratories, including donkey anti-mouse (715-005-151), donkey anti-rat (712-005-153), donkey anti-rabbit (711-005-152) and bovine anti-goat (805-005-180).

3.3.2 Fluorescence microscope

All single-molecule and imaging measurements were performed on an Olympus IX-71 inverted microscope configured for either total internal reflection fluorescence (TIRF) or oblique incidence excitation. In the oblique incidence geometry, the incidence angle was slightly below the critical angle such that the excitation light illuminated 1–2 μm deep into the sample in the

field of view. The samples were continuously illuminated using different excitation sources depending on the fluorophore used. Blue- and red-absorbing dyes were excited using the 488 nm and 647 nm lines of a mixed-gas argon-krypton laser (Innova 70C; Coherent), respectively; yellow-absorbing dyes were excited with a 561-nm diode-pumped solid-state laser (Sapphire 561; Coherent); NIR-absorbing dyes were excited using the 752-nm line of a krypton laser (Innova 302C; Coherent). A 405-nm solid-state laser (Cube 405-100C; Coherent) was used for activation of dyes as described below. Beam selection and modulation of the laser intensities were controlled in several ways, depending on the laser wavelength. For 488-nm and 647-nm lines, an acousto-optic tunable filter (AA Optoelectronic) was used. The 561-nm and 752-nm lines were controlled using mechanical shutters (Uniblitz). For the 405-nm line, the light was controlled by direct digital modulation of the laser power supply. In all cases, neutral density filter wheels were used for coarse adjustments of laser power along with a combination of a half-wave plate and polarizer for fine power adjustments.

The following longpass dichroic mirrors were used to reflect the excitation sources listed above: T495LP (Chroma) for 488 nm, Di01-R561 (Semrock) for 561 nm, Z660DCXRU (Chroma) for 647 nm and Q770DCXR (Chroma) for 752 nm. Fluorescence was collected using an Olympus UPlanSApo 100×, 1.4 numerical aperture (NA) oil-immersion objective lens and passed through either one of the following bandpass emission filters: 535/50 (ET535/50m; Chroma) for blue-absorbing dyes; 617/73 (FF01-617/73-25; Semrock) for yellow-absorbing dyes; 700/75 (ET700/75m; Chroma) for red-absorbing dyes; or 800/60 (HQ800/60m; Chroma) for NIR-absorbing dyes. All movies were recorded onto a 256 × 256 pixel region of an electron-multiplying charge coupled device (EMCCD) camera (iXon 897; Andor).

During data acquisition, a home-built focus lock was used to maintain a constant focal plane. A 975-nm infrared laser diode (PL980P330J; Thorlabs) was directed toward the sample using a dichroic mirror (900DCSP; Chroma). The beam reflected off of the sample was detected by a quadrant photodiode (QPD). Changes in the axial position of the sample resulted in a shift of the reflected beam position on the QPD. A feedback system using custom LabVIEW software then adjusted the sample z-dimension position using a piezo stage (NZ100CE; Prior Scientific) until the original beam position was restored.

3.3.3 Single molecule experiments

Single molecule characterization of fluorophores was performed using dye-labeled donkey anti-rat secondary antibodies. Antibodies were labeled using the manufacturers' protocols. In brief, 1 M aqueous NaHCO3 pH 8.0 is added to 50-100 μg of antibody to a final concentration of 100 mM. Dyes were dissolved in dimethyl sulfoxide and added to the labeling mixture at varying concentrations (typically at a final concentration of ≈4 μM to achieve the desired dye to protein ratios). Labeled antibodies were purified by gel filtration using Nap-5 columns (17-0853; GE Healthcare). The labeling ratio was determined using a Beckman-Coulter UV-Vis spectrophotometer, and the absorption spectrum indicated a labeling ratio of 0.2 to 0.8 dyes per antibody such that individual antibody molecules would be expected to be labeled with at most one dye. Labeled antibodies were immobilized onto the surface of a flow chamber assembled from a glass slide and a #1.5 coverslip that were separated by double-sided tape. Slides and coverslips were cleaned by sonication in ethanol for 30 minutes followed by sonication in a 1 M aqueous potassium hydroxide for an additional 30 minutes. Coverslips were then washed with Mili-Q water and blow-dried with compressed nitrogen prior to assembly of

the flow chamber. The labeled antibodies were adsorbed to the coverslip at a low density (typically 0.04-0.12 dye/μm2) such that individual dye molecules could be clearly resolved. Control experiments showed that labeled antibodies do not detach from the surface on the time scale of the experiments (data not shown). The flow chambers were rinsed extensively with phosphate buffered saline (PBS) to remove unbound antibodies prior to imaging.

During measurements, an imaging buffer was used which contained TN buffer (50 mM Tris, pH 8.0, 10 mM NaCl), an oxygen scavenging system (0.5 mg/mL glucose oxidase (G2133, Sigma-Aldrich), 40 μg/mL catalase (106810, Roche Applied Science), and 5% (w/v) glucose), and either 143 mM 2-hydroxy-1-ethanethiol (βME; Fluka) or 10 mM 2-aminoethanethiol (MEA; Fluka), except where indicated otherwise. Blue-, yellow-, red-, and NIR-absorbing dyes were illuminated with 488 nm (1.2 kW/cm2), 561 nm (2.2 kW/cm2), 647 nm (0.8 kW/cm2), or 752 nm (1.3 kW/cm2) laser light in a total-internal reflection geometry. Slight variations in laser power are due to different transmission percentages for the excitation optics. Under constant illumination, the dyes typically switched to a dark state and spontaneously recovered to a fluorescent state several times before photobleaching. For each dye, several movies of ≈1500-2000 seconds duration were recorded, and 50–150 single molecule fluorescence time traces were obtained per movie. All movies were collected at a frame rate of 33 Hz. Samples lacking fluorescently labeled antibodies but otherwise prepared in the same way yielded virtually no detected fluorescent molecules in each of the four spectral ranges studied (data not shown). The low density of adsorbed fluorescent antibodies (<0.1 μm2) and the low number of dyes per antibody molecule (<0.8/IgG) was chosen to minimize the probability that fluorescent signals being studied originated from more than one dye molecule. Additionally, control experiments

were performed under nonswitching buffer conditions and showed single-step photobleaching for >90% of fluorescent spots evaluated.

3.3.4 Single molecule data analysis

In the initial frame of each single molecule characterization movie, the positions of adsorbed molecules were determined by identifying fluorescent signals above a threshold of 4 times the standard deviation of the background signal. Any fluorescent molecule within 10 pixels (1600 nm) of another fluorescent molecule was omitted from analysis. Fluorescence time traces were calculated for each molecule by analyzing the total integrated signal within a 7x7 pixel region (160 nm/pixel) during each frame of the movie. Fluorescence switching events were identified within the time traces by identifying fluorescent signals above a background threshold of 5-10 times the standard deviation of the fluorescence time trace background fluctuations. The duty cycle, D.C., of the dyes was calculated within a sliding window of 100 sec as

$$D.C. = \left\langle \frac{\sum_{i} \tau_{on,i}}{100sec} \right\rangle \tag{x}$$

where $\tau_{on,i}$ denotes the on-time of the i^{th} switching event for a given molecule within the time window, and the brackets indicate averaging over all molecules not yet photobleached. The equilibrium duty cycle reported was the mean of the duty cycle value between 400–600 s. Photobleached molecules were not included in the duty cycle calculation for times past photobleaching. Note that we only report the duty cycle and survival fraction, that is, the fraction of dyes not yet photobleached, for the first 700 s of illumination although the total movie was 2,000 s long. Therefore, only molecules that photobleached within the first 700 s need to be identified. Here we consider a molecule to be bleached after the final on event identified for that

molecule in the movie. The molecules that were identified to be bleached during the first 700 s based on this criterion most likely have in fact been photobleached given that the remaining 1,300 s of the movie is typically 4–20 times longer in duration than the mean off time of each fluorophore.

3.3.5 Immunofluorescence staining of cellular structures

Immunostaining was performed using BS-C-1 cells (American Type Culture Collection) cultured with Eagle's Minimum Essential Medium fortified with 10% fetal bovine serum, penicillin and streptomycin, and incubated at 37oC with 5% CO_2. Cells were typically plated in LabTek 8-well coverglass chambers (154534; Nunc) at ~20,000 cells per well on the day prior to fixation.

The protocol for immunostaining clathrin consisted of: fixation for 20 min with 4% paraformaldehyde (Electron Microscopy Sciences) in PBS; washing with PBS; permeabilization for 2 min with 0.5% Triton X-100 and 3% bovine syrum albumin (BSA; Jackson ImmunoResearch Laboratories) in PBS; blocking for 60 min with 0.2% Triton X-100 and 3% BSA in PBS; incubation for 60 min with mouse anti-clathrin heavy chain monoclonal antibody (Abcam ab2731) diluted to 2 μg/mL in blocking buffer; washing with blocking buffer; incubation for 30 min with dye-labeled donkey anti-mouse secondary antibody (~1-2 dyes per antibody) at a concentration of ~0.25 μg/mL in blocking buffer; washing with blocking buffer; washing with PBS; postfixation for 10 min with 3% paraformaldehyde and 0.1% glutaraldehyde (Electron Microscopy Sciences) in PBS; and finally washing with PBS.

The protocol for immunostaining of microtubules consisted of: fixation for 10 min with 3% paraformaldehyde and 0.1% glutaraldehyde in PBS; washing with PBS; reduction for 5 min

with 0.1% sodium borohydride in PBS; washing with PBS; blocking for 30 min with 3% bovine serum albumin and 0.25% v/v Triton X-100 in PBS; staining for 45 min with primary antibody against tubulin (rat or mouse anti-tubulin) diluted in block buffer to a concentration of ~10 μg/mL; washing with PBS; incubation for 45 min with secondary antibodies (~1-2 dyes per antibody) at a concentration of ~2.5 μg/mL in blocking buffer; washing with PBS; postfixation for 10 min with 3% paraformaldehyde and 0.1% glutaraldehyde in PBS; and finally washing with PBS.

The procedure for multicolor immunostaining of tubulin, detyrosinated tubulin, endoplasmic reticulum, and mitochondria was nearly identical to the microtubule immunostaining protocol described above. Prior to seeding, cells had been nucleofected with myc-ATL1 (gift of Dr. Craig Blackstone) using an Amaxa Nucleofector II (Lonza) with Amaxa Cell Line Nucleofector Kit V (Lonza) approximately 24 hours before fixation. Primary antibodies were used at concentrations of ~10 μg ml^{-1} (rat anti-tubulin, goat anti-myc), ~20 μg ml^{-1} (mouse anti-acetylated tubulin) and ~2 μg ml^{-1} (rabbit anti-TOM20) in blocking buffer 2. Secondary antibodies were labeled at a dye-to-protein ratio of ~2.6 Atto 488 dyes per donkey anti-rat antibody, ~1.7 DyLight 750 dyes per donkey anti-mouse antibody, ~1.5 Alexa 647 (or ~1.7 Dyomics 654) dyes per bovine anti-goat antibody and ~3.1 Cy3B dyes per donkey anti-rabbit antibody.

3.3.6 *In vitro* assembled microtubules

Microtubules were grown and labeled according to the following. Two prechilled 1 mg tubulin aliquots (Cytoskeleton T240) were carefully dissolved in a total of 500 μL of a prechilled solution consisting of PEM (100 mM PIPES pH 7.0, 1 mM EGTA, 1 mM MgCl2) with 1 mM

GTP (Cytoskeleton BST06) and 10% glycerol v/v. After centrifugation for 10 min at 18,000 g at 4 C, the supernatant was incubated at 37 C for 20 min to polymerize the microtubules. The microtubules were then stabilized by addition of paclitaxel (TXD01; Cytoskeleton) to 1 mM and incubation for 5 min at 37 C, and then stored at 25 C. Polymerized microtubules were labeled with dye by incubation of 60 μL of ~4 mg/mL polymerized microtubules with either 0.13 mM Atto488-NHS, 0.16 mM Cy3B-NHS, 0.77 mM Alexa Fluor 647-NHS, 0.45 mM Dyomics 654, or 0.64 mM DyLight 750-NHS for 60 min. Dye-labeled microtubules were purified from free dye by centrifugation through a dense layer consisting of PEM with 60% glycerol (v/v) and 1 mM paclitaxel for 2 hours at 300,000 g. Labeled microtubule pellets were washed carefully and then gently resuspended in 50 μL of buffer consisting of PEM with 1 mM paclitaxel. From absorption spectra, we estimated the labeling densities to be 0.26-0.62 dyes per tubulin heterodimer. A set of four labeled microtubule samples was mixed (either Atto 488, Cy3B, Alexa Fluor 647, and DyLight 750; or Atto 488, Cy3B, Dyomics 654, and DyLight 750), diluted 1:10 in PEM with 1 mM paclitaxel, and incubated for 10 min in LabTek 8-well chambers which had been cleaned by sonication in 1 M aqueous potassium hydroxide and silanized by incubation with 1% N-(2-aminoethyl)-3-aminopropyl trimethoxysilane (UCT Specialties LLC), 5% acetic acid, and 94% methanol for 20 min to facilitate microtubule adsorption to coverglass. The immobilized microtubules were fixed by incubation for 10 min with 0.5% glutaraldehyde in PEM with 1 mM paclitaxel followed by washing with PEM.

3.3.7 STORM imaging of immunostained cells and *in vitro* microtubules

STORM imaging of immunostained cells was performed using one of five different buffer conditions described above. Samples were illuminated with oblique incidence and

detected using different excitation sources, dichroic mirrors and emission filters depending on the fluorophore, as described above. For 3D STORM imaging, a 1-m cylindrical lens was inserted into the detection path to introduce astigmatism such that the images from individual molecules typically appeared elliptical. The x–y and z positions of the molecules were determined from the centroid position and the ellipticity of the image, respectively. To determine the z position of the molecule from the ellipticity of the single-molecule image, a calibration curve of the change in ellipticity for different z positions is needed. The calibration curve was determined before imaging by recording fluorescence from single immobilized beads (580/605 nm, 200 nm diameter microspheres; Molecular Probes) while scanning the z-piezo stage over an 800-nm range (400 nm above and below the central focal plane where the molecule images are circular rather than elliptical).

All STORM movies for blue-, yellow- and red-absorbing dyes were collected at a frame rate of 61 Hz and those for NIR-absorbing dyes were collected at 33 Hz. An illumination intensity of 1–3 kW cm^{-2} was used for these movies. Under this imaging condition, all dye molecules are typically in the fluorescent state initially but rapidly switch to a dark state. Recording of STORM movies was started once a low density of fluorescent molecules was attained (after ~10–300 s illumination) such that the images from individual fluorescence molecules were largely not overlapped with each other. All movie durations were 30,000–80,000 frames and were collected using constant illumination at the corresponding excitation wavelength for the dye.

For multicolor imaging, dyes were imaged sequentially: DyLight 750, then Alexa 647, then Cy3B and Atto 488 last. Filters were switched between datasets for each color. About 70,000 frames were recorded in each color channel before switching to another channel.

Simultaneous imaging of differently colored probes is in principle possible by properly separating the fluorescence emissions in the different color channels. However, owing to substantial activation of the red and NIR dyes by the blue light used to image Atto 488, the imaging quality may be compromised if all four color ranges were to be imaged simultaneously. For alignment between different color channels, bright-field images of 250 nm Au beads (Corpuscular) were taken before each STORM movie and used to register each channel into a common reference frame. We determined that the alignment error was <20 nm, based on control experiments performed with beads that emit in all four spectral bands.

Individual frames were convolved with a zero-sum Gaussian to suppress image noise. Molecules were identified in each frame and fit with a 2D rotating cylindrical Gaussian to determine the (x,y) position of each molecule. Molecules that were within a distance of 160 nm (one pixel) in successive frames were linked together and the mean position of the molecule in successive frames was used to report the position of that molecule. Lateral drift in the sample during acquisition was corrected computationally. The STORM image was divided into different time slices and the time correlation of each slice relative to the start of the movie was calculated. The image was then corrected, with typically <160 nm drift occurring during the entire movie. STORM images were rendered with each localization plotted as a Gaussian whose width is weighted by the inverse square root of the number of detected photons for that switching event. All images were filtered to reject molecules that are too dim (<300 photons), too elliptical (axial ratio >1.5), or that are “on” for >8 consecutive frames. For four-color immunofluorescence images, a density filter is applied to remove spurious localizations by removing a localization if there are fewer than 1-3 localizations in a 60x60 nm region around the localization.

3.3.8 Crosstalk Analysis.

To determine crosstalk between spectral channels, we obtained super-resolution images of in vitro microtubules labeled with one of the four probes in each of the four channels. For example, to determine the crosstalk of Alexa Fluor 647, we imaged microtubules labeled only with Alexa Fluor 647 in the DyLight Fluor 750 channel, followed by the Alexa Fluor 647 channel, the Cy3B channel, and lastly the Atto 488 channel. The percent of localizations contributing to the crosstalk was then determined from the STORM image coming from the DyLight Fluor 750, Cy3B, and Atto 488 channels relative to the correct Alexa Fluor 647 channel. This was done for each of the dyes used in multicolor imaging.

Chapter 4

Conclusion

Acquiring the complete picture of gene expression in an organism would entail measuring the copy number as well as the location of every transcript in every cell. This ability to measure genome wide expression profile and its spatial organization within single cells in their native environment would dramatically improve our abilities to understand biology and diseases. However, no previously reported technique offers such a capability. Most sequencing based methods lack positional information of the identified transcripts. Fluorescence *In Situ* Hybridization (FISH) based technologies, on the other hand, has been the workhorse for measuring transcripts localization, but is usually done for just a few genes at a time.

This thesis described two projects: 1) a new FISH technology to label and image thousands of cellular targets (in this case, RNA) and 2) a screening method to identify best performing dyes for multi-color super-resolution imaging. We believe that with the continual improvement of our massively multiplexed FISH approach, and in combination with advanced multi-color super-resolution microscopy, one could potentially perform spatially resolved gene expression measurement of the entire human transcriptome inside tissues or even a whole organism.

4.1 MERFISH for RNA imaging

Fluorescence microscopy has been a powerful tool for biological research. One of the reasons why light microscopy is so useful for probing cellular functions is its ability to achieve molecule-specific contrast. The first work presented in this thesis pushes the boundary on the

number of different molecules that could be probed simultaneously in a biological sample. By using the highly specific base pairing ability of DNA, sequential imaging with combinatorial labeling, and error robust encoding schemes to combat accumulating errors, thousands of RNA species could be tagged and readout sequentially with oligonucleotide probes.

In chapter 2 of this thesis, we demonstrated simultaneous imaging of 140 RNA species in human fibroblast cells with 16 rounds of hybridization and a modified Hamming code capable of both error detection and correction. We observed excellent correlation of RNA copy numbers determined by our method with both bulk RNA sequencing data and conventional smFISH measurements. Next, we used an alternative encoding scheme, which is capable of detecting but not correcting errors, to simultaneously image 1001 RNA species in individual cells using only 14 rounds of hybridization. The observed RNA copy numbers again correlate well with bulk sequencing data. However, the detection efficiency of this encoding scheme is only ~1/3 of that of the error-correcting encoding scheme.

We performed correlation analysis of the $10^4 – 10^6$ pairs of measured genes and identified many co-varying gene groups that likely share common regulatory elements. Such grouping allowed us to hypothesize potential functions of ~100 unannotated or partially annotated genes of unknown functions. We further analyzed correlations in the spatial distributions of different RNA species and identified groups of RNAs with different distribution patterns in the cell.

This highly multiplexed imaging approach enables analyses based on the variation and correlation of copy numbers, and spatial distributions, of a massive number of RNA species within single cells. Such analyses should facilitate the delineation of regulatory networks and the in situ identification of cell types. We envision that this technology will open up the possibility of performing gene expression profiling of single cells in native tissues at the transcriptome

scale, thus allowing spatially resolved transcriptomes or genomes to be determined. This ability to reveal the transcriptomes of different cell types in a complex tissue will allow researchers to probe the underlying mechanism of how cells are organized in tissues and how biological complexity can arise from cells having essentially the same genome.

4.2 Multi-color super-resolution microscopy

A key factor affecting the quality of STORM images is the choice of switchable probes. In chapter 3 of this thesis, we showed that probe performance could be quantified in terms of the number of photons per switching event, the on-off duty cycle, the survival fraction when the probe reaches on-off equilibrium and the number of switching cycles. In general, fluorophores with high photon yield per switching event, low on-off duty cycle, high survival fraction and large number of switching cycles are desired for producing high-quality STORM images. From this screen, we identified dyes suitable for STORM in each of four separate spectral bands (blue, yellow, red and NIR) and demonstrated four-color STORM imaging using these probes.

Continual advancement of super-resolution imaging technologies promises to bring fluorescence microscopy and all its advantages, like multi-color and live-cell imaging capability, into the molecular realm, where scientists can watch dynamic molecular processes with nanometer precision happening in their native environment. The ability to resolve multiple cellular targets at the molecular length scale will also allow researchers to image overlapping structures or interacting molecular machineries in the crowded cells. Indeed, one of the exciting applications of super-resolution microscopy is the imaging of our transcriptome.

4.3 Imaging of single cell transcriptome with multi-color super-resolution microscopy

To extend multiplexed FISH measurement to the transcriptome scale will require the ability to distinguish tens of thousands of RNA species. As discussed in chapter 2, by using encoding schemes with more number of bits while keeping or increasing the Hamming distance between codewords, our multiplexed labeling scheme can be extended to the transcriptome scale without the lost of accuracy or precision. The most straightforward way to readout more bits is to incorporate additional color channels into our measurement. For example, by using 4 color channels, 64 bits encoding schemes could be readout in 16 hybridizations.

There are a few foreseeable challenges to implement multiplexed FISH of the entire transcriptome. The synthesis or hybridization of encoding probes at such complexity needs to be tested and optimized in future experiments. Additionally, the enormous density of RNA inside a cell poses a fundamental optical limitation on the number of fluorescence spots that can be reliably resolved. This challenge may be overcome by employing super-resolution imaging. With ~20nm imaging resolution, the density of resolvable spots should allow the entire human transcriptome to be measured. Our extensive effort of screening dyes for four-color super-resolution imaging would now be useful for the implementation of multi-color, multiplexed FISH at the transcriptome scale.

Our multiplexed FISH technology has the potential to transform many areas of biological research. For example, the spatial patterning of transcription in a multi-cellular organism is known to play a crucial role in development. In neuroscience, transport of RNAs to the axon or dendrites may be important for the polarization of neurons. Additionally, the high spatial resolution of our approach should also allow transcriptomic analysis in important sub-cellular locations, such as synapses in our brain. The high spatial resolution of our method will also facilitate transcriptomic based mapping of cell types in our brain. Finally, we envision that this

spatially resolved RNA profiling technology would find broad applications in clinical setting. For example, many disease processes, such as the development of cancerous tumor, may involve spatially heterogeneous gene expression. The ability to spatially resolve the expression profile of tumor biopsy may be helpful for guiding the treatment program or choosing drugs appropriate for the different tumor types.

Appendix A: Oligo sequences

Chemically synthesized oligos purchased from Integrated DNA Technologies (IDT):

Directly labeled *ACTB* sequences

GGTGTTGAAGGTCTCAAACATGATCTGGGTCATCTTCTCG/3Cy3Sp/
GCTCAGTGAGGATCTTCATGAGGTAGTCAGTCAGGTCCCG/3Cy5Sp/
CGACGTAGCACAGCTTCTCCTTAATGTCACGCACGATTTC/3Cy3Sp/
TCACACTTCATGATGGAGTTGAAGGTAGTTTCGTGGATGC/3Cy5Sp/
CAGACAGCACTGTGTTGGCGTACAGGTCTTTACGGATGTC/3Cy3Sp/
GTAATCTCCTTCTGCATCCTGTCAGCAATGCCAGGGTACA/3Cy5Sp/
CAGACTCGTCATACTCCTGCTTGCTGATCCACATCTGCTG/3Cy3Sp/
TTAGGTTTTGTCAAGAAAGGGTGTAACGCAACTAAGTCAC/3Cy5Sp/
AAACAAATAAAGCCATGCCAATCTCATCTTGTTTTCTGCG/3Cy3Sp/
GTTCCAGTTTTTAAATCCTGAGTCAAGCCAAAAAAAAAAC/3Cy5Sp/

GAPDH encoding probes

GCGACGCAAAAGAAGATGCGGCTGACTGTCGAACAGGAGGGATCCGATTGGAACCGTCCCAAGCGTTGCGGCTATCG
TTCGTTCGAGGCCAGAGCATTCG
CTTCACCTTCCCCATGGTGTCTGAGCGATGTGGCTCGGCTCCCATGATCGTCCGATCTGGTCGGATTTGTAACGAGA
CGGTCAATCGCGCTGCATACTTG
GTGACCAGGCGCCCAATACGACCAAATCCGTTGACTCCGAATTTGCGTGCCGTGCTAGCCAGACGACGCAAGTGAAT
TGCTGCGCATCGGCGTACGACTT
GACAAGCTTCCCGTTCTCAGCCTTGACGGTGCCATGGAATGATCCGATTGGAACCGTCCCAAGCGTTGCGGCTATCG
TTCGTTCGAGGCCAGAGCATTCG
AGTGGACTCCACGACGTACTCAGCGCCAGCATCGCCCCACCCCATGATCGTCCGATCTGGTCGGATTTGTAACGAGA
CGGTCAATCGCGCTGCATACTTG
TGCAAATGAGCCCCAGCCTTCTCCATGGTGGTGAAGACGCATTTGCGTGCCGTGCTAGCCAGACGACGCAAGTGAAT
TGCTGCGCATCGGCGTACGACTT
GGTGCAGGAGGCATTGCTGATGATCTTGAGGCTGTTGTCAGATCCGATTGGAACCGTCCCAAGCGTTGCGGCTATCG
TTCGTTCGAGGCCAGAGCATTCG
GTCCTTCCACGATACCAAAGTTGTCATGGATGACCTTGGCCCCATGATCGTCCGATCTGGTCGGATTTGTAACGAGA
CGGTCAATCGCGCTGCATACTTG
CCATCCACAGTCTTCTGGGTGGCAGTGATGGCATGGACTGATTTGCGTGCCGTGCTAGCCAGACGACGCAAGTGAAT
TGCTGCGCATCGGCGTACGACTT
CAGCCTTGGCAGCGCCAGTAGAGGCAGGGATGATGTTCTGGATCCGATTGGAACCGTCCCAAGCGTTGCGGCTATCG
TTCGTTCGAGGCCAGAGCATTCG
CTAGACGGCAGGTCAGGTCCACCACTGACACGTTGGCAGTCCCATGATCGTCCGATCTGGTCGGATTTGTAACGAGA
CGGTCAATCGCGCTGCATACTTG
CAGCCCCAGCGTCAAAGGTGGAGGAGTGGGTGTCGCTGTTATTTGCGTGCCGTGCTAGCCAGACGACGCAAGTGAAT
TGCTGCGCATCGGCGTACGACTT
CATACCAGGAAATGAGCTTGACAAAGTGGTCGTTGAGGGCGATCCGATTGGAACCGTCCCAAGCGTTGCGGCTATCG
TTCGTTCGAGGCCAGAGCATTCG
TCTTACTCCTTGGAGGCCATGTGGGCCATGAGGTCCACCACCCATGATCGTCCGATCTGGTCGGATTTGTAACGAGA
CGGTCAATCGCGCTGCATACTTG
ACTGAGTGTGGCAGGGACTCCCCAGCAGTGAGGGTCTCTCATTTGCGTGCCGTGCTAGCCAGACGACGCAAGTGAAT
TGCTGCGCATCGGCGTACGACTT

Appendix B: Codebooks

The "Genes" columns contain the name of the gene. The "Codewords" columns contain the specific binary word assigned to each gene.

12-genes experiments:

Genes	Codewords
THBS1	0 0 1 1 1 0 0 1
FLNA	1 0 1 0 0 0 1 1
TNC	0 1 0 0 1 0 1 1
DYNC1H1	1 1 1 0 1 0 0 0
SPTBN1	0 1 1 0 0 1 0 1
SPTAN1	0 0 1 0 1 1 1 0
PCNX	1 0 0 1 1 0 1 0
ASCC3	0 0 0 1 0 1 1 1
PRKCA	0 1 1 1 0 0 1 0
GTF3C4	1 0 1 1 0 1 0 0
POLQ	0 1 0 1 1 1 0 0
STARD9	1 0 0 0 1 1 0 1
notarget013	1 1 0 0 0 1 1 0
blank014	1 1 0 1 0 0 0 1

32-genes experiments:

Codebook 1		Codebook 2	
Genes	Codewords	Genes	Codewords
CACNA1A	0 0 0 0 1 1 0 1 0 0 0 1	CACNA1A	0 1 1 0 0 0 1 0 0 1 0 0
AFF4	0 1 0 0 1 1 0 0 0 0 1 0	AFF4	0 0 1 0 1 0 1 1 0 0 0 0
THBS1	1 1 0 1 0 0 0 0 0 1 0 0	THBS1	1 1 1 0 0 1 0 0 0 0 0 0
FLNA	1 0 0 0 0 0 1 1 0 0 1 0	FLNA	1 0 0 1 1 0 0 1 0 0 0 0
TNC	0 1 1 0 0 0 1 0 0 1 0 0	TNC	1 0 0 1 0 1 0 0 0 0 0 1
DYNC1H1	0 1 1 1 0 0 0 0 0 0 0 1	DYNC1H1	1 0 1 0 1 0 0 0 0 0 1 0
SPTBN1	0 1 1 0 1 0 0 0 1 0 0 0	SPTBN1	0 0 0 0 0 0 1 0 1 0 1 1
SPTAN1	0 0 1 1 0 1 0 0 0 1 0 0	SPTAN1	0 0 0 1 0 1 0 1 1 0 0 0
PCNX	0 0 1 0 1 0 1 1 0 0 0 0	PCNX	0 1 0 0 1 0 0 1 0 1 0 0
ASCC3	0 0 0 0 1 0 1 0 1 1 0 0	ASCC3	0 1 1 0 1 0 0 0 1 0 0 0
PRKCA	0 0 1 0 0 1 0 0 1 0 1 0	PRKCA	1 0 0 0 1 1 0 0 1 0 0 0
GTF3C4	0 0 0 1 0 0 0 0 1 1 1 0	GTF3C4	0 0 0 1 0 0 0 0 1 1 1 0
POLQ	1 0 0 0 0 1 1 0 0 1 0 0	POLQ	1 0 0 0 0 0 0 1 1 0 0 1
STARD9	0 1 0 0 0 1 0 0 0 1 0 1	STARD9	0 0 1 0 0 0 0 1 1 1 0 0
TLN1	0 0 0 0 0 1 0 1 0 1 1 0	TLN1	0 0 1 1 0 0 0 1 0 0 1 0
FBN1	0 0 0 1 1 0 0 0 1 0 0 1	FBN1	1 0 0 0 0 0 1 1 0 0 1 0
MALAT1	1 1 1 0 0 1 0 0 0 0 0 0	MALAT1	0 1 1 1 0 0 0 0 0 0 0 1
NUP98	0 0 0 1 1 0 1 0 0 0 1 0	NUP98	1 1 0 1 0 0 0 0 0 1 0 0
KIAA1199	0 0 1 0 0 0 0 1 1 1 0 0	KIAA1199	0 0 1 0 0 1 1 0 0 0 0 1
GTF3C1	0 1 0 0 1 0 0 1 0 1 0 0	GTF3C1	1 1 0 0 0 0 1 0 0 0 0 1
NOTCH2	0 0 1 0 0 1 1 0 0 0 0 1	NOTCH2	0 0 0 1 1 0 1 0 0 0 1 0
IQGAP1	1 0 0 1 0 1 0 0 0 0 0 1	IQGAP1	0 1 0 0 1 1 0 0 0 0 1 0
DOCK7	0 0 0 0 1 0 0 0 0 1 1 1	DOCK7	0 0 0 0 0 1 0 1 0 1 1 0
FLNB	1 0 0 1 1 0 0 1 0 0 0 0	FLNB	0 1 0 0 0 1 0 0 0 1 0 1
TEAD1	0 1 0 0 0 0 0 1 0 0 1 1	TEAD1	1 0 0 0 0 1 1 0 0 1 0 0
TPR	1 0 1 1 0 0 1 0 0 0 0 0	TPR	0 1 0 0 0 0 1 1 1 0 0 0
SRRM2	0 0 1 1 0 0 0 1 0 0 1 0	SRRM2	0 0 0 0 1 0 0 0 0 1 1 1
SSH1	0 1 0 1 0 1 1 0 0 0 0 0	SSH1	0 0 0 1 0 0 1 0 0 1 0 1
DNAJC13	1 0 1 0 0 0 0 0 0 1 0 1	DNAJC13	0 1 0 0 0 0 0 1 0 0 1 1
KIAA1147	0 0 0 1 0 0 1 0 0 1 0 1	KIAA1147	0 0 0 0 1 0 1 0 1 1 0 0
FRY	0 0 0 1 0 1 0 1 1 0 0 0	FRY	0 0 1 0 0 1 0 0 1 0 1 0
BRCA2	0 1 0 0 0 0 1 1 1 0 0 0	BRCA2	0 0 1 1 0 1 0 0 0 1 0 0
notarget033	1 0 0 0 1 1 0 0 1 0 0 0	notarget033	0 1 0 1 0 1 1 0 0 0 0 0
notarget034	1 0 0 0 0 0 0 1 1 0 0 1	notarget034	1 0 1 1 0 0 1 0 0 0 0 0
notarget035	1 1 0 0 0 0 1 0 0 0 0 1	notarget035	0 0 0 0 1 1 0 1 0 0 0 1
blank036	1 0 1 0 1 0 0 0 0 0 1 0	blank036	0 0 0 1 1 0 0 0 1 0 0 1
blank037	0 0 0 0 0 0 1 0 1 0 1 1	blank037	1 1 0 0 0 0 0 0 1 0 1 0
blank038	1 1 0 0 0 0 0 0 1 0 1 0	blank038	1 0 1 0 0 0 0 0 0 1 0 1

140-genes experiments:

Codebook 1		Codebook 2	
Genes	Codewords	Genes	Codewords
THBS1	0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0	THBS1	0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 1
FLNA	0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 1	FLNA	0 0 0 0 1 1 0 0 0 1 0 0 1 0 0 0
TNC	1 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0	TNC	1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1
FBN2	0 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0	FBN2	0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1
KIAA1199	0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0	KIAA1199	0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1
COL5A1	0 0 0 0 0 1 0 0 1 1 0 0 0 0 1 0	COL5A1	0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1 0
MALAT1	0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0	MALAT1	0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1
TLN1	0 0 0 0 1 0 0 0 0 0 1 0 0 1 1 0	TLN1	0 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0 0
FBN1	0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 0	FBN1	1 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0
SRRM2	0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1	SRRM2	0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 1
TEAD1	0 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0	TEAD1	1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0
IQGAP1	0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1 0	IQGAP1	0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 1
MYH10	0 0 1 0 1 0 0 0 0 0 0 1 0 0 1 0	MYH10	0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 1
PAPPA	0 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0	PAPPA	0 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0
NR4A1	0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0	NR4A1	0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 1
LRP1	0 0 0 0 0 1 1 0 0 0 0 1 0 0 1 0	LRP1	0 0 0 0 0 0 0 0 1 0 1 0 1 1 0 0
SLC38A1	1 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0	SLC38A1	0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 1
SMARCA5	0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1	SMARCA5	0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 1
CKAP5	0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0	CKAP5	0 0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0
PIEZO1	1 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0	PIEZO1	0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 1
VCAN	0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0	VCAN	0 0 0 0 0 0 1 0 0 0 0 1 1 1 0 0
PRRC2B	0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0	PRRC2B	1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1
SON	1 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0	SON	0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 0
SLC7A11	1 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0	SLC7A11	0 1 0 0 1 0 1 0 0 0 0 0 0 1 0 0
GPR107	0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 0	GPR107	0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 1
KIAA1462	0 1 0 1 1 0 0 0 0 0 0 0 0 0 1 0	KIAA1462	0 0 1 1 1 0 0 0 0 0 0 0 0 1 0 0 0
notarget001	0 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0	notarget001	1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1
blank001	0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0	blank001	1 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0
ROCK2	0 1 0 0 1 0 1 0 0 0 0 0 0 1 0 0	ROCK2	0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1
SPTBN1	0 0 0 1 1 1 0 0 0 0 0 0 0 1 0 0	SPTBN1	0 0 0 0 1 0 0 0 0 0 1 0 0 1 1 0
FLNC	0 0 0 1 0 0 1 0 0 0 0 0 0 1 1 0	FLNC	0 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0
PRPF8	0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 1	PRPF8	0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1
IGF2R	0 0 0 0 1 1 0 0 0 1 0 0 1 0 0 0	IGF2R	0 0 0 1 1 1 0 0 0 0 0 0 0 1 0 0
FOSB	1 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0	FOSB	0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 1
UBR5	0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 1	UBR5	0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0
SVEP1	0 0 0 0 0 0 1 0 0 1 0 0 1 0 1 0	SVEP1	0 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0
CBX5	0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 1	CBX5	0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 1
PDS5A	1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1	PDS5A	1 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0
NUMA1	0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 1	NUMA1	0 1 0 0 0 1 0 1 0 0 0 0 1 0 0 0
PLXNA1	0 0 0 1 0 0 0 0 0 0 0 1 1 0 1 0	PLXNA1	0 1 0 1 0 0 0 0 0 0 1 0 0 1 0 0
ALPK2	0 1 0 0 0 1 0 0 0 1 0 1 0 0 0 0	ALPK2	0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0
AFAP1	0 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0	AFAP1	1 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0
SPTAN1	0 0 1 1 1 0 0 0 0 0 0 0 1 0 0 0	SPTAN1	0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0
DOCK7	0 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0	DOCK7	0 0 1 0 0 0 0 0 0 0 1 1 0 1 0 0
NOTCH2	0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1	NOTCH2	0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0
DYNC1H1	1 0 0 0 1 0 0 0 0 0 0 1 0 1 0 0	DYNC1H1	0 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0
ITGA2	1 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0	ITGA2	0 0 0 1 1 0 0 0 1 0 0 1 0 0 0 0
PRKDC	0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 1	PRKDC	0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1
GTF3C1	0 0 1 0 0 1 0 0 0 0 0 0 1 1 0 0	GTF3C1	0 0 1 0 1 0 0 0 0 0 0 1 0 0 1 0
THSD4	0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 1	THSD4	1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1
PTPN14	0 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0	PTPN14	0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0
LUZP1	0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 1	LUZP1	1 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0
AMOTL1	1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0	AMOTL1	0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 0
MAN1A2	1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0	MAN1A2	0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1
notarget002	1 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0	notarget002	1 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0
blank002	1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0	blank002	1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0
USP9X	0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0	USP9X	0 0 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0
CREBBP	0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 1	CREBBP	0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1
KPNA4	0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0	KPNA4	1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1
EGFR	0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 0	EGFR	1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
ZC3H13	0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 0	ZC3H13	1 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0
TNRC6A	0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0	TNRC6A	0 0 0 0 1 0 0 0 0 1 1 1 0 0 0 0
TPR	0 0 0 0 0 1 0 0 1 0 0 1 0 1 0 0	TPR	1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0
COL7A1	0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 1	COL7A1	0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1
ASCC3	0 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0	ASCC3	0 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0
USP34	0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1	USP34	0 0 1 0 0 0 1 0 0 0 1 0 1 0 0 0
CENPF	0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0	CENPF	0 0 1 0 0 0 0 0 1 0 0 1 1 0 0 0
KIDINS220	0 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0	KIDINS220	0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1
SCUBE3	1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0	SCUBE3	0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 1
NRIP1	0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0	NRIP1	0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0
NUP98	1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1	NUP98	0 0 0 0 0 1 0 0 0 0 1 1 1 0 0 0
ZFC3H1	0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0	ZFC3H1	0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1
FAM208B	0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 1	FAM208B	0 1 0 1 1 0 0 0 0 0 0 0 0 0 1 0
RAB3B	0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0	RAB3B	0 0 0 1 0 0 0 0 0 0 0 1 1 0 1 0
AFF4	1 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0	AFF4	0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 1
PRKCA	0 0 1 0 0 0 1 0 0 0 1 0 1 0 0 0	PRKCA	0 1 0 0 0 1 0 0 0 0 0 0 0 1 1 0
FLNB	1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1	FLNB	0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0
CHD8	1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1	CHD8	0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 1
ANKRD52	0 1 0 0 1 0 0 0 0 0 0 1 1 0 0 0	ANKRD52	0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 1
MKI67	0 0 0 0 0 0 1 0 0 0 0 1 1 1 0 0	MKI67	0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 0
DIP2C	0 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0	DIP2C	1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0
SSH1	0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 1	SSH1	0 1 0 1 0 0 0 0 1 0 0 0 1 0 0 0
notarget003	0 0 1 0 0 1 0 0 1 0 1 0 0 0 0 0	notarget003	0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 1
blank003	1 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0	blank003	0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 0

DNAJC13	0 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0	DNAJC13	0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 0
DSEL	0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0	DSEL	0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 0
PIK3CA	0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0	PIK3CA	0 0 0 0 0 1 1 0 0 0 1 0 0 1 0 0
MED14	1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0	MED14	0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1
HERC2	0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1	HERC2	1 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0
H6PD	0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0	H6PD	0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0
C14orf132	0 0 0 0 0 0 0 1 0 0 0 0 1 1 1 0	C14orf132	0 0 0 0 0 1 0 0 1 0 0 1 0 1 0 0
SIPA1L3	0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0	SIPA1L3	1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1
C17orf51	0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0	C17orf51	0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0
NF1	0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1	NF1	1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0
VPS13A	1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0	VPS13A	0 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0
ABCA2	1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0	ABCA2	0 0 0 0 0 0 1 0 0 1 0 0 1 0 1 0
PCNX	0 1 0 0 0 1 0 1 0 0 0 0 1 0 0 0	PCNX	0 0 0 1 0 0 1 0 0 0 0 0 0 1 1 0
NKTR	0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 0	NKTR	0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0
FYCO1	0 0 0 1 1 0 0 0 1 0 0 1 0 0 0 0	FYCO1	0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0
ATP11A	0 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0	ATP11A	1 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0
REV3L	0 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0	REV3L	1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0
CASC5	0 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0	CASC5	0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0
AKAP11	0 1 0 1 0 0 0 0 0 0 1 0 0 1 0 0	AKAP11	0 0 1 0 0 1 0 0 1 0 1 0 0 0 0 0
CEP250	0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1	CEP250	1 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0
USP24	0 0 0 0 0 0 0 0 1 0 1 0 1 1 0 0	USP24	1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0
AGPS	1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1	AGPS	1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0
CBL	1 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0	CBL	0 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0
RNF169	0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1	RNF169	0 0 1 1 0 0 0 1 0 0 1 0 0 0 0 0
TMOD2	0 0 0 1 0 0 0 0 0 1 0 0 1 1 0 0	TMOD2	0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 0
FASN	0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1	FASN	0 0 0 0 0 0 1 1 0 0 0 1 0 0 1 0
notarget004	1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0	notarget004	0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0
blank004	0 0 1 0 0 0 0 0 1 0 0 1 1 0 0 0	blank004	1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0
UBR2	0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 1	UBR2	0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0
PHIP	1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1	PHIP	0 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0
GTF3C4	0 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0	GTF3C4	0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0
QSER1	1 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0	QSER1	1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0
VPS13D	0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0	VPS13D	1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0
FRY	0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0	FRY	1 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0
DIP2B	0 1 0 1 0 0 0 0 1 0 0 0 1 0 0 0	DIP2B	0 1 0 0 0 1 0 0 0 1 0 1 0 0 0 0
HIVEP2	1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0	HIVEP2	1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0
DOPEY1	1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0	DOPEY1	0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0
ALMS1	0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0	ALMS1	0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0
HEATR5B	0 0 1 1 0 0 0 1 0 0 1 0 0 0 0 0	HEATR5B	0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0
HIVEP1	0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1	HIVEP1	0 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0
ARL10	0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 1	ARL10	0 0 0 0 0 1 0 0 1 1 0 0 0 0 1 0
HELZ2	0 0 0 0 0 0 0 1 1 0 0 0 1 0 0 1	HELZ2	0 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0
POLQ	1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1	POLQ	0 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0
CACNA1A	1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0	CACNA1A	1 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0
KIAA2018	0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1	KIAA2018	0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0
LYST	1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0	LYST	0 0 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0
ZKSCAN2	0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1	ZKSCAN2	0 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0
DIEXF	0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 0	DIEXF	0 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0
FZD4	1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0	FZD4	0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 0
SLC5A3	0 1 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0	SLC5A3	1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0
KIAA1147	1 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0	KIAA1147	0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 1 0
LMTK2	0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0	LMTK2	0 1 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0
STARD9	0 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0	STARD9	0 0 0 1 0 0 0 0 1 1 1 0 0 0 0 0 0
BRCA2	0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 1	BRCA2	1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0
notarget005	1 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0	notarget005	1 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0
blank005	0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 0	blank005	1 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0

1001-genes experiments:

Genes	Codewords
VCAN	0 0 0 0 0 0 0 0 0 1 0 1 1 1
PRRC2B	0 0 0 0 0 0 1 0 0 0 0 1 1 1
SON	0 0 1 0 0 0 0 0 0 1 0 0 1 1
SLC7A11	0 0 1 0 0 0 0 0 0 1 0 1 1 0
GPR107	0 0 0 0 0 0 0 0 0 0 1 1 1 1
KIAA1462	0 0 1 0 0 0 0 0 0 0 0 1 1 1
ROCK2	1 0 0 0 0 0 0 0 0 1 0 1 1 0
SPTBN1	0 0 0 1 0 0 0 0 0 1 0 1 1 0
FLNC	0 0 0 0 0 0 0 1 0 1 0 1 0 1
PRPF8	0 0 0 0 0 0 1 0 0 1 0 0 1 1
IGF2R	0 0 0 0 0 1 0 0 0 0 0 1 1 1
FOSB	0 0 0 1 0 0 0 0 0 0 0 1 1 1
UBR5	0 0 0 0 0 0 0 1 0 1 0 0 1 1
SVEP1	0 1 0 0 0 0 0 0 0 1 0 1 1 0
CBX5	0 0 0 1 0 0 0 0 0 1 0 1 0 1
PDS5A	0 0 0 0 0 0 0 0 1 1 0 1 1 0
NUMA1	0 0 0 0 0 0 0 1 0 0 0 1 1 1
PLXNA1	0 1 0 0 0 0 0 0 0 1 0 0 1 1
ALPK2	0 0 0 0 0 0 0 0 0 1 1 1 1 0
AFAP1	0 0 0 0 0 1 0 0 0 1 0 0 1 1
SPTAN1	0 0 0 0 0 0 1 0 0 1 0 1 1 0
DOCK7	0 0 0 0 1 0 0 0 0 1 0 1 1 0
NOTCH2	1 0 0 0 0 0 0 0 0 0 0 1 1 1
DYNC1H1	0 0 0 0 1 0 0 0 0 0 0 1 1 1
ITGA2	0 0 0 1 0 0 0 0 0 1 0 0 1 1
PRKDC	0 0 0 0 0 1 0 0 0 1 0 1 1 0
GTF3C1	0 0 0 0 0 0 0 0 0 1 1 1 0 1
THSD4	1 0 0 0 0 0 0 0 0 1 0 1 0 1
PTPN14	0 1 0 0 0 0 0 0 0 0 0 1 1 1
LUZP1	1 0 0 0 0 0 0 0 0 1 0 0 1 1
AMOTL1	0 0 0 0 1 0 0 0 0 1 0 0 1 1
MAN1A2	0 1 0 0 0 0 0 0 0 1 0 1 0 1
USP9X	0 0 0 0 1 0 0 0 0 1 0 1 0 1
CREBBP	0 0 0 0 0 0 0 0 1 1 0 1 0 1
KPNA4	0 0 0 0 0 0 0 1 0 1 0 1 1 0
EGFR	0 0 0 0 0 0 1 0 0 1 0 1 0 1
ZC3H13	0 0 0 0 0 1 0 0 0 1 0 1 0 1
TNRC6A	0 0 0 0 0 0 0 0 0 1 1 0 1 1
TPR	0 0 0 0 0 0 0 0 1 0 0 1 1 1
COL7A1	0 0 1 0 0 0 0 0 0 1 0 1 0 1
ASCC3	0 0 0 0 0 0 0 0 1 1 0 0 1 1
USP34	0 0 0 0 0 1 0 1 0 1 0 0 1 0
CENPF	0 0 0 0 0 0 1 0 1 0 0 0 1 1
KIDINS220	0 0 1 0 0 0 0 0 0 0 1 1 0 1
SCUBE3	0 0 0 0 0 1 0 0 1 0 0 1 0 1
NRIP1	0 0 1 0 0 0 0 0 0 0 1 0 1 1
NUP98	1 0 0 0 0 0 0 0 1 0 0 1 1 0
ZFC3H1	0 1 0 0 0 0 1 0 0 1 0 1 0 0
FAM208B	0 0 0 1 0 1 0 0 0 0 0 1 0 1
RAB3B	0 0 0 0 0 1 0 1 0 1 0 0 0 1
AFF4	0 0 0 0 0 0 0 1 1 1 0 1 0 0
PRKCA	0 0 1 0 0 0 0 0 0 1 1 0 1 0
FLNB	0 1 0 0 0 0 1 0 0 0 0 1 0 1
CHD8	1 0 0 0 1 0 0 0 0 0 0 1 1 0
ANKRD52	1 0 0 0 0 0 0 0 1 1 0 1 0 0
MKI67	1 0 0 1 0 0 0 0 0 1 0 1 0 0
DIP2C	0 0 1 0 0 0 0 0 1 1 0 0 0 1
SSH1	0 0 0 0 0 0 0 0 1 1 1 1 0 0
DNAJC13	0 0 0 0 0 0 1 1 0 0 0 0 1 1
B4GALT5	0 0 1 0 0 0 0 1 0 0 0 1 1 0
NAV1	0 0 1 0 1 0 0 0 0 1 0 0 0 1
EHMT1	1 0 0 0 0 0 0 0 0 1 1 1 0 0
CHD6	1 1 0 0 0 0 0 0 0 1 0 0 1 0
RAPGEF1	0 0 0 0 0 0 0 1 0 0 1 1 1 0
DSEL	0 0 0 0 1 1 0 0 0 0 0 1 1 0
PI4KA	0 0 1 0 1 0 0 0 0 0 0 0 1 1
TMX4	0 0 0 0 1 0 0 0 0 1 1 0 1 0
POGZ	0 0 0 1 0 0 1 0 0 0 0 1 1 0
VPS39	0 0 1 0 0 1 0 0 0 1 0 0 0 1
CDC42BPA	0 0 0 1 1 0 0 0 0 1 0 0 1 0
EIF4EBP2	0 0 0 0 1 0 0 0 0 0 1 1 1 0
FAM13B	0 0 0 0 1 0 0 0 1 1 0 0 0 1
ZMAT3	0 1 0 0 1 0 0 0 0 0 0 1 0 1
MCM3AP	0 1 1 0 0 0 0 0 0 0 0 0 1 1
SMARCC2	0 0 1 0 1 0 0 0 0 0 0 1 1 0
NPC1	0 0 1 1 0 0 0 0 0 1 0 0 0 1
ATP13A1	1 0 0 0 0 0 1 0 0 1 0 1 0 0
KRAS	0 0 1 0 0 0 1 0 0 1 0 0 0 1
SUGP2	0 0 0 0 1 0 0 0 1 0 0 1 1 0
PLD2	0 1 0 0 0 0 0 0 0 0 1 0 1 1
SMC6	1 0 0 0 0 1 0 0 0 1 0 1 0 0
HPS5	0 0 0 0 0 0 0 0 1 0 1 0 1 1
BROX	0 0 0 0 0 0 0 1 0 1 1 1 0 0
NOTCH3	0 1 0 1 0 0 0 0 0 0 0 1 0 1
SIDT2	1 0 0 1 0 0 0 0 0 0 0 1 0 1

Genes	Codewords
UBE4A	0 0 1 0 0 0 0 0 0 1 1 1 0 0
PCNX	0 0 0 0 0 1 1 0 0 0 0 0 1 1
ZNF516	0 1 0 0 0 1 0 0 0 0 0 0 1 1
DHX8	1 0 0 0 0 1 0 0 0 0 0 1 0 1
MEGF8	0 1 0 1 0 0 0 0 0 0 0 1 1 0
SBNO1	0 1 0 0 0 1 0 0 0 1 0 0 1 0
FNBP1L	0 0 0 1 0 0 1 0 0 0 0 1 0 1
LRP5	0 1 0 0 0 0 0 0 0 0 1 1 1 0
CES2	0 0 0 1 0 0 0 0 0 0 1 1 0 1
SEC16A	1 0 1 0 0 0 0 0 0 1 0 1 0 0
IKZF5	0 0 0 1 0 0 0 0 1 0 0 1 1 0
ATG16L1	0 0 0 1 0 0 0 0 0 0 1 1 1 0
PLCB3	0 0 1 0 0 1 0 0 0 1 0 1 0 0
SYNJ2	0 1 0 0 0 0 0 0 0 1 1 0 1 0
FAM13A	0 0 0 0 0 1 1 0 0 1 0 1 0 0
FAM178A	0 1 0 0 0 0 0 1 0 1 0 1 0 0
KIAA0247	0 0 0 0 1 0 1 0 0 0 0 1 0 1
MED15	1 0 0 0 0 0 0 0 0 0 1 0 1 1
AREL1	1 0 0 1 0 0 0 0 0 0 0 0 1 1
ZBTB21	1 0 1 0 0 0 0 0 0 0 0 1 1 1
SMURF1	0 0 0 0 0 1 0 0 1 1 0 0 1 0
TRPC6	1 0 0 0 0 0 0 0 0 1 1 0 1 0
ARHGAP18	0 0 0 0 0 0 1 1 0 1 0 1 0 0
blank003	0 0 0 0 1 0 0 0 0 0 1 0 1 1
KLF3	0 0 0 1 0 0 0 0 0 1 1 0 0 1
TOR4A	0 1 0 0 1 0 0 0 0 0 0 0 1 1
COG3	0 1 0 0 0 0 0 1 0 1 0 0 1 0
MCM8	0 0 0 0 1 0 0 1 0 1 0 0 1 0
FYCO1	0 0 0 1 0 0 0 1 0 1 0 0 1 0
SEMA5A	1 0 0 0 0 1 0 0 0 0 0 1 1 1
RP11-473I1.10	0 0 0 0 1 0 0 0 0 1 1 1 0 0
MTR	0 1 0 0 1 0 0 0 0 1 0 0 1 0
SMCR7L	0 0 0 0 1 0 0 0 0 1 1 0 0 1
ZFPM2	0 0 1 0 0 0 1 0 0 0 0 1 0 1
AMMECR1L	1 0 0 0 1 0 0 0 0 1 0 0 1 0
STAT5B	0 1 0 0 0 0 0 0 1 0 0 0 1 1
UST	0 0 0 0 1 0 0 0 0 0 1 1 0 1
PRDM2	0 0 1 1 0 0 0 0 0 0 0 1 0 1
GBF1	0 0 0 0 0 0 1 0 1 1 1 0 1 0
AKAP9	1 0 0 0 1 0 0 0 0 1 0 1 0 0
CSPP1	0 0 0 1 1 0 0 0 0 0 0 1 1 0
GUF1	0 0 0 0 0 0 0 0 1 1 1 0 0 0
ATP11A	0 0 1 0 0 0 0 0 0 1 1 0 0 1
POLD3	0 1 0 0 0 0 0 1 0 1 0 0 0 1
CD2AP	0 0 0 1 0 0 0 1 0 0 0 1 1 0
EPC2	0 0 0 0 1 0 0 0 1 1 0 0 1 0
BRD3	0 0 0 0 1 0 1 0 0 1 0 1 0 0
CDCA2	0 0 0 0 0 0 0 0 1 0 1 1 0 1
WDR44	0 0 1 0 0 1 0 0 0 0 0 0 1 0
ATF7IP	0 1 0 0 0 0 1 0 0 1 0 0 0 1
EVC	0 0 0 0 0 0 1 0 0 0 1 1 0 1
CDK5RAP2	1 0 0 0 0 0 0 0 1 1 0 0 0 1
NFAT5	0 0 1 0 0 1 0 0 0 0 0 0 1 1
KDM4A	0 1 1 0 0 0 0 0 0 1 0 1 0 0
TTLL4	0 0 1 1 0 0 0 0 0 1 0 0 1 0
AVL9	0 0 0 1 0 0 0 0 1 0 0 0 1 1
UBN1	0 0 0 1 0 1 0 0 0 1 0 0 0 1
BRWD1	0 0 0 1 0 0 0 0 1 0 0 0 1 1
TRIM37	0 0 0 1 0 0 1 0 0 1 0 0 1 0
ACSL1	1 0 0 0 0 0 0 0 0 0 1 1 1 0
MRAS	0 0 0 0 1 1 0 0 0 1 0 1 0 0
REV3L	0 0 0 0 1 0 0 0 1 1 0 1 0 0
LRRC8C	0 0 0 0 0 0 1 0 1 1 1 0 0 0
LLGL1	0 0 0 1 0 1 1 0 0 0 0 0 0 1
KLF13	0 1 1 0 0 0 1 0 0 0 0 1 0 0
KIAA0754	1 0 0 0 0 0 0 1 0 0 1 0 1 0
DDX60	1 0 0 0 0 0 0 0 1 0 1 1 0 0
DMD	0 0 0 0 0 1 0 1 0 0 1 1 0 0
LSM3	1 0 1 0 0 1 0 0 0 0 0 1 0 0
PCNXL3	0 0 1 0 1 0 0 0 1 0 0 0 1 0
ETV5	0 1 0 0 1 0 0 0 0 1 0 0 1 0
MLLT6	0 0 1 1 0 1 0 0 0 0 0 0 1 0
ARFGEF1	1 0 0 0 1 0 0 0 1 1 0 0 0 0
CASC5	0 0 0 0 0 0 1 1 0 0 1 0 1 0
AKAP11	0 1 1 0 0 1 0 0 0 0 0 0 1 0
CSRNP2	0 0 1 1 0 0 1 0 0 1 0 0 0 0
TWISTNB	0 0 1 0 0 0 0 0 1 0 1 1 0 0
NOL6	0 0 0 1 0 1 0 1 0 1 0 0 0 0
C17orf70	0 0 1 0 0 1 1 0 0 1 0 0 0 0
ENTPD7	0 1 0 1 0 0 0 0 1 1 0 0 0 0
LSS	1 0 1 0 1 0 0 0 0 0 0 0 1 0
FAM129A	0 1 0 1 0 1 0 0 0 1 0 0 0 0
MON2	0 0 0 0 0 0 0 1 1 0 1 0 0 1
TMEM194A	0 0 0 0 0 0 1 0 1 0 1 0 0 1
EXOC2	0 0 1 1 0 0 0 0 1 1 0 0 0 0

Genes	Codewords
HEATR6	0 0 0 1 0 0 1 0 1 0 0 0 0 1
LRCH2	0 1 0 1 0 0 0 0 1 0 0 0 0 1
ASH1L	0 0 0 1 0 0 1 1 0 0 0 0 0 1
TSPYL4	0 1 1 0 0 0 0 0 0 0 1 1 0 0
ADAMTS15	0 0 0 0 0 1 1 1 0 0 0 1 0 0
ZMYND8	0 1 0 1 0 0 0 0 1 0 0 1 0 0
ZSWIM6	0 0 0 1 0 0 0 0 1 0 1 0 0 1
ZHX3	0 0 1 0 0 0 1 0 0 1 0 0 1 0
ARFGEF2	0 1 0 0 1 0 1 0 0 0 0 0 1 0
ELF4	0 0 1 0 0 1 1 0 0 0 0 1 0 0
ZSCAN29	0 0 0 0 0 1 0 1 1 0 1 0 0 0
SENP1	0 0 0 0 0 1 0 1 0 1 1 0 0 0
TDRD7	0 1 0 0 0 1 0 0 1 0 0 1 0 0
TBC1D9	0 1 0 0 0 0 0 0 1 0 1 0 1 0
CRY2	0 0 0 0 0 0 0 1 1 0 1 0 1 0
SHC4	0 1 0 0 0 1 1 0 0 0 0 0 1 0
FAM168A	0 1 0 1 1 0 0 0 0 0 0 0 0 1
ABHD13	0 1 1 0 1 0 0 0 0 0 0 0 1 0
SLC30A1	0 1 0 1 1 0 0 0 0 1 0 0 0 0
DHX57	0 0 0 1 0 0 0 0 1 0 1 1 0 0
AP4E1	0 0 1 0 0 0 1 0 0 0 1 1 0 0
ZNF480	1 0 1 1 0 0 0 0 0 0 0 1 0 0
CEP170B	0 0 0 1 1 0 0 0 1 0 0 0 1 0
RNF169	1 0 0 1 0 0 0 0 0 0 1 0 1 0
USPL1	1 0 0 0 0 0 0 0 1 0 1 0 0 1
ANKRD50	0 0 0 0 0 1 0 0 1 1 1 0 0 0
APH1B	0 1 0 0 0 0 1 1 0 0 0 1 0 0
TMOD2	0 1 0 0 0 0 0 1 1 0 0 0 0 1
LYRM7	0 0 0 0 0 1 1 0 0 1 1 0 0 0
PHRF1	0 1 1 0 0 1 0 0 0 0 0 0 0 1
KAT6A	0 0 0 0 0 0 1 1 1 0 0 1 0 0
LIG3	0 0 0 1 0 0 0 1 1 1 0 0 0 0
KIAA1551	0 1 0 1 0 0 0 0 0 0 0 1 0 1
HPS4	0 0 0 0 1 0 1 1 0 0 0 1 0 0
LPIN1	0 1 0 0 1 0 0 0 1 0 0 0 0 1
C11orf95	0 0 1 1 0 0 0 0 0 0 0 1 0 1
WASH5P	1 0 1 0 0 0 0 0 1 1 0 0 0 0
BRD1	0 1 0 0 1 0 0 0 0 0 1 1 0 0
INTS2	0 1 1 0 0 0 0 1 0 0 0 0 0 1
MEGF9	1 0 0 1 0 0 0 0 0 0 1 1 0 0
C2CD3	0 0 0 1 0 0 1 0 0 0 1 0 0 1
FNIP1	0 1 1 0 1 0 0 0 0 1 0 0 0 0
FASN	0 0 0 1 1 0 0 1 0 0 0 0 0 1
RRP12	0 1 1 0 0 0 0 0 1 1 0 0 0 0
TCF20	0 0 0 1 1 0 0 0 0 0 0 1 0 1
BAHD1	1 1 0 0 0 0 0 0 0 1 0 0 1 0
CEP170P1	1 0 0 0 0 1 0 0 0 1 1 0 0 0
blank006	0 0 0 1 0 0 0 1 1 0 0 0 1 0
SLC12A2	0 0 0 0 0 0 0 1 1 1 1 0 0 0
KIF16B	1 0 1 0 0 0 1 0 0 0 0 0 0 1 0
USP37	0 0 1 1 1 0 0 0 0 0 0 0 0 1 0
FBXO30	0 1 1 0 0 0 0 0 1 0 0 1 0 0
USP6NL	0 0 0 1 0 0 1 0 0 1 1 0 0 0
VANGL1	1 0 0 0 0 1 1 0 0 0 0 0 1 0
KIAA2026	0 0 1 0 0 0 1 0 1 0 0 1 0 0
AHCTF1P1	1 0 0 0 0 0 0 1 1 0 0 0 0 1
KCNMA1	0 1 0 0 0 1 0 1 0 0 0 1 0 0
TGS1	0 1 1 1 0 0 0 0 0 0 0 0 1 0
NOM1	1 0 1 1 0 0 0 0 0 0 0 0 0 1
SLC12A7	0 0 0 1 0 0 1 0 0 0 1 0 1 0
UBR2	1 0 0 1 0 0 0 1 0 0 0 1 1 0
TANGO6	1 0 0 0 1 0 0 0 1 0 0 1 0 0
WDR60	1 0 0 0 1 0 1 0 0 0 0 0 0 1
FOXJ2	0 0 0 0 0 1 0 1 1 1 0 0 0 0
RMND5A	0 1 1 0 0 0 1 0 0 1 0 0 0 0
KDM1B	0 1 0 0 0 0 0 1 1 0 0 0 1 0
CSPG4	0 1 0 1 1 0 0 0 0 0 0 0 1 0
INPP4A	0 0 0 1 1 0 1 0 0 0 0 0 1 0
ATP9A	1 0 0 0 0 0 1 0 0 1 1 0 0 0
DZIP3	0 0 1 0 0 0 0 0 1 1 1 0 0 0
CAPN15	1 0 0 0 1 1 0 0 0 0 0 0 1 0
ITGA11	0 1 0 0 0 0 0 0 1 0 1 0 0 1
CYFIP2	1 0 0 0 1 0 1 0 0 0 0 0 1 0
SUSD5	0 1 0 0 1 0 0 0 0 0 0 1 0 1
ZNF654	1 1 0 0 1 0 0 0 0 0 0 0 0 1
ZBTB41	0 0 0 0 1 1 1 0 0 0 0 0 1 0
RWDD2A	0 0 0 0 1 0 0 1 0 0 1 0 0 1
NCOA3	0 1 0 0 0 0 0 1 0 1 0 0 0 1
MED23	0 0 0 0 1 0 0 0 1 0 1 1 0 0
CSTF2T	0 0 0 0 1 0 0 0 1 0 1 0 0 1
SIPA1L2	0 0 1 0 0 0 0 1 0 0 0 1 0 1
CYP2U1	0 0 0 1 1 0 1 0 0 0 0 1 0 0
HECTD2	1 0 1 0 0 1 0 0 0 0 0 0 0 1
SMAD9	0 0 1 0 1 0 1 0 0 0 0 0 1 0
HIP1	0 1 0 1 0 0 1 0 0 0 0 0 0 1

blank011	1 0 1 0 1 0 0 0 0 0 1 0 0 0
NEK1	0 0 0 0 1 1 0 1 0 1 0 0 0 0
CNTN3	0 1 0 0 0 0 0 1 1 0 0 1 0 0
RFX7	0 0 1 0 0 1 0 0 0 0 1 0 1 0
NUAK1	1 0 0 1 0 1 0 0 0 1 0 0 0 0
NAPEPLD	1 0 0 0 0 1 0 1 0 1 0 0 0 0
POLR3B	1 0 0 1 0 0 0 0 1 0 0 1 0 0
B3GALTL	0 1 0 0 0 0 0 1 0 0 1 0 1 0
HAS2	1 0 0 0 1 1 0 0 0 1 0 0 0 0
RAP1GAP2	1 1 0 0 0 0 0 1 0 0 0 0 1 0
CIT	0 0 1 0 0 1 0 1 0 0 0 0 0 1
SOX11	0 0 1 0 1 1 0 0 0 1 0 0 0 0
MTHFR	0 1 0 0 1 0 0 1 0 0 0 0 1 0
DUSP16	0 0 0 0 0 0 1 1 1 1 0 0 0 0
SLC9A5	0 1 0 0 0 0 1 0 0 0 1 1 0 0
HEATR5B	0 0 1 1 0 0 1 0 0 0 0 0 1 0
VCPIP1	0 0 0 0 1 1 0 0 0 0 1 0 0 1
JARID2	0 1 1 0 0 0 0 0 0 0 1 0 1 0
MEX3A	0 0 0 1 0 0 0 0 1 1 1 0 0 0
GPR155	0 0 0 1 0 1 0 1 0 0 0 0 0 1
ZNF526	0 1 0 0 0 0 0 0 1 0 0 1 0 1
TAB3	1 0 0 0 0 1 0 0 1 0 0 0 0 1
FLRT3	1 0 0 0 1 0 0 0 0 1 1 0 0 0
ZNF436	0 1 0 0 1 0 0 0 1 1 0 0 0 0
WDR7	1 0 0 1 0 0 1 0 0 1 0 0 0 0
WDR81	1 1 0 0 0 0 1 0 0 0 0 1 0 0
SLC41A2	1 1 0 0 0 0 0 0 0 0 1 0 1 0
WDR37	0 1 0 0 1 0 0 1 0 1 0 0 0 0
NLRX1	0 0 1 0 1 0 0 0 1 0 0 0 0 1
BRIP1	0 1 0 0 0 1 1 0 0 0 0 0 0 1
PHKA2	0 0 0 1 0 1 0 0 0 1 1 0 0 0
POLR1A	0 0 1 0 0 0 1 1 0 0 0 0 1 0
ZNF445	0 0 1 1 0 0 1 0 0 0 0 0 0 1
UBE2O	0 0 0 0 0 1 0 0 1 0 1 0 0 1
PTGFR	0 0 0 1 1 0 0 0 1 0 0 0 0 1
GMEB2	0 0 0 1 0 0 0 1 1 0 0 0 0 1
CCNF	0 0 1 1 1 0 0 0 0 1 0 0 0 0
OTUD3	0 1 0 0 0 0 1 0 1 0 0 0 0 1
CACNA1C	0 1 1 1 0 0 0 0 0 0 0 0 0 1
BRPF1	0 0 1 0 0 0 0 1 1 0 0 1 0 0
ARID2	0 0 1 0 1 0 0 1 0 0 0 0 0 1
BLM	0 1 0 0 1 1 0 0 0 0 0 0 0 1
PDE10A	0 0 0 0 1 1 0 0 0 0 1 0 1 0
AGL	0 1 1 0 1 0 0 0 0 0 0 0 1 0
TRPS1	0 0 0 0 1 1 0 0 1 0 0 0 0 1
RASAL2	0 1 0 0 0 1 1 1 0 0 0 0 0 0
ENDOD1	0 1 0 0 1 1 1 0 0 0 0 0 0 0
GREM2	0 1 1 0 0 1 1 0 0 0 0 0 0 0
ANKRD26	0 1 0 1 1 0 0 0 1 0 0 0 0 0
NHLRC2	1 1 0 0 1 0 1 0 0 0 0 0 0 0
ZBTB5	0 0 0 1 0 0 1 1 1 0 0 0 0 0
GPRIN1	0 1 0 0 1 1 0 1 0 0 0 0 0 0
TCEANC2	0 0 1 0 0 1 1 0 0 0 1 0 0 0
notarget004	0 0 0 0 1 1 0 0 1 0 1 0 0 0
MAGI3	1 0 1 0 0 0 0 1 0 0 1 0 0 0
SHROOM2	0 0 1 0 1 1 0 0 1 0 0 0 0 0
FMNL3	0 1 1 1 0 0 0 0 1 0 0 0 0 0
TRIM62	1 0 0 0 0 1 0 1 1 0 0 0 0 0
TNFAIP3	0 1 1 0 1 0 1 0 0 0 0 0 0 0
HIP1R	0 1 0 1 0 0 1 0 0 0 1 0 0 0
DCLRE1A	1 1 0 0 0 0 1 0 0 0 1 0 0 0
DNA2	0 0 0 0 1 1 1 0 0 0 1 0 0 0
VWA8	0 1 0 0 1 0 0 1 0 0 1 0 0 0
PHF16	1 0 1 0 0 1 0 0 0 0 1 0 0 0
TARBP1	0 0 1 0 1 1 0 0 0 0 1 0 0 0
RNASEL	1 0 1 1 0 1 0 0 0 0 0 0 0 0
DHX34	0 1 0 0 0 1 1 0 0 0 1 0 0 0
GLTSCR1L	1 0 0 1 1 0 0 0 1 0 0 0 0 0
UBXN2B	0 0 1 1 1 1 0 0 0 0 0 0 0 0
PAPOLG	0 1 1 0 1 1 0 0 0 0 0 0 0 0
WDR35	1 0 0 0 1 0 1 1 0 0 0 0 0 0
WDR62	1 0 0 1 0 1 1 0 0 0 0 0 0 0
blank009	0 1 1 1 0 0 0 0 0 0 1 0 0 0
BRI3BP	0 0 1 0 1 1 1 0 0 0 0 0 0 0
IQGAP3	0 0 1 1 0 0 0 1 0 0 1 0 0 0
PIK3CD	0 0 1 1 1 0 0 0 0 1 0 0 0 0
ERMAP	0 1 0 0 0 0 1 1 0 0 1 0 0 0
PTPLAD2	0 0 1 1 0 0 1 1 0 0 0 0 0 0
CUL9	0 0 1 1 1 0 0 0 0 0 1 0 0 0
ANKS1A	0 1 0 1 0 0 0 0 1 0 1 0 0 0
TRERF1	0 0 0 0 1 0 1 0 1 0 1 0 0 0
URB2	1 0 0 1 0 1 0 0 0 0 1 0 0 0
TMED8	1 0 0 0 0 1 0 1 0 0 1 0 0 0
ARHGEF10L	1 1 0 0 0 0 0 1 1 0 0 0 0 0
GPRC5B	0 0 0 0 0 0 1 1 1 0 1 0 0 0
HSPA12A	0 0 1 0 0 0 0 1 0 1 0 1 0 0

FAM198B	0 0 1 0 0 0 0 1 0 0 0 0 0 1 1
SENP5	0 0 0 0 0 0 0 0 0 1 1 1 0 0 1
CLOCK	0 0 0 0 1 1 0 0 0 0 0 0 1 0 1
ARID4A	0 0 1 1 0 0 0 0 0 0 0 1 1 0
PIK3CA	0 0 0 0 0 1 0 0 0 0 1 0 1 1
blank001	0 0 0 0 0 1 0 0 0 1 1 1 0 0
IPO9	0 0 0 1 0 0 0 0 0 0 1 0 1 1
WDR48	0 0 1 0 0 0 0 1 0 1 0 0 0 1
UNC119B	0 0 1 0 1 0 0 0 0 1 0 1 0 0
MAP3K5	0 0 0 0 0 1 0 0 0 0 1 1 1 0
LPGAT1	1 0 0 1 0 0 0 0 0 0 1 1 0
MED14	0 0 0 1 0 0 0 0 0 1 1 0 1 0
DCBLD1	0 0 0 0 1 1 0 0 0 1 0 0 1 0
TP53BP1	1 0 0 0 0 0 0 1 0 0 0 1 1 0
DENND6A	0 0 0 0 0 0 1 0 1 1 0 0 1 0
TBCD	0 0 0 0 0 0 1 0 0 0 1 0 1 1
EXOC6B	0 0 0 0 0 0 1 0 0 0 1 1 1 0
HERC2	0 0 0 0 1 0 0 1 0 1 0 0 0 1
AQR	0 0 0 1 1 0 0 0 0 0 0 1 0 1
SLC1A4	0 0 0 0 0 0 1 1 0 1 0 0 0 1
CENPJ	0 0 0 0 0 0 0 1 0 0 1 0 1 1
TJP2	0 0 0 0 0 1 0 0 1 1 0 0 0 1
SCAP	0 0 0 1 0 0 0 0 1 1 0 0 0 1
EXO1	0 0 0 0 0 0 0 0 1 0 1 1 1 0
REST	0 1 1 0 0 0 0 0 0 0 0 1 1 0
BICD2	1 0 0 0 0 0 0 1 0 1 0 1 0 0
ARIH1	0 1 0 0 1 0 0 0 0 0 0 1 1 0
ZCCHC11	0 0 0 0 1 1 0 0 0 0 0 0 1 1
RFX5	0 0 0 0 0 0 0 1 1 1 0 0 1 0
NUFIP2	0 1 1 0 0 0 0 0 0 1 0 0 0 1
TCEB3	0 1 0 1 0 0 0 0 0 0 0 0 1 1
SACS	0 0 0 0 0 0 0 1 0 1 1 0 1 0
TRO	0 0 0 0 1 0 1 0 0 1 0 0 0 1
ERI1	0 1 0 0 0 0 0 1 0 0 0 0 1 1
GTF3C2	0 0 0 0 0 0 0 0 1 1 1 0 1 0
PDE7B	0 0 0 0 1 0 1 0 0 0 0 0 1 1
GSE1	0 0 0 0 0 1 0 0 1 0 0 0 1 1
DIS3L	0 0 1 0 0 0 1 0 0 1 0 0 1 0
SESN2	0 1 0 0 0 0 0 0 1 0 0 0 1 1
WDR59	0 0 0 0 0 0 1 0 1 0 0 1 1 0
RBM26	1 0 0 0 0 0 0 0 0 0 1 1 0 1
ATM	0 0 0 0 0 1 0 1 0 0 0 1 0 1
H6PD	1 0 0 1 0 0 0 0 0 1 0 0 1 0
MPP5	0 0 0 0 0 1 1 0 0 0 0 1 1 0
ZCCHC6	0 1 0 0 0 0 0 0 1 1 0 0 0 1
OSBPL5	0 0 0 0 0 0 1 0 1 1 0 0 0 1
PDS5B	0 0 0 0 1 0 0 1 0 0 0 1 1 0
NUP133	1 0 1 0 0 0 0 0 0 1 0 0 0 1
SMIM13	1 0 1 0 0 0 0 0 0 0 1 0 0 1 0
PHF20	1 0 0 0 0 0 0 1 0 0 0 1 0 1
ADAM17	0 0 1 0 0 0 1 0 0 0 0 0 1 1
ST3GAL2	0 0 0 0 0 1 0 0 1 1 0 1 0 0
KNTC1	0 0 1 0 0 0 0 0 1 0 0 1 1 0
HIPK1	0 1 0 1 0 0 0 0 0 1 0 1 0 0
DHX16	1 0 0 0 0 0 1 0 0 1 0 0 0 1
PDGFD	0 0 0 1 0 0 0 1 0 1 0 0 0 1
KCTD2	0 1 0 0 1 0 0 0 0 1 0 1 0 0
ASAP2	1 0 0 0 0 0 0 1 0 0 0 0 1 1
TTL	0 0 0 0 0 0 1 0 1 0 0 1 0 1
TOPORS	0 0 0 1 0 0 0 1 0 1 0 1 0 0
KDM3B	0 1 0 0 0 1 0 0 0 1 0 1 0 0
MYCBP2	0 0 0 0 0 0 1 0 0 1 1 1 0 0
NSMAF	0 1 0 1 0 0 0 0 0 1 0 0 0 1
MPDZ	0 0 1 0 1 0 0 0 0 0 0 1 0 1
MMGT1	1 0 0 0 0 0 0 0 1 1 0 0 1 0
MED1	1 0 0 0 0 0 0 1 0 1 0 0 0 1
UBR1	0 0 1 0 0 0 1 0 0 0 0 1 1 0
FOXK2	0 0 0 0 0 1 0 1 0 0 0 1 1 0
LRIG3	1 0 0 0 0 0 0 0 1 0 0 0 1 1
AGO4	0 1 0 1 0 0 0 0 0 1 0 0 1 0
CLSPN	0 0 0 1 0 0 0 1 0 0 0 1 0 1
PDPK1	0 0 1 0 0 0 0 0 0 1 1 0 0 1
DNAJC18	0 0 0 0 1 0 0 0 1 0 0 0 1 1
FAM63B	0 0 0 0 1 1 0 0 0 1 0 0 0 1
ADNP	0 0 0 0 1 0 0 1 0 0 0 0 1 1
APC	0 0 1 0 0 0 0 1 0 1 0 1 0 0
RHOBTB1	1 0 0 0 0 0 1 0 0 0 0 1 0 1
MORC3	1 0 0 0 1 0 0 0 0 0 0 1 0 1
PLXNA3	0 0 0 1 0 1 0 0 0 0 0 0 1 1
NES	1 0 0 0 1 0 0 0 0 0 0 0 1 1
ARID1A	0 0 0 0 1 0 0 1 0 1 0 1 0 0
HCFC1	0 1 0 0 0 0 1 0 0 0 0 0 1 1
C14orf132	0 0 0 0 0 1 1 0 0 1 0 0 1 0
YEATS2	1 1 0 0 0 0 0 0 0 0 0 1 1 0
LDLR	0 1 0 0 0 1 0 0 0 1 0 0 0 1
EHBP1L1	0 0 0 0 0 1 0 1 0 1 0 1 0 0
RNF213	0 0 0 1 0 0 1 0 0 0 0 0 1 1
NUP205	0 0 0 0 0 0 1 1 0 0 0 1 0 1
ITPRIPL2	0 0 1 1 0 0 0 0 0 0 0 0 1 1
SIRT1	1 1 0 0 0 0 0 0 0 1 0 0 0 1

MCM2	1 0 0 0 1 1 0 0 0 0 0 0 0 0 1
UPF2	0 0 0 0 1 1 0 1 0 0 0 0 0 0 1
LPHN3	0 1 0 0 0 1 0 1 0 0 0 0 1 0
KANK1	0 0 0 0 1 1 0 0 1 0 0 1 0 0
HERC1	0 1 0 0 0 1 1 0 0 1 0 0 0 0
EPHA2	0 0 1 0 0 1 0 0 0 0 0 1 1 0 0
TBC1D16	0 0 0 0 0 0 0 0 1 1 0 1 1 0 0
MIB1	0 1 0 1 1 0 0 0 0 0 0 1 0 0
SNX30	0 0 0 0 0 0 1 1 1 0 0 0 1 0
MAFG	0 0 1 0 1 0 0 1 0 0 0 1 0 0
RASGRF2	1 0 0 0 0 1 1 0 0 0 0 1 0 0
MAP7D3	0 1 0 0 1 0 0 1 0 0 0 1 0 0
SMG1	0 1 0 0 1 0 0 1 0 0 0 0 0 1
USP19	0 0 0 0 0 1 1 0 1 0 0 1 0 0
STXBP1	1 0 0 1 0 0 0 1 0 0 0 1 0 0
PRDM4	1 0 1 0 0 0 0 1 0 0 0 0 0 1
URGCP	0 0 1 0 0 1 0 1 0 0 0 0 1 0
FBXO42	0 0 1 0 0 0 0 0 1 0 0 1 0 1
TADA2B	1 0 0 0 0 0 1 0 0 0 1 0 1 0
PHF13	0 0 0 0 0 1 1 0 1 1 0 0 0 0
TUBGCP5	0 0 0 1 1 0 1 0 0 0 0 0 0 1
ANKFY1	1 0 1 0 1 0 0 0 0 0 0 0 0 1
CTCF	1 0 0 1 0 0 0 1 0 1 0 1 0 0 0
CUL7	0 1 0 0 0 0 1 0 1 0 0 1 0 0
NUP155	0 0 1 0 0 1 1 0 0 0 0 0 1 0
C11orf30	0 1 0 0 0 1 0 0 1 0 0 0 1 0
TCOF1	0 0 0 1 0 0 1 0 1 0 0 1 0 0
LARS2	0 1 0 1 0 0 0 0 0 1 1 0 0 0
RIPK1	1 1 1 0 0 0 0 0 0 0 1 0 0 0
blank004	0 0 0 0 0 1 1 0 0 0 1 0 1 0
CEP250	1 0 0 0 0 0 1 0 1 0 0 0 0 1
ZNF317	1 0 0 0 0 1 0 0 1 0 0 1 0 0
LEMD3	0 0 1 0 0 0 0 0 0 1 0 1 0 1 0
AGO1	0 0 0 0 1 0 1 0 0 0 1 1 0 0
TIMELESS	0 1 1 0 0 0 0 1 0 0 0 1 0 0
MORC2	1 0 0 1 1 0 0 0 0 1 0 0 0 0
LIG4	0 0 1 1 0 0 0 0 1 0 0 1 0 0
PIKFYVE	1 0 0 0 0 1 0 0 1 1 0 0 0 0
MLLT11	1 0 0 0 0 0 0 0 1 0 1 0 1 0
SETD2	1 0 1 0 0 1 0 0 0 0 0 0 1 0
YTHDC2	0 0 1 0 1 0 0 0 0 1 1 0 0 0
TECPR2	1 0 0 0 1 0 0 0 1 0 0 1 0 0
HEXIM1	0 0 1 0 0 0 0 1 0 0 1 1 0 0
USP24	0 0 0 0 1 1 0 0 1 0 0 0 1 0
AGPS	1 0 0 0 1 0 0 0 0 0 1 1 0 0
EEF2K	0 1 0 0 1 0 0 0 0 0 0 1 0 1 0
WWC3	1 0 1 0 0 0 0 0 1 0 0 0 0 1
ASPM	0 0 0 0 1 0 0 1 0 1 1 0 0 0
RC3H1	0 1 0 0 0 0 0 1 1 1 0 0 0 0
KIF13A	0 1 0 0 0 0 0 1 0 0 1 1 0 0
NPR2	0 0 1 0 0 1 0 0 1 0 0 1 0 0
CEP192	0 0 0 0 0 0 1 0 1 1 0 0 0 1
NOL9	0 1 0 1 0 0 0 1 0 1 0 0 0 0
SFXN5	1 1 0 0 0 1 0 0 0 1 0 0 0 0
ARHGEF9	0 0 1 0 0 0 1 1 0 0 0 0 0 1
SPRY4	0 1 0 0 0 1 0 0 0 0 1 1 0 0
ULK2	0 1 0 1 0 1 0 0 0 0 0 1 0 0
MTMR4	0 0 0 0 1 0 1 0 1 1 0 0 1 0
CABIN1	0 0 1 1 0 0 0 1 0 0 0 0 1 0
KREMEN1	0 0 1 0 0 1 1 0 0 0 0 0 0 1
WDHD1	0 0 0 0 1 1 1 0 0 0 0 1 0 0
TSPYL5	0 1 0 1 0 0 0 0 1 0 0 1 0
USP38	1 0 0 0 0 0 1 0 0 0 1 0 0 1
NRDE2	1 1 1 0 0 0 0 0 0 0 0 0 0 1
PPRC1	0 0 0 0 1 0 1 0 1 1 0 0 0 1
notarget002	1 0 1 0 0 0 0 0 0 1 1 0 0 0
RCOR1	1 1 0 0 0 1 0 0 0 0 0 0 0 1
ETV3	0 0 1 0 0 1 0 1 0 0 1 0 0 0
TOX	1 0 1 0 0 0 0 0 0 0 1 0 1 0
INTS1	0 1 0 0 1 1 0 0 0 0 0 1 0 0
PARP14	0 0 0 0 0 1 1 0 0 0 1 0 0 1
VAT1L	0 1 1 1 0 0 0 0 0 1 0 0 0 0
MECOM	0 1 0 1 0 0 1 0 0 1 0 0 0 0
IFT122	0 0 0 1 1 1 0 0 0 0 0 0 0 1
SOCS4	0 1 0 0 1 0 1 0 0 0 0 1 0 0
ZNF512B	1 0 0 0 0 0 1 0 1 1 0 0 0 0
C12orf4	1 1 0 0 1 0 0 0 0 1 0 0 0 0
PAPD7	0 0 1 0 0 0 1 1 0 1 0 0 0 0
SLC35B4	0 0 0 1 0 0 1 1 0 1 0 0 0 0
SIN3B	0 0 1 0 1 0 0 1 0 1 0 0 0 0
TMEM57	0 1 0 0 1 0 1 0 0 0 0 0 0 1
TBC1D8B	0 0 1 0 0 1 0 0 0 1 1 0 0 0
LRRC58	0 0 0 1 0 1 1 0 0 1 0 0 0 0
ZNF45	0 0 0 0 1 1 0 1 0 0 0 0 1 0
FMNL2	0 0 0 0 1 0 1 0 0 1 1 0 0 0
RAPGEF6	0 0 1 0 0 0 1 0 0 0 1 0 0 1
FOXK1	1 0 0 0 0 1 0 0 0 1 0 1 0 1
MYBBP1A	1 1 1 0 0 0 0 0 0 0 0 0 1 0
PCMTD2	0 0 0 1 0 1 0 0 0 0 1 1 0 0
ITSN2	1 0 1 0 0 0 1 0 0 0 0 0 0 1

CDC25A	1 1 0 1 0 0 0 0 0 0 0 0 0 1
SERTAD2	0 0 0 1 1 1 0 0 0 0 0 0 1 0
NPAT	0 1 0 1 0 0 1 0 0 0 0 0 1 0
DPYD	0 0 0 1 1 0 0 0 0 1 1 0 0 0
KIAA0753	1 0 0 1 0 0 0 0 1 0 0 1 1 0
VPS13C	0 0 0 1 0 1 0 1 0 0 0 1 0 0
ZNF592	0 0 1 1 1 0 0 0 0 0 0 0 0 1
CLUH	1 1 0 0 0 0 1 0 0 1 0 0 0 0
WDR55	0 1 0 0 1 0 0 0 0 1 1 0 0 0
ZNF71	1 1 0 0 0 0 0 1 0 0 0 1 0 0
CRYBG3	1 1 0 0 0 0 0 0 0 0 0 1 0 1
FCHSD2	1 0 0 1 0 0 1 0 0 0 0 0 0 1 0
EFNB2	1 0 1 0 0 0 0 1 0 0 0 0 1 0
TMEM260	0 0 1 1 0 0 0 1 0 0 0 1 0 0
TTF2	1 0 0 1 0 1 0 0 0 0 0 1 0 0
NCKAP5L	1 0 0 1 0 0 1 0 0 0 0 1 0 0
notarget003	0 0 1 0 1 1 1 0 0 0 0 1 0 0
ZNF652	1 0 0 0 0 0 0 1 0 1 1 0 0 0
C10orf118	1 0 0 0 1 0 0 1 0 0 0 0 0 1
DGKQ	0 0 0 0 0 1 1 1 0 1 0 0 0 0
MKL2	1 0 0 0 0 0 0 1 0 0 1 1 0 0
CD109	0 0 0 0 1 0 0 1 0 0 1 0 1 0
DIO2	0 0 0 1 0 0 1 1 0 0 0 0 1 0
KIF15	0 0 1 1 0 1 0 0 0 0 0 1 0 0
GTF3C4	0 0 0 0 0 0 1 1 0 0 1 0 0 1
ZNF687	0 0 1 0 1 0 0 0 0 1 0 0 1 0 0
IQSEC1	0 0 0 1 1 0 0 1 0 0 0 1 0 0
TRIM14	0 0 0 1 1 0 1 0 0 1 0 0 0 0
PIK3R1	1 1 0 0 0 0 0 0 0 1 1 0 0 0
RHOBTB2	0 1 1 0 0 1 0 0 0 0 0 1 0 0
ZNF776	0 0 1 0 0 1 0 0 1 1 0 0 0 0
CTPS2	0 0 1 0 1 0 1 0 0 1 0 0 0 0
PRMT10	0 1 1 0 0 0 0 1 0 1 0 0 0 0
LTBP2	0 1 0 0 1 1 0 0 0 0 0 0 1 0
SLC9A6	1 0 1 0 1 0 0 0 0 0 0 0 1 0
TBC1D13	0 0 0 1 0 0 0 0 1 0 0 1 1 0
QTRTD1	0 0 0 0 1 0 1 1 0 0 0 0 0 1
TRIM35	0 1 1 0 1 0 0 0 0 0 0 0 0 1
SIN3A	0 1 0 0 0 1 0 0 0 1 1 0 0 0
ZBTB6	0 0 0 0 0 1 0 1 1 0 0 0 1 0
QSER1	0 0 1 0 0 0 1 0 1 0 0 0 0 1
FEM1C	0 1 0 0 0 0 0 0 1 1 1 0 0 0
PLEKHG3	0 0 0 0 1 1 0 0 0 0 1 1 0 0
FBXO45	0 0 1 0 0 1 0 0 0 0 1 0 0 1
GPR180	1 1 0 1 0 0 0 0 0 0 0 0 1 0
SPICE1	0 0 0 1 0 0 0 0 0 1 0 1 0 1
ANKRD32	0 0 0 1 0 0 1 1 0 0 0 1 0 0
SNIP1	1 0 0 1 0 0 0 1 0 0 0 0 0 1
NSD1	0 0 1 1 0 0 0 0 0 1 1 0 0 0
GIGYF1	0 0 0 1 0 0 0 1 0 0 1 0 0 1
DTX3L	0 1 0 1 0 0 0 0 0 0 0 1 1 0 0
CCSAP	0 0 1 1 0 1 0 0 0 0 0 0 0 1
ADCY7	1 0 0 1 1 0 0 0 0 0 0 0 0 1
blank007	0 0 1 0 1 0 0 0 0 0 0 1 0 1 0
MSH3	1 0 1 0 0 0 1 0 0 1 0 0 0 0
PHF2	0 1 1 0 0 0 0 0 1 0 0 0 0 1
VTI1A	0 0 1 0 0 0 0 0 1 1 1 0 0 0 0
AP1G2	1 1 0 0 0 0 0 0 0 0 1 1 0 0
ANPEP	0 0 1 0 1 0 0 0 0 0 0 1 1 0 0
TMEM104	0 0 1 0 0 0 0 0 1 0 0 1 0 1 0
DOCK5	1 0 0 0 0 0 1 0 1 0 0 0 1 0 0
MAP3K6	0 0 0 1 0 0 0 1 0 1 1 0 0 0
TTI1	0 0 1 1 0 0 0 0 0 1 0 0 0 1 0
ASAP3	0 0 1 0 0 0 1 1 0 0 0 1 0 0
TRRAP	0 1 0 1 0 0 0 0 1 0 0 0 0 0 1
CEP41	0 0 0 0 1 0 1 0 1 0 0 1 0 0
KIF18B	0 1 0 0 0 0 0 1 0 1 1 0 0 0
LMBRD2	0 0 0 0 0 0 1 1 1 0 0 0 0 1
ZDHHC21	0 0 0 1 0 1 1 0 0 0 0 0 1 0
CNTLN	0 1 1 0 0 0 1 0 0 0 0 0 1 0
ERCC6L	0 0 1 1 1 0 0 0 0 0 0 1 0 0
VPS13D	0 0 0 1 0 1 0 0 1 0 0 1 0 0
ITPR1	1 0 0 0 0 1 1 0 0 1 0 0 0 0
POLR3A	0 0 0 0 0 1 0 1 1 0 0 1 0 0
FAM214A	0 1 1 0 0 0 0 1 0 0 0 1 0 0
FRY	0 1 0 1 0 1 0 0 0 0 0 0 1 0
ZFAT	0 0 0 1 1 0 0 1 0 1 0 0 0 0
FAN1	0 0 1 1 0 0 1 0 0 0 0 1 0 0
SOWAHC	0 0 1 0 1 0 1 0 0 0 0 0 0 1
RBM12B	0 0 0 0 1 1 1 0 0 1 0 0 0 0
MCM9	1 1 0 0 0 0 0 0 1 1 0 0 0 0
KIF14	0 0 0 0 1 0 1 0 1 1 0 0 0 0
GPD1L	1 0 0 1 0 0 0 0 0 0 0 1 0 1
ZMYM3	0 0 0 1 1 0 0 0 1 1 0 0 0 0
SMCR8	0 1 0 0 1 1 0 0 0 1 0 0 0 0
GM2A	1 0 1 0 0 0 0 0 1 0 0 0 1 0
TTPAL	0 1 0 1 0 0 0 0 1 0 0 0 1 0
OFD1	1 1 0 0 1 0 0 0 0 0 0 0 1 0
LNPEP	0 1 1 0 0 0 1 0 0 0 0 0 0 1
TSTD2	0 0 1 0 1 1 0 0 0 0 0 0 0 1

UHRF1BP1	0 1 1 0 0 0 1 1 0 0 0 0 0 0
SYTL4	0 1 0 0 1 0 1 0 0 0 1 0 0 0
ESCO1	0 1 0 0 1 0 0 1 1 0 0 0 0 0
TULP4	0 1 0 0 0 1 0 1 1 0 0 0 0 0
MAPKBP1	1 0 0 0 1 1 1 0 0 0 0 0 0 0
GPAM	0 0 0 1 1 0 1 0 1 0 0 0 0 0
ATG14	0 0 0 0 1 1 0 1 1 0 0 0 0 0
DMXL2	0 1 0 1 0 0 0 1 0 0 1 0 0 0
HIVEP1	0 0 1 0 1 0 0 0 1 0 1 0 0 0
SNPH	1 0 1 0 1 0 0 0 0 1 0 0 0 0
SSH2	0 0 1 0 1 0 1 0 0 0 1 0 0 0
ZNF202	0 0 1 0 1 0 1 0 1 0 0 0 0 0
SALL2	0 1 0 1 1 0 1 0 0 0 0 0 0 0
OSBPL6	1 0 0 1 1 0 0 0 0 0 1 0 0 0
ERN1	0 1 0 0 1 0 0 0 1 0 1 0 0 0
ARL10	1 0 0 1 0 0 1 0 0 0 1 0 0 0
RP11-350F4.2	0 1 0 1 1 0 0 1 0 0 0 0 0 0
MAP3K14	1 0 0 1 0 0 0 0 1 0 1 0 0 0
FKBP5	0 0 0 0 1 0 0 1 1 0 1 0 0 0
HELZ2	0 1 1 0 0 1 0 0 0 0 0 0 0 0
PTPRU	1 1 1 0 0 0 1 0 0 0 0 0 0 0
RALGAPA1P	0 0 0 1 0 1 0 1 0 0 1 0 0 0
FITM2	1 0 0 1 0 0 1 0 1 0 0 0 0 0
PER3	0 0 0 1 1 1 1 0 0 0 0 0 0 0
ARL4C	0 0 0 1 0 0 1 1 0 0 1 0 0 0
DDX60L	0 1 0 0 0 0 0 1 0 1 0 1 0 0 0
ZNF41	0 0 1 1 0 0 0 0 1 0 1 0 0 0
KLHL15	0 1 1 0 1 0 0 0 1 0 0 0 0 0
PHLPP2	1 0 0 0 1 0 1 0 1 0 0 0 0 0
ZNF142	0 0 1 0 0 1 1 0 1 0 0 0 0 0
KLF11	0 1 0 1 0 1 1 0 0 0 0 0 0 0
MBLAC2	0 0 0 1 0 1 1 1 0 0 0 0 0 0
NIPAL2	1 1 1 0 0 1 0 0 0 0 0 0 0 0
LRP4	0 0 1 0 1 0 1 1 0 0 0 0 0 0
N6AMT1	1 0 1 0 0 1 1 0 0 0 0 0 0 0
SLC12A6	0 0 1 0 0 1 1 1 0 0 0 0 0 0
SLC4A8	0 1 1 0 1 0 0 1 0 0 0 0 0 0
PINK1-AS	0 0 1 0 0 0 0 1 1 0 1 0 0 0
VDR	1 1 0 0 1 0 0 0 0 0 1 0 0 0
THADA	1 1 0 1 0 0 0 0 0 0 1 0 0 0
URB1	0 0 0 0 0 1 1 1 0 0 1 0 0 0
KLHL29	0 0 1 1 0 1 0 0 0 0 1 0 0 0
RP11-566E18.3	0 0 0 1 0 0 0 1 1 0 1 0 0 0
POLQ	1 1 1 1 0 0 0 0 0 0 0 0 0 0
HMCN1	0 0 1 1 0 0 1 0 1 0 0 0 0 0
ZNF230	0 1 1 0 0 0 0 0 1 0 1 0 0 0
YOD1	1 1 0 0 0 0 0 1 0 0 1 0 0 0
DGKI	1 0 0 1 1 0 0 1 0 0 0 0 0 0
CACNA1A	0 0 1 0 0 1 0 1 1 0 0 0 0 0
ITGA7	1 0 0 0 0 0 1 1 0 0 1 0 0 0
ZNF175	0 1 0 1 0 1 0 0 1 0 0 0 0 0
LINC00641	1 0 0 0 0 0 0 1 0 1 0 1 0 0
ZNF646	0 1 1 0 0 1 0 1 0 0 0 0 0 0
FAM131B	0 0 0 1 0 1 1 0 0 0 1 0 0 0
MOK	1 1 0 0 0 0 1 1 0 0 0 0 0 0
ZNF354C	0 0 0 1 1 1 0 0 0 0 1 0 0 0
NCR3LG1	0 0 0 1 0 1 0 0 1 0 1 0 0 0
TBC1D22B	1 0 0 0 1 0 0 1 0 0 1 0 0 0
KIAA2018	0 1 0 0 1 0 1 0 1 0 0 0 0 0
SEMA4F	1 1 1 0 0 0 0 1 0 0 0 0 0 0
RALGAPA2	0 0 1 1 1 0 0 1 0 0 0 0 0 0
CDKN2B	1 0 0 1 1 0 1 0 0 0 0 0 0 0
C10orf54	0 1 1 0 0 0 0 1 1 0 0 0 0 0
TMPPE	1 0 1 0 0 0 1 0 1 0 0 0 0 0
CDKL5	1 1 0 0 1 0 0 1 0 0 0 0 0 0
AHNAK2	1 0 1 0 1 0 1 0 0 0 0 0 0 0
SLC2A12	0 1 0 0 0 1 0 1 0 0 1 0 0 0
ZKSCAN2	1 0 0 0 0 1 1 1 0 0 0 0 0 0
BAIAP2L1	1 0 1 0 0 0 0 1 1 0 0 0 0 0
C10orf12	0 1 0 0 1 0 1 1 0 0 0 0 0 0
ITGB8	0 0 1 1 0 0 0 1 1 0 0 0 0 0
DIEXF	1 1 0 0 0 1 1 0 0 0 0 0 0 0
NINL	1 1 1 0 0 0 0 0 0 0 1 0 0 0
WDR91	0 1 0 1 0 0 1 0 1 1 0 0 0 0
DOCK11	0 0 0 1 0 1 1 0 1 0 0 0 0 0
IGHMBP2	1 1 0 0 0 0 0 0 1 0 1 0 0 0
ATAD2B	0 0 1 1 1 0 1 0 0 0 0 0 0 0
blank010	1 0 1 0 0 0 1 0 0 1 0 0 0 0
VSIG10	0 1 1 0 0 0 0 1 0 0 1 0 0 0
FZD4	1 0 0 0 0 0 1 1 1 0 0 0 0 0
LOXL4	1 0 0 0 1 1 0 0 0 0 1 0 0 0
ZNF304	1 0 1 0 0 1 0 0 1 0 0 0 0 0
CDAN1	0 1 0 0 1 1 0 0 1 0 0 0 0 0
ARSD	1 0 0 0 1 1 0 0 1 0 0 0 0 0
LPPR4	0 1 0 0 1 1 0 0 0 0 1 0 0 0
SH3BGRL2	1 1 0 0 0 1 0 0 0 0 1 0 0 0
LRRK1	1 0 1 1 0 0 0 1 0 0 0 0 0 0
DDX26B	1 1 0 1 1 0 0 0 0 0 0 0 0 0
RP11-18F14.2	1 0 1 0 1 1 0 0 0 0 0 0 0 0
TONSL	1 1 0 0 1 1 0 0 0 0 0 0 0 0

CRNKL1	0 0 1 0 0 1 0 0 0 0 0 0 1 1 0
NUP153	0 1 0 0 0 0 0 0 1 1 0 1 0 0
POLG	1 0 1 0 0 0 0 0 0 0 0 1 0 1
MANEA	0 1 0 0 1 0 0 0 0 1 0 0 0 1
EHD3	1 0 0 0 0 0 1 0 0 0 0 0 1 1
BMPER	1 0 0 0 0 0 1 0 0 1 0 0 1 0
blank002	0 0 0 0 1 0 0 0 1 0 0 1 0 1
DIAPH1	0 0 0 0 0 1 1 0 0 0 0 1 0 1
TUBGCP4	0 0 1 0 0 0 0 0 1 0 0 0 1 1
CCDC117	0 0 1 0 0 0 0 0 0 0 1 1 1 0
MUT	0 0 0 1 1 0 0 0 0 1 0 1 0 0
CDC5L	0 0 0 1 0 0 1 0 0 1 0 0 0 1
ELMSAN1	0 1 0 0 0 0 0 0 0 1 1 1 0 0
SLC25A44	0 0 0 1 1 0 0 0 0 0 0 0 1 1
POMT2	0 0 0 0 0 0 0 1 0 1 1 0 0 1
INTS3	0 1 1 0 0 0 0 0 0 0 0 1 0 1
NUP214	0 0 0 1 0 1 0 0 0 1 0 1 0 0
ZNF252P	0 0 0 1 0 0 0 0 1 0 0 1 0 1
SIPA1L3	0 0 0 0 0 0 0 1 1 0 0 0 1 1
C16orf72	1 0 0 0 0 1 0 0 0 1 0 0 1 0
RYBP	0 1 0 0 0 0 1 0 0 0 0 1 1 0
CHML	0 0 1 0 0 0 0 1 0 1 0 0 1 0
ADCY6	1 0 0 1 0 0 0 0 0 1 0 0 0 1
C2CD5	0 1 0 0 0 0 0 0 0 1 1 0 0 1
notarget001	0 0 0 1 0 0 0 0 0 1 1 1 0 0
C17orf51	0 1 0 0 0 0 0 0 1 1 0 0 1 0
REV1	1 1 0 0 0 0 0 0 0 1 0 1 0 0
TAF2	0 0 0 1 0 1 0 0 0 1 0 0 1 0
PRKD1	0 0 1 0 0 0 0 1 0 0 0 1 0 1
E2F3	0 1 0 0 0 0 0 0 1 0 0 1 1 0
ADO	0 0 0 0 0 1 0 1 0 0 0 0 1 1
AGGF1	0 0 0 0 0 1 0 0 0 1 1 0 1 0
C5orf51	0 0 0 0 0 0 0 1 1 0 0 1 0 1
ABCA1	0 1 0 0 0 0 0 0 1 0 0 1 0 1
EIF4ENIF1	0 0 0 0 0 1 0 0 0 1 1 0 0 1
ZER1	0 0 1 0 0 0 0 0 1 0 0 1 0 1
ITPRIP	0 0 0 0 0 1 0 0 0 0 1 1 0 1
DMXL1	0 1 1 0 0 0 0 0 0 1 0 0 1 0
ZNF521	0 0 0 0 1 0 1 0 0 0 0 1 1 0
MCUR1	0 0 0 0 1 0 1 0 0 1 0 0 1 0
FAM171B	0 0 0 1 1 0 0 0 0 1 0 0 0 1
NUDT3	0 0 0 1 0 1 0 0 0 0 0 1 1 0
SRGAP1	0 0 1 0 0 0 0 0 1 1 0 1 0 0
EYA3	0 0 0 0 0 0 1 0 0 1 1 0 1 0
UBL3	0 0 0 0 0 1 0 0 1 0 0 1 1 0
ZNF280D	1 0 0 0 0 0 0 0 1 0 0 1 0 1
ARL3	0 0 0 0 0 0 1 0 0 1 1 0 0 1
LRRC2	1 0 0 0 0 1 0 0 0 1 0 0 0 1
ITGA6	0 0 0 0 0 0 1 1 0 1 0 0 1 0
STYX	0 1 0 0 0 0 1 0 0 1 0 0 1 0
NF1	0 0 0 0 0 0 1 1 0 0 0 1 1 0
JAG1	1 0 0 0 0 1 0 0 0 0 0 1 1 0
TOP3A	0 1 0 0 0 0 0 0 0 0 1 1 0 1
VPS13A	0 0 0 1 0 0 0 0 1 1 0 0 1 0
BOD1L1	1 0 0 0 1 0 0 0 0 1 0 0 0 1
NFX1	1 0 1 0 0 0 0 0 0 0 0 1 1 0
C18orf25	1 0 0 0 0 0 1 0 0 0 0 1 1 0
ZNF160	0 0 0 0 0 0 0 1 1 0 0 1 1 0
CCSER2	0 0 0 0 1 0 0 1 0 0 0 1 0 1
MAVS	1 0 0 0 0 0 0 0 0 1 1 0 0 1
FHOD1	0 1 0 0 0 1 0 0 0 0 0 1 0 1
RNF123	0 0 1 1 0 0 0 0 0 1 0 1 0 0
DHX38	0 0 1 0 1 0 0 0 0 1 0 0 1 0
FAM8A1	0 0 0 0 0 0 0 1 0 0 1 1 0 1
NCAPD3	0 0 0 0 0 1 1 0 0 1 0 0 0 1
ABCA2	0 1 0 0 0 1 0 0 0 0 0 1 1 0
CREBL2	1 1 0 0 0 0 0 0 0 0 0 1 0 1
DOCK10	0 1 0 0 0 0 0 1 0 0 0 1 1 0
KLF9	0 0 0 1 0 0 1 0 0 1 0 1 0 0
DTX4	1 0 0 0 0 0 0 1 0 1 0 0 1 0
MARCH4	0 0 1 0 0 1 0 0 0 1 0 0 1 0
SEC24B	1 1 0 0 0 0 0 0 0 0 0 0 1 1
BAG5	0 0 0 1 0 0 0 0 1 1 0 1 0 0
FOXM1	0 0 1 0 0 0 1 0 0 1 0 1 0 0
CBX2	1 0 0 1 0 1 0 1 0 0 0 0 0 0

RPAP1	0 1 1 1 0 0 0 0 0 0 0 0 1 0 0
FTSJD2	1 0 0 0 0 0 1 1 0 0 0 0 1 0
SCAF8	1 0 0 1 1 0 0 0 0 0 0 0 1 0 0
FAM53C	0 0 1 1 0 0 0 0 0 0 1 1 0 0
UBR3	1 1 0 0 0 0 0 1 0 1 0 0 0 0
KIAA0355	0 1 0 0 0 1 0 0 0 0 1 0 0 1
ZNF25	1 0 0 0 0 0 0 1 0 0 1 0 0 1
HIRA	0 0 0 1 1 0 0 0 0 0 1 1 0 0
ARHGEF11	0 0 0 0 1 0 0 0 1 0 1 0 1 0
SKIV2L	1 0 0 0 0 1 0 0 1 0 0 0 1 0
NRIP3	0 0 0 0 0 1 1 1 0 0 0 0 0 1
RNMT	1 0 0 0 0 0 0 1 1 0 0 0 1 0
TMEM110	0 1 0 1 0 1 0 0 0 0 0 0 0 1
RAB22A	0 0 1 0 1 0 0 0 0 0 1 0 0 1
WDFY2	0 0 0 0 1 1 1 0 0 0 0 0 0 1
DENND4C	0 0 0 1 0 0 1 0 1 1 0 0 0 0
STK10	1 0 0 0 0 1 0 1 0 0 0 0 0 1
SV2A	0 0 0 0 0 1 1 0 1 0 0 0 0 1
SPRED2	0 1 0 0 0 1 0 0 1 0 0 0 0 1
ABCC10	1 1 0 0 1 0 0 0 0 0 0 1 0 0
RAI1	0 0 0 0 0 1 0 1 0 0 1 0 1 0
MLLT4	0 0 0 0 1 0 0 1 1 1 0 0 0 0
MICAL1	1 0 0 1 0 1 0 0 0 0 0 0 0 1
UBE4B	1 0 0 0 0 1 0 0 0 0 1 1 0 0
USP20	0 0 0 1 1 1 0 0 0 1 0 0 0 0
NACC2	0 0 0 1 1 0 0 0 0 0 1 0 1 0
DGKZP1	0 0 1 0 0 0 0 0 1 0 1 0 0 1
PEX1	0 0 0 1 0 1 0 0 1 1 0 0 0 0
CASC7	1 0 0 1 0 0 0 0 1 0 0 0 0 1
ARHGAP29	0 0 1 1 0 0 0 1 0 1 0 0 0 0
blank005	1 0 0 0 0 1 0 1 0 0 0 0 0 1
ZNFX1	0 1 0 0 0 0 1 0 0 0 1 0 0 1
VASH1	0 0 1 1 0 0 0 0 0 0 1 0 1 0
FAM179B	1 0 0 1 0 0 0 0 0 1 1 0 0 0
SOGA1	0 0 0 1 0 0 0 1 0 0 1 0 1 0
HMGXB3	0 0 0 0 0 1 0 0 1 0 1 0 1 0
ATR	0 1 1 0 0 0 0 0 0 1 1 0 0 0
PDCD11	1 0 0 0 0 0 1 1 0 1 0 0 0 0
TEX2	0 0 0 1 0 1 0 0 0 0 1 0 1 0
CIC	0 1 0 1 0 0 0 1 0 0 0 0 1 0
CAD	0 1 0 0 1 0 0 0 1 0 0 0 1 0
GDF11	1 0 0 1 1 0 0 0 0 0 0 0 1 0
CLN5	0 1 0 0 0 0 1 0 0 0 1 0 1 0
CBL	1 0 0 0 1 0 0 0 1 0 0 0 1 0
SMO	0 0 0 1 1 0 0 0 1 0 0 1 0 0
TPCN1	1 1 0 0 0 0 0 0 1 0 0 0 0 1
LRIF1	1 1 0 0 0 1 0 0 0 0 0 1 0 0
RPRD1B	0 0 1 1 0 1 0 0 0 1 0 0 0 0
C20orf194	0 0 0 1 0 0 1 0 0 0 1 1 0 0
ZBED6	0 1 1 0 0 1 0 0 0 1 0 0 0 0
PTPN23	1 0 0 0 0 0 0 1 1 1 0 0 0 0
CTDP1	1 0 1 0 0 0 0 0 1 0 0 1 0 0
HECTD3	0 0 0 0 1 1 0 0 1 1 0 0 0 0
NUDT16	1 0 0 0 0 1 0 0 0 0 1 0 0 1
C2orf44	0 1 0 1 0 0 0 0 0 0 1 0 1 0
INPP5F	0 0 1 0 1 0 1 0 0 0 0 1 0 0
SERAC1	0 1 0 0 1 0 1 0 0 1 0 0 0 0
AKAP8	0 0 0 1 0 0 0 1 1 0 0 1 0 0
LAMA1	0 0 1 0 0 0 1 0 1 1 0 0 0 0
MAML2	1 1 0 0 0 0 1 0 0 0 0 0 1 0
BDP1	0 0 0 0 0 1 1 0 0 0 1 1 0 0
ATXN1L	0 0 0 1 0 1 0 1 0 0 0 0 1 0
CHST3	1 0 0 1 0 0 0 0 1 1 0 0 0 0
DARS2	1 0 1 0 0 0 0 1 0 1 0 0 0 0
MAP3K1	0 1 1 0 0 0 0 0 1 0 0 0 1 0
XPO5	0 0 1 0 0 0 0 0 1 1 0 0 0 1
ABCC5	0 0 1 0 1 0 0 1 0 0 0 0 1 0
TARSL2	1 1 1 0 0 0 0 0 0 0 0 1 0 0
CLCN7	1 0 0 0 0 0 1 1 0 0 0 0 0 1
GID4	0 0 0 0 0 0 1 0 1 0 1 1 0 0
PRKAB2	0 0 0 0 0 0 0 1 0 1 0 1 0 1
RBM19	0 0 0 1 0 1 0 0 1 0 0 0 1 0
OSBPL2	0 0 0 0 0 1 0 0 0 1 1 1 0 0
C2orf49	1 0 1 0 0 0 1 0 0 0 0 1 0 0
SYT1	0 0 0 1 1 0 0 1 0 0 1 0 0 0

NCOA1	0 1 0 0 0 1 0 1 0 0 0 0 0 1
C7orf41	0 0 1 0 0 0 1 0 0 1 1 0 0 0
DIP2A	0 0 0 0 0 1 0 1 0 0 1 0 0 1
ZNF248	1 0 0 0 1 1 0 0 0 0 0 1 0 0
PIGM	0 1 0 0 0 0 1 1 0 0 0 0 1 0
FAM46A	0 0 1 0 1 0 0 0 1 1 0 0 0 0
ZNF697	1 1 0 0 0 1 0 0 0 0 0 0 1 0
NCOA6	1 1 0 0 0 0 0 1 0 0 0 0 0 1
FAM212B	0 0 1 0 0 1 0 0 1 0 0 0 0 1
CEP135	0 0 0 0 1 1 0 1 0 0 0 1 0 0
WDR5B	0 0 0 1 0 1 0 0 0 0 1 0 0 1
ERMP1	1 0 1 0 1 0 0 0 0 1 0 0 0 0
ZSCAN30	1 0 0 0 0 0 1 0 0 0 1 1 0 0
SH2D5	1 0 0 0 0 0 0 1 1 0 0 1 0 0
ZNF275	1 0 0 0 0 0 1 1 0 0 0 1 0 0
DIP2B	0 0 1 1 0 0 0 0 1 0 0 0 0 1
CEP85	0 1 0 0 0 0 1 0 1 1 0 0 0 0
UNC13B	1 0 0 0 1 0 1 0 0 1 0 0 0 0
ALPK1	0 0 1 0 0 1 0 1 0 1 0 0 0 0
TRAPPC9	0 1 0 0 0 1 0 1 0 1 0 0 0 0
MTF1	0 0 1 1 0 0 0 1 0 0 0 0 0 1
GOLGA8A	0 1 0 0 0 1 0 0 0 0 1 0 1 0
WRN	0 0 0 0 1 0 0 1 0 0 1 1 0 0
ZBED1	0 0 0 1 1 1 0 0 0 0 0 1 0 0
AGTPBP1	0 0 0 0 1 0 1 1 0 0 0 0 1 0
ZZEF1	1 0 1 0 0 0 0 1 0 0 0 1 0 0
UBOX5	0 1 0 0 0 0 0 0 1 0 1 1 0 0
AKAP13	0 0 0 0 1 0 0 1 1 0 0 0 1 0
C4orf21	0 0 0 0 0 1 0 0 1 0 1 1 0 0
RP11-159D12.9	1 0 0 1 0 1 0 0 0 0 0 0 1 0
NOTCH1	1 0 0 0 1 0 0 0 0 0 0 1 0 1
FAM46C	1 0 0 0 0 0 1 0 1 0 0 0 1 0
AMPD3	0 0 0 0 0 0 1 1 0 1 1 0 0 0
SPAST	0 0 0 0 1 0 0 1 1 0 0 0 0 1
TBC1D12	0 0 0 1 0 1 0 0 1 0 0 0 0 1
FGD1	0 0 0 1 1 0 0 1 0 0 0 0 1 0
ZNF614	1 0 1 0 0 0 0 0 0 0 1 1 0 0
FAM83G	1 0 1 0 0 1 0 0 0 1 0 0 0 0
FAM126B	1 0 0 0 0 0 1 0 1 0 0 1 0 0
RAP2B	0 1 0 1 0 0 1 0 0 0 0 1 0 0
ZNF407	1 0 0 0 1 0 0 1 0 1 0 0 0 0
UTP20	0 0 1 0 0 0 0 1 0 1 1 0 0 0
RELN	1 0 0 0 1 0 1 0 0 0 0 1 0 0
ORC2	1 0 0 0 1 0 0 1 0 0 0 0 1 0
GEMIN4	1 1 0 1 0 0 0 0 0 1 0 0 0 0
ZFP28	0 0 1 0 0 0 0 1 1 0 0 0 1 0
IL17RA	0 0 0 1 0 1 1 0 0 0 0 1 0 0
PAXIP1	1 0 0 0 1 0 0 0 1 0 0 0 0 1
CNTRL	0 0 0 0 1 0 1 0 0 0 1 0 0 1
HIVEP2	1 0 0 1 0 0 1 0 0 0 0 0 0 1
NEDD9	0 0 0 1 0 0 1 0 1 0 0 0 1 0
PIK3R3	0 0 0 0 1 0 0 1 1 0 0 1 0 0
TMEM246	0 1 0 0 0 1 0 0 1 1 0 0 0 0
FAM171A1	1 0 1 1 0 0 0 0 0 1 0 0 0 0
blank008	1 0 1 1 0 0 0 0 0 0 0 0 1 0
SLC24A1	0 1 0 0 0 0 1 0 0 1 1 0 0 0
SLC7A6OS	1 1 0 0 0 0 1 0 0 0 0 0 0 1
ZBTB34	0 0 0 0 0 0 1 1 0 0 1 1 0 0
ULK4	0 0 1 0 1 1 0 0 0 0 0 0 1 0
ARHGAP20	1 1 0 0 0 0 0 0 1 0 0 0 1 0
CCDC93	0 1 0 0 0 0 1 1 0 1 0 0 0 0
LPCAT2	1 0 0 0 0 0 0 0 1 1 1 0 0 0
KIF7	0 0 0 0 0 1 1 1 0 0 0 0 1 0
RAD54L2	0 1 1 0 0 0 0 0 0 0 0 1 0 0 1
DHX37	0 0 0 0 1 0 1 0 0 0 1 0 1 0
ZBTB11	1 0 1 0 0 0 0 0 0 0 1 0 0 1
DNAH5	0 0 0 0 0 1 1 0 1 0 0 0 1 0
ARHGEF6	0 1 0 0 0 0 1 1 0 0 0 0 0 1
RCAN3	0 0 0 0 1 1 0 0 0 1 1 0 0 0
DOPEY1	0 0 1 0 0 0 1 0 1 0 0 0 1 0
FAM217B	1 0 0 0 0 1 1 0 0 0 0 0 0 1
BCL9	1 1 0 1 0 0 0 0 0 0 0 1 0 0
IQCE	1 0 0 0 1 0 0 0 0 0 0 1 0 0 1
ALMS1	0 1 0 0 0 1 1 0 0 0 0 1 0 0
notarget005	0 1 1 1 1 0 0 0 0 0 0 0 0 0

ZBTB39	0 0 0 0 1 0 1 1 1 0 0 0 0 0
GJA3	0 1 0 1 0 1 0 0 0 0 1 0 0 0
SWT1	0 1 0 1 0 0 0 1 1 0 0 0 0 0
MDN1	1 0 0 0 1 0 0 1 1 0 0 0 0 0
ZNF280C	0 0 0 0 1 1 0 1 0 0 1 0 0 0
FAM26E	0 0 0 1 1 0 1 1 0 0 0 0 0 0
CDON	1 0 1 0 0 0 0 0 0 1 0 1 0 0
ATP10A	0 0 1 0 1 0 0 1 1 0 0 0 0 0
SLC30A4	0 1 0 1 0 1 0 1 0 0 0 0 0 0
ZSCAN20	1 1 0 1 0 0 0 1 0 0 0 0 0 0
ZNF813	1 0 0 1 0 0 0 0 1 0 0 1 0 0
SLC5A3	0 0 0 0 0 1 1 1 1 0 0 0 0 0
ZBTB24	1 0 1 0 0 1 0 1 0 0 0 0 0 0
FAM120C	0 0 0 1 1 0 0 0 1 0 1 0 0 0
ZFYVE26	1 0 0 0 1 1 0 1 0 0 0 0 0 0
LPIN3	1 1 0 1 0 0 0 0 1 0 0 0 0 0
RP11-496N12.9	1 1 0 0 0 0 1 0 1 0 0 0 0 0
CAND2	0 0 1 0 0 1 0 0 1 0 1 0 0 0
LSM11	1 0 1 1 0 0 0 0 0 0 1 0 0 0
C1orf112	0 1 1 1 0 0 1 0 0 0 0 0 0 0
CEP128	0 1 0 0 0 1 1 0 1 0 0 0 0 0
OXTR	0 1 0 0 0 1 0 0 1 0 1 0 0 0
SLC1A1	1 0 0 0 0 1 1 0 1 0 0 0 0 0
KIAA1147	1 0 0 1 0 0 0 1 1 0 0 0 0 0
LMTK2	0 1 1 0 0 0 1 0 1 0 0 0 0 0
ZNF549	1 1 0 0 0 1 0 0 1 0 0 0 0 0
CLTCL1	1 1 0 1 0 0 1 0 0 0 0 0 0 0
ANGEL1	0 0 0 1 1 1 0 0 1 0 0 0 0 0
CARD10	1 0 0 0 0 1 0 0 1 0 1 0 0 0
APOLD1	0 1 0 1 1 0 0 0 0 0 1 0 0 0
WDR66	1 1 0 0 1 0 0 0 1 0 0 0 0 0
STARD9	0 0 0 1 1 0 1 0 0 0 1 0 0 0
ZNF33B	1 1 0 0 0 1 0 1 0 0 0 0 0 0
INSR	1 1 0 1 0 1 0 0 0 0 0 0 0 0
RP11-958N24.2	0 0 1 1 0 1 1 0 0 0 0 0 0 0
MYPN	0 1 0 0 0 0 0 1 1 0 1 0 0 0
PRDM15	0 0 1 1 0 1 0 0 1 0 0 0 0 0
GPATCH1	0 0 1 0 1 0 0 1 0 0 1 0 0 0
GAN	0 0 0 1 0 1 0 1 1 0 0 0 0 0
QRICH2	1 0 0 1 1 1 0 0 0 0 0 0 0 0
BRCA2	1 1 1 0 0 0 0 0 0 1 0 0 0 0
EVC2	0 0 0 1 1 0 0 1 1 0 0 0 0 0
DISP1	1 0 1 1 0 0 1 0 0 0 0 0 0 0
ADAMTS3	0 0 1 0 0 0 0 1 1 1 0 0 0 0
PHLPP1	1 0 1 0 0 0 0 1 1 0 0 0 0 0
RUNX2	1 0 0 0 0 1 1 0 0 0 1 0 0 0
ERVK13-1	0 1 1 1 0 1 0 0 0 0 0 0 0 0
SAMD9L	0 0 1 0 1 1 0 1 0 0 0 0 0 0
ARSG	0 0 0 0 1 0 1 1 0 0 1 0 0 0
SEMA6D	0 0 1 0 0 1 0 1 0 0 1 0 0 0
ARHGEF3	1 1 1 0 1 0 0 0 0 0 0 0 0 0
UBN2	0 0 0 0 1 1 1 0 1 0 0 0 0 0
CERS6	0 0 0 0 0 1 0 1 1 0 1 0 0 0
JRK	1 0 0 0 1 0 1 0 0 0 1 0 0 0
DAAM2	0 0 1 0 0 0 0 1 1 0 0 1 0 0 0
ATP7B	1 0 0 0 1 0 0 0 1 0 1 0 0 0
ZNF783	0 0 1 1 0 1 0 1 0 0 0 0 0 0
ZNF470	0 1 1 0 0 1 0 0 0 0 1 0 0 0
AFAP1L1	1 0 1 1 0 0 0 1 0 0 0 0 0 0
DSP	1 0 0 0 0 0 0 1 1 0 1 0 0 0
CARD6	0 0 0 1 1 1 0 1 0 0 0 0 0 0
HSPA4L	1 0 1 1 1 0 0 0 0 0 0 0 0 0
SLC18A2	0 1 0 1 0 0 1 1 0 0 0 0 0 0
SP4	0 1 1 0 1 0 0 0 0 0 1 0 0 0
SCN3A	0 0 1 1 0 0 1 0 0 0 1 0 0 0
KRBA1	0 0 0 1 0 0 1 0 1 0 1 0 0 0
RP11-121C2.2	0 0 0 0 1 1 1 1 0 0 0 0 0 0
BIRC3	1 0 0 1 0 0 1 1 0 0 0 0 0 0
COL11A1	0 0 0 0 0 1 1 0 1 0 1 0 0 0
EPB41	0 1 1 1 0 0 0 1 0 0 0 0 0 0
ZSCAN22	0 1 0 1 1 1 0 0 0 0 0 0 0 0
KLHL25	1 0 0 1 0 1 0 0 1 0 0 0 0 0
PLCL2	1 0 1 0 1 0 0 1 0 0 0 0 0 0
PHKA1	0 1 1 0 0 0 1 0 0 0 1 0 0 0
TICRR	0 1 0 0 0 0 1 1 1 0 0 0 0 0

Appendix C: GO analysis of gene groups for 140 genes

The 'Genes' column contains the names of the genes for each identified group. "Correlation_difference" is the difference in the average correlation coefficient between the specified gene and all others in that group and the average correlation coefficient between the specified gene and all others genes not in that group. "P_value_correlation" represents the significance (p-value) of this difference in average correlation coefficient as determined by a student's t-test. All of grouped genes have pvalues that are substantially smaller than most of p-values of the ungrouped genes. "GO_ID" lists the GO IDs associated with the enriched GO terms for each group. "GO_terms" lists the names of these GO terms. "GO_enrichment" provides the measured enrichment, defined as the ratio of the fraction of genes within each group that have this term to the fraction of all measured genes that have this term. "P_value_GO" lists the significance (p-values) of the enrichment of these terms, calculated via the hypergeometric function. Only the top 10 statistically significantly enriched GO terms are listed for each groups.

Genes	Correlation_difference	P_value_correlation	GO_ID	GO_terms	GO_enrichment	P_value_GO
Group 1						
COL5A1	0.28767	4.32E-13	GO:0005578	proteinaceous extracellular matrix	8.6154	4.74E-08
COL7A1	0.16413	1.36E-10	GO:0048196	plant extracellular matrix	8.6154	4.74E-08
FBN1	0.35189	5.54E-17	GO:0097311	biofilm matrix	8.6154	4.74E-08
FBN2	0.23948	6.08E-11	GO:0005604	basement membrane	8.6154	7.17E-07
H6PD	0.14712	2.89E-06	GO:0048046	apoplast	5.7436	1.10E-06
IGF2R	0.28859	1.90E-10	GO:0098595	perivitelline space	5.7436	1.10E-06
KIAA1199	0.14252	1.47E-06	GO:0031012	extracellular matrix	6.7009	1.52E-06
LRP1	0.32458	2.91E-13	GO:0044420	extracellular matrix part	6.7009	1.52E-06
NUMA1	0.12483	0.00051351	GO:0044421	extracellular region part	4.3077	4.36E-06
SVEP1	0.21406	6.74E-09	GO:0010215	cellulose microfibril organization	6.0308	4.76E-06
THBS1	0.12914	3.63E-05				
TNC	0.15274	2.03E-10				
VCAN	0.1583	3.13E-07				
Group 2						
PAPPA	0.24221	4.45E-05	GO:0007566	embryo implantation	11.2	0.0091308
PCNX	0.15326	0.00016447	GO:0009319	cytochrome o ubiquinol oxidase complex	3.95294	0.022632
PIEZO1	0.22648	3.29E-06	GO:0031224	intrinsic component of membrane	3.95294	0.022632
PLXNA1	0.12552	0.0061191	GO:0031301	integral component of organelle membrane	3.95294	0.022632
SCUBE3	0.20343	0.00025839	GO:0031361	integral component of thylakoid membrane	3.95294	0.022632
			GO:0034702	ion channel complex	3.95294	0.022632
			GO:0045203	integral component of cell outer membrane	3.95294	0.022632
			GO:0046696	lipopolysaccharide receptor complex	3.95294	0.022632
			GO:1902495	transmembrane transporter complex	3.95294	0.022632
			GO:0005887	integral component of plasma membrane	3.73333	0.026587
Group 3						
ALMS1	0.10234	0.0037949	GO:0005929	cilium	8.4	0.0022358
C17orf51	0.098302	0.00641	GO:0021549	cerebellum development	14	0.0045045
CACNA1A	0.10906	0.00011113	GO:0021953	central nervous system neuron differentiation	14	0.0045045
DIEXF	0.16799	5.07E-06	GO:0044093	positive regulation of molecular function	14	0.0045045
FZD4	0.18559	2.19E-08	GO:0001931	uropod	9.33333	0.012776
GTF3C4	0.12059	0.0014645	GO:0005902	microvillus	9.33333	0.012776
HIVEP1	0.18011	2.52E-06	GO:0005903	brush border	9.33333	0.012776
NRIP1	0.10659	0.0036223	GO:0005937	mating projection	9.33333	0.012776
			GO:0007214	gamma-aminobutyric acid signaling pathway	9.33333	0.012776
			GO:0009288	bacterial-type flagellum	9.33333	0.012776
Group 4						
KIAA2018	0.20621	1.28E-05	GO:0007422	peripheral nervous system development	14.9333	0.0046946
RNF169	0.18948	0.00017517	GO:0007399	nervous system development	11.2	0.0091308
SLC5A3	0.17784	1.15E-05	GO:0048731	system development	6.4	0.029344
TMOD2	0.22141	8.56E-08	GO:0000983	RNA polymerase II core promoter sequence-specific DNA binding	22.4	0.044643
ZKSCAN2	0.17443	0.00013402		transcription factor activity		
			GO:0001200	metal ion regulated sequence-specific DNA binding RNA polymerase II transcription factor activity	22.4	0.044643

			GO:0001228	RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	22.4	0.044643
			GO:0001963	synaptic transmission, dopaminergic	22.4	0.044643
			GO:0003096	renal sodium ion transport	22.4	0.044643
			GO:0003705	RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity	22.4	0.044643
			GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	22.4	0.044643
Group 5						
CBX5	0.30298	2.82E-06	GO:0005721	centromeric heterochromatin	21	6.96E-05
CENPF	0.41298	6.45E-09	GO:0000775	chromosome, centromeric region	16.8	0.0001723
MED14	0.18338	0.00025891	GO:0000776	kinetochore	7.63636	0.0026832
MKI67	0.45536	7.38E-11	GO:0044427	chromosomal part	7.63636	0.0026832
			GO:0005643	nuclear pore	18.6667	0.0028431
			GO:0007068	negative regulation of transcription during mitosis	18.6667	0.0028431
			GO:0010620	negative regulation of transcription by transcription factor catabolism	18.6667	0.0028431
			GO:0010621	negative regulation of transcription by transcription factor localization	18.6667	0.0028431
			GO:0010629	negative regulation of gene expression	18.6667	0.0028431
			GO:0010678	negative regulation of cellular carbohydrate metabolic process by negative regulation of transcription, DNA-templated	18.6667	0.0028431
Group 6						
AFAP1	0.1085	8.21E-07	GO:0032320	positive regulation of Ras GTPase activity	3.1111	0.0028101
AFF4	0.062857	0.0031913	GO:0043087	regulation of GTPase activity	3.1111	0.0028101
AMOTL1	0.15865	1.23E-13	GO:0043547	positive regulation of GTPase activity	3.1111	0.0028101
ASCC3	0.09096	1.58E-06	GO:0005681	spliceosomal complex	3.1111	0.0094843
CHD8	0.14712	1.69E-12	GO:0007043	cell-cell junction assembly	3.1111	0.0094843
CKAP5	0.16855	2.73E-09	GO:0008285	negative regulation of cell proliferation	3.1111	0.0094843
CREBBP	0.12337	4.83E-09	GO:0017048	Rho GTPase binding	3.1111	0.0094843
DIP2C	0.10597	1.89E-07	GO:0030424	axon	3.1111	0.0094843
DNAJC13	0.15033	8.07E-14	GO:0030981	cortical microtubule cytoskeleton	3.1111	0.0094843
DOCK7	0.17244	3.93E-15	GO:0042757	giant axon	3.1111	0.0094843
DYNC1H1	0.17865	6.78E-12				
FLNA	0.17725	2.45E-11				
FLNC	0.11215	1.20E-09				
GTF3C1	0.1054	9.72E-09				
HERC2	0.13889	3.22E-12				
IQGAP1	0.12496	3.23E-08				
KIAA1462	0.19804	1.42E-12				
MYH10	0.20647	4.41E-13				
NF1	0.11755	2.21E-09				
PDS5A	0.122	1.67E-09				
PRKCA	0.10097	2.68E-08				
PRKDC	0.088163	9.11E-06				
PRPF8	0.18189	7.93E-12				
PTPN14	0.091215	1.21E-07				
RAB3B	0.13823	6.27E-13				
ROCK2	0.18626	6.01E-13				
SIPA1L3	0.10274	1.18E-06				
SMARCA5	0.12014	3.76E-07				
SPTAN1	0.21915	1.29E-18				
SPTBN1	0.14832	2.80E-08				
SRRM2	0.098781	5.10E-05				
TEAD1	0.16176	3.45E-11				
TLN1	0.25499	1.46E-18				
TPR	0.22228	6.11E-16				
USP9X	0.077136	0.00058002				
ZC3H13	0.17975	5.98E-12				
Group 7						
DIP2B	0.16088	0.0001923	GO:0004843	ubiquitin-specific protease activity	28	0.0009653
TNRC6A	0.12268	0.025353	GO:0004221	ubiquitin thiolesterase activity	18.6667	0.0028431
USP24	0.10502	0.039184	GO:0008234	cysteine-type peptidase activity	18.6667	0.0028431
USP34	0.17678	0.006807	GO:0016787	hydrolase activity	18.6667	0.0028431
			GO:0019783	small conjugating protein-specific protease activity	18.6667	0.0028431
			GO:0036459	ubiquitinyl hydrolase activity	18.6667	0.0028431
			GO:0004197	cysteine-type endopeptidase activity	14	0.0055819

			GO:0019941	modification-dependent protein catabolic process	11.2	0.0091308
			GO:0030579	ubiquitin-dependent SMAD protein catabolic process	11.2	0.0091308
			GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	11.2	0.0091308
Ungrouped						
ABCA2	-0.026182	0.074984				
AKAP11	0.036444	0.036699				
ALPK2	0.022796	0.13529				
ANKRD52	0.0041061	0.81754				
ARL10	0.015274	0.32813				
ATP11A	0.0022604	0.88044				
CBL	0.028711	0.092434				
CEP250	0.024328	0.20724				
DOPEY1	0.031118	0.050935				
EGFR	0.015905	0.36613				
FAM208B	0.018509	0.27401				
FASN	-0.0070651	0.70997				
FLNB	-0.0087654	0.63698				
FRY	0.023029	0.15026				
FYCO1	-0.0195	0.30366				
GPR107	0.019456	0.27644				
HEATR5B	0.026001	0.12272				
HELZ2	0.031234	0.064398				
HIVEP2	0.01736	0.23825				
ITGA2	0.039459	0.050376				
KIDINS220	0.0060024	0.71134				
KPNA4	0.0037153	0.84442				
LMTK2	0.051681	0.00061633				
LUZP1	0.018951	0.24286				
MALAT1	0.045716	0.010815				
MAN1A2	0.0052039	0.73434				
NKTR	0.0096182	0.59453				
PHIP	0.0045215	0.77665				
PRRC2B	0.013221	0.43553				
REV3L	0.0042432	0.77678				
SON	0.0074876	0.66633				
SSH1	0.028114	0.13818				
THSD4	-0.011103	0.48796				
UBR2	0.028553	0.043559				
UBR5	-0.012633	0.5658				
VPS13D	0.014763	0.35202				
ZFC3H1	0.017008	0.27928				

Appendix D: GO analysis of gene groups for 1001 genes

The columns are defined as those listed in Appendix C. Here the p-values of the average correlation difference is greater than 0.05 for a small number of the grouped genes (marked red) suggesting that assignment to that group may not be statistically significant. The statistically most significantly enriched GO terms (maximum 10) are listed for each group. Several groups do not have statistically significant enrichment of GO terms and these groups are labeled blue.

Genes	Correlation_difference	P_value_correlation	GO_ID	GO_terms	GO_enrichment	P_value_GO
Group 1						
UBL3	0.29014	3.18E-06				
IQSEC1	0.248	1.66E-05				
C11orf30	0.32738	4.94E-07				
LPHN3	0.36901	6.26E-10				
Group 2						
ATG16L1	0.079698	0.13601	GO:0031931	TORC1 complex	23.9	0.0023659
SCAP	0.21932	0.00031808	GO:0031932	TORC2 complex	23.9	0.0023659
CEP170B	0.15104	0.009668	GO:0043234	protein complex	4.93349	0.011689
TTI1	0.11104	0.055541	GO:0005874	microtubule	9.43421	0.014818
Group 3						
SMURF1	0.25896	2.07E-06	GO:0015822	ornithine transport	71.7	0.00023179
VPS13D	0.097011	0.035912	GO:0002522	leukocyte migration involved in immune	57.36	0.00038469
SLC7A11	0.26293	2.43E-06		response		
KRAS	0.2043	0.00025785	GO:0036336	dendritic cell migration	57.36	0.00038469
ATP13A1	0.2959	7.64E-08	GO:0050904	diapedesis	57.36	0.00038469
			GO:0071674	mononuclear cell migration	57.36	0.00038469
			GO:0072676	lymphocyte migration	57.36	0.00038469
			GO:0097529	myeloid leukocyte migration	57.36	0.00038469
			GO:0006812	cation transport	47.8	0.00057461
			GO:0045123	cellular extravasation	47.8	0.00057461
			GO:0002376	immune system process	13.4438	0.0007462
Group 4						
UST	0.33201	6.19E-09	GO:0005622	intracellular	12.8036	0.0082044
RNF213	0.37531	1.30E-09	GO:1990302	Bre1-Rad6 ubiquitin ligase complex	11.2031	0.010641
TTPAL	0.33117	5.02E-08	GO:1990303	UBR1-RAD6 ubiquitin ligase complex	11.2031	0.010641
CDC25A	0.19814	0.00025963	GO:1990304	MUB1-RAD6-UBR2 ubiquitin ligase complex	11.2031	0.010641
			GO:1990305	RAD6-UBR2 ubiquitin ligase complex	11.2031	0.010641
			GO:1990306	RSP5-BUL ubiquitin ligase complex	11.2031	0.010641
Group 5						
TBCD	0.17873	0.00099039	GO:1990188	euchromatin binding	7.7514	0.022314
BDP1	0.24421	1.50E-06	GO:0031490	chromatin DNA binding	7.3538	0.024609
PDE7B	0.28148	5.16E-08	GO:0003682	chromatin binding	7.17	0.025789
MTF1	0.23285	1.65E-06	GO:0006412	translation	6.2348	0.033317
TICRR	0.20015	8.63E-05	GO:0005730	nucleolus	3.5262	0.033505
			GO:0044452	nucleolar part	3.5262	0.033505
			GO:0031491	nucleosome binding	5.975	0.035987
Group 6						
TJP2	0.29921	3.20E-17	GO:0005829	cytosol	3.4416	0.0025577
AGO1	0.21068	4.15E-09	GO:0044445	cytosolic part	3.4416	0.0025577
AKAP8	0.27529	5.13E-15	GO:0097342	ripoptosome	3.4416	0.0025577
TRIM37	0.12963	0.00011766	GO:1990133	molybdopterin cofactor (Moco) biosynthesis	3.4416	0.0025577
ZMYM3	0.19948	3.25E-09		adenylyltransferase complex		
SYNJ2	0.26611	8.88E-14	GO:1990140	MPT synthase complex	3.4416	0.0025577
NUP98	0.229	1.13E-10	GO:1990198	ModE complex	3.4416	0.0025577
DHX16	0.28303	3.75E-17	GO:1990201	alkanesulfonate monooxygenase complex	3.4416	0.0025577
WDR35	0.2795	1.11E-16	GO:1990202	FMN reductase complex	3.4416	0.0025577
WWC3	0.23102	8.59E-11	GO:1990205	taurine dioxygenase complex	3.4416	0.0025577
			GO:1990220	GroEL-GroES complex	3.4416	0.0025577
Group 7						
CENPF	0.27782	3.84E-17	GO:0072089	stem cell proliferation	15.042	0.00068868
DIAPH1	0.19238	4.92E-10	GO:0008283	cell proliferation	13.0364	0.0010708
CLSPN	0.34648	1.67E-22	GO:0006259	DNA metabolic process	7.04668	0.0014303
C20orf194	0.26823	1.57E-15	GO:0000086	G2/M transition of mitotic cell cycle	10.8636	0.0018555
ZFAT	0.14514	5.29E-06	GO:0031660	regulation of cyclin-dependent protein	10.8636	0.0018555

FOXM1	0.34196	1.07E-23		serine/threonine kinase activity involved in		
FAM178A	0.16874	3.22E-07		G2/M transition of mitotic cell cycle		
ANKRD52	0.13903	3.54E-05	GO:0010972	negative regulation of G2/M transition of	26.0727	0.0020627
NES	0.23968	6.25E-11		mitotic cell cycle		
MKI67	0.30513	5.13E-20	GO:0044839	cell cycle G2/M phase transition	10.2919	0.0021781
TPCN1	0.22917	2.29E-11	GO:0034506	chromosome, centromeric core region	21.7273	0.0030549
			GO:0034507	chromosome, centromeric outer repeat region	21.7273	0.0030549
			GO:0034776	response to histamine	21.7273	0.0030549
Group 8						
CREBBP	0.17485	0.004846	GO:0001214	metal ion regulated core promoter proximal	71.7	0.00023179
FAM13A	0.2584	1.24E-05		region sequence-specific DNA binding RNA		
CDC5L	0.25651	7.47E-06		polymerase II transcription factor activity		
REST	0.21331	0.000152		involved in negative regulation of transcription		
			GO:0001078	RNA polymerase II core promoter proximal region	59.75	0.00034671
				sequence-specific DNA binding transcription		
				factor activity involved in negative regulation of		
				transcription		
			GO:1990188	euchromatin binding	14.5338	0.00048384
			GO:0016607	nuclear speck	51.2143	0.00048403
			GO:0031490	chromatin DNA binding	13.7885	0.00056742
			GO:0003682	chromatin binding	13.4438	0.00061252
			GO:0001159	core promoter proximal region DNA binding	39.8333	0.00082509
			GO:0031491	nucleosome binding	11.2031	0.0010596
			GO:0001227	RNA polymerase II transcription regulatory region	27.5769	0.0017675
				sequence-specific DNA binding transcription factor		
				activity involved in negative regulation of		
				transcription		
			GO:0001025	RNA polymerase III transcription factor binding	23.9	0.0023659
Group 9						
RAPGEF1	0.24708	6.37E-05	GO:0007167	enzyme linked receptor protein signaling	32.5909	0.0012534
CABIN1	0.16819	0.0037013		pathway		
BOD1L1	0.21301	0.00074941	GO:0038179	neurotrophin signaling pathway	10.5441	0.011965
ZZEF1	0.21729	0.00035229				
Group 10						
ABCA1	0.19203	0.00056861				
DTX3L	0.28923	1.22E-09				
JARID2	0.22145	5.21E-06				
SCAF8	0.23844	6.29E-06				
NRDE2	0.16835	0.00049977				
Group 11						
EIF4ENIF1	0.13259	0.0011128	GO:0006668	sphinganine-1-phosphate metabolic process	18.0756	0.00035935
CDAN1	0.071597	0.097691	GO:0006673	inositolphosphoceramide metabolic process	18.0756	0.00035935
KIAA1551	0.091223	0.033346	GO:0006684	sphingomyelin metabolic process	18.0756	0.00035935
MAVS	0.14893	0.00011719	GO:0032760	positive regulation of tumor necrosis factor	51.2143	0.00048403
ARSD	0.095812	0.025759		production		
GPD1L	0.1368	0.00074702	GO:0032680	regulation of tumor necrosis factor production	40.9714	0.00080105
PIK3R1	0.29307	1.23E-13	GO:0001819	positive regulation of cytokine production	18.6234	0.0042227
			GO:0061423	positive regulation of sodium ion transport by positive regulation of transcription from RNA polymerase II promoter	7.14618	0.0056035
			GO:0006650	glycerophospholipid metabolic process	15.7582	0.005904
			GO:0006661	phosphatidylinositol biosynthetic process	15.7582	0.005904
			GO:0009395	phospholipid catabolic process	15.7582	0.005904
Group 12						

ADAMTS15	0.15519	5.01E-07	GO:0003735	structural constituent of ribosome	19.9167	0.0037011
NCR3LG1	0.25697	3.86E-14	GO:0005199	structural constituent of cell wall	19.9167	0.0037011
ELF4	0.15221	6.38E-05	GO:0005212	structural constituent of eye lens	19.9167	0.0037011
TSTD2	0.25591	2.67E-11	GO:0005213	structural constituent of chorion	19.9167	0.0037011
FAN1	0.23748	6.23E-11	GO:0008147	structural constituent of bone	19.9167	0.0037011
CNTN3	0.20319	5.09E-08	GO:0016490	structural constituent of peritrophic membrane	19.9167	0.0037011
ZMAT3	0.080625	0.026889	GO:0019911	structural constituent of myelin sheath	19.9167	0.0037011
DSP	0.31295	1.70E-18	GO:0030280	structural constituent of epidermis	19.9167	0.0037011
TECPR2	0.20669	8.36E-10	GO:0030527	structural constituent of chromatin	19.9167	0.0037011
			GO:0035804	structural constituent of egg coat	19.9167	0.0037011
Group 13						
ASAP3	0.27126	1.55E-07	GO:0042221	response to chemical	14.0588	0.0072999
PAPD7	0.2151	6.38E-06	GO:0001010	sequence-specific DNA binding transcription	5.975	0.036619
MAFG	0.1643	0.00017411		factor recruiting transcription factor activity		
TULP4	0.18762	0.00021202	GO:0001011	sequence-specific DNA binding RNA polymerase	5.975	0.036619
KCTD2	0.17631	0.00026788		recruiting transcription factor activity		
SH2D5	0.29034	6.69E-11	GO:0001034	sequence-specific DNA binding RNA polymerase	5.975	0.036619
				III transcription factor activity		
			GO:0001073	DNA binding transcription antitermination factor	5.975	0.036619
				activity		
			GO:0001130	sequence-specific DNA binding bacterial-type RNA polymerase transcription factor activity	5.975	0.036619
			GO:0001142	sequence-specific DNA binding mitochondrial RNA polymerase transcription factor activity	5.975	0.036619
			GO:0001167	sequence-specific DNA binding RNA polymerase I transcription factor activity	5.975	0.036619
			GO:0001199	metal ion regulated sequence-specific DNA binding transcription factor activity	5.975	0.036619
			GO:0098531	direct ligand regulated sequence-specific DNA binding transcription factor activity	5.975	0.036619
Group 14						
PIK3R3	0.17188	2.43E-07	GO:0051091	positive regulation of sequence-specific DNA	16.5462	0.00050804
SMO	0.15418	6.42E-06		binding transcription factor activity		
TRIM14	0.15788	3.14E-06	GO:0007494	midgut development	47.8	0.00052005
KIAA2026	0.20777	1.56E-09	GO:0008528	G-protein coupled peptide receptor activity	35.85	0.0010284
ZNF142	0.15584	4.23E-06	GO:0051240	positive regulation of multicellular organismal	35.85	0.0010284
OXTR	0.2349	1.53E-11		process		
ZMYND8	0.14056	1.01E-05	GO:0055123	digestive system development	35.85	0.0010284
EHMT1	0.25326	1.90E-14	GO:1901227	negative regulation of transcription from RNA	6.82857	0.0015478
TAB3	0.23341	3.27E-12		polymerase II promoter involved in heart		
CTDP1	0.18052	9.50E-07		development		
			GO:0048565	digestive tract development	28.68	0.0016948
			GO:0048566	embryonic digestive tract development	28.68	0.0016948
			GO:0044464	cell part	2.95235	0.0020977
			GO:0061053	somite development	23.9	0.0025137
Group 15						
PRDM2	0.16896	4.50E-05	GO:0007532	regulation of mating-type specific transcription,	3.532	0.036715
HIP1R	0.24447	5.30E-09		DNA-templated		
ZNF512B	0.22181	8.32E-08	GO:0009373	regulation of transcription by pheromones	3.532	0.036715
CAPN15	0.25076	2.82E-10	GO:0019056	modulation by virus of host transcription	3.532	0.036715
FCHSD2	0.23059	6.44E-08	GO:0031554	regulation of DNA-templated transcription,	3.532	0.036715
TOX	0.080033	0.031335		termination		
PHLPP1	0.25873	1.68E-10	GO:0032784	regulation of DNA-templated transcription,	3.532	0.036715
				elongation		
			GO:0043620	regulation of DNA-templated transcription in	3.532	0.036715

				response to stress		
			GO:0044324	regulation of transcription involved in anterior/posterior axis specification	3.532	0.036715
			GO:0045896	regulation of transcription during mitosis	3.532	0.036715
			GO:0045990	carbon catabolite regulation of transcription	3.532	0.036715
			GO:0060194	regulation of antisense RNA transcription	3.532	0.036715
Group 16						
HPS5	0.18655	7.92E-23	GO:0035257	nuclear hormone receptor binding	12.8036	0.00010219
WDR59	0.11629	3.59E-10	GO:0046966	thyroid hormone receptor binding	6.40179	0.00015132
TMEM194A	0.14581	2.55E-14	GO:0050681	androgen receptor binding	16.8047	0.00031113
GSE1	0.16816	5.51E-18	GO:0042809	vitamin D receptor binding	5.89638	0.00087076
PCNX	0.14247	3.31E-14	GO:0042974	retinoic acid receptor binding	5.60156	0.0011161
MPP5	0.12796	1.55E-12	GO:0030521	androgen receptor signaling pathway	11.2031	0.0014316
COL11A1	0.10185	8.41E-09	GO:0060520	activation of prostate induction by androgen	11.2031	0.0014316
FOSB	0.13443	1.28E-12		receptor signaling pathway		
MED14	0.075568	6.58E-06	GO:0060614	negative regulation of mammary gland	11.2031	0.0014316
DHX57	0.086921	9.83E-07		development in males by androgen receptor		
PLXNA3	0.15277	3.17E-15		signaling pathway		
PEX1	0.12593	7.68E-11	GO:0007031	peroxisome organization	22.4063	0.0019323
FAM120C	0.13095	6.91E-14	GO:0048672	positive regulation of collateral sprouting	22.4063	0.0019323
NRIP1	0.14931	1.00E-15				
TUBGCP4	0.11031	3.42E-08				
ZER1	0.11886	7.63E-11				
ADAM17	0.11262	3.11E-09				
HSPA12A	0.1562	1.12E-17				
C11orf95	0.17016	6.58E-16				
NSD1	0.054439	0.0025004				
DIP2B	0.10118	1.22E-07				
NCOA3	0.12265	3.69E-11				
CIC	0.087915	1.16E-07				
TRPC6	0.10416	1.35E-09				
LRIG3	0.14795	5.07E-16				
EHD3	0.14016	2.38E-14				
C18orf25	0.11342	5.11E-09				
FAM46C	0.13887	2.69E-14				
AREL1	0.12765	3.73E-13				
ITSN2	0.11091	6.02E-09				
SETD2	0.10784	7.26E-10				
INSR	0.13775	3.59E-13				
Group 17						
SENP5	0.22595	2.02E-06	GO:0006487	protein N-linked glycosylation	47.8	0.00057461
GMEB2	0.21633	5.32E-06	GO:0018410	C-terminal protein amino acid modification	47.8	0.00057461
NPC1	0.25115	5.54E-07	GO:0061606	N-terminal protein amino acid propionylation	47.8	0.00057461
MAN1A2	0.18951	0.00025608	GO:0071587	CAAX-box protein modification	47.8	0.00057461
RBM26	0.24563	9.77E-08	GO:0034421	post-translational protein acetylation	40.9714	0.00080105
			GO:0043687	post-translational protein modification	35.85	0.0010636
			GO:0006124	ferredoxin metabolic process	23.9	0.0024647
			GO:0030091	protein repair	23.9	0.0024647
			GO:0043335	protein unfolding	23.9	0.0024647
			GO:0046154	rhodopsin metabolic process	23.9	0.0024647
Group 18						
INPP4A	0.25181	0.00057645	GO:0002330	pre-B cell receptor expression	15.4194	0.0052142
TCEB3	0.23817	0.00054757	GO:0043170	macromolecule metabolic process	14.9375	0.0055537
AGO4	0.18142	0.0068489	GO:0051604	protein maturation	14.4848	0.0059034
			GO:0006406	mRNA export from nucleus	12.5789	0.0078026
			GO:1901148	gene expression involved in extracellular matrix organization	11.1163	0.0099486
			GO:0006412	translation	10.3913	0.011352
			GO:0006396	RNA processing	9.95833	0.012335
			GO:0019080	viral gene expression	8.38596	0.017218
			GO:0007166	cell surface receptor signaling pathway	4.92784	0.047186
Group 19						
SSH2	0.23635	6.63E-06				
RC3H1	0.18807	0.0011092				
FAM171A1	0.11742	0.052608				
ABCC10	0.26814	1.46E-06				

Group 20						
SMCR7L	0.20975	4.75E-09	GO:0006575	cellular modified amino acid metabolic process	26.5556	0.0020224
ZSCAN29	0.14158	0.00014586	GO:0005509	calcium ion binding	2.50919	0.046129
FASN	0.19729	5.32E-08				
NLRX1	0.21029	5.51E-09				
C2CD5	0.19013	1.22E-06				
SMCR8	0.086051	0.014448				
ZNF652	0.18973	1.56E-07				
MAPKBP1	0.21183	1.14E-09				
YOD1	0.23728	2.11E-11				
Group 21						
PHF13	0.22424	2.24E-18	GO:0003677	DNA binding	2.5158	0.0030573
MLLT4	0.18931	1.80E-14	GO:0000497	base pairing with DNA	2.7445	0.0033229
ZBTB39	0.22378	1.56E-20	GO:0003689	DNA clamp loader activity	2.7445	0.0033229
HECTD3	0.1863	1.89E-12	GO:0010844	recombination hotspot binding	2.7445	0.0033229
NEK1	0.15937	1.81E-10	GO:0008301	DNA binding, bending	2.7198	0.0035161
EVC2	0.26724	1.85E-26	GO:0003676	nucleic acid binding	2.4259	0.0039267
ANGEL1	0.26869	2.89E-28	GO:0003684	damaged DNA binding	2.6716	0.0039284
ZNF776	0.24681	1.90E-22	GO:0000975	regulatory region DNA binding	2.6252	0.0043767
GPATCH1	0.22236	1.49E-15	GO:0043566	structure-specific DNA binding	2.6252	0.0043767
ZSCAN22	0.14714	1.67E-09	GO:0043565	sequence-specific DNA binding	2.495	0.0059576
ATR	0.16426	2.76E-13				
LMTK2	0.14663	1.36E-08				
ZBED6	0.22238	1.99E-17				
ZNF646	0.14517	8.08E-10				
ZNF407	0.2617	4.97E-24				
GLTSCR1L	0.15852	1.09E-11				
DARS2	0.19943	1.09E-16				
BAIAP2L1	0.17932	1.04E-11				
ZBTB24	0.25372	3.03E-24				
Group 22						
LRRC58	0.1827	8.14E-07	GO:0019065	receptor-mediated endocytosis of virus by host	35.85	0.0010636
UTP20	0.23111	1.63E-09		cell		
XPO5	0.22235	6.73E-09	GO:0072584	caveolin-mediated endocytosis	35.85	0.0010636
C17orf70	0.11319	0.0017475	GO:0031623	receptor internalization	29.875	0.0015819
LDLR	0.20922	5.84E-08	GO:0072583	clathrin-mediated endocytosis	29.875	0.0015819
WDR55	0.25386	8.09E-12	GO:0090118	receptor-mediated endocytosis of low-density	29.875	0.0015819
PIKFYVE	0.22982	3.14E-09		lipoprotein particle involved in cholesterol		
AFAP1L1	0.22865	4.60E-10		transport		
			GO:0000154	rRNA modification	25.6071	0.002196
			GO:0000469	cleavage involved in rRNA processing	25.6071	0.002196
			GO:0000470	maturation of LSU-rRNA	25.6071	0.002196
			GO:0000475	maturation of 2S rRNA	25.6071	0.002196
			GO:0000476	maturation of 4.5S rRNA	25.6071	0.002196
Group 23						
TMEM57	0.12243	0.012479	GO:0022892	substrate-specific transporter activity	16.8706	0.0049713
CSPG4	0.21898	6.74E-06	GO:0044463	cell projection part	14.34	0.0068562
PDCD11	0.30303	2.29E-09	GO:0007227	signal transduction downstream of smoothened	9.88966	0.01409
DISP1	0.26554	1.23E-06	GO:0043005	neuron projection	8.43529	0.019047
SEC24B	0.27602	3.90E-07	GO:0044767	single-organism developmental process	4.1767	0.021373
			GO:0016021	integral component of membrane	4.13654	0.021936
			GO:0005887	integral component of plasma membrane	4.05849	0.023087
			GO:0044699	single-organism process	7.17	0.025789
			GO:0005215	transporter activity	6.51818	0.030726
			GO:0060089	molecular transducer activity	6.51818	0.030726
Group 24						
HMGXB3	0.24028	6.40E-06	GO:0006796	phosphate-containing compound metabolic	10.6222	0.012288
NEDD9	0.19271	0.0003007		process		
FAM212B	0.13015	0.0099344	GO:0022610	biological adhesion	6.37333	0.032011
TCEANC2	0.1508	0.0036946	GO:0030054	cell junction	6.23478	0.033317
NUDT16	0.17735	0.00068575	GO:0023014	signal transduction by phosphorylation	5.12143	0.047387
Group 25						
VPS39	0.237	0.0018001	GO:0005524	ATP binding	5.2527	0.041903
LARS2	0.20485	0.0016637	GO:0032559	adenyl ribonucleotide binding	5.1957	0.042765

MSH3	0.30722	6.45E-05	GO:0035639	purine ribonucleoside triphosphate binding	4.8776	0.048092
			GO:0032550	purine ribonucleoside binding	4.78	0.049923
Group 26						
POLD3	0.2009	1.40E-06	GO:0043247	telomere maintenance in response to DNA	12.5789	0.0090739
TMOD2	0.31336	2.19E-11		damage		
EEF2K	0.26095	6.71E-10	GO:0022008	neurogenesis	9.19231	0.016563
MECOM	0.15475	0.0003147	GO:0006974	cellular response to DNA damage stimulus	6.63889	0.030284
DMXL1	0.33	1.57E-12	GO:0044451	nucleoplasm part	3.73438	0.030713
RNF169	0.30397	2.87E-12	GO:0006281	DNA repair	6.28947	0.033397
			GO:0006975	DNA damage induced protein phosphorylation	5.31111	0.045117
Group 27						
SV2A	0.13604	1.35E-05	GO:0006900	membrane budding	11.95	0.0014035
FNBP1L	0.14955	1.31E-06	GO:0044464	cell part	2.81176	0.0015372
C2CD3	0.17449	9.15E-09	GO:0030662	coated vesicle membrane	23.9	0.0024647
PLD2	0.10853	0.00020919	GO:0007099	centriole replication	19.9167	0.0036451
FHOD1	0.11773	8.78E-05	GO:0030665	clathrin-coated vesicle membrane	19.9167	0.0036451
LNPEP	0.15236	5.26E-07	GO:0031623	receptor internalization	19.9167	0.0036451
RHOBTB1	0.15084	4.61E-07	GO:0072583	clathrin-mediated endocytosis	19.9167	0.0036451
TTF2	0.13666	1.45E-06	GO:1990176	MalFGK2 complex	3.14474	0.0049436
ADCY7	0.10363	0.00058111	GO:0007565	female pregnancy	17.0714	0.0050314
ZNF697	0.19098	1.57E-09	GO:0030135	coated vesicle	17.0714	0.0050314
CLTCL1	0.19351	7.46E-10				
BRCA2	0.13697	7.88E-06				
Group 28						
C10orf12	0.24237	3.24E-06				
LTBP2	0.21319	1.53E-05				
MORC3	0.17092	0.00084745				
POLR3B	0.24083	2.77E-06				
MICAL1	0.28261	1.39E-07				
Group 29						
MON2	0.21661	5.34E-12	GO:1901652	response to peptide	19.9167	0.0036451
TBC1D16	0.21429	4.09E-10	GO:0071417	cellular response to organonitrogen compound	17.0714	0.0050314
NUP214	0.14318	2.53E-06	GO:0044212	transcription regulatory region DNA binding	7.96667	0.022446
ERMAP	0.175	2.64E-09	GO:0006886	intracellular protein transport	4.26786	0.024829
SIN3A	0.099639	0.0015019	GO:0005730	nucleolus	2.44877	0.029832
NOL9	0.14632	2.11E-06	GO:0044452	nucleolar part	2.44877	0.029832
KLF11	0.16584	2.63E-08	GO:0031981	nuclear lumen	2.1213	0.030813
FAM217B	0.17058	9.83E-08	GO:0000976	transcription regulatory region sequence-specific	5.43182	0.044642
HIVEP2	0.21287	9.22E-12		DNA binding		
NUAK1	0.15658	8.32E-07	GO:0017038	protein import	5.19565	0.048191
WDR62	0.19877	1.30E-09				
QRICH2	0.13136	2.16E-06				
Group 30						
TRERF1	0.22266	4.46E-16	GO:0016790	thiolester hydrolase activity	17.925	0.0044053
BRWD1	0.11591	8.48E-06	GO:0008610	lipid biosynthetic process	12.8036	0.0088907
ZNF354C	0.12608	3.71E-07	GO:0005815	microtubule organizing center	4.07386	0.011253
USP20	0.25201	2.29E-20	GO:0000083	regulation of transcription involved in G1/S	4.80134	0.018792
CSRNP2	0.15355	6.51E-09		transition of mitotic cell cycle		
CEP85	0.22827	1.97E-17	GO:0007038	endocytosed protein transport to vacuole	8.14773	0.021499
SYTL4	0.22123	1.77E-17	GO:0015869	protein-DNA complex transport	8.14773	0.021499
SERAC1	0.11205	9.23E-05	GO:0030970	retrograde protein transport, ER to cytosol	8.14773	0.021499
GPRIN1	0.2364	6.49E-16	GO:0032594	protein transport within lipid bilayer	8.14773	0.021499
RMND5A	0.22369	2.68E-14	GO:0033369	establishment of protein localization to mast	8.14773	0.021499
PHRF1	0.14137	2.94E-07		cell secretory granule		
FNIP1	0.13526	1.13E-07	GO:0033376	establishment of protein localization to T cell	8.14773	0.021499
AHCTF1P1	0.12478	4.24E-06		secretory granule		
MCM2	0.13452	1.83E-07				
NCKAP5L	0.097125	0.00041957				
RP11-496N12.9	0.15377	5.23E-09				

Group 31						
TUBGCP5	0.19547	2.24E-06	GO:0045598	regulation of fat cell differentiation	29.2653	0.0016587
ZBED1	0.16591	7.71E-05	GO:0045599	negative regulation of fat cell differentiation	29.2653	0.0016587
LPIN1	0.12502	0.004001	GO:0000117	regulation of transcription involved in G2/M	7.49477	0.0048981
ETV5	0.21052	6.81E-07		transition of mitotic cell cycle		
ANKRD26	0.14727	0.0001217	GO:0044237	cellular metabolic process	5.79784	0.010018
PAPOLG	0.26263	4.25E-08	GO:0000429	carbon catabolite regulation of transcription	8.90683	0.017825
WDR81	0.31488	1.45E-12		from RNA polymerase II promoter		
			GO:0001079	nitrogen catabolite regulation of transcription	8.90683	0.017825
				from RNA polymerase II promoter		
			GO:0001196	regulation of mating-type specific transcription	8.90683	0.017825
				from RNA polymerase II promoter		
			GO:0010672	regulation of transcription from RNA polymerase II promoter during meiosis	8.90683	0.017825
			GO:0034243	regulation of transcription elongation from RNA polymerase II promoter	8.90683	0.017825
			GO:0034395	regulation of transcription from RNA polymerase II promoter in response to iron	8.90683	0.017825
Group 32						
CNTRL	0.22236	4.49E-06				
OTUD3	0.16731	0.0017419				
RRP12	0.25185	2.22E-07				
AHNAK2	0.20509	7.49E-05				
ARHGEF10L	0.19564	3.21E-05				
Group 33						
HEATR6	0.27084	0.00017815	GO:0044822	poly(A) RNA binding	6.9275	0.024849
CLCN7	0.20917	0.0025951	GO:0003723	RNA binding	5.2527	0.041903
RNASEL	0.17886	0.0062238	GO:0005524	ATP binding	5.2527	0.041903
			GO:0032559	adenyl ribonucleotide binding	5.1957	0.042765
			GO:0035639	purine ribonucleoside triphosphate binding	4.8776	0.048092
			GO:0032550	purine ribonucleoside binding	4.78	0.049923
Group 34						
DNAH5	0.31248	1.03E-10	GO:0044421	extracellular region part	5.975	0.008713
MTMR4	0.22358	2.78E-07	GO:0005615	extracellular space	9.1923	0.016563
UBR1	0.18509	3.66E-05	GO:0016942	insulin-like growth factor binding protein	9.1923	0.016563
VASH1	0.24377	3.45E-07		complex		
CEP128	0.20884	5.58E-06	GO:0042571	immunoglobulin complex, circulating	9.1923	0.016563
TMPPE	0.23707	1.09E-07	GO:0072562	blood microparticle	9.1923	0.016563
			GO:0034358	plasma lipoprotein particle	8.8519	0.017784
			GO:0005577	fibrinogen complex	5.975	0.036619
			GO:0005601	classical-complement-pathway C3/C5 convertase	5.975	0.036619
				complex		
			GO:0005602	complement component C1 complex	5.975	0.036619
			GO:0005616	larval serum protein complex	5.975	0.036619
Group 35						
ATXN1L	0.23998	2.10E-05	GO:0001059	transcription from RNA polymerase IV promoter	6.9612	0.00040488
HIPK1	0.3148	1.71E-07	GO:0001060	transcription from RNA polymerase V promoter	6.9612	0.00040488
NPAT	0.3859	8.74E-12	GO:0001121	transcription from bacterial-type RNA	6.9612	0.00040488
ZNF549	0.30417	5.53E-08		polymerase promoter		
			GO:0006390	transcription from mitochondrial promoter	6.9612	0.00040488
			GO:0009300	antisense RNA transcription	6.9612	0.00040488
			GO:0009301	snRNA transcription	6.9612	0.00040488
			GO:0009302	snoRNA transcription	6.9612	0.00040488
			GO:0009304	tRNA transcription	6.9612	0.00040488
			GO:0039695	DNA-templated viral transcription	6.9612	0.00040488
			GO:0042793	transcription from plastid promoter	6.9612	0.00040488
Group 36						
DIP2A	0.26422	0.00010475				
KLHL29	0.341	2.63E-06				
ARHGAP20	0.24458	0.00011163				

Group 37						
NRIP3	0.29101	3.97E-05				
MYPN	0.4072	9.78E-10				
TONSL	0.29308	2.57E-05				
Group 38						
POLR3A	0.19644	0.00094014	GO:0005783	endoplasmic reticulum	11.95	0.00087323
CLOCK	0.21097	0.00021052	GO:0005790	smooth endoplasmic reticulum	15.587	0.0055713
CES2	0.26581	1.19E-06	GO:0009511	plasmodesmatal endoplasmic reticulum	15.587	0.0055713
TRAPPC9	0.31637	2.70E-07	GO:1990007	membrane stack	15.587	0.0055713
			GO:0005791	rough endoplasmic reticulum	14.9375	0.0060603
			GO:0016529	sarcoplasmic reticulum	14.34	0.0065683
			GO:0048515	spermatid differentiation	12.3621	0.0087865
			GO:0044432	endoplasmic reticulum part	9.43421	0.014818
			GO:0031490	chromatin DNA binding	9.19231	0.015573
			GO:0003682	chromatin binding	8.9625	0.016345
Group 39						
USP19	0.13855	0.015205	GO:0030968	endoplasmic reticulum unfolded protein	71.7	0.00023179
QTRTD1	0.22695	4.80E-05		response		
CUL7	0.16441	0.005697	GO:0034976	response to endoplasmic reticulum stress	59.75	0.00034671
RCOR1	0.19038	0.00097637	GO:0045840	positive regulation of mitosis	51.2143	0.00048403
			GO:0010564	regulation of cell cycle process	25.6071	0.0020563
			GO:0019941	modification-dependent protein catabolic	25.6071	0.0020563
				process		
			GO:0043162	ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway	23.9	0.0023659
			GO:0030579	ubiquitin-dependent SMAD protein catabolic	22.4063	0.0026961
				process		
			GO:0071947	protein deubiquitination involved in ubiquitin-dependent protein catabolic process	22.4063	0.0026961
			GO:0016598	protein arginylation	21.0882	0.0030469
			GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	17.0714	0.0046511
Group 40						
HIRA	0.16091	0.0037827	GO:1902064	regulation of transcription from RNA polymerase	15.8162	0.00037427
WDR48	0.1562	0.0090939		II promoter involved in spermatogenesis		
ETV3	0.24553	1.21E-05	GO:0060349	bone morphogenesis	44.8125	0.00064355
ZNF304	0.2347	1.05E-05	GO:0009887	organ morphogenesis	35.85	0.0010284
			GO:0042692	muscle cell differentiation	18.8684	0.0038092
			GO:0017053	transcriptional repressor complex	17.0714	0.0046511
			GO:0000429	carbon catabolite regulation of transcription from RNA polymerase II promoter	15.587	0.0055713
			GO:0001079	nitrogen catabolite regulation of transcription from RNA polymerase II promoter	15.587	0.0055713
			GO:0001196	regulation of mating-type specific transcription from RNA polymerase II promoter	15.587	0.0055713
			GO:0010672	regulation of transcription from RNA polymerase II promoter during meiosis	15.587	0.0055713
			GO:0034243	regulation of transcription elongation from RNA polymerase II promoter	15.587	0.0055713
Group 41						
UNC119B	0.16451	0.0006184	GO:0046939	nucleotide phosphorylation	31.8667	0.0013616
RALGAPA2	0.17377	0.00046874	GO:0036038	TCTN-B9D complex	19.12	0.0038712
SMG1	0.18851	0.0001355	GO:0097537	Y-shaped link	19.12	0.0038712
STXBP1	0.24975	1.40E-05	GO:0097538	ciliary necklace	17.925	0.0044053
MAGI3	0.11513	0.017647	GO:0042953	lipoprotein transport	15.0947	0.0061972
			GO:1902113	nucleotide phosphorylation involved in DNA	11.95	0.0097888
				repair		
			GO:0015031	protein transport	8.69091	0.018005
			GO:0016323	basolateral plasma membrane	4.13654	0.021936
			GO:0042383	sarcolemma	4.13654	0.021936
			GO:0097524	sperm plasma membrane	4.13654	0.021936
Group 42						
RALGAPA1P	0.17891	9.75E-05	GO:0003833	beta-alanyl-dopamine synthase activity	13.0364	0.0082641

RNF123	0.20081	0.0002179	GO:0008754	O antigen ligase activity	13.0364	0.0082641
SNPH	0.16486	0.0017245	GO:0016877	ligase activity, forming carbon-sulfur bonds	13.0364	0.0082641
UBR3	0.15057	0.0019124	GO:0016885	ligase activity, forming carbon-carbon bonds	13.0364	0.0082641
CLUH	0.13164	0.0069444	GO:0045438	delta-(L-alpha-aminoadipyl)-L-cysteinyl-D-valine synthetase activity	13.0364	0.0082641
			GO:0051002	ligase activity, forming nitrogen-metal bonds	13.0364	0.0082641
			GO:0016875	ligase activity, forming carbon-oxygen bonds	12.4696	0.009012
			GO:0016886	ligase activity, forming phosphoric ester bonds	12.4696	0.009012
			GO:0016879	ligase activity, forming carbon-nitrogen bonds	11.95	0.0097888
			GO:0007014	actin ubiquitination	10.6222	0.012288
Group 43						
OSBPL5	0.11814	0.00012884	GO:0030724	testicular fusome organization	14.9375	0.00070533
LIG3	0.15825	8.33E-07	GO:0030726	male germline ring canal formation	14.9375	0.00070533
USP6NL	0.19619	5.74E-13	GO:0048136	male germ-line cyst formation	14.9375	0.00070533
CIT	0.15982	1.62E-07	GO:0048140	male germ-line cyst encapsulation	14.9375	0.00070533
SHROOM2	0.15421	5.56E-08	GO:0007285	primary spermatocyte growth	13.7885	0.00090523
WDR37	0.18706	3.40E-09	GO:0048133	male germ-line stem cell division	13.7885	0.00090523
ULK2	0.19552	4.53E-11	GO:0048232	male gamete generation	13.7885	0.00090523
VAT1L	0.090157	0.0014606	GO:0007283	spermatogenesis	12.8036	0.0011374
ITPR1	0.23149	5.22E-14	GO:0007284	spermatogonial cell division	12.8036	0.0011374
AKAP9	0.21858	1.89E-11	GO:0048137	spermatocyte division	12.8036	0.0011374
DGKI	0.17701	2.09E-10				
LPIN3	0.14117	1.07E-06				
Group 44						
TIMELESS	0.24508	4.05E-05	GO:0044768	NMS complex interaction involved in	48.8864	1.07E-05
CDK5RAP2	0.22647	8.17E-05		chromosome segregation		
MDN1	0.29418	1.11E-06	GO:0008608	attachment of spindle microtubules to	71.7	0.00023179
CTCF	0.20384	0.00029177		kinetochore		
			GO:0051304	chromosome separation	71.7	0.00023179
			GO:0051310	metaphase plate congression	71.7	0.00023179
			GO:0007059	chromosome segregation	59.75	0.00034671
			GO:0000819	sister chromatid segregation	51.2143	0.00048403
			GO:0045132	meiotic chromosome segregation	51.2143	0.00048403
			GO:0001013	RNA polymerase I regulatory region DNA binding	44.8125	0.00064355
			GO:0001016	RNA polymerase III regulatory region DNA	44.8125	0.00064355
				binding		
			GO:0001017	bacterial-type RNA polymerase regulatory region DNA binding	44.8125	0.00064355
Group 45						
CBX5	0.18463	0.0062315	GO:0034708	methyltransferase complex	119.5	6.99E-05
BAG5	0.35011	1.88E-07	GO:0044666	MLL3/4 complex	119.5	6.99E-05
PAXIP1	0.28907	1.71E-05	GO:0070311	nucleosomal methylation activator complex	119.5	6.99E-05
			GO:0048188	Set1C/COMPASS complex	95.6	0.00011638
			GO:0070688	MLL5-L complex	95.6	0.00011638
			GO:0031396	regulation of protein ubiquitination	79.66667	0.00017433
			GO:0035097	histone methyltransferase complex	79.66667	0.00017433
			GO:0035098	ESC/E(Z) complex	79.66667	0.00017433
			GO:0044665	MLL1/2 complex	68.28571	0.00024372
			GO:0046787	viral DNA repair	9.372549	0.01388
Group 46						
AP4E1	0.18726	0.0051833				
SPRED2	0.095936	0.14602				
BRD1	0.22747	0.00094952				
Group 47						
NCAPD3	0.23663	3.18E-08	GO:0035064	methylated histone residue binding	15.7582	0.005904
HPS4	0.2124	2.97E-06	GO:0006342	chromatin silencing	6.82857	0.029128
DIP2C	0.22661	4.57E-08	GO:0007068	negative regulation of transcription during	6.82857	0.029128
NAV1	0.23632	1.43E-07		mitosis		
ATF7IP	0.18765	3.73E-05	GO:0010620	negative regulation of transcription by	6.82857	0.029128
KDM4A	0.29712	4.73E-11		transcription factor catabolism		
SEC16A	0.19928	1.00E-06	GO:0010678	negative regulation of cellular carbohydrate	6.82857	0.029128

				metabolic process by negative regulation of		
				transcription, DNA-templated		
			GO:0010944	negative regulation of transcription by	6.82857	0.029128
				competitive promoter binding		
			GO:0016480	negative regulation of transcription from RNA	6.82857	0.029128
				polymerase III promoter		
			GO:0019043	establishment of viral latency	6.82857	0.029128
			GO:0019044	maintenance of viral latency	6.82857	0.029128
			GO:0019045	latent virus replication	6.82857	0.029128
Group 48						
KIF13A	0.22632	0.0022254	GO:0016874	ligase activity	15.9333	0.0048849
MIB1	0.24066	0.0012763	GO:0000242	pericentriolar material	14.4848	0.0059034
LSS	0.21999	0.0016119	GO:0031592	centrosomal corona	14.4848	0.0059034
			GO:0034451	centriolar satellite	14.4848	0.0059034
			GO:0005814	centriole	13.2778	0.0070129
			GO:0009058	biosynthetic process	12.9189	0.0074028
			GO:0010166	wax metabolic process	12.9189	0.0074028
			GO:0044033	multi-organism metabolic process	12.9189	0.0074028
			GO:1901902	tyrocidine metabolic process	12.9189	0.0074028
			GO:1902421	hydrogen metabolic process	12.9189	0.0074028
Group 49						
USP9X	0.17489	0.0011094	GO:0048846	axon extension involved in axon guidance	22.6421	0.00015137
BRD3	0.26036	5.05E-07	GO:0007409	axonogenesis	17.925	0.00031166
SRGAP1	0.18428	0.00039494	GO:0045211	postsynaptic membrane	57.36	0.00038469
MPDZ	0.23776	1.60E-05	GO:0097060	synaptic membrane	47.8	0.00057461
SEMA4F	0.20389	1.85E-05	GO:0030426	growth cone	22.0615	0.0029004
			GO:0001764	neuron migration	20.4857	0.0033694
			GO:0045177	apical part of cell	19.12	0.0038712
			GO:0016198	axon choice point recognition	17.925	0.0044053
			GO:0021966	corticospinal neuron axon guidance	17.925	0.0044053
			GO:0021967	corticospinal neuron axon guidance through the cerebral cortex	17.925	0.0044053
Group 50						
EGFR	0.27868	0.00019471	GO:0044802	single-organism membrane organization	47.8	0.00052005
GBF1	0.21838	0.0021989	GO:0005925	focal adhesion	28.1176	0.0015561
FLNB	0.26817	0.00020147	GO:0038109	Kit signaling pathway	28.1176	0.0015561
			GO:0060706	cell differentiation involved in embryonic	28.1176	0.0015561
				placenta development		
			GO:0005924	cell-substrate adherens junction	25.1579	0.001951
			GO:0060479	lung cell differentiation	25.1579	0.001951
			GO:0003779	actin binding	20.7826	0.0028701
			GO:0019221	cytokine-mediated signaling pathway	20.7826	0.0028701
			GO:0001649	osteoblast differentiation	17.7037	0.0039589
			GO:0051015	actin filament binding	14.4848	0.0059034
Group 51						
GTF3C1	0.21493	5.10E-06	GO:0005768	endosome	6.9951	0.026992
ZFPM2	0.20307	1.20E-05	GO:0044425	membrane part	2.6679	0.027907
WDR44	0.13728	0.0082852	GO:0006412	translation	6.2348	0.033317
EXOC2	0.24797	1.36E-07	GO:0005628	prospore membrane	3.4694	0.034957
TANGO6	0.25826	3.20E-07	GO:0034357	photosynthetic membrane	3.4694	0.034957
			GO:0036362	ascus membrane	3.4694	0.034957
			GO:0055036	virion membrane	3.4694	0.034957
			GO:0098533	ATPase dependent transmembrane transport	3.4694	0.034957
				complex		
			GO:1990195	macrolide transmembrane transporter complex	3.4694	0.034957
			GO:1990196	MacAB-TolC complex	3.4694	0.034957
Group 52						
TRO	0.32765	1.12E-05	GO:0005737	cytoplasm	3.2443	0.029008
ASH1L	0.32334	3.25E-06	GO:0045495	pole plasm	3.2443	0.029008
PHF2	0.24918	0.00043579	GO:0016528	sarcoplasm	3.2297	0.029406
			GO:0098593	goblet cell theca	3.1867	0.030619
			GO:0008270	zinc ion binding	5.6905	0.03607
Group 53						
ST3GAL2	0.18303	0.00036332				
RAI1	0.20915	3.29E-06				
DIS3L	0.18899	0.00044983				

C16orf72	0.26051	1.06E-07				
YEATS2	0.20576	3.24E-05				
Group 54						
NUMA1	0.20666	9.84E-06	GO:0033206	meiotic cytokinesis	44.8125	1.79E-05
EYA3	0.18362	6.04E-05	GO:0051294	establishment of spindle orientation	79.6667	0.00017433
ARHGEF11	0.15524	0.00079964	GO:0000912	assembly of actomyosin apparatus involved in	34.1429	0.0011931
SON	0.21923	9.17E-06		cytokinesis		
NFAT5	0.16762	0.00066217	GO:0000916	actomyosin contractile ring contraction	34.1429	0.0011931
SVEP1	0.22156	3.06E-06	GO:0000917	barrier septum assembly	34.1429	0.0011931
			GO:0007105	cytokinesis, site selection	34.1429	0.0011931
			GO:0010069	zygote asymmetric cytokinesis in embryo sac	34.1429	0.0011931
			GO:0031566	actomyosin contractile ring maintenance	34.1429	0.0011931
			GO:0032187	actomyosin contractile ring localization	34.1429	0.0011931
			GO:0043093	cytokinesis by binary fission	34.1429	0.0011931
Group 55						
MAP3K5	0.17314	0.00024059	GO:0042803	protein homodimerization activity	7.31633	0.004931
NACC2	0.27447	5.98E-10	GO:0033392	actin homodimerization activity	11.95	0.010024
COL7A1	0.25594	3.61E-09	GO:0019538	protein metabolic process	8.85185	0.017784
UBE4A	0.2703	1.41E-09	GO:0044432	endoplasmic reticulum part	6.28947	0.033397
B4GALT5	0.2124	7.91E-07				
LRRC2	0.2683	2.94E-09				
Group 56						
ITGA2	0.12191	0.037904	GO:0001707	mesoderm formation	89.625	0.00013947
ABCA2	0.16067	0.01032	GO:0006928	cellular component movement	51.2143	0.00048403
NFX1	0.15926	0.0057238	GO:0007018	microtubule-based movement	51.2143	0.00048403
KIF16B	0.19963	0.00066367	GO:0002246	wound healing involved in inflammatory	44.8125	0.00064355
				response		
			GO:0009725	response to hormone	39.8333	0.00082509
			GO:0071383	cellular response to steroid hormone stimulus	39.8333	0.00082509
			GO:0009611	response to wounding	32.5909	0.0012534
			GO:0014070	response to organic cyclic compound	29.875	0.0014999
			GO:0015893	drug transport	29.875	0.0014999
			GO:0036270	response to diuretic	29.875	0.0014999
Group 57						
HERC1	0.25306	0.0004622	GO:0019787	small conjugating protein ligase activity	22.7619	0.0023891
ARIH1	0.34277	4.53E-07	GO:0050372	ubiquitin-calmodulin ligase activity	22.7619	0.0023891
REV1	0.34463	4.55E-07	GO:0090302	mitotic anaphase-promoting complex activity	22.7619	0.0023891
			GO:0004842	ubiquitin-protein ligase activity	20.7826	0.0028701
			GO:0034450	ubiquitin-ubiquitin ligase activity	20.7826	0.0028701
			GO:0007014	actin ubiquitination	17.7037	0.0039589
			GO:0018276	isopeptide cross-linking via N6-glycyl-L-lysine	17.7037	0.0039589
			GO:0016574	histone ubiquitination	17.0714	0.0042572
			GO:0032446	protein modification by small protein conjugation	17.0714	0.0042572
			GO:0051865	protein autoubiquitination	17.0714	0.0042572
Group 58						
DMD	0.3288	1.12E-07	GO:0045892	negative regulation of transcription, DNA-templated	5.736	0.00088748
PER3	0.39252	1.29E-11				
HIVEP1	0.23019	6.09E-05	GO:0006357	regulation of transcription from RNA polymerase	4.62581	0.0021181
ZKSCAN2	0.25103	3.12E-06		II promoter		
			GO:0000122	negative regulation of transcription from RNA	8.02612	0.0028515
				polymerase II promoter		
			GO:0045202	synapse	14.34	0.0065683
			GO:0001059	transcription from RNA polymerase IV promoter	5.22087	0.0099438
			GO:0001060	transcription from RNA polymerase V promoter	5.22087	0.0099438
			GO:0001121	transcription from bacterial-type RNA polymerase promoter	5.22087	0.0099438
			GO:0006390	transcription from mitochondrial promoter	5.22087	0.0099438
			GO:0009300	antisense RNA transcription	5.22087	0.0099438
			GO:0009301	snRNA transcription	5.22087	0.0099438
Group 59						

FAM63B	0.18711	0.011732	GO:0044421	extracellular region part	7.9667	0.019009
ZNF45	0.18568	0.0065738	GO:0001071	nucleic acid binding transcription factor activity	6.8286	0.025541
WDR60	0.14736	0.026885				
Group 60						
WDHD1	0.27251	0.0001469				
FAM171B	0.09434	0.17125				
RP11-958N24.2	0.19585	0.0028029				
Group 61						
GPR107	0.2916	5.69E-16	GO:0035774	positive regulation of insulin secretion involved	71.7	0.00017531
EVC	0.29959	1.10E-15		in cellular response to glucose stimulus		
TTL	0.18327	1.92E-08	GO:0060350	endochondral bone morphogenesis	71.7	0.00017531
PRKAB2	0.30497	3.83E-17	GO:0061179	negative regulation of insulin secretion involved	35.85	0.0010284
SCUBE3	0.16656	3.51E-07		in cellular response to glucose stimulus		
PDGFD	0.26059	7.17E-13	GO:0048046	apoplast	11.95	0.0013787
AQR	0.1914	1.01E-08	GO:0098595	perivitelline space	11.95	0.0013787
LRP5	0.14599	1.87E-05	GO:0032844	regulation of homeostatic process	28.68	0.0016948
NOTCH3	0.20609	4.37E-09	GO:0048583	regulation of response to stimulus	28.68	0.0016948
SIDT2	0.29256	7.77E-16	GO:0061448	connective tissue development	28.68	0.0016948
			GO:0002947	tumor necrosis factor receptor superfamily	20.4857	0.0034796
				complex		
			GO:0008230	ecdysone receptor holocomplex	20.4857	0.0034796
Group 62						
ZBTB41	0.32069	2.36E-06	GO:0007532	regulation of mating-type specific transcription,	5.4943	0.038525
ZNFX1	0.29225	3.08E-05		DNA-templated		
TGS1	0.29062	2.46E-05	GO:0009373	regulation of transcription by pheromones	5.4943	0.038525
			GO:0019056	modulation by virus of host transcription	5.4943	0.038525
			GO:0031554	regulation of DNA-templated transcription,	5.4943	0.038525
				termination		
			GO:0032784	regulation of DNA-templated transcription,	5.4943	0.038525
				elongation		
			GO:0043620	regulation of DNA-templated transcription in response to stress	5.4943	0.038525
			GO:0044324	regulation of transcription involved in anterior/posterior axis specification	5.4943	0.038525
			GO:0045896	regulation of transcription during mitosis	5.4943	0.038525
			GO:0045990	carbon catabolite regulation of transcription	5.4943	0.038525
			GO:0060194	regulation of antisense RNA transcription	5.4943	0.038525
Group 63						
DSEL	0.2036	0.0058011				
USP24	0.16134	0.028077				
POGZ	0.034178	0.64001				
Group 64						
C5orf51	0.21485	0.00010961				
FKBP5	0.22057	2.75E-05				
PCMTD2	0.17074	0.0038017				
ZNF687	0.08065	0.17439				
Group 65						
PARP14	0.18938	0.00036007	GO:0007267	cell-cell signaling	11.95	0.0097888
CSTF2T	0.18011	0.00038463	GO:0005628	prospore membrane	3.46935	0.034957
KREMEN1	0.25373	6.89E-07	GO:0034357	photosynthetic membrane	3.46935	0.034957
USP38	0.33489	3.54E-11	GO:0036362	ascus membrane	3.46935	0.034957
ADCY6	0.25099	1.38E-06	GO:0055036	virion membrane	3.46935	0.034957
			GO:0098533	ATPase dependent transmembrane transport	3.46935	0.034957
				complex		
			GO:1990195	macrolide transmembrane transporter complex	3.46935	0.034957
			GO:1990196	MacAB-TolC complex	3.46935	0.034957
			GO:0019867	outer membrane	3.4416	0.035696
			GO:0048475	coated membrane	3.4416	0.035696
Group 66						

ITGA11	0.20002	5.18E-05				
TSPYL4	0.14524	0.002086				
KLF13	0.17479	0.00070649				
TRIM62	0.16987	0.0012177				
RP11-159D12.9	0.17903	0.00016301				
Group 67						
ANKRD50	0.18186	2.41E-06	GO:0070307	lens fiber cell development	68.2857	0.00024372
REV3L	0.14946	0.00026029	GO:0070309	lens fiber cell morphogenesis	51.2143	0.00048403
EPHA2	0.1316	0.00079411	GO:0002089	lens morphogenesis in camera-type eye	18.6234	0.0042227
NUP153	0.21726	4.26E-07	GO:0044339	canonical Wnt signaling pathway involved in	14.6327	0.0068391
TDRD7	0.20162	8.11E-07		osteoblast differentiation		
TMEM246	0.21416	3.42E-07	GO:0009653	anatomical structure morphogenesis	10.2429	0.013679
BCL9	0.14438	0.00024164	GO:0048468	cell development	9.31169	0.016393
			GO:0019043	establishment of viral latency	6.82857	0.029128
			GO:0019044	maintenance of viral latency	6.82857	0.029128
			GO:0019045	latent virus replication	6.82857	0.029128
			GO:0019061	uncoating of virus	6.82857	0.029128
Group 68						
UBE2O	0.27696	2.75E-05	GO:0016197	endosomal transport	34.1429	0.0010457
ALMS1	0.30081	1.20E-05	GO:0042147	retrograde transport, endosome to Golgi	31.8667	0.0012049
PHF16	0.37509	1.79E-08				
Group 69						
PDS5A	0.17347	0.0014322	GO:0031518	CBF3 complex	14.34	0.0068562
PRRC2B	0.1512	0.0073809	GO:0000776	kinetochore	10.2429	0.013175
KIDINS220	0.26316	8.32E-08	GO:0034995	SC5b-7 complex	8.43529	0.019047
ANPEP	0.22004	4.92E-05	GO:0034997	alphav-beta5 integrin-vitronectin complex	8.43529	0.019047
ZNF516	0.19902	6.39E-05	GO:1990270	platelet-derived growth factor receptor-ligand	8.43529	0.019047
				complex		
			GO:0005577	fibrinogen complex	7.17	0.025789
			GO:0005601	classical-complement-pathway C3/C5 convertase	7.17	0.025789
				complex		
			GO:0005602	complement component C1 complex	7.17	0.025789
			GO:0005616	larval serum protein complex	7.17	0.025789
			GO:0031395	bursicon neuropeptide hormone complex	7.17	0.025789
Group 70						
UBN1	0.2747	5.70E-05	GO:0019080	viral gene expression	8.386	0.017218
RFX7	0.38405	2.06E-07	GO:0005654	nucleoplasm	5.9012	0.033683
GEMIN4	0.32654	9.24E-07	GO:0044451	nucleoplasm part	4.9792	0.046288
Group 71						
DCBLD1	0.18062	4.13E-06	GO:0001113	transcriptional open complex formation at RNA	14.9375	0.0066141
ZSWIM6	0.22366	4.53E-09		polymerase II promoter		
TAF2	0.11788	0.0019934	GO:0001174	transcriptional start site selection at RNA	14.9375	0.0066141
PLXNA1	0.14078	0.00018419		polymerase II promoter		
FAM129A	0.16626	7.99E-06	GO:0051123	RNA polymerase II transcriptional preinitiation	14.9375	0.0066141
RHOBTB2	0.13503	0.00022612		complex assembly		
SLC7A6OS	0.1548	6.15E-05	GO:0030334	regulation of cell migration	10.5441	0.013057
MAML2	0.19528	9.99E-07	GO:0097485	neuron projection guidance	10.5441	0.013057
			GO:0007409	axonogenesis	7.46875	0.024942
Group 72						
EIF4EBP2	0.16972	0.00054146	GO:0003713	transcription coactivator activity	10.6222	0.012288
MED23	0.15602	0.0015524	GO:0070940	dephosphorylation of RNA polymerase II C-	8.9625	0.016988
PTPN14	0.15643	0.0022106		terminal domain		
CACNA1C	0.20726	1.53E-05	GO:0003712	transcription cofactor activity	8.19429	0.020112
NHLRC2	0.19182	4.97E-05	GO:0006412	translation	6.23478	0.033317
			GO:0010467	gene expression	3.18667	0.043537
Group 73						
FAM13B	0.29239	4.73E-10	GO:0070577	histone acetyl-lysine binding	29.875	0.0015819
OSBPL2	0.1899	0.00015073	GO:0031493	nucleosomal histone binding	26.5556	0.0020224
IPO9	0.1916	1.33E-05	GO:0035064	methylated histone residue binding	18.3846	0.0042834
C17orf51	0.21847	7.04E-07	GO:0060055	angiogenesis involved in wound healing	14.9375	0.0064781
MAP3K1	0.18753	4.50E-05	GO:0051019	mitogen-activated protein kinase binding	10.3913	0.013119
SIRT1	0.17921	1.83E-05	GO:0019901	protein kinase binding	8.85185	0.017784
			GO:0019900	kinase binding	6.63889	0.030284

			GO:0006915	apoptotic process	6.12821	0.034995
			GO:0006975	DNA damage induced protein phosphorylation	5.31111	0.045117
			GO:0097411	hypoxia-inducible factor-1alpha signaling pathway	5.08511	0.048679
Group 74						
SLC12A2	0.25233	2.08E-06	GO:0000242	pericentriolar material	10.8636	0.011294
WRN	0.29401	2.01E-08	GO:0031592	centrosomal corona	10.8636	0.011294
FAM168A	0.18812	0.0012081	GO:0034451	centriolar satellite	10.8636	0.011294
OFD1	0.22028	5.43E-05	GO:0005814	centriole	9.95833	0.013358
			GO:0005813	centrosome	9.43421	0.014818
			GO:0031616	spindle pole centrosome	9.43421	0.014818
			GO:0005815	microtubule organizing center	8.14773	0.019589
			GO:1902589	single-organism organelle organization	6.76415	0.027776
Group 75						
CEP192	0.24863	3.38E-09	GO:0019902	phosphatase binding	11.381	0.011174
SUGP2	0.17132	0.0001743	GO:0007049	cell cycle	5.53668	0.042363
KANK1	0.20559	1.21E-06	GO:0043229	intracellular organelle	5.53668	0.042363
FAM208B	0.20695	1.87E-06				
ENTPD7	0.22226	7.69E-08				
STK10	0.2005	2.47E-07				
NCOA6	0.24689	6.95E-10				
Group 76						
ZCCHC11	0.18008	0.0018678	GO:0009966	regulation of signal transduction	32.5909	0.0012534
TRPS1	0.27062	1.30E-05	GO:0048731	system development	9.19231	0.015573
LRCH2	0.22008	0.0002795	GO:0065010	extracellular membrane-bounded organelle	5.12143	0.046217
SEMA5A	0.27022	1.65E-05	GO:0070062	extracellular vesicular exosome	5.12143	0.046217
			GO:0031988	membrane-bounded vesicle	4.91096	0.049827
Group 77						
PIK3CA	0.15267	8.90E-06	GO:0002523	leukocyte migration involved in inflammatory	17.7037	0.0047116
RWDD2A	0.19227	2.39E-08		response		
PDE10A	0.22224	1.08E-09	GO:1901215	negative regulation of neuron death	14.4848	0.0070565
LLGL1	0.16604	1.39E-06	GO:0051259	protein oligomerization	11.381	0.011331
ATP11A	0.14906	2.99E-05	GO:0006461	protein complex assembly	9.95833	0.014646
HELZ2	0.21828	2.47E-09	GO:0032559	adenyl ribonucleotide binding	3.46377	0.016653
TNFAIP3	0.17494	2.49E-07	GO:0005765	lysosomal membrane	7.96667	0.022275
FOXK1	0.18562	4.31E-07	GO:0007599	hemostasis	7.5873	0.024372
VANGL1	0.24669	6.93E-12	GO:0050817	coagulation	7.5873	0.024372
			GO:0071892	thrombocyte activation	7.5873	0.024372
			GO:0072378	blood coagulation, fibrin clot formation	7.5873	0.024372
Group 78						
SNX30	0.24293	0.00072885	GO:0097190	apoptotic signaling pathway	6.2078	0.03061
TEX2	0.29305	1.17E-05	GO:0009319	cytochrome o ubiquinol oxidase complex	5.2527	0.041903
TOR4A	0.16954	0.010936	GO:0031361	integral component of thylakoid membrane	5.2527	0.041903
			GO:0034702	ion channel complex	5.2527	0.041903
			GO:0045203	integral component of cell outer membrane	5.2527	0.041903
			GO:1902495	transmembrane transporter complex	5.2527	0.041903
			GO:0031224	intrinsic component of membrane	5.1957	0.042765
			GO:0031301	integral component of organelle membrane	5.1398	0.043635
			GO:0046696	lipopolysaccharide receptor complex	5.0851	0.044512
Group 79						
IGF2R	0.22815	1.33E-07	GO:0035166	post-embryonic hemopoiesis	38.41071	3.11E-05
RYBP	0.18617	2.91E-06	GO:0001709	cell fate determination	102.4286	8.18E-05
KIAA0355	0.1356	0.00088795	GO:0001713	ectodermal cell fate determination	102.4286	8.18E-05
TMEM110	0.15156	9.13E-05	GO:0003137	Notch signaling pathway involved in heart	102.4286	8.18E-05
NOTCH2	0.25493	2.04E-08		induction		
DHX8	0.17055	3.44E-05	GO:0003177	pulmonary valve development	102.4286	8.18E-05
JAG1	0.22592	8.61E-08	GO:0003184	pulmonary valve morphogenesis	102.4286	8.18E-05
			GO:0003193	pulmonary valve formation	102.4286	8.18E-05
			GO:0003270	Notch signaling pathway involved in regulation of secondary heart field cardioblast proliferation	102.4286	8.18E-05
			GO:0007403	glial cell fate determination	102.4286	8.18E-05
			GO:0007425	epithelial cell fate determination, open tracheal system	102.4286	8.18E-05

Group 80						
ARID2	0.21282	0.00046048	GO:0046914	transition metal ion binding	3.9396	0.0040499
KDM1B	0.31269	2.80E-07	GO:0003824	catalytic activity	5.3507	0.042706
TRIM35	0.25779	1.30E-05				
ANKFY1	0.3302	1.09E-06				
Group 81						
DAAM2	0.30082	3.61E-09	GO:0071910	determination of liver left/right asymmetry	22.0615	0.0029004
TBC1D9	0.27459	3.50E-08	GO:1900094	regulation of transcription from RNA polymerase	10.6222	0.012288
STAT5B	0.30809	6.44E-10		II promoter involved in determination of		
FBXO30	0.24378	1.86E-06		left/right symmetry		
CRYBG3	0.28559	3.46E-08				
Group 82						
SIPA1L3	0.32062	1.27E-26				
EXOC6B	0.16019	1.29E-07				
ADNP	0.16986	1.83E-10				
RBM19	0.26453	8.58E-19				
TCF20	0.19702	5.53E-12				
INPP5F	0.28947	2.68E-23				
TMEM260	0.16373	5.10E-10				
E2F3	0.14055	7.98E-07				
HCFC1	0.20538	2.13E-12				
FRY	0.27592	6.46E-22				
INTS3	0.1454	2.82E-07				
AKAP11	0.28845	1.06E-22				
RPAP1	0.19626	3.35E-11				
UBE4B	0.2238	7.90E-16				
Group 83						
CCSER2	0.30071	2.06E-09	GO:0001180	transcription initiation from RNA polymerase I	119.5	5.84E-05
TBC1D13	0.31105	4.46E-12		promoter for nuclear large rRNA transcript		
TCOF1	0.17903	0.00010342	GO:0001183	transcription elongation from RNA polymerase I	119.5	5.84E-05
POLR1A	0.31328	4.02E-11		promoter for nuclear large rRNA transcript		
KIAA0754	0.24761	4.85E-07	GO:0001185	termination of RNA polymerase I transcription	119.5	5.84E-05
DOCK5	0.32864	5.41E-12		from promoter for nuclear large rRNA transcript		
			GO:0042790	transcription of nuclear large rRNA transcript	79.66667	0.00017433
				from RNA polymerase I promoter		
			GO:0032320	positive regulation of Ras GTPase activity	6.128205	0.034995
Group 84						
PLEKHG3	0.36965	4.92E-08	GO:0005089	Rho guanyl-nucleotide exchange factor activity	53.1111	0.00041663
FGD1	0.27297	0.0002624	GO:0030676	Rac guanyl-nucleotide exchange factor activity	53.1111	0.00041663
POLG	0.31493	2.01E-05	GO:0043089	positive regulation of Cdc42 GTPase activity	47.8	0.00052005
			GO:0032319	regulation of Rho GTPase activity	39.8333	0.00076058
			GO:0032862	activation of Rho GTPase activity	39.8333	0.00076058
			GO:0032321	positive regulation of Rho GTPase activity	36.7692	0.00089759
			GO:0032855	positive regulation of Rac GTPase activity	36.7692	0.00089759
			GO:0005088	Ras guanyl-nucleotide exchange factor activity	29.875	0.001375
			GO:0010259	multicellular organismal aging	19.12	0.0033934
			GO:0009835	fruit ripening	18.3846	0.0036709
Group 85						
AMOTL1	0.25316	0.0010285				
ARID4A	0.16293	0.023184				
ASAP2	0.12943	0.076309				
Group 86						
C14orf132	0.22167	2.99E-05	GO:0046930	pore complex	5.16757	0.0023307
MARCH4	0.17859	0.00035853	GO:0033211	adiponectin-activated signaling pathway	19.12	0.0038712
VWA8	0.21665	4.11E-06	GO:0019221	cytokine-mediated signaling pathway	12.4696	0.009012
ACSL1	0.19401	0.00013585	GO:0016874	ligase activity	9.56	0.01503
NUP133	0.13405	0.0070374	GO:0009319	cytochrome o ubiquinol oxidase complex	4.72747	0.015262
			GO:0031361	integral component of thylakoid	4.72747	0.015262

				membrane		
			GO:0034702	ion channel complex	4.72747	0.015262
			GO:0045203	integral component of cell outer membrane	4.72747	0.015262
			GO:1902495	transmembrane transporter complex	4.72747	0.015262
			GO:0031224	intrinsic component of membrane	4.67609	0.015726
Group 87						
DHX37	0.20348	0.00074864	GO:0016070	RNA metabolic process	29.875	0.0014999
CDC42BPA	0.28637	4.11E-06	GO:0010737	protein kinase A signaling	16.2955	0.0051015
PRKCA	0.26686	3.84E-05	GO:0038008	TRAF-mediated signal transduction	16.2955	0.0051015
ZFC3H1	0.23218	0.00016927	GO:0038061	NIK/NF-kappaB signaling	16.2955	0.0051015
			GO:0039585	PKR signal transduction	16.2955	0.0051015
			GO:0070528	protein kinase C signaling	16.2955	0.0051015
			GO:2000803	endosomal signal transduction	16.2955	0.0051015
			GO:0006397	mRNA processing	15.587	0.0055713
			GO:0007249	I-kappaB kinase/NF-kappaB signaling	15.587	0.0055713
			GO:0019932	second-messenger-mediated signaling	15.587	0.0055713
Group 88						
ARID1A	0.21034	0.00039951	GO:0001726	ruffle	24.4432	9.80E-05
APC	0.22125	0.00035173	GO:0050772	positive regulation of axonogenesis	71.7	0.00023179
CACNA1A	0.34025	4.87E-08	GO:0097458	neuron part	17.3468	0.00028238
CDKL5	0.22483	0.00022154	GO:0030516	regulation of axon extension	59.75	0.00034671
			GO:0048856	anatomical structure development	13.7885	0.00056742
			GO:0031256	leading edge membrane	44.8125	0.00064355
			GO:0048667	cell morphogenesis involved in neuron differentiation	44.8125	0.00064355
			GO:0031253	cell projection membrane	39.8333	0.00082509
			GO:0032839	dendrite cytoplasm	39.8333	0.00082509
			GO:0044292	dendrite terminus	39.8333	0.00082509
Group 89						
DUSP16	0.27519	0.00010692	GO:0004722	protein serine/threonine phosphatase activity	68.2857	0.00024372
ALPK1	0.28486	4.81E-05	GO:0035970	peptidyl-threonine dephosphorylation	53.1111	0.00041663
PHLPP2	0.23151	0.0023188	GO:0035971	peptidyl-histidine dephosphorylation	53.1111	0.00041663
			GO:0070262	peptidyl-serine dephosphorylation	53.1111	0.00041663
			GO:0098628	peptidyl-N-phospho-arginine dephosphorylation	53.1111	0.00041663
			GO:0006464	cellular protein modification process	12.5789	0.00047829
			GO:0016576	histone dephosphorylation	47.8	0.00052005
			GO:0035335	peptidyl-tyrosine dephosphorylation	39.8333	0.00076058
			GO:0004721	phosphoprotein phosphatase activity	36.7692	0.00089759
			GO:0016311	dephosphorylation	31.8667	0.0012049
Group 90						
CARD6	0.23982	3.08E-07	GO:0012501	programmed cell death	11.627	0.0011519
ZNF592	0.32022	7.05E-10	GO:0016265	death	28.68	0.0016948
B3GALTL	0.16705	0.00091052	GO:0044763	single-organism cellular process	5.41132	0.0019492
CNTLN	0.26542	5.44E-08	GO:0008219	cell death	26.0727	0.0020627
BICD2	0.22128	5.37E-06	GO:0070265	necrotic cell death	26.0727	0.0020627
			GO:0070997	neuron death	26.0727	0.0020627
Group 91						
KPNA4	0.21	4.30E-09	GO:0045165	cell fate commitment	10.755	0.0018877
PRKDC	0.18857	3.48E-08	GO:0035051	cardiocyte differentiation	10.2429	0.0021802
USP34	0.21378	3.62E-10	GO:0000245	spliceosomal complex assembly	20.4857	0.0034796
ZNF521	0.20146	4.61E-09	GO:0000365	mRNA trans splicing, via spliceosome	20.4857	0.0034796
SLC25A44	0.12786	8.31E-05	GO:0000377	RNA splicing, via transesterification reactions	20.4857	0.0034796
CSPP1	0.16813	1.71E-07	GO:0000380	alternative mRNA splicing, via spliceosome	20.4857	0.0034796
CRNKL1	0.23213	4.36E-11	GO:0000390	spliceosomal complex disassembly	20.4857	0.0034796
ZHX3	0.20435	1.33E-08	GO:0000393	spliceosomal conformational changes to	20.4857	0.0034796
DHX38	0.19715	2.26E-08		generate catalytic conformation		
GM2A	0.083797	0.0086792	GO:0045292	mRNA cis splicing, via spliceosome	20.4857	0.0034796
			GO:0071007	U2-type catalytic step 2 spliceosome	20.4857	0.0034796
Group 92						
GTF3C2	0.14761	0.025606	GO:0031981	nuclear lumen	4.2426	0.012917
URB1	0.18418	0.0095262	GO:0005654	nucleoplasm	5.9012	0.033683
MCM8	0.16125	0.029833	GO:0044451	nucleoplasm part	4.9792	0.046288
Group 93						
CAD	0.14775	0.01102	GO:0061174	type I terminal bouton	179.25	2.34E-05
MAP7D3	0.13987	0.013659	GO:0061175	type II terminal bouton	179.25	2.34E-05

DTX4	0.19613	0.000745	GO:0097467	type III terminal bouton	179.25	2.34E-05
UNC13B	0.20448	0.00050529	GO:1990024	C bouton	179.25	2.34E-05
			GO:1990025	F bouton	179.25	2.34E-05
			GO:1990026	hippocampal mossy fiber expansion	179.25	2.34E-05
			GO:1990027	S bouton	179.25	2.34E-05
			GO:0043195	terminal bouton	119.5	6.99E-05
			GO:0043679	axon terminus	59.75	0.00034671
			GO:0033267	axon part	32.59091	0.0012534
Group 94						
SPAST	0.32232	2.53E-06	GO:0005737	cytoplasm	3.2443	0.029008
PHF20	0.34208	8.53E-07	GO:0045495	pole plasm	3.2443	0.029008
MORC2	0.27755	6.71E-05	GO:0016528	sarcoplasm	3.2297	0.029406
			GO:0098593	goblet cell theca	3.1867	0.030619
			GO:0008270	zinc ion binding	5.6905	0.03607
Group 95						
FLNC	0.16171	4.42E-05	GO:0060864	positive regulation of floral organ abscission by	8.9625	0.017776
ARHGAP18	0.17866	3.94E-06		small GTPase mediated signal transduction		
AMPD3	0.17298	7.37E-06	GO:0060868	regulation of vesicle-mediated transport	8.9625	0.017776
EHBP1L1	0.14229	0.00077434		involved in floral organ abscission by small		
HERC2	0.28294	1.32E-12		GTPase mediated signal transduction		
MRAS	0.095296	0.009466	GO:0007265	Ras protein signal transduction	8.5357	0.019478
KDM3B	0.14194	0.00024655	GO:0032156	septin cytoskeleton	7.7935	0.023063
ENDOD1	0.18172	1.24E-06	GO:0043087	regulation of GTPase activity	7.7935	0.023063
			GO:0001533	cornified envelope	7.4688	0.024942
			GO:0030863	cortical cytoskeleton	7.4688	0.024942
			GO:0014731	spectrin-associated cytoskeleton	7.17	0.026877
			GO:0043547	positive regulation of GTPase activity	7.17	0.026877
			GO:0005628	prospore membrane	2.8911	0.028724
Group 96						
FOXK2	0.15043	0.0027261	GO:0044306	neuron projection terminus	22.0615	0.0029004
AKAP13	0.20227	0.00022303	GO:0006611	protein export from nucleus	15.9333	0.0055688
VTI1A	0.18903	0.00033156	GO:0048471	perinuclear region of cytoplasm	13.6571	0.0075453
FTSJD2	0.096861	0.058975	GO:0097038	perinuclear endoplasmic reticulum	13.6571	0.0075453
RAP1GAP2	0.15764	0.0024612	GO:0097458	neuron part	9.25161	0.015997
			GO:0043005	neuron projection	8.43529	0.019047
			GO:0032320	positive regulation of Ras GTPase activity	7.35385	0.024609
			GO:0005768	endosome	6.99512	0.026992
			GO:0033011	perinuclear theca	6.66977	0.02946
			GO:0005794	Golgi apparatus	5.62353	0.040131
Group 97						
KCNMA1	0.30029	8.57E-05	GO:0000182	rDNA binding	39.8333	0.00076058
CREBL2	0.19085	0.0077867	GO:0003680	AT DNA binding	39.8333	0.00076058
MYBBP1A	0.14267	0.03904	GO:0003693	P-element binding	39.8333	0.00076058
			GO:0003696	satellite DNA binding	39.8333	0.00076058
			GO:0031634	replication fork barrier binding	39.8333	0.00076058
			GO:0045322	unmethylated CpG binding	39.8333	0.00076058
			GO:0071443	tDNA binding	39.8333	0.00076058
			GO:0003688	DNA replication origin binding	36.7692	0.00089759
			GO:0019237	centromeric DNA binding	36.7692	0.00089759
			GO:0042162	telomeric DNA binding	36.7692	0.00089759
Group 98						
NUP205	0.24523	3.20E-06	GO:0005088	Ras guanyl-nucleotide exchange factor activity	17.925	0.0044053
PCNXL3	0.21956	2.14E-06	GO:0016482	cytoplasmic transport	13.0364	0.0082641
ARFGEF2	0.24122	7.45E-07	GO:0048489	synaptic vesicle transport	12.4696	0.009012
INTS1	0.33026	1.05E-09	GO:0046578	regulation of Ras protein signal transduction	11.472	0.010594
RASGRF2	0.3532	3.10E-12	GO:0005886	plasma membrane	2.85373	0.021869
			GO:0097193	intrinsic apoptotic signaling pathway	7.75135	0.022314
			GO:0042981	regulation of apoptotic process	7.54737	0.02345
			GO:0071806	protein transmembrane transport	7.54737	0.02345
			GO:0006915	apoptotic process	7.35385	0.024609
			GO:0032320	positive regulation of Ras GTPase activity	7.35385	0.024609
Group 99						
ASCC3	0.24069	3.99E-06	GO:0006307	DNA dealkylation involved in DNA repair	18.7043	0.00027349
UBR5	0.30206	3.37E-07	GO:0006281	DNA repair	11.3211	0.001247
EPC2	0.31632	5.94E-09	GO:0001834	trophectodermal cell proliferation	26.0727	0.0020627

PI4KA	0.2507	9.15E-06	GO:0002174	mammary stem cell proliferation	26.0727	0.0020627
MTHFR	0.16437	0.0019919	GO:0002941	synoviocyte proliferation	26.0727	0.0020627
			GO:0003419	growth plate cartilage chondrocyte proliferation	26.0727	0.0020627
			GO:0014009	glial cell proliferation	26.0727	0.0020627
			GO:0014841	satellite cell proliferation	26.0727	0.0020627
			GO:0033002	muscle cell proliferation	26.0727	0.0020627
			GO:0033687	osteoblast proliferation	26.0727	0.0020627
Group 100						
COG3	0.20148	0.0040206				
TRRAP	0.23564	0.00095304				
AMMECR1L	0.18073	0.0081692				
Group 101						
RFX5	0.17136	0.0027482	GO:0006915	apoptotic process	9.1923	0.015573
WDFY2	0.2102	0.00075771				
ARHGEF9	0.151	0.011693				
CYFIP2	0.17996	0.0033501				
Group 102						
TPR	0.21677	8.49E-13	GO:0005874	microtubule	8.25493	5.14E-06
PRPF8	0.17467	3.26E-10	GO:0005875	microtubule associated complex	10.755	5.84E-06
SPTAN1	0.24291	4.49E-15	GO:0008180	COP9 signalosome	14.0039	1.03E-05
ITGA6	0.088382	0.0015939	GO:0035658	Mon1-Ccz1 complex	13.1801	1.44E-05
ZC3H13	0.18387	9.01E-11	GO:0097149	centralspindlin complex	13.1801	1.44E-05
KIF7	0.12721	2.85E-06	GO:0005819	spindle	7.46875	5.47E-05
DYNC1H1	0.36143	9.34E-30	GO:0031080	nuclear pore outer ring	10.1847	5.65E-05
DOCK7	0.20413	5.79E-10	GO:0031262	Ndc80 complex	10.1847	5.65E-05
SPTBN1	0.2387	8.45E-15	GO:0031617	NMS complex	10.1847	5.65E-05
FYCO1	0.22526	4.93E-14	GO:0044611	nuclear pore inner ring	10.1847	5.65E-05
KIAA1462	0.16454	3.63E-09				
PLCB3	0.13817	1.30E-06				
SMARCC2	0.20186	3.15E-12				
ROCK2	0.10418	0.00035289				
CHD8	0.11864	2.28E-05				
ZFYVE26	0.11419	1.05E-05				
Group 103						
MYCBP2	0.16991	8.09E-05	GO:0007532	regulation of mating-type specific transcription,	3.532	0.036715
KAT6A	0.21337	1.46E-07		DNA-templated		
FMNL2	0.21619	3.28E-07	GO:0009373	regulation of transcription by pheromones	3.532	0.036715
PPRC1	0.18332	8.43E-06	GO:0019056	modulation by virus of host transcription	3.532	0.036715
PINK1-AS	0.26207	1.01E-09	GO:0031554	regulation of DNA-templated transcription,	3.532	0.036715
ARHGEF6	0.23135	7.03E-09		termination		
AGL	0.23335	3.06E-08	GO:0032784	regulation of DNA-templated transcription,	3.532	0.036715
				elongation		
			GO:0043620	regulation of DNA-templated transcription in response to stress	3.532	0.036715
			GO:0044324	regulation of transcription involved in anterior/posterior axis specification	3.532	0.036715
			GO:0045896	regulation of transcription during mitosis	3.532	0.036715
			GO:0045990	carbon catabolite regulation of transcription	3.532	0.036715
			GO:0060194	regulation of antisense RNA transcription	3.532	0.036715
Group 104						
AFF4	0.21011	4.73E-06	GO:0035327	transcriptionally active chromatin	29.875	0.0015819
PDS5B	0.16431	0.00020018	GO:0044093	positive regulation of molecular function	23.9	0.0025137
NUDT3	0.26239	2.51E-09	GO:0000775	chromosome, centromeric region	17.0714	0.004969
MTR	0.15202	0.0021987	GO:0000785	chromatin	13.2778	0.0081655
NSMAF	0.17288	0.00020527	GO:0000776	kinetochore	8.53571	0.019041
TP53BP1	0.23007	8.59E-07	GO:0001015	snoRNA transcription from an RNA polymerase II	7.96667	0.021659
				promoter		
			GO:0001111	promoter clearance from RNA polymerase II	7.96667	0.021659
				promoter		
			GO:0042789	mRNA transcription from RNA polymerase II	7.96667	0.021659
				promoter		
			GO:0042795	snRNA transcription from RNA	7.96667	0.021659

				polymerase II		
				promoter		
			GO:0061614	pri-miRNA transcription from RNA polymerase II	7.96667	0.021659
				promoter		
Group 105						
VCAN	0.29121	1.15E-07	GO:0044431	Golgi apparatus part	5.5154	0.041548
ALPK2	0.22155	2.55E-05				
AFAP1	0.19694	0.00025152				
DOPEY1	0.26156	1.01E-06				
LUZP1	0.21328	6.02E-05				
Group 106						
NOL6	0.2961	5.35E-10				
NUP155	0.30268	1.76E-10				
TTLL4	0.18633	4.40E-05				
HMCN1	0.31204	1.46E-11				
KIAA2018	0.23236	5.61E-07				
FMNL3	0.22852	3.36E-07				
Group 107						
RAB3B	0.25906	9.59E-05	GO:0031489	myosin V binding	89.625	0.00013947
BRPF1	0.19506	0.00094824	GO:0017022	myosin binding	51.2143	0.00048403
NCOA1	0.18959	0.0010911	GO:0016573	histone acetylation	35.85	0.0010284
ARFGEF1	0.27858	9.85E-06	GO:0043966	histone H3 acetylation	35.85	0.0010284
			GO:0033571	lactoferrin transport	19.9167	0.003418
			GO:0034436	glycoprotein transport	19.9167	0.003418
			GO:0036372	opsin transport	19.9167	0.003418
			GO:0046752	viral capsid precursor transport to host cell	19.9167	0.003418
				nucleus		
			GO:0051808	translocation of peptides or proteins into other organism involved in symbiotic interaction	19.9167	0.003418
			GO:0071693	protein transport within extracellular region	19.9167	0.003418
Group 108						
MAP3K6	0.33408	1.26E-09	GO:0023019	signal transduction involved in regulation of	7.17	0.024916
FAM46A	0.36823	4.96E-11		gene expression		
CTPS2	0.32154	1.28E-08				
PTPN23	0.25045	2.33E-05				
	0.25045					
Group 109						
NF1	0.22122	0.0028062				
CCDC93	0.21526	0.0014707				
RAD54L2	0.13896	0.034175				
Ungrouped						
TNRC6A	-0.013257	0.2546				
SSH1	0.0019034	0.86964				
SACS	-0.0090666	0.40738				
DNAJC13	-0.018734	0.15967				
SLC1A4	-0.0013685	0.90418				
ADO	0.023736	0.048381				
FOXJ2	0.014768	0.18979				
DNAJC18	0.011168	0.35303				
KIAA0247	0.0095399	0.46245				
MCUR1	0.020895	0.07774				
CEP41	0.0069656	0.56712				
UPF2	0.011486	0.3406				
ATG14	0.0087021	0.45597				
AVL9	0.034289	0.0020319				
CD2AP	0.014723	0.19852				
SOGA1	-0.0069301	0.47182				
ZBTB5	0.0069753	0.50359				
GAN	0.0033942	0.72716				
IL17RA	0.02979	0.012973				
MUT	-0.014396	0.24714				
IFT122	0.00010259	0.99367				
SERTAD2	0.018855	0.10069				
PDPK1	0.0033045	0.78499				
DGKZP1	0.01102	0.30345				
PRKD1	0.014264	0.21945				

LAMA1	-0.0017916	0.86945				
NPR2	-0.010071	0.3974				
URGCP	-0.016348	0.1494				
SLC12A6	0.016564	0.13457				
SIN3B	0.022823	0.078752				
SMAD9	-0.0040589	0.73894				
LRP4	0.0021317	0.85637				
ITPRIPL2	-0.0063684	0.61107				
ZNF445	0.017061	0.1496				
MLLT6	0.013805	0.23699				
ERCC6L	0.0095868	0.45434				
CUL9	-0.01226	0.28403				
CCNF	-0.0027282	0.81827				
PIK3CD	0.0084563	0.47869				
TOP3A	0.0011177	0.92163				
ZCCHC6	0.00019739	0.98631				
STYX	0.0048059	0.67821				
SBNO1	0.00085405	0.94023				
SHC4	0.025346	0.040622				
RASAL2	0.027027	0.034652				
ERN1	-0.0079987	0.45961				
MEGF8	0.0064573	0.57011				
TSPYL5	0.023184	0.038196				
ANKS1A	0.015191	0.18175				
DMXL2	0.010121	0.3657				
HIP1	0.0019466	0.86211				
RP11-350F4.2	0.012336	0.2811				
MCM3AP	-0.012519	0.29363				
NUFIP2	0.0074357	0.51675				
UHRF1BP1	0.004169	0.71674				
GREM2	0.016264	0.18455				
ABHD13	0.0024327	0.84415				
EPB41	0.0053095	0.63883				
THSD4	-0.0058653	0.63922				
MED1	-0.0016096	0.88553				
BMPER	0.0025224	0.82564				
CEP250	-0.0057881	0.60847				
SMC6	0.022353	0.06614				
CEP170P1	-0.0067616	0.55158				
SKIV2L	-0.0071541	0.56223				
ZNF317	0.0025181	0.81767				
IQCE	-0.008594	0.41964				
CBL	0.010651	0.32248				
TBC1D22B	0.020754	0.051421				
RELN	0.0073262	0.50544				
H6PD	0.039238	0.00047371				
GDF11	-0.015125	0.20916				
WASH5P	0.018734	0.11402				
PRDM4	-0.0021741	0.8563				
NOM1	-0.0073963	0.52898				
CHD6	-0.0020649	0.86826				
BAHD1	0.010315	0.35263				
MCM9	0.0064276	0.58309				
TARSL2	0.015783	0.16292				
RIPK1	0.0077443	0.51523				
PTPRU	0.0056845	0.61205				
POLQ	0.017782	0.12096				

Bibliography

1. P. Liberali, B. Snijder, L. Pelkmans, Single-cell and multivariate approaches in genetic perturbation screens. *Nat. Rev. Genet.* **16**, 18–32 (2014).

2. K. Schneeberger, Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat. Rev. Genet.* **15**, 662–676 (2014).

3. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

4. J. Eberwine, J.-Y. Sul, T. Bartfai, J. Kim, The promise of single-cell sequencing. *Nat. Methods* **11**, 25–27 (2013).

5. E. Shapiro, T. Biezuner, S. Linnarsson, Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).

6. P. C. Blainey, S. R. Quake, Dissecting genomic diversity, one cell at a time. *Nat. Methods* **11**, 19–21 (2013).

7. N. Crosetto, M. Bienko, A. van Oudenaarden, Spatially resolved transcriptomics and beyond. *Nat Rev Genet*. **16**, 57–66 (2015).

8. A. J. Rodriguez, K. Czaplinski, J. S. Condeelis, R. H. Singer, Mechanisms and cellular roles of local protein synthesis in mammalian cells. *Curr. Opin. Cell Biol.* **20**, 144–9 (2008).

9. V. Balagopal, R. Parker, Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs. *Curr. Opin. Cell Biol.* **21**, 403–8 (2009).

10. H. Jung, C. G. Gkogkas, N. Sonenberg, C. E. Holt, Remote Control of Gene Function by Local Translation. *Cell* **157**, 26–40 (2014).

11. A. R. Buxbaum, G. Haimovich, R. H. Singer, In the right place at the right time: visualizing and understanding mRNA localization. *Nat. Rev. Mol. Cell Biol.* (2014).

12. W. A. Bickmore, B. van Steensel, Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–84 (2013).

13. W. A. Bickmore, The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.* **14**, 67–84 (2013).

14. T. Cremer, C. Cremer, Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2**, 292–301 (2001).

15. I. J. Cajigas *et al.*, The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging. *Neuron* **74**, 453–66 (2012).

16. H. Y. Park, T. Trcek, A. L. Wells, J. A. Chao, R. H. Singer, An unbiased analysis method to quantify mRNA localization reveals its correlation with cell motility. *Cell Rep.* **1**, 179–84 (2012).

17. Z. B. Katz *et al.*, β-Actin mRNA compartmentalization enhances focal adhesion stability and directs cell migration. *Genes Dev.* **26**, 1885–90 (2012).

18. E. Lécuyer *et al.*, Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. *Cell* **131**, 174–187 (2007).

19. Y. Oleynikov, R. H. Singer, Real-Time Visualization of ZBP1 Association with β-Actin mRNA during Transcription and Localization. *Curr. Biol.* **13**, 199–207 (2003).

20. A. Sanchez, I. Golding, Genetic determinants and cellular constraints in noisy gene expression. *Science* **342**, 1188–93 (2013).

21. A. Raj, A. van Oudenaarden, Single-molecule approaches to stochastic gene expression. *Annu. Rev. Biophys.* **38**, 255–70 (2009).

22. B. Munsky, G. Neuert, A. van Oudenaarden, Using gene expression noise to understand gene regulation. *Science* **336**, 183–7 (2012).

23. M. Lagha, J. P. Bothma, M. Levine, Mechanisms of transcriptional precision in animal development. *Trends Genet.* **28**, 409–16 (2012).

24. A. M. Femino, F. S. Fay, K. Fogarty, R. H. Singer, Visualization of single RNA transcripts in situ. *Science* **280**, 585–90 (1998).

25. A. Raj, P. van Den Bogaard, S. Rifkin, A. van Oudenaarden, S. Tyagi, Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).

26. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).

27. D. Y. Vargas *et al.*, Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell* **147**, 1054–65 (2011).

28. M. N. Cabili *et al.*, Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol.* **16**, 20 (2015).

29. O. Padovan-Merhar *et al.*, Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Mol. Cell*, doi:10.1016/j.molcel.2015.03.005.

30. A. N. Boettiger, M. Levine, Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science* **325**, 471–3 (2009).

31. M. W. Perry, A. N. Boettiger, M. Levine, Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13570–5 (2011).

32. A. Raj, S. A. Rifkin, E. Andersen, A. van Oudenaarden, Variability in gene expression underlies incomplete penetrance. *Nature* **463**, 913–8 (2010).

33. A. N. Boettiger, M. Levine, Rapid transcription fosters coordinate snail expression in the Drosophila embryo. *Cell Rep.* **3**, 8–15 (2013).

34. S. C. Little, M. Tikhonov, T. Gregor, Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell* **154**, 789–800 (2013).

35. a.-E. Saliba, a. J. Westermann, S. a. Gorski, J. Vogel, Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **42**, 8845–8860 (2014).

36. D. Zenklusen, D. R. Larson, R. H. Singer, Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol.* **15**, 1263–71 (2008).

37. J. M. Levsky, S. M. Shenoy, R. C. Pezo, R. H. Singer, Single-cell gene expression profiling. *Science* **297**, 836–840 (2002).

38. M. R. Speicher, S. Gwyn Ballard, D. C. Ward, Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat. Genet.* **12**, 368–75 (1996).

39. E. Schrock *et al.*, Multicolor Spectral Karyotyping of Human Chromosomes. *Science (80).* **273**, 494–497 (1996).

40. M. J. Levesque, A. Raj, Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat. Methods* **10**, 246–8 (2013).

41. E. Lubeck, L. Cai, Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9**, 743–748 (2012).

42. M. Bates, B. Huang, G. T. Dempsey, X. Zhuang, Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science* **317**, 1749–53 (2007).

43. B. Huang, H. Babcock, X. Zhuang, Breaking the diffraction barrier: Super-resolution imaging of cells. *Cell* **143**, 1047–1058 (2010).

44. M. J. Rust, M. Bates, X. Zhuang, Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–5 (2006).

45. E. Betzig *et al.*, Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–5 (2006).

46. S. T. Hess, T. P. K. Girirajan, M. D. Mason, Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91**, 4258–72 (2006).

47. A. Dani, B. Huang, J. Bergan, C. Dulac, X. Zhuang, Superresolution imaging of chemical synapses in the brain. *Neuron* **68**, 843–56 (2010).

48. P. Kanchanawong *et al.*, Nanoscale architecture of integrin-based cell adhesions. *Nature* **468**, 580–4 (2010).

49. S. A. Jones, S.-H. Shim, J. He, X. Zhuang, Fast, three-dimensional super-resolution imaging of live cells. *Nat. Methods* **8**, 499–508 (2011).

50. K. Xu, G. Zhong, X. Zhuang, Actin, spectrin, and associated proteins form a periodic cytoskeletal structure in axons. *Science* **339**, 452–6 (2013).

51. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80)*, science.aaa6090 (2015).

52. J. Harrow *et al.*, GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).

53. A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, S. Tyagi, Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–9 (2008).

54. E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, L. Cai, Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).

55. D. Y. Zhang, G. Seelig, Dynamic DNA nanotechnology using strand-displacement reactions. *Nat. Chem.* **3**, 103–13 (2011).

56. B. J. Beliveau *et al.*, Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 21301–21306 (2012).

57. Y. E. Murgha, J.-M. Rouillard, E. Gulari, Methods for the Preparation of Large Quantities of Complex Single-Stranded Oligonucleotide Libraries. *PLoS One* **9**, 1–10 (2014).

58. A. Raj, A. van Oudenaarden, Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–26 (2008).

59. L.-H. So *et al.*, General properties of transcriptional time series in Escherichia coli. *Nat. Genet.* **43**, 554–60 (2011).

60. M. Safran *et al.*, GeneCards Version 3: the human gene integrator. *Database (Oxford)*. **2010**, baq020 (2010).

61. O. Padovan-Merhar, A. Raj, Using variability in gene expression as a tool for studying gene regulation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **5**, 751–9.

62. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **95**, 14863–14868 (1998).

63. N. Gehlenborg *et al.*, Visualization of omics data for systems biology. *Nat. Methods* **7**, S56–68 (2010).

64. M. Ashburner, C. Ball, J. Blake, D. Botstein, Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

65. H. Yoshida *et al.*, KIAA1199, a deafness gene of unknown function, is a new hyaluronan binding protein involved in hyaluronan depolymerization. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5612–7 (2013).

66. D. A. Lauffenburger, A. F. Horwitz, Cell migration: a physically integrated molecular process. *Cell* **84**, 359–69 (1996).

67. T. A. Rapoport, Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* **450**, 663–9 (2007).

68. C. H. Jan, C. C. Williams, J. S. Weissman, Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* **346**, 1257521–1257521 (2014).

69. L. A. Mingle *et al.*, Localization of all seven messenger RNAs for the actin-polymerization nucleator Arp2/3 complex in the protrusions of fibroblasts. *J. Cell Sci.* **118**, 2425–33 (2005).

70. J. B. Lawrence, R. H. Singer, Intracellular localization of messenger RNAs for cytoskeletal proteins. *Cell* **45**, 407–15 (1986).

71. H. Babcock, Y. M. Sigal, X. Zhuang, A high-density 3D localization algorithm for stochastic optical reconstruction microscopy. *Opt. Nanoscopy*. **1**, 6 (2012).

72. L. Zhu, W. Zhang, D. Elnatan, B. Huang, Faster STORM using compressed sensing. *Nat. Methods* **9**, 721–3 (2012).

73. H. Babcock, J. Moffitt, Y. Cao, X. Zhuang, Fast compressed sensing analysis for super-resolution imaging using L1-homotopy. *Opt. Express* **21**, 28583–28596 (2013).

74. S. W. Hell, Microscopy and its focal switch. *Nat. Methods* **6**, 24–32 (2009).

75. J. H. Lee *et al.*, Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).

76. J. H. Lee *et al.*, Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, 442–458 (2015).

77. R. Sandberg, Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **11**, 22–24 (2013).

78. D. A. Jaitin *et al.*, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–9 (2014).

79. Q. Deng, D. Ramsköld, B. Reinius, R. Sandberg, Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–6 (2014).

80. A. R. Wu *et al.*, Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–6 (2014).

81. A. M. Streets *et al.*, Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 7048–53 (2014).

82. L. Wang, K. A. Janes, Stochastic profiling of transcriptional regulatory heterogeneities in tissues, tumors and cultured cells. *Nat. Protoc.* **8**, 282–301 (2013).

83. P. Brennecke *et al.*, Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–5 (2013).

84. D. Grün, L. Kester, A. van Oudenaarden, Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–40 (2014).

85. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).

86. B. J. Beliveau *et al.*, Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 21301–6 (2012).

87. Q. Xu, M. R. Schlabach, G. J. Hannon, S. J. Elledge, Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci.* . **106** , 2289–2294 (2009).

88. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

89. C. Trapnell *et al.*, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

90. I. Dunham *et al.*, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

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

92. M. Batish, A. Raj, S. Tyagi, in *Methods in molecular biology*, J. E. Gerst, Ed. (Humana Press, Totowa, NJ, 2011), vol. 714 of *Methods in Molecular Biology*, pp. 3–13.

93. A. Buxbaum, B. Wu, R. Singer, Single β-actin mRNA detection in neurons reveals a mechanism for regulating its translatability. *Science* **343**, 419–423 (2014).

94. T. K. Moon, *Error Correction Coding: Mathematical Methods and Algorithms* (Wiley, ed. 1st, 2005).

95. G. T. Dempsey, J. C. Vaughan, K. H. Chen, M. Bates, X. Zhuang, Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nat. Methods* **8** (2011), pp. 1027–1036.

96. S. W. Hell, J. Wichmann, Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Opt. Lett.* **19**, 780 (1994).

97. M. G. L. Gustafsson, Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13081–6 (2005).

98. R. E. Thompson, D. R. Larson, W. W. Webb, Precise nanometer localization analysis for individual fluorescent probes. *Biophys. J.* **82**, 2775–83 (2002).

DECLARATION

I, **Heather Stewart,** declare that the following statements are true and correct to the best of my knowledge, information and belief:

1. I am over 18 years of age and have personally reviewed the facts set forth in this declaration, based upon practices and training encountered in the course of my work overseeing a team of Technical Support Analysts, including my personal review and confirmation of available business records and documentation maintained by my employer, ProQuest LLC ("ProQuest").

2. I am a **Manager, Content Curation** employed by ProQuest in Ann Arbor, Michigan. I have worked for ProQuest since **1998** (+25 years). ProQuest is an information company that collects, organizes publishes and distributes for sale content from a wide range of sources, including dissertations and master's theses. I am familiar with ProQuest's sales and publication practices with respect to the publication of dissertations and master's theses as well as the related Abstract Database (defined in in paragraph 3).

3. ProQuest receives dissertations and master's theses from participating degree granting institutions. Shortly after these works are received, ProQuest makes them available for sale to the public, unless specifically restricted from sales by the dissertation author or the degree granting institution. Prior to 2013, a "Record" of each dissertation and each master's thesis submitted to ProQuest was published in ProQuest's print publication Dissertations Abstracts International or Master's Abstracts International, respectively (the print editions ceased being distributed in print as of volume 73-06 published in December 2012). Beginning in 1985, each Record is also published in the Dissertation Abstracts International Database ("Abstract Database").[1] A "Record" consists of an

[1] The Abstract Database has been maintained exclusively by ProQuest under various corporate names, including Bell & Howell Information and Learning Company, ProQuest Information and Learning Company, UMI Company and University Microfilms, Inc. and has been licensed to subscribers under various product names including "Dissertation Abstracts Online" and "ProQuest® Digital Dissertations."

abstract, index record, and citation information including title, author name, degree granting institution and degree date.

4. Prior to 1985, Records were published only in the print publications Dissertations Abstracts International and Master's Abstracts International. In 1985, the Records in these print publications were combined to create the Abstract Database which was made available online through third party vendors, such as Dialog, Data-Star, and OCLC. In 1987, ProQuest began to offer the Abstract Database on CD ROM. In 1997, ProQuest made available the contents of the Abstract Database in its ProQuest® Digital Dissertations and Theses database ("PQDT"), which is accessible via the world wide web to authenticated subscribers. The Abstract Database allows for text searching of the Records, including the title, author, index terms and the abstract.

5. The **PQDT** database includes the **Abstract Database** and is supplemented by the addition of certain select dissertations and master's theses in PDF format. Whether a dissertation is represented in full PDF format varies based on a number of factors, including the format in which the dissertation or theses was originally submitted for publication (digital, paper or microfiche) or whether the dissertation or theses was part of a retrospective digitization effort. PQDT allows for text searching of the Records as well as full text where the full PDF is properly formatted.

6. At the time of publication of the Dissertation (the full citation is provided in paragraph 9 below), ProQuest published in the **Abstract Database** a Record of each dissertation and each master's thesis that ProQuest offers for sale with new Records added on a **monthly** basis. Once a Record is published on the Abstract Database, the Record is available to the students, faculty, staff and/or patrons of any institutions that subscribe to the Abstract Database. At the time of publication the Abstract Database was available throughout the United States, through third party vendors as well as in CD-ROM format directly from ProQuest.

7. At the time of the Dissertation's publication, the Record and full PDF of the Dissertation was added to PQDT as specified below.

ProQuest LLC (and its predecessor entities) is and at all times has been a U.S. Company based in Ann Arbor, MI.

8. I have reviewed ProQuest's records regarding the Dissertation cited as:

Dissertation No.	3738727
Title:	Spatially resolved, highly multiplexed RNA profiling in single cells
Author:	Chen, Kok Hao
Degree Date:	2015
University:	Harvard University

Full Text Availability

First Received by UMI/ProQuest:	12/2015
First Made Available for Sale (print or electronic)	12/4/2016
DAI volume issue:	DAI-B 77/04(E), Dissertation Abstracts International
DAI publication month and year:	12/2015
Date full text is loaded into PQDT:	12/10/2015 (Abstract and Index loaded, PDF suppressed from viewing or sales)
Information available on restrictions:	Dissertation was submitted with a 1 year embargo. Start date: 12/4/2015 - End date: 12/4/2016

9. At the time of the Dissertation's publication, the Record of the Dissertation was made available directly from ProQuest in electronic format and online via third party vendors, including Dialog, as part of the Abstract Database which was made available via such vendors in the ordinary course of business.

10. The attached print copy of the Dissertation is a true and complete copy of the Record and the full Dissertation as originally published and kept by ProQuest in the ordinary course of its business.

I make this Declaration on personal knowledge and, if called as a witness at trial, I am competent to testify.

I declare under penalty of perjury that the forgoing is true and correct.

Executed on 12/5/2024

Heather Stewart

Heather Stewart
Manager, Content Curation